Title

A rare inter-comparison of nutrient analysis at sea: lessons learned and recommendations to enhance comparability of open ocean nutrient data

Authors

Triona McGrath[1], Margot Cronin[2], Elizabeth Kerrigan[3], Douglas Wallace[3], Clynton Gregory[1], Claire Normandeau[3], and Evin McGovern[2]

Affiliations

1; National University of Ireland, Galway, University Road, Galway, Ireland

2; The Marine Institute, Ireland, Rinville, Oranmore, Galway, Ireland

3; Dalhousie University, Steele Ocean Sciences Building, 1355 Oxford St., PO Box 15000, Halifax, Nova Scotia, Canada B3H 4R2

Correspondence Author;

Triona McGrath; triona_mcgrath@hotmail.com/ triona.mcgrath@marine.ie

Evin McGovern; evin.mcgovern@marine.ie

Abstract

An inter-comparison study has been carried out on the analysis of inorganic nutrients at sea following the operation of two nutrient analysers simultaneously on the GO-SHIP A02 trans-Atlantic survey in May 2017. Both instruments were Skalar San^{++} Continuous Flow Analysers, one from the Marine Institute, Ireland and the other from Dalhousie University, Canada, each operated by their own laboratory analysts following GO-SHIP guidelines, while adopting their existing laboratory methods. There was high comparability between the two datasets and vertical profiles of nutrients also compared well with those collected in 1997 along the same A02 transect by the World Ocean Circulation Experiment. The largest differences between datasets were observed in the low nutrient surface waters and results highlight the value of using three reference materials (low, mid and high concentration) to cover the full range of expected nutrients and identify bias and non-linearity in the calibrations. The inter-comparison also raised some interesting questions on the comparison of nutrients analysed by different systems and a number of recommendations have been suggested that we feel will enhance the existing GO-SHIP guidelines to improve the comparability of global nutrient datasets. A key recommendation is for specification of clearly-defined data quality objectives for oceanic nutrient measurements and a flagging method for reported data that do not meet these criteria.

The A02 nutrient dataset is currently available at the National Oceanographic Data Centre of Ireland; http://dx.doi.org/10.20393/CE49BC4C-91CC-41B9-A07F-D4E36B18B26F and http://dx.doi.org/10.20393/EAD02A1F-AAB3-4F4E-AD60-6289B9585531.

1. Introduction

Dissolved nutrients such as nitrate, nitrite, silicate and phosphate can be a critical limiting factor constraining growth of phytoplankton, which in turn form the base of the marine food web. They also provide useful chemical signatures (e.g. ratios of preformed nutrients) that can distinguish water masses and their origins (Broecker and Peng, 1982) as well as act as tracers for biogeochemical processes such as nitrogen fixation and denitrification (Deutsch and Weber, 2012). There is growing evidence for significant variability including long-term trends in nutrient levels in both coastal (Kim et al., 2011) and open ocean surface (Yasunaka et al., 2014), and deep waters (Kim et al., 2014). These changes reflect both direct human intervention in the global environment, especially the effects of the massive ongoing perturbation of the nitrogen cycle (Yang and Gruber, 2016) as well as changes in ocean circulation and biogeochemical cycling that may or may not be anthropogenically influenced (e.g. Di Lorenzo et al., 2008).

Identification and attribution of variability of nutrient concentrations has been complicated by the existence of systematic analytical errors in datasets collected by different groups at different times. This can lead to controversy over the significance of observed long-term changes (e.g. Zhang et al., 2001) and generally requires empirical correction of historical data, using a variety of ad hoc approaches and principles (Keller et al., 2002; Moon et al., 2016; Pahlow and Riebesell, 2000; Tanhua et al., 2009b). Recognition of such systematic errors within and between datasets led to a series of international comparison studies and the introduction of Certified Reference Materials for dissolved nutrients (Aoyama et al., 2016; Aoyama et al., 2007), as well as recommendations concerning standard protocols for sampling, sample preservation and analysis (Hydes et al., 2010). These steps have undoubtedly contributed to a general improvement in inter-laboratory comparability of field-collected data. However, it is notable that most inter-comparison studies rely on either: a) shore-based laboratory-based analysis of replicate samples in the context of specially organised inter-comparison studies; or b) crossover analysis of measurements made at nearby locations in the ocean where temporal and spatial variability is expected to be small.

The former approach is valuable, but most analysts are aware that conditions during an actual research cruise do not always match the stable, controlled conditions of a shore-based laboratory where a group can prepare carefully for their measurement of inter-comparison samples. On the other hand, the latter approach works well in oceanic regions where stable, unchanging nutrient concentrations can be expected. However, in regions such as the surface open ocean of the North Atlantic, or the Northwest Pacific and in coastal regions everywhere, temporal and/or spatial variations can be expected which complicates the interpretation of crossover comparisons.

In this paper we report the results, findings and lessons learned from a rare opportunity in which two independent nutrient analysis teams participated jointly in a deep ocean hydrographic section as part of the international GO-SHIP program (Talley et al., 2016). Both teams followed standard protocols (Hydes et al., 2010) and both groups used Certified Reference Materials during the cruise. As such, the cruise provided an opportunity to assess the likely comparability of nutrient data collected following such protocols as well as helping to identify a number of issues affecting data quality that could be of general relevance to groups conducting such measurements elsewhere. The inter-comparison illustrates how lab-based performance assessment can be compared to at-sea assessment. We are not aware of any other report of such an extensive, at-sea inter-comparison of nutrient measurement systems.

The GO-SHIP A02 survey was completed in April/May 2017 on the RV Celtic Explorer, travelling from St. John's, Newfoundland, Canada, across the North Atlantic to Galway, Ireland with on-board teams from Ireland, Canada, Germany, the UK, and the USA. The survey provided an

unusual opportunity for cross-comparison of methods, data quality procedures and exchange of technical expertise between the international scientific groups. The Marine Institute (MI) and Dalhousie University (Dal) teams brought separate nutrient Skalar San++ auto analysers on the survey to provide contingency against technical failures and allow for on-board inter-comparison of data as well as exploration of the impact on data quality of subtle differences in laboratory methods, procedures and instrument configurations that ostensibly conform to the same (GO-SHIP) guidelines and quality assurance criteria.

A total of 67 stations were occupied along the A02 transect (Fig. 1), with 1231 nutrient samples analysed for total oxidised nitrogen (TOxN), nitrite, phosphate and silicate on the MI nutrient system. Of these, 12 stations were sampled and analysed on both the MI and Dal nutrient systems, allowing the comparison of 291 samples between the two systems. The 12 stations were also compared with historical data from the A02 transect completed on a World Ocean Circulation Experiment survey in 1997.

2. Methods

Sampling, sample preservation and analytical procedures on both systems followed methods outlined in the GO-SHIP guidelines for nutrient analysis at sea (Hydes et al., 2010), while both groups also incorporated their existing laboratory quality control (QC), which was specifically adapted to their individual instruments. Note, a draft revised version of the GO-SHIP nutrients manual available at time of writing, Becker et al. (in prep.), was not available ahead of the 2017 A02 survey.

2.1 Sampling Procedures

Both groups collected nutrient samples directly from the Niskin bottles into falcon tubes (details in Table 1) and as per GO-SHIP guidelines, the samples were not filtered. Samples were analysed on board typically within 12 hours of sampling.

2.2 Analytical Methods

Analysis was carried out on two separate Skalar San^{++} Continuous Flow Analysers, setup in two separate on-board containerised laboratories brought by each team. Both analysers run four channels of nutrients simultaneously; total-oxidised nitrogen, nitrite, silicate and phosphate. The Dal system also measures ammonia, however contamination issues were encountered during the survey and therefore, there is no further discussion of this method. Both instruments consisted of an auto-sampler, where a needle draws the sample into the analyser, which is then split into the four channels. Each channel had its own set of reagents, where the stream of reagents and samples is pumped through the manifold to undergo treatment such as mixing and heating before entering a flow cell to be detected. The air-segmented flow promotes mixing of the sample and prevents contamination between samples. The reagents react to develop a colour, which is measured as an absorbance through a flow cell at a given wavelength. The Skalar Interface transmits all the data to the Skalar Flow Access software.

Reagents for both systems were made using high-purity chemicals, pre-weighed using high-precision calibrated balances prior to the survey, stored in acid-washed polyethylene (PE)

containers and mixed to final volume on board using ultrapure water. See reagent compositions in Table 1. The ultrapure water was generated using a Smart2Pure water purification system. Reagent storage time was in accordance with the Skalar methods: most can be stored for 1 week, the silicate ammonium heptamolybdate and oxalic acid reagents for 1 month, however fresh reagents were typically made every 2-3 days due to the volume required during the survey.

The analytical procedures for all nutrients were similar between the Dal and MI systems, but with some differences in the chemical composition of reagents and volumes of reagents/sample through the instruments (Table 1). For the determination of nitrite, the diazonium compounds formed by diazotizing of sulfanilamide by nitrite in water under acidic conditions (due to phosphoric acid in the reagent) is coupled with N-(1-naphthyl) ethylenediamine dihydrochloride to produce a reddish-purple colour, is measured at 540 nm.

For silicate determination the sample is acidified with sulphuric acid and mixed with an ammonium heptamolybdate solution forming molybdosilicic acid. This acid is reduced with L(+)ascorbic acid to a blue dye, and measured at 810 nm. Oxalic acid is added to avoid phosphate interference.

For the determination of phosphate, ammonium heptamolybdate and potassium antimony(III) oxide tartrate react in an acidic medium (with sulphuric acid) with diluted solutions of phosphate to form an antimony-phospho-molybdate complex. This complex is reduced to an intensely blue-coloured complex by L(+)ascorbic acid and is measured at 880 nm.

For the determination of total oxidised nitrogen (TOxN) both methods buffer the sample to pH of 8.2, which is then passed through a column containing granulated copper-cadmium to reduce nitrate to nitrite. The nitrite originally present, plus the reduced nitrate, is determined by diazotizing with sulfanilamide and coupling with N-(1-naphthyl) ethylenediamine dihydrochloride to form a strong reddish-purple dye which is measured at 540nm. MI uses a ammonium chloride and ammonium hydroxide buffer solution, while the Dal buffer solution is made of imidazole and hydrochloric acid (Table 1). The MI uses a cadmium column where no air bubbles are allowed through, while the Dal system allows air bubbles though their column but monitors the efficiency of the reduction process daily, re-activating the cadmium column with 1M hydrochloric acid and a copper sulfate solution if the efficiency falls below 95%. It should be noted that above 95%, the reduction efficiency is consistent throughout a run and therefore does not have to be corrected for; below 95% the reduction efficiency may be variable, so the column must be reactivated to ensure there is no impact on the samples; this follows GO-SHIP protocol (Hydes et al., 2010).

Both instruments were calibrated daily using a suite of calibration standards (see calibration range in Table 2). The primary standards for each nutrient was made by each team immediately prior to the survey using calibrated balances and high purity chemicals diluted to 1L with ultrapure water, as per Skalar methods. The primary stocks were stored in a refrigerator for the duration of the survey. Two batches of primary stocks were used on the MI system to ensure no bias from an individual batch, while one batch of primary stock was used on the Dal system. Weekly secondary stocks were diluted from the primary stocks into 100ml polypropylene (PP) flasks and stored in the fridge when not in use. These could be used for one week. Daily standards were made from secondary stock into 100ml PP volumetric flasks.

MI calibration standards were made using calibrated fixed volume pipettes while Dal standards were made using calibrated adjustable volume pipettes (0.1 – 1 ml, 0.5 – 5 ml) and one calibrated fixed volume pipette (10 ml). All pipettes were tested prior to the start of the survey to ensure that the volumes delivered were accurate. The MI secondary stocks were made using ultrapure

water, while the daily standards were made using artificial seawater (ASW) with salinity of 35. Both secondary and daily standards on the Dal system were made using ASW (salinity 33-35). Concentrations of daily standards for each system are in Table 2, where first-order (linear) calibration curves were fitted; neither group forced their calibrations through zero. An R^2 >0.99 was deemed acceptable for goodness-of-fit, as recommended by Skalar methods. Additional details on the primary and secondary stock solutions can be found in Table 1 in the Supplementary Material.

A notable difference between the two systems was the composition of the baseline wash; the MI analyser used ASW – a sodium chloride solution with a similar salinity to the expected samples (salinity 35), as the baseline wash for all channels. Batches of sodium chloride used were tested prior to the survey to ensure no contamination with any of the nutrients. The MI system runs its baseline wash as the first (zero) standard. The Dal system used ultrapure water as the baseline wash and ran a sample of ASW (effectively a blank, i.e. no nutrients) as the first standard, which was set to 0 for each standard curve (e.g. Standard 1 in Table 2). The GOSHIP manual recognises both ASW and ultrapure water as suitable baseline washes for nutrient analysis at sea.

2.3 Quality Control

The Certified Reference Materials (CRMs) used on the survey by both groups were supplied from KANSO (Aoyama et al., 2016; Aoyama et al., 2007). Two batches (Batch CD and Batch BW, Table 3) were used on the MI system to cover the full range of nutrients expected on the survey, with a CD and BW analysed at the beginning of a run and another CD at the end of the run. While Dal primarily analysed Batch CD, they also analysed a BW CRM on three runs, as a comparison. The KANSO certified values are in µmol/kg (Table 3), which were converted to µmol/l for the QC charts since the Skalar results are in µmol/l. The density for this conversion was calculated as per Millero and Poisson (1981), where the CRM salinity and analysis temperature (laboratory temperature, of 20°C for both the MI and Dal containers) was used. The BW CRM for silicate has a concentration (61.47 µmol/l) higher than the highest standard (60 µmol/l) used by both groups, and is therefore only used as an indication of QC variations for higher levels of silicate.

Prior to GO-SHIP, the MI laboratory developed acceptance criteria for CRMs based on the standard deviation of CRM results. The MI had primarily used Eurofins seawater and estuarine CRMs[1] in the daily nutrient runs, with good results. The MI also participates in the QUASIMEME marine and estuarine proficiency testing schemes; between 2008 and 2017, the average absolute z-scores |Z| from 84 test samples at the MI laboratory were 0.5 for TOxN, 0.4 for nitrite, 0.5 for silicate and 0.4 for phosphate. In that period, |Z|-scores were satisfactory for all results >LOQ, with the exception of a single silicate result (Z = 2.04).

With no history of KANSO CRM results prior to the A02 survey, the Quasimeme z-score assessment criteria were used where a z-score < 2 is considered satisfactory. The z-score is calculated as:

[1] (https://www.eurofins.dk/miljoe/vores-ydelser/certificerede-vki-referencematerialer/information-in-english/)

Equation 1; $$z - score = \frac{Measured\ value - Certified\ value}{Total\ error}$$

(Cofino and Wells, 1994)

Total error is calculated as;

Equation 2; $$Total\ error = \frac{Assigned\ value\ x\ Proportional\ Error\ (6\%)}{100} + 0.5\ x\ Constant\ error$$

Constant errors are 0.05, 0.01, 0.1 and 0.05 µmol/l for TOxN, nitrite, silicate and phosphate, respectively, which are defined by the Scientific Advisory Board of Quasimeme. These constant errors are similar to accuracy/uncertainty levels called for by the Global Ocean Observing System's (GOOS) Biogeochemistry Expert Panel,

(http://www.goosocean.org/index.php?option=com_oe&task=viewDocumentRecord&docID=17474). (We note that the GOOS Panel does not follow Quasimeme in also specifying a proportional error: see Discussion section).

On the MI system, every sample was analysed twice and relative percentage differences (RPD_{REP}) were calculated for replicates using Equation 3. Samples with RPD_{REP} were >10%, were re-analysed.

Equation 3; $$RPD_{REP} = \frac{Replicate\ A - Replicate\ B\ concentration}{Average\ nutrient\ concentration} \times 100\%$$

On the Dal system, every sample was measured in triplicate and a coefficient of variation (CV(%)) was calculated (Eq. 4). For samples with concentrations of 0.5 to 10 µmol/l and >10 µmol/l, an outlier replicate was removed if the CV(%) was >5% or >3%, respectively. If the remaining two replicates differed by more than these amounts, both were rejected and the sample re-analysed during the following run. For samples with lower concentrations (<0.5 µmol/l), the CV(%) test was not used.

Equation 4; $$CV(\%) = \frac{\text{Standard deviation of replicates}}{\text{Average of replicates}}\ \text{x}\ 100\%$$

For both systems, limits of detection (LOD) and quantification (LOQ) were calculated as 3*standard deviation (LOD) and 10*standard deviation (LOQ) based on 10 replicate analyses of low nutrient seawater solution (see Table 4). Concentrations falling between the LOD and LOQ value were reported as <LOQ, while concentrations lower than the detection limit were reported as <LOD.

Drift samples were analysed after every four samples on both systems, to correct for instrumental drift during a run. The drift samples were prepared from secondary stock and artificial seawater (see concentrations in Table 2).

System Suitability Standards (SSS) were made daily by the MI group using secondary stock standards and artificial seawater. These were not used to correct for drift but instead analysed as

an internal reference material every four samples to ensure drift correction was accurate and to identify any problems during the course of a run. All SSS were checked in post processing: any falling > ±10% of the SSS value were marked as failed QC. The four samples on either side of a failed SSS were then re-analysed. The Dal group analysed their drift solution as an internal reference material every 4 samples; this "drift check" was monitored during a run but was not used for post-processing rejection/flagging.

2.4 Comparison of data

To compare final nutrient concentrations analysed on the two instruments, the sample relative percentage difference ($RPD_{MI\text{-}DAL}$) was also calculated based on the MI and Dal nutrient concentrations;

Equation 5. $$RPD_{MI\text{-}DAL} = \frac{Average\ MI\ concentration - Average\ Dal\ concentration}{Average\ nutrient\ (MI+Dal)\ concentration} \times 100\%$$

While nitrite was analysed on both instruments, there were issues with nitrite contamination in both systems, potentially due to the ultrapure water quality on board. Whereas all frozen samples were re-analysed at the MI after the cruise, this was not possible for the Dal samples so a comparison of nitrite methods and data cannot be carried out in this study.

3. Results

3.1 Sample-to-sample comparisons including vertical profiles

The MI and Dal data are both available on the MI database (see links in data availability). It is important to note that the MI data used in this comparison is calculated using split calibration curves; any TOxN and silicate data <5umol/l was calculated from a calibration range of 0-10 µmol/l while all other data was calculated using the 0-50 µmol/l calibration range. The reason for this split calibration is discussed in section 3.2 and 3.3.

Overall, without any adjustments based on CRM analysis results, there was relatively good agreement between vertical profiles of nutrients measured with the two systems, as can be seen from vertical profiles presented in Fig. 2 and Supplementary Material (Fig. 1). The mean percentage differences ($RPD_{MI\text{-}DAL}$) for all of the comparison samples measured during the cruise (n = 278-284) are shown in Table 5 and are -1.4±0.6%, -1.1±1.1% and +2.3±1.2% for TOxN, silicate and phosphate, respectively, where uncertainties are 95% confidence intervals. This gives general confidence in the overall comparability of the data and individual methods, standardization and analysis protocols used by each group.

For silicate, 70% of samples had $RPD_{MI\text{-}DAL}$ < 5%. The largest differences are in the top 400m which typically had < 3 µmol/l silicate, where 8% of all the samples have RPDs between 11 – 117%, with the highest RPDs in stations with lowest silicate values (see vertical profiles of $RPD_{MI\text{-}DAL}$ in Fig. 3). In contrast, for samples >400m, there was no significant difference between silicate concentrations measured on the two systems with an average $RPD_{MI\text{-}DAL}$ of 0.3±0.7%, where the uncertainty is the 95% confidence interval.

TOxN vertical profiles also compare reasonably well, with 77% of all $RPD_{MI\text{-}DAL}$ < 5%. Virtually all TOxN samples with $RPD_{MI\text{-}DAL}$ > 10% are within the top 200m where TOxN concentrations are low (Fig. 3). However, Fig. 3 shows that MI values of TOxN from deeper than 400m are significantly lower, by 2.1±0.4% (95% CI), than concentrations measured on the Dal system. This is consistent

with the difference in mean values reported for CRM analyses on the two systems (see Section 3.2 and Table 6).

There was less agreement between the two systems for phosphate; with only 38% of samples having $RPD_{MI\text{-}DAL}$ < 5% (79% of all samples had $RPD_{MI\text{-}DAL}$ < 10%,). Almost half of the samples with $RPD_{MI\text{-}DAL}$ > 10% were in the top 400m (Fig. 3). The remaining samples with larger differences deeper in the water column were from early stations of the cruise when the Dal system had problems with its phosphate channel. These problems were resolved and, in addition, the calibration range was altered from Station 46 onwards. If the earlier stations are excluded from the comparison, the average $RPD_{MI\text{-}DAL}$ for samples >400m showed an average $RPD_{MI\text{-}DAL}$ of 6.4±0.8% (95% CI). The negative bias of Dal's phosphate results, relative to MI's, is also consistent with the difference of ca. 4% in CRM results measured on the two systems from station 46 onwards (see Section 3.2; Fig. 4; Table 6).

A comparison was also performed between analyses of frozen replicate samples conducted in the MI laboratory after the survey with MI samples analysed at sea. The $RPD_{SEA\text{-}LAB}$ [($conc_{sea}$-$conc_{lab}$)/average $conc_{sea\&lab}$ x100%], was 4(±8)% for TOxN, 8(±14)% for silicate and 13(±16)% for phosphate (where uncertainties are given as 1 standard deviation). The frozen samples were defrosted at the MI overnight prior to analysis, which was carried out within two months of sample collection. The $RPD_{SEA\text{-}LAB}$ was typically positive, so that nutrient concentrations were lower in the frozen samples. This was also observed in a number of frozen samples that were analysed while at sea during the A02 survey. Of the nitrite samples that passed QC early in the survey, the frozen re-runs had differences within the limit of quantification (<LOQ = 0.04 µmol/l) of the method.

3.2 Comparison of QC results at-sea and on-shore

Both systems used the z-score criteria used by Quasimeme (with a proportional error of 6%) for assessment of the CRM results during the survey; all CRMs had $|Z|$-scores within 2, as shown on the QC charts in Fig. 4.

Table 6 presents summary statistics for differences between measured and Certified values as measured on both systems, expressed as percentages of Certified values, together with the coefficient of variation, CV(%), of these differences. Overall, coefficients of variation for CRM analyses made on both systems were in the range of 3-5% for all three nutrients. Early results for phosphate on the Dal system showed higher variation (10%), but this improved later in the cruise following modifications to calibration procedures (Table 2).

For TOxN there were statistically significant biases of order -3% (95% confidence interval of ±1) (Dal) and -5% (±1.5) (MI) for the lower concentration CRM (CD), with apparently smaller bias at the higher concentration (BW). For silicate, the Dal and MI analyses were not statistically distinguishable from Certified values. For phosphate, the high scatter of the Dal analyses at earlier stations (before Stn. 46), precluded useful estimation of bias for the cruise as a whole. The later analyses on the Dal system, with reduced scatter, suggested a bias of order -6% (±3) for the mid-range CRM, whereas the MI phosphate analyses showed a smaller bias of ca. -2.5% with the mid-range CRM.

Comparison of the QC results of the MI system during the A02 cruise with those from shore-based analyses conducted before and afterwards suggests a considerable reduction in the precision of CRM analyses conducted at-sea. Between 2013 and 2017 the Eurofins CRMs (n=67) were measured with a CV(%) of 1.9% for TOxN, 3.0% for silicate and 2.6% for phosphate. Following the survey, the CV(%) of KANSO CD CRM (n=20) was 2.2% for TOxN, 1.7% for silicate and 4.4% for phosphate; whereas the CV(%) of the KANSO CJ CRM (n=18) was 1.7% for TOxN, 3.0% for silicate and 2.8% for phosphate. Hence the variability of CRM analyses for TOxN and silicate during the A02 cruise (Table 6) is almost a factor of two larger than that of corresponding shore-based analyses whereas phosphate variability was largely unchanged. This, together with the bias in the TOxN data, has been noted in the metadata for the dataset.

At-sea QC results with the Dal system on the A02 cruise were comparable to subsequent on-shore analyses (September, 2017) which had a CV(%) for the KANSO CD CRM (n=21) of 2.7% for TOxN, 3.3% for silicate and 4% for phosphate (these values can be compared with Table 6). Analyses conducted at-sea one year later (on cruise MSM74, May – June 2018) were also comparable, with CV(%) of 2.5% for TOxN, 2.8% for silicate and 5.4% for phosphate.

3.3 Comparison of instrument calibrations

Both groups carried out testing of instrument calibrations prior to the A02 survey to determine optimal calibration range. Tests indicated that the optimal calibration range for TOxN on the MI instrument was 0-30 µmol/l. However, early in the cruise, a negative bias was observed in the MI QC charts for the higher TOxN CRM (Batch BW, 25.19 µmol/l) while, at the same time, comparison of the MI and Dal datasets also identified a negative bias in the MI TOxN data relative to Dal data for samples at concentrations > 15µmol/l). In an attempt to correct the bias while at-sea, the TOxN calibration range on the MI system was increased from 0 – 30 µmol/l to 0 – 50 µmol/l to match the Dal system's calibration range. This change appeared to reduce the negative bias in the BW CRM, without substantially affecting the CD CRM results (Fig. 2 Supplementary Material). The reason for the negative bias was, and remains, unclear since on return of the instrument to the laboratory following the cruise, standards up to 30 µmol/l resulted in better performance with greater precision and with less bias evident for TOxN.

A positive bias in the CD CRM was noted on the Dal phosphate channel early on in the cruise. This was corrected for by adding three new standards were between 0 and 0.8 µmol/l to help with standard curve fit (Table 2). This change in the calibration range removed the positive bias (Figure 4), and as such, stations 46 – 59, measured after the curve was changed, are primarily considered in the phosphate intercomparison. This change in the calibration curve and use in the inter-comparison is noted throughout the text.

Following the cruise, a calibration test was carried out in the MI laboratory, in which two sets of 14 Quasimeme Proficiency test materials with a wide range of nutrient concentrations were analysed, together with three batches of KANSO CRMs. The full suite of calibration standards (Table 2) was analysed during the run, while in the post-processing, results were calculated after selecting different standards and calibration coefficients (either first or second order calibration). This test was repeated a number of times and the results illustrate that the range of calibration standards used can indeed have an appreciable effect on the final reported value, particularly for

lower nutrient concentrations (Table 7). While nitrite and phosphate were also analysed during this experiment, the range used on the A02 cruise did not extend beyond 2.2 µmol/l and adjusting the lower calibration standards had minimal effect on the final reported concentrations. Therefore, only results for TOxN and silicate are discussed in this section.

For silicate, the use of different calibration standard ranges had only a marginal effect on samples with mid- to high-concentrations, for which almost all Z-scores were $|Z| < 1$ (all <4% bias). The samples that illustrated a significant difference were those with concentrations < 2 µmol/l, where $|Z|$ scores increased to 2 if the higher concentration calibration standards were included. For example, in the QNU 300 sample (Table 7), the measured value had a difference of 7% from the assigned value when using standards ≤10 µmol/l, whereas the difference increased to 21% with use of standards up to 60 µmol/l.

There was greater variation in the TOxN results depending on which standards were used, but again it is clear that inclusion of the highest concentration standards (≤50 µmol/l) results in larger bias in the accuracy of low concentration TOxN samples. With the QNU 307 sample, the measured value was exactly the same as the assigned value (0% difference) when standards ≤10 µmol/l were used, while the difference increased to ±19% if standards up to 50 µmol/l were included.

Based on this experiment's finding that the lowest TOxN and silicate concentrations showed reduced bias when calculated with a smaller range of calibration standards, the MI GO-SHIP A02 data with TOxN and silicate concentrations ≤ 5 µmol/l were recalculated using standards of ≤10 µmol/l (Table 2). The TOxN CD values (5.65 µmol/l) were also plotted using the calibration range of 0 – 10 µmol/l to illustrate the accuracy of this method (Supplementary Material; Fig. 2). This is a key finding in this inter-comparison, which illustrates that it could potentially reduce bias and CV(%) in CRMs and samples across a broad concentration range, to split up a sample run into two (or more) components that are linear, which will be specific to individual instruments and configurations.

3.4 Comparison with earlier WOCE data on the A02 section

Nutrient analysis on the WOCE A02 survey in 1997 was also carried out using a Skalar Continuous Flow Auto-Analyser (SA 4000) for photometric determination of nitrate, nitrite, phosphate and silicate. Analytical methods were similar to the MI and Dal systems, with nutrients measured at the same wavelengths, while calibrated flasks and pipettes were also used for the daily calibration standards. There were no CRMs available for the 1997 cruise, instead the internal consistency of the nutrient measurements between cruises were assessed by comparison of quality controlled dissolved inorganic carbon (DIC) data, where any inaccuracies in the nutrient measurements would show up as offsets or slope changes in the DIC-nutrient plots derived from various cruises. The "estimated accuracy on the WOCE survey, was 0.02 µmol/kl for nitrite, 0.1 µmol/l for nitrate, 0.05 µmol/l for phosphate and 0.5 µmol/l for silicate" https://cchdo.ucsd.edu/cruise/06MT39_3. There was no information provided in the cruise report, and no articles published (that we know of) which states the calibration ranges used on this survey. The vertical profiles of nutrient data compared quite well with the 2017 data (Fig. 2 and Supplementary Material; Fig. 1). Not every

station on the 2017 survey could be compared directly with the 1997 survey due to small differences in some station positions, which sometimes resulted in with bottom depth differences of over 500m between the two surveys.

4. Discussion

The comparison of the MI and Dal datasets from the A02 survey highlights the importance and effectiveness of following standard protocols. Both groups followed the GO-SHIP manual (Hydes et al., 2010) for the sampling and determination of nutrients in seawater, while also incorporating their existing laboratory QC methods that were specifically adapted to their instruments.

4.1 MI vs Dal Station-by-Station Comparison

Figure 5 presents differences between samples that were measured on both the MI and Dal systems on a station-by-station basis. Summary statistics for the station-by-station comparisons are shown in Table 5. Because most of the stations plotted and listed were measured on different autoanalyzer runs, these plots and statistics also give an indication of run-to-run differences in the level of agreement between the systems. $RPD_{MI\text{-}Dal}$ values are shown for three subsets: all data (upper panels); samples from >400m only (middle panels) and samples from <400m only (lower panels).

The plots show the larger RPDs and greater number of outliers for comparisons made on shallower (<400m) samples with deeper concentrations, which is also evident from the depth profiles (Fig. 3). Figure 5 and Table 5 also show good overall agreement between MI and Dal measurements of TOxN and silicate as determined on a cruise-wide basis (average bias of ca. 1-2%; see section 3.1), However, the difference is variable from station to station, with individual stations having average differences as large as 3-4%; this is likely due to run-to-run variations in measurement calibration on both systems. For phosphate, there was a clear improvement in the variability and magnitude the between-system agreement later in the cruise.

Figure 5 and Table 5 show that, on a cruise-wide basis, average differences (MI-Dal) determined on the water samples and CRMs are similar. The respective differences of MI-Dal results for water samples and CRMs are: -1.4% and -2.2% (TOxN); -1.1% and +1.3% (silicate) and +2.3 and +3.6% (phosphate). Figure 5 also shows that the station-by-station means of differences measured on the water samples generally fall within ±1 standard deviation of the cruise-wide average RPD that was determined from analyses of CRMs.

We regressed the station-to-station differences of sample analyses with the corresponding differences of CRM analyses but found no significant correlation. This implies that for this data set at least, we cannot use run-by-run analyses of CRMs to correct sample data from individual stations. This is likely due to the limited number of CRMs that were analysed per station/ run relative to the within-run precision.

Overall, the results suggest that average levels of agreement between independent nutrient data sets should be interpreted with caution. Clearly, comparisons of data collected in deepwater with high concentrations risk not being applicable directly to samples from shallower depths with lower concentration ranges where percentage errors are generally larger. Perhaps more significantly, our results also show that station-to-station variations in data quality and bias can be considerably larger (by several percent) than the mean bias between two cruise-wide data sets. These station-to-station variations in bias arise from short-term differences in instrument calibration that are difficult to identify without very detailed monitoring of system performance. This observation is relevant to “secondary quality control” (Tanhua et al., 2009a,b) of nutrient

data in which adjustments to entire cruise data sets might potentially be recommended on the basis of offsets between deepwater measurements made on different cruises at a limited number of crossover or co-located stations. "Drifting or variable measurement precision and accuracy during a cruise" (Tanhua et al., 2009b) is a recognised potential pitfall of this approach and the A02 Survey provides a rare example of a "crossover cruise" from which its impact on between-cruise data comparisons can be estimated.

4.2 At sea versus on shore measurement: potential sources of error.

A key observation from this study was demonstration of the potential for reduced precision and increased bias of CRM results analysed at sea, relative to those analysed onshore. This was evident for TOxN, silicate and nitrite analyses on the MI system – with almost a doubling in the CV(%) of CRMs analysed on the A02 survey, while phosphate QC was similar for land- and sea-based analyses. This implies that shore-based intercomparisons and QC tests, where samples are measured under stable conditions and where there may be a tendency to analyse test samples when instruments are working "normally", do not necessarily reflect the quality of data collected at-sea under more difficult conditions and, often, when analysts are under time pressure. It is likely not possible to pinpoint exact cause(s) for the increased scatter in MI's silicate and TOxN CRM results relative to shore-based analyses or for the negative bias in the TOxN results from both systems that was observed during A02. However a number of potential sources of error associated with at-sea analysis can be speculated on;

- Ship vibrations: These were particularly evident in the MI container during A02. Unlike the other container labs, which were lined along the middle of the aft deck, the MI container was located along the starboard aft deck, in contact with the ship's hull and appeared to suffer greater vibration at higher speeds and during dynamic positioning of the ship (when the thrusters were in action) than noticed in other containers. The vibrations even caused the instrument to crash a number of times when the auto-sampler syringe could not address the cup correctly. These vibrations had not been encountered on previous surveys on which the onboard laboratory was deployed and analysis undertaken. Vibration could potentially disrupt the light path of the instrument photometers, which could ultimately affect the measured nutrient concentrations.
  During a transit westward across the Atlantic immediately prior to the A02 survey, during which sea-state was calmer and dynamic positioning was not used, two trial runs on the MI system showed little bias and better precision (CV(%) in CD - TOxN <2.8, phosphate <2.2, silicate <2, n=6; CV(%) in BW – TOxN 0.6, phosphate <1.2, silicate <1.5n=5; see QC charts in Supplementary Material). The trial runs on the westward leg used the same reagents, stock solutions, pipettes, glassware, as used on the survey proper. A vibration-related error affecting the MI system more than the Dal system, could lead to variable differences between the measurements made on the two systems during the cruise.
- Water purification unit: Although the ultrapure water from the RV Celtic Explorer was tested ahead of the A02 survey to ensure no nutrient contamination, problems arose for both groups during the survey with their nitrite channels, and this appeared to be due to varying levels of nitrite in different batches of ultrapure water. This was sometimes seen as a shift in the nitrite baseline when a new batch of ultrapure water was used. If, in fact, there were nitrite in the ultrapure water used to make reagents, standards and baseline wash, then it would contribute to the negative bias observed in the TOxN measurements with both systems, as it would raise the baseline due to higher levels of nitrite present. It was noted that on the westward leg, there were no such issues with the nitrite analysis on the

MI system. Anecdotal reports of problems with pure-water supplies on research vessels are common. Such a contamination issue on a shared water supply might lead to bias with TOxN measurements on both systems, as observed.

- Standard preparation: A key difference between shore-based and at-sea analysis by the MI group was the use of pipettes rather than balances for preparation of daily calibration standards. However, all pipettes used by the MI on the A02 survey were calibrated ahead of the survey and should not have influenced the final results. There also did not appear to be any bias in the results between the two analysts using the MI system. The Dal system used the same pipettes to make secondary and work standards on land as were used on the survey. This source of error might be expected to be result in constant (rather than variable) differences between the two systems.
- Reagent preparation: All reagent chemicals were pre-weighed and stored in acid-cleaned containers until use. Tests were carried out at the MI and Dal prior to the A02 survey to ensure there were no issues of contamination in the pre-weighed chemicals. The accuracy and precision measured on the test runs on the westward transit prior to the A02 survey also indicated no contamination in the MI chemicals. The Dal team had extra pre-weighed reagents which they continued to use for up to 9 months after the survey, indicating there were no contamnation issues with storage time of the reagents.
- CRM use: The latest revision of the GO-SHIP guidelines (Becker et al., in prep.) recommends that a new CRM bottle should be opened for every run, or at least every 2 days (Becker et al., in prep.). This protocol was not followed on the GO-SHIP A02 survey and CRMs were generally used until they ran out. Similarly, this was not done during shore-based analysis, and therefore is unlikely to have contributed to the difference between at-sea and shore-based analysis. Changes in CRM concentrations after opening could impact the comparison of CRM results between the two systems, and the CV(%) of the CRM measurements. However there is no reason for this to impact the differences observed between MI and Dal analyses of water samples.

Based on this difference in overall method performance between the lab based and at-sea analysis, the z-score acceptance criteria were re-calculated following the survey reducing the proportional error from 6% to 2% in Eqn. 2, to better quantify the land-based instrument capability. This narrowed the CRM assessment criteria, (see both limits in; Fig. 4), to levels which we feel are more suitable for oceanic nutrient samples. This was also closer to the CV(%) results of international laboratories from the recent JAMSTEC Inter-comparison exercise, which was typically less than 2% for both TOxN and silicate (Aoyama et al., 2018).

4.3 Quality control, including reference materials

The results from this inter-comparison exercise highlight the need for using low, mid and top range reference materials covering the full range of the expected nutrient concentrations for ocean surveys. This is recommended by Hydes et al. (2010), and also in JAMSTEC I/C report (Aoyama et al., 2018). If solely the CD CRM had been used by both groups on the A02 survey, the negative bias in the MI TOxN at high concentrations would not have been apparent. Without confirmation from the higher concentration CRM (Batch BW), it would not have been clear whether there was a negative bias in the MI data or a positive bias in the Dal data, since both were producing similar values for the lower (CD) CRM. Similarly a low concentration CRM would have improved comparison of surface waters where nutrient concentrations were close to the detection limit, and where the largest differences between the two datasets were observed. The low nutrient KANSO CRM available at the time of the survey (BY), similar to the current low

nutrient batch (Batch CE by KANSO or Batch 7601a from NMIJ), have nutrient levels below our limits of quantification and therefore they are not useful as a low concentration CRM for the MI/Dal methods. For future surveys if a low KANSO batch is still not suitable, alternatives could be used to check precision and accuracy at low levels, such as low concentration materials remaining from intercalibration/proficiency testing or in-house materials used to check precision.

With availability of a range of CRMs for nutrients in seawater, there remains a need for clearly-defined data quality objectives for oceanic nutrient measurements to meet GO-SHIP objectives as well as clear criteria for flagging acceptable and questionable data. Such criteria exist for other biogeochemical parameters; for example, for dissolved inorganic carbon (DIC) and total alkalinity (TA) in the open ocean, a level of uncertainty of 2 µmol/kg , (~ 0.1%), is recommended to assess long-term anthropogenic trends in the marine carbonate system (referred to as "climate" level objectives) although for short-term changes and spatial variability less stringent objectives are specified ("Weather") (Newton et al., 2015.). In coastal waters, the level of accuracy required would be less since the range of carbonate parameters observed would be much wider than those in the open ocean. If clear criteria for nutrient measurements were set, laboratories could flag reported data where these were not attained. The metadata supplied with published datasets should include all of the related QC information, including calibration ranges, batches of CRMs used, CRM assessment criteria, accuracy of CRMs achieved, sample storage prior to analysis, etc.

In a 2015 I/C exercise, Aoyama et al. (2016) reported CV(%) of 1% for TOxN, 2% for silicate and 6% for phosphate with the reference material batch BU (which is similar to Batch CD used on the A02 survey), and 2% for all nutrients for batch CA (similar to Batch BW). These CV(%) are lower than those produced by the MI and Dal groups on the A02 survey (Table 6). The CV(%) for the participating laboratories of the 2015 I/C exercise were, however, calculated from measurements carried out in shore-based laboratories, a much more stable and less pressured environment than during a research cruise. Our comparison of QC before and after the A02 survey with performance at sea illustrated an increase in CV(%) during A02 in all parameters for the MI group as well as systematic bias for TOxN with both groups and variable performance of the Dal phosphate analyses during the cruise. These observations highlight the difficulty and nature of problems associated with carrying out ship-based nutrient analysis of open ocean samples. A key question is whether accuracy goals/ targets for sea-going analyses should acknowledge that at-sea analytical performance may not always attain the standards that can be reached in shore-based studies.

Hydes et al. (2010) suggest that use of CRMs along with best practices in using analysis equipment and internal standardisation, should make it "commonly possible to achieve comparability of nutrient analysis to a level better than 1%". The draft revised guidelines for nutrients state that accuracy of 1% should be aimed at in order to be able to quantify decadal trends in the deep ocean. Based on inter-calibration performance during A02 and into international I/C exercises, a target *proportional error* of 2% for analysis of nutrients might instead be reasonable and achievable. The associated narrower z-score limits (Fig.4) calculated with a PE% of 2% could be considered as oceanic nutrient CRM acceptance criteria for future surveys. However, additionally, specification of an appropriate total error combining *proportional* and *constant error* components, as applied by the QUASIMEME system, may be appropriate to allow for a wider allowable total error for concentrations extending closer to the LOQs. We note that the GOOS Essential Ocean Variables specifications list accuracy goals for nutrients in terms of *constant*

errors that are similar to those specified for QUASIMEME.

4.4 Quality of data

The largest differences between the MI and Dal datasets were observed in the low nutrient surface waters, where the $RPD_{MI\text{-}DAL}$ of all nutrients were considerably higher than the rest of the water column. In the 2015 I/C exercise (Aoyama et al., 2016), poorer comparability between the participating laboratories was also observed in the low nutrient reference materials, which yielded CV(%) of up to 60%. This was confirmed in the I/C 2018 exercise (Aoyama et al., 2018), where CV(%) for the low nutrient sample was 50% for TOxN, and 120% for silicate, compared to CV(%) <2 (TOxN) and <2.3 (silicate) for all higher concentration samples. Larger differences in low nutrient waters would be expected since any error in calibration standards, instrument baselines and detection limits would more strongly impact concentrations close to the limit of detection. The larger differences in the low nutrient concentrations could be sensitive to the sample:reagent ratio of each system, where the instruments have different capabilities of measuring low nutrient concentrations. Also the low nutrient surface samples (concentrations <5 µmol/l for TOxN and silicate) were measured with a restricted calibration curve (0-10 µmol/l) on the MI system whereas the Dal group used their full calibration range (0-50 µmol/l) for their entire dataset. The calibration tests carried out in the MI laboratory following the survey illustrate how low concentration measurements can be significantly affected by the higher concentration standards. This will vary between instruments depending on the linearity of the calibration curves over different ranges. The JAMSTEC I/C 2018 report indicates that non-linearity of calibration curves is a significant source of reduced comparability of nutrient data, and recommends the use of CRMs of concentrations covering the whole range of measurements (Aoyama et al., 2018).

Accurate, intercomparable measurement of nutrient concentrations in the upper ocean, with lower concentrations, is important for a range of applications. Inaccurate measurement of nutrient concentrations in the euphotic zone would lead to large discrepancies in primary production estimation, or estimation of near-surface N:P ratios and indices of nutrient limitation. Hence our interpretation of ocean function can be directly related to the quality of the measurements. In the entire GO-SHIP A02 survey, 32% of all samples are from the upper 400m of the water column. Clearly, achieving high accuracy measurements across the large concentration ranges encountered from surface to deep waters remains an analytical challenge. It is generally not possible to compare upper water column nutrient data quality using cross-over analyses between different cruises from the same geographic area due to the greater "real" variability on short spatial and temporal scales (Tanhua et al., 2009a; Tanhua et al., 2009b). This inter-comparison study therefore identifies a key issue in the comparability of nutrient data in lower nutrient upper ocean waters and suggests the need for in-house testing on the impact of higher standards on low nutrient samples. It may, for example, be useful to split calibration curve into low and high ranges, as was done on the MI system during the A02 survey.

In an inter-comparison study carried out in 2005 and 2006 (Sahlsten and Håkansson, 2006), five different laboratories from monitoring institutes of Denmark, Norway and Sweden, compared nutrient concentrations from identical sets of natural seawater sub-samples (as opposed to prepared reference materials) that were analysed ashore in individual laboratories. Results for the deep water samples indicated precision generally better than than 5% CV(%) between laboratories. The study indicated that variations between laboratories could be explained by improper storage of the nutrient samples between sampling and analysis. Tanhua et al. (2009b)

and Tanhua et al. (2009a) carried out crossover analyses as a secondary QC on nutrient data from the Atlantic (CARINA), where an offset and standard deviation were calculated for nutrients at depths >1500m. They found nitrate data showed the largest consistency with RMSE of 2.9%, with a RMSE of 4.2% for phosphate and 7% for silicate, and suggested the larger differences in the reported data were likely due to analytical difficulties.

The results of this inter-comparison strongly support the recommendation of Hydes et al. (2010) that individual laboratories or groups must carry out extensive internal testing on their own instruments to understand the full capability of their instruments and ensure their laboratory methods achieve the highest level of accuracy for the samples being measured. Ahead of bringing a laboratory based instrument to sea, scientists must take account of the different requirements of analysis at sea and be aware that if analytical problems arise, analysts may have limited time and resources to troubleshoot compared to a shore based laboratory; a constant throughput of samples requiring analysis leaves little time for investigative work in the event of problems.

Despite carrying out extensive testing ahead of the survey (including testing the ships' ultrapure water and batches of pre-weighed reagents), along with a contingency plan for almost all foreseeable problems that may arise at sea (including a back-up of all equipment used during analysis, and a second Skalar system), there were unresolved changes in the QC of the ship-based analysis, illustrating the challenges that can occur during analysis at sea. Results also highlighted the value of carrying out a between-laboratory testing exercise, which in this case, helped both groups to identify quality assurance issues in their internal procedures which would otherwise not have been evident. All laboratory groups should ensure they incorporate additional QC into their methods, including extra calibration standards, extra reference materials and internal standards, to allow for post-correction of data if some unforeseen changes to their instrument occurs while at sea.

5. Data Availability

The GO-SHIP A02 nutrient dataset (analysed on the Marine Institute Skalar nutrient analyser) is currently available at the National Oceanographic Data Centre of Ireland;
http://data.marine.ie/publication/dataset/ce49bc4c-91cc-41b9-a07f-d4e36b18b26f.html.

http://dx.doi.org/10.20393/CE49BC4C-91CC-41B9-A07F-D4E36B18B26F

The Dalhousie Nutrient dataset is also available at the National Oceanographic Data Centre of Ireland;

http://data.marine.ie/geonetwork/srv/eng/catalog.search#/metadata/ie.marine.data:dataset.2932

http://dx.doi.org/10.20393/EAD02A1F-AAB3-4F4E-AD60-6289B9585531

6. Conclusions and Recommendations

For data to be of use to the scientific community, oceanographic data collected by different groups at different times must be comparable in order that true changes in the marine environment can be quantified. The presence of biases or imprecision in the measurement of nutrients in seawater reduces our ability to understand spatial and temporal trends in nutrient concentrations in the ocean. The comparison of two nutrient datasets from the 2017 A02 survey illustrated how analysis at sea can change the method performance relative to the analytical ability of a system and expectations of data accuracy and precision in shore-based laboratories. This study illustrates the importance of including extra QC checks (e.g. higher number of calibration and internal standards) should post-processing of the data be necessary. The cross-comparison of

laboratory methods, quality control and instrument configurations allowed the MI and Dal groups to scrutinize their laboratory procedures in order to identify reasons for analytical bias while carrying out nutrient analysis at sea. The GO-SHIP hydro-manual provides essential guidelines to analytical teams undertaking onboard nutrient analysis. Following this study, some additional suggestions/recommendations were identified which could enhance those in the GO-SHIP manual (Hydes et al., 2010) for improved quality of global nutrient datasets;

- Agreed and clearly-defined data quality objectives and acceptance criteria for flagging ocean observation nutrient measurement would aid in improving data quality and support flagging of reported data that doesn't meet these criteria. Such criteria could include proportional and constant error components.
- Additional information could be provided to indicate how CRMs can be used to correct data from a cruise if a bias is observed. This should factor in station-to-station variability, which was found to be several percent larger than cruise-wide average bias.
- If low nutrient CRMs are below limits of detection, an alternative low nutrient reference material should be considered, for example an internal reference solution or past proficiency test material. Extensive testing must be carried out ahead of a survey to understand individual instrument capabilities and additional QC checks should be included to allow for changes to the methods due to unforeseen changes while carrying out analysis at sea.
- Depending on individual auto-analysers, it may be necessary and effective to use two (or more) separate calibration curves to cover different nutrient concentration ranges.
- Metadata should include all information related to QC, including calibration ranges and CRM performance, so to increase comparability and traceability between different nutrient datasets.

Acknowledgements

Financial support for the survey was provided by the Marine Institute, under the Irish Government's Marine Research Programme 2014-2020, with support from the AtlantOS project - funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 633211 and the Canada Excellence Research Chair in Ocean Science and Technology. The survey represents an initial activity of the newly-formed Ocean Frontier Institute of which the Irish Marine Institute and Dalhousie are both partners. We would like to thank the crew and scientists on board the RV Celtic Explorer on the A02 survey, and the various support teams at the Marine Institute.

Competing interests

The authors declare that they have no conflict of interest.

References

Aoyama, M., Abad, M., Anstey, C., Ashraf, P.M., Bakir, A., Becker, S., Bell, S., Berdalet, E., Blum, M., Briggs, R., Caradec, F., Cariou, T., Church, M.J., Coppola, L., Crump, M., Curless, S., Dai, M., Daniel, A., Davis, C., de Santis Braga, E., Solis, M.E., Ekern, L., Faber, D., Fraser, T., Gundersen, K., Jacobsen, S., Knockaert, M., Komada, T., Kralj, M., Kramer, R., Kress, N., Lainela, S., Ledesma, J., Li, X., Lim, J.-H., Lohmann, M., Lønborg, C., Ludwichowski, K.-U., Mahaffey, C., Malien, F., Margiotta, F., McCormack, T., Murillo, I., Naik, H., Nausch, G., Ólafsdóttir, S.R., van Ooijen , J., Paranhos, R., Payne, C., Pierre-Duplessix, O., Prove, G.,

Rabiller, E., Raimbault, P., Reed, L., Rees, C., Rho, T., Roman, R., Woodward, E.M.S., Sun, J., Szymczycha, B., Takatani, S., Taylor, A., Thamer, P., Torres-Valdés, S., Trahanovsky, K., Waldron, H., Walsham, P., Wang, L., Wang, T., White, L., Yoshimura, T. and Zhang, J.-Z., 2016. IOCCP-JAMSTEC 2015 Inter-laboratory Calibration Exercise of a Certified Reference Material for Nutrients in Seawater. International Ocean Carbon Coordination Project (IOCCP) Report Number 1/2016., Japan Agency for Marine-Earth Science and Technology, Tokosuka, Japan

Aoyama, M., Abad, M., Aguilar-Islas, A., Ashraf, P.M., Azetsu-Scott, K., Bakir, A., Becker, S., Benoit-Cattin-Breton, A., Berdalet, E., Björkman, K., Blum, M., de Santis Braga, E., Caradec, F., Cariou, T., Chiozzini, V.G., Collin, K., Coppola, L., Crump, M., Dai, M., Daniel, A., Davis, C., Solis, M. E., Edelvang, K., Faber, D., Fidel, R., Fonnes, L.L., Frank, J., Frew, P., Funkey, C., Gallia, R., Giani, M., Gkritzalis, T., Grage, A., Greenan, B., Gundersen, K., Hashihama, F., Ibar, V.F.C., Jung, J., Kang, S.H., Karl, D., Kasai, H., Kerrigan, L.A., Kiyomoto, Y., Knockaert, M., Kodama, T., Koo, J., Kralj, M., Kramer, R., Kress, N., Lainela, S., Ledesma, J., Lewandowska, J., López, M.C.A., López Garcia, P., Ludwichowski, K., Mahaffey, C., Malien, F., Margiotta, F., Márquez, A., Mawji, E.W., McCormack, T., McGrath, T., Le Merrer, Y., Møgster, J.S., Nagai, N., Naik, H., Normandeau, C., Ogawa, H., Ólafsdottir, S.R., van Ooijen, J., Paranhos, R., Park, M., Parmentier, K., Passarelli, A., Payne, C., Pierre-Duplessix, O., Povazhnyi, V., Quesnel, S., Raimbault, P., Rees, C., Rember, R., Rho, T.K., Ringuette, M., Riquier, E.D., Rodriguez, A., Roman, R.E., Rosero, C., Woodward, E.M.S., Saito, S., Schuller, D., Segal, Y., Silverman, J., Sørensen, D., Stedmon, C.A., Stinchcombe, M., Sun, J., Thamer, P., Urbini, L., Wallace, D., Walsham, P., Wang, L., Waniek, J., Yamamoto, H., Yoshimura, T., and Zhang, J.-Z., 2018. IOCCP-JAMSTEC 2018 Inter-laboratory Calibration Exercise of a Certified Reference Material for Nutrients in Seawater. International Ocean Carbon Coordination Project (IOCCP) Report Number 1/2018., Japan Agency for Marine-Earth Science and Technology, Tokosuka, Japan.

Aoyama, M., Becke, S., Dai, M., Daimon, H., Gordon, L.I., Kasai, H., Kerouel, R., Kress, N., Masten, D., Murata, A., Nagai, N., Ogawa, H., Ota, H., Saito, H., Saito, K., Shimizu, T., Takano, H., Tsuda, A., Yokouchi, K. and Youenou, A., 2007. Recent comparability of Oceanographic Nutrients Data: Results of a 2003 Intercomparison Exercise using Reference Materials. Analytical Science, 23: 1151-1154.

Becker, S., Aoyama, M., Woodward, E.M.S., Bakker, K., Coverly, S., Mahaffey, C., and Tanhua, T., In preparation. GO-SHIP Repeat Hydrography Nutrient Manual 2019: The precise and accurate determination of dissolved inorganic nutrients in seawater; Continuous Flow Analysis methods and laboratory practices. Draft Report.

Broecker, W.S. and Peng, T.H., 1982. Tracers in the Sea, 2543. Lamont-Doherty Geological Observatory, Columbia University.

Cofino, W.P. and Wells, D.E., 1994. Design and Evaluation of the QUASIMEME Inter- Laboratory Performance Studies: A Test Case for Robust Statistics. Marine Pollution Bulletin, 29: 149-158.

Deutsch, C. and Weber, T., 2012. Nutrient Ratios as a Tracer and Driver of Ocean Biogeochemistry. Annual Review of Marine Science, 4(1): 113-141.

Di Lorenzo, E., Schneider, N., Cobb, K.M., Franks, P.J.S., Chhak, K., Miller, A.J., McWilliams, J.C., Bograd, S.J., Arango, H., Curshitser, E., Powell, T.M. and Riviere, P., 2008. North Pacific Gyre Oscillation links ocean climate and ecosystem change. Geophysical Research Letters, 35.

Dickson, A.G., 2010. Guide to best practices for ocean acidifcation research and data reporting, Publications Office of the European Union, Luxembourg.

Hydes, D.J., Aoyama, M., Aminot, A., Bakker, K., Becker, S., Coverly, S., Daniel, A., Dickson, A.G., Grosso, O., Kerouel, R., van Ooijen, J., Sato, K., Tanhua, T., Woodward, E.M.S. and Zhang, J.Z., 2010. Determination of dissolved nutrients (N, P, Si) in seawater with high precision and inter-comparability using das-segmented continuous flow analysers, UNESCO-IOC, Paris, France.

ICES, 2014. Report of the Joint OSPAR/ICES Ocean Acidification Study Group (SGOA),, International Council for the Exploration of the Sea, Copenhagen, Denmark.
Keller, K., Slater, R.D., Bender, M. and Key, R.M., 2002. Possible biological or physical explanations for decadal scale trends in North Pacific nutrient concentrations and oxygen utilization. Deep Sea Research Part II: Topical Studies in Oceanography, 49(1-3): 345-362.
Kim, T.-W., Lee, K., Duce, R. and Liss, P., 2014. Impact of atmospheric nitrogen deposition on phytoplankton productivity in the South China Sea. Geophysical Research Letters, 41: 3156–3162.
Kim, T.-W., Lee, K., Najjar, R.G., Jeong, H.-D. and Jeong, H.J., 2011. Increasing N Abundance in the Northwestern Pacific Ocean Due to Atmospheric Nitrogen Deposition. Science, 334(6055): 505-509.
Moon, J.-Y., Lee, K., Tanhua, T., Kress, N. and Kim, I.-N., 2016. Temporal nutrient dynamics in the Mediterranean Sea in response to anthropogenic inputs. Geophysical Research Letters, 43: 5243–5251.
Newton J.A., Feely R. A., Jewett E. B., Williamson P. & Mathis J., 2015. Global Ocean Acidification Observing Network: Requirements and Governance Plan. Second Edition, GOA-ON, http://goa-on.org/documents/resources/GOA-ON_2nd_edition_final.pdf.
Pahlow, M. and Riebesell, U., 2000. Temporal trends in deep ocean Redfield ratios. Science, 287(5454): 831-833.
Sahlsten, E. and Håkansson, J., 2006. Intercomparison exercise of eutrophication related parameters in sea water March 2005 and January 2006, Swedish Meteorological and Hydrological Institute, Uddevalla, Sweden.
Talley, L.D., Feely, R.A., Sloyan, B.M., Wanninkhof, R., Baringer, M.O., Bullister, J.L., Carlson, C.A., Doney, S.C., Fine, R.A., Firing, E., Gruber, N., Hansell, D.A., Ishii, M., Johnson, G.C., Katsumata, K., Key, R.M., Kramp, M., Langdon, C., Macdonald, A.M., Mathis, J.T., McDonagh, E.L., Mecking, S., Millero, F.J., Mordy, C.W., Nakano, T., Sabine, C.L., Smethie, W.M., Swift, J.H., Tanhua, T., Thurnherr, A.M., Warner, M.J. and Zhang, J.-Z., 2016. Changes in Ocean Heat, Carbon Content, and Ventilation: A Review of the First Decade of GO-SHIP Global Repeat Hydrography. Annual Review of Marine Science, 8(1): 185-215.
Tanhua, T., Brown, P.J. and Key, R.M., 2009a. CARINA: nutrient data in the Atlantic Ocean. Earth Science Systems Data, 1: 7–24.
Tanhua, T., van Heuven, S., Key, R.M., Velo, A., Olsen, A. and Schirnick, C., 2009b. Quality control procedures and methods of the CARINA database. Earth Science Systems Data Discussions, 2(205-240).
Yang, S. and Gruber, N., 2016. The anthropogenic perturbation of the marine nitrogen cycle by atmospheric deposition: Nitrogen cycle feedbacks and the 15N Haber-Bosch effect. Global Biogeochemical Cycles, 30(10): 1418-1440.
Yasunaka, S., Nojiri, Y., Nakaoka, S.-i., Ono, T., Whitney, F.A. and Telszewski, M., 2014. Mapping of sea surface nutrients in the North Pacific: Basin-wide distribution and seasonal to interannual variability. Journal of Geophysical Research: Oceans, 119(11): 7756-7771.
Zhang, J.Z., Wanninkhof, R. and Lee, K., 2001. Enhanced new production observed from the diurnal cycle of nitrate in an oligotrophic anticyclonic eddy. Geophysical Research Letters, 28: 1579–1582.

Table 1. A comparison of sampling, instrument configurations (including sample and reagent tubing sizes) and reagent compositions for each nutrient from the Marine Institute, Ireland (MI) and Dalhousie University, Canada (Dal) systems.

	MI	Dal
Sampling		

Sample tubes	50ml falcon tubes	15 ml falcon tubes
Primary sample analysis	Within 12 hours of sampling	Within 12 hours of sampling
Replicate samples	Frozen immediately to -20°C	Stored at 4°C and analysed within 36 hours if necessary
Analysis		
Auto-sampler size	300 cups	50 cups (can be re-filled during a run)
Auto-sampler cup size	10ml	4ml
Baseline wash	Artificial Seawater	Ultrapure water
Analysis Lab Temperature	20°C	20°C
Reagents (Chemicals g/L or ml/L)		
Artificial Seawater	35g Sodium Chloride	35g Sodium Chloride
	0.5g Sodium hydrogen carbonate	
TOxN		
Sample tubing size	1.02 ml/min	0.16 ml/min
Colour Reagent	150ml Phosphoric Acid	150 ml Phosphoric acid
	10g Sulfanimide	10 g Sulfanilamide
	0.5g N-(1-Naphthyl)ethylene diamine dihydrochloride (NEDD)	0.5 g NEDD
		6 ml Brij solution
Reagent tubing size	0.42 ml/min	0.42 ml/min
Buffer Solution (pH 8.2)	80g Ammonium Chloride	17.5 g Imidazole
	~3ml Ammonia Solution	~25 ml 1M Hydrochloric Acid
	3ml Brij solution (surfactant)	1 ml Brij solution
Reagent tubing size	0.8 ml/min	1.6 ml/min
Cadmium column	Skalar 5358 activated Cd column	Skalar 5347 nitrate reduction coil
Copper Sulfate Solution		12 g copper sulfate
Nitrite		
Sample tubing size	0.42 ml/min	1.20 ml/min
Colour Reagent	150ml Phosphoric Acid	150 ml Phosphoric acid
	10g Sulfanilamide	10 g Sulfanilamide
	0.5g NEDD	0.5 g NEDD
		6 ml Brij solution
Reagent tubing size	0.23 ml/min	0.23 ml/min
Wash Solution	3ml Brij solution	NA
Reagent tubing size	1.00 ml/min	
Silicate		
Sample tubing size	1.40 ml/min	0.42 ml/min
Sulfuric Acid Solution	20ml Sulfuric Acid	5 ml Sulfuric acid
		1 g Lauryl sulfate
Reagent tubing size	0.23 ml/min	0.42 ml/min
Ammonium heptamolybdate	20g Ammonium heptamolybdate	10 g Ammonium heptamolybdate
Reagent tubing size	0.42 ml/min	0.42 ml/min
Oxalic Acid	44g Oxalic Acid	44 g Oxalic acid
Reagent tubing size	0.42 ml/min	0.42 ml/min
L(+) Ascorbic Acid	40g Ascorbic Acid	40 g Ascorbic acid
Reagent tubing size	0.32 ml/min	0.32 ml/min
Phosphate		
Sample tubing	1.40 ml/min	1.60 ml/min

Ammonium heptamolybdate	0.23g Potassium antimony (III)	0.23 g Potassium antimony (III) oxide
	70ml Sulfuric Acid	70 ml Sulfuric acid
	6g Ammonium heptamolybdate	6 g Ammonium heptamolybdate
	2ml FFD6 (Skalar Surfactant)	5 ml FFD6
Reagent tubing size	0.42 ml/min	0.32 ml/min
L(+) Ascorbic Acid	11g Ascorbic Acid	11 g Ascorbic acid
	60ml Acetone	60 ml Acetone
	2ml FFD6	5 ml FFD6
Reagent tubing size	0.42 ml/min	0.32 ml/min

890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911

912 Table 2. Concentrations of daily calibration standards in µmol/l on the MI and Dal systems. Standard 1 is
913 the blank made of artificial seawater (sal 35). Following discussions with the MI group after the first 7 runs,
914 standards 2-4 (indicated with a *) on the Dal system were added to the Dal systems's standard curve for
915 the last 5 days of analysis. SSS are the system suitability standards that were analysed during a run as
916 internal quality standards.

	MI	Dal

STD #	TOxN µmol/l	Silicate µmol/l	PO4 µmol/l	NO2 µmol/l	TOxN µmol/l	Silicate µmol/l	PO4 µmol/l	NO2 µmol/l
1	0	0	0	0	0	0	0	0
2	0.26	0.26	0.05	0.05	1.25 *	1.25 *	0.1 *	0.15 *
3	0.5	0.5	0.15	0.15	2.5 *	2.5 *	0.2 *	0.3 *
4	2.5	2.5	0.25	0.25	5 *	5 *	0.4 *	0.6
5	5	5	0.5	0.5	10	10	0.8	1.2
6	10	10	1	1	20	20	1.6	1.8
7	15	15	1.5	1.5	30	30	2.4	2.4
8	22.5	22.5	2.25	2.25	40	40	3.2	3.0
9	30	30			50	50	4.0	
10	40	40						
11	50	50						
12		60						
SSS	10	10	1	1	40	40	3.2	2.4
Drift	10	10	1	1	40	40	3.2	2.4

Table 3. Certified values in µmol/kg for the two batches of KANSO CRMs used on the survey. These were converted to µmol/l for comparison with Skalar data using a laboratory temperature of 20°C and CRM salinity.

Certified Values KANSO CRMs				
	CD	BW	CD	BW
	µmol/kg		µmol/l	
Nitrate	5.498	24.59	5.63	25.19
Nitrite	0.018	0.067	0.02	0.07
TOxN	5.516	24.66	5.65	25.26
Silicate	13.93	60.01	14.27	61.47
Phosphate	0.446	1.541	0.46	1.58

Table 4. The limit of detection (LOD) and limit of quantification (LOQ) in µmol/l, for both instruments.

	MI				Dal			
	TOxN	**Nitrite**	**Silicate**	**Phosphate**	**TOxN**	**Nitrite**	**Silicate**	**Phosphate**
LOD	0.02	0.01	0.03	0.01	0.14	0.02	0.13	0.04
LOQ	0.26	0.04	0.38	0.16	0.48	0.07	0.43	0.13

Table 5. :Relative percentage difference (RPD_{MI-DAL}) calculated as (MI conc - Dal conc)/average conc * 100% for each station in the inter-comparison. N represents the number of samples, and sd-RPD is the standard deviation. Grey shading represents the stations analyzed before the phosphate standard curve was altered (additional details in Table 2).

	TOxN			Silicate			Phosphate		
Station	**RPD**	**N**	**sd-RPD**	**RPD**	**N**	**sd-RPD**	**RPD**	**N**	**sd-RPD**
14	-0.63	24	4.60	0.29	24	5.00	*10.80*	*23*	*9.20*
22	-4.82	24	3.40	-6.36	24	9.10	*-14.69*	*24*	*10.00*
23	-2.59	24	3.50	2.61	24	2.60	*-9.19*	*24*	*8.80*
29	-0.93	24	3.60	2.38	24	3.80	*1.05*	*24*	*7.00*
33	1.56	22	6.60	-0.44	24	5.27	*2.12*	*23*	*5.95*
37	4.81	24	5.62	-6.48	22	14.86	*0.84*	*24*	*11.47*
42	-3.36	23	2.68	-1.08	22	8.75	*5.53*	*23*	*6.86*
46	-2.73	24	6.67	4.39	22	8.13	7.97	24	3.82
49	-4.25	24	2.31	-6.36	24	7.51	3.01	24	3.58
52	-4.30	24	4.94	-1.23	24	8.70	6.05	24	2.99
56	-0.11	23	3.13	1.64	21	6.04	8.98	23	3.98
59	0.29	24	2.71	-2.77	23	14.42	5.64	24	3.50
All Data	-1.44	284	5.08	-1.14	278	9.14	2.28	284	10.30

Table 6. Mean differences from certified values, and coefficients of variation of the differences (CV(%)) for the KANSO CRMs analysed by the Marine Institute (MI) and Dalhousie University (Dal). The CV(%) were calculated as the (standard deviation/mean*100%). The KANSO batches CD and BW were used by both groups, where n is the number of measurements. Dal results for phosphate do not include analyses prior to Station 46 (see text).

Nutrient	MI			Dal		
	Mean	CV(%)	N	Mean	CV(%)	N
TOxN (CD)	-5.6	3.7	*27*	-2.8	2.6	*27*
Silicate (CD)	0.9	4.6	*27*	-0.4	3.7	*27*
Phosphate (CD)	-2.5	3.8	*27*	-6.1	4	*10*
TOxN (BW)	-3.1	3.3	*16*	-1.0	0.7	*4*
Silicate (BW)	-2.9	4.7	*16*	0.8	3.0	*4*
Phosphate (BW)	0.9	2.8	*16*	-3.2		*1*

Table 7. Results from a laboratory experiment testing the effect of using different calibration ranges, where STD in the first column of the table indicates the top standard included in the calibration. The second column (Order) indicates whether the first or second order calibration coefficient was used in the calibration. The samples are either Quasimeme test materials (QNU) or KANSO CRMs; MV is the measured value; AV is the assigned (or certified value); TE is the total error used for calculating the z-score; Z is the calculated z-score as per Eq. 1 and RPD is the relative % difference (MV-AV/AV*100%). LOD and LOQ are the limit of quantification and detection, respectively.

			TOxN					Silicate				
STD	**Order**	**Sample**	**MV**	**AV**	**TE**	**Z**	**RPD**	**MV**	**AV**	**TE**	**Z**	**RPD**
10	1st	QNU 304 EW	**<LOD**	0.07	0.03			1.97	2.17	0.18	**-1.1**	-9
22	1st	QNU 304 EW	**<LOD**	0.07	0.03			1.97	2.17	0.18	**-1.1**	-9
30	1st	QNU 304 EW	**<LOD**	0.07	0.03			1.94	2.17	0.18	**-1.3**	-11
50	1st	QNU 304 EW	**<LOD**	0.07	0.03			1.96	2.17	0.18	**-1.2**	-10
50	2nd	QNU 304 EW	**<LOQ**	0.07	0.03			1.81	2.17	0.18	**-2.0**	-17
60	1st	QNU 304 EW	Failed Calibration					1.95	2.17	0.18	**-1.2**	-10
60	2nd	QNU 304 EW	0.43	0.07	0.03	**11.6**	552	1.97	2.17	0.18	**-1.1**	-9
10	1st	QNU 307 SW	2.16	2.16	0.16	**0.0**	0	1.91	2.00	0.17	**-0.5**	-4
22	1st	QNU 307 SW	2.15	2.16	0.16	**-0.1**	-1	1.91	2.00	0.17	**-0.5**	-5
30	1st	QNU 307 SW	2.15	2.16	0.16	**-0.1**	-1	1.90	2.00	0.17	**-0.6**	-5
30	2nd	QNU 307 SW	2.15	2.16	0.16	**-0.1**	-1	1.90	2.00	0.17	**-0.6**	-5
50	1st	QNU 307 SW	1.75	2.16	0.16	**-2.6**	-19	1.82	2.00	0.17	**-1.0**	-9
50	2nd	QNU 307 SW	2.18	2.16	0.16	**0.1**	1	1.91	2.00	0.17	**-0.5**	-4
60	1st	QNU 307 SW	Failed Calibration					1.72	2.00	0.17	**-1.6**	-14
60	2nd	QNU 307 SW	2.22	2.16	0.16	**0.4**	3	1.92	2.00	0.17	**-0.4**	-4
10	1st	QNU 300 SW	2.92	2.75	0.19	**0.9**	6	1.46	1.57	0.15	**-0.8**	-7
22	1st	QNU 300 SW	2.91	2.75	0.19	**0.8**	6	1.45	1.57	0.15	**-0.8**	-8
30	1st	QNU 300 SW	2.91	2.75	0.19	**0.8**	6	1.43	1.57	0.15	**-0.9**	-9
50	1st	QNU 300 SW	2.57	2.75	0.19	**-0.9**	-7	1.35	1.57	0.15	**-1.5**	-14
50	2nd	QNU 300 SW	2.87	2.75	0.19	**0.6**	4	1.46	1.57	0.15	**-0.8**	-7
60	1st	QNU 300 SW	Failed Calibration					1.25	1.57	0.15	**-2.2**	-21
60	2nd	QNU 300 SW	2.89	2.75	0.19	**0.7**	5	1.47	1.57	0.15	**-0.7**	-6
10	1st	QNU 299 SW	6.69	6.75	0.43	**-0.2**	-1	5.36	5.36	0.37	**0.0**	0
22	1st	QNU 299 SW	6.66	6.75	0.43	**-0.2**	-1	5.37	5.36	0.37	**0.0**	0
30	1st	QNU 299 SW	6.50	6.75	0.43	**-0.6**	-4	5.34	5.36	0.37	**-0.1**	0
50	1st	QNU 299 SW	6.70	6.75	0.43	**-0.1**	-1	5.31	5.36	0.37	**-0.2**	-1
50	2nd	QNU 299 SW	6.30	6.75	0.43	**-1.1**	-7	5.35	5.36	0.37	**0.0**	0
60	1st	QNU 299 SW	Failed Calibration					5.31	5.36	0.37	**-0.1**	-1

60	2nd	QNU 299 SW	6.08	6.75	0.43	**-1.5**	-10	5.28	5.36	0.37	**-0.2**	-2
10	1st	KANSO CD	5.55	5.50	0.35	**0.2**	1		13.93	0.89		
22	1st	KANSO CD	5.53	5.50	0.35	**0.1**	0	14.30	13.93	0.89	**0.4**	3
30	1st	KANSO CD	5.53	5.50	0.35	**0.1**	1	14.34	13.93	0.89	**0.5**	3
50	1st	KANSO CD	5.39	5.50	0.35	**-0.3**	-2	14.45	13.93	0.89	**0.6**	4
50	2nd	KANSO CD	5.30	5.50	0.35	**-0.6**	-4	14.24	13.93	0.89	**0.3**	2
60	1st	KANSO CD	Failed Calibration					14.51	13.93	0.89	**0.7**	4
60	2nd	KANSO CD	5.24	5.50	0.35	**-0.7**	-5	14.18	13.93	0.89	**0.3**	2
22	1st	KANSO CJ	16.08	16.2	1.00	**-0.1**	-1		38.5	2.360		
30	1st	KANSO CJ	16.22	16.2	1.00	**0.0**	0		38.5	2.360		
50	1st	KANSO CJ	17.16	16.2	1.00	**1.0**	6	39.36	38.5	2.360	**0.4**	2
50	2nd	KANSO CJ	15.59	16.2	1.00	**-0.6**	-4	39.32	38.5	2.360	**0.3**	2
60	1st	KANSO CJ	Failed Calibration					39.62	38.5	2.360	**0.5**	3
60	2nd	KANSO CJ	15.29	16.2	1.00	**-0.9**	-6	39.33	38.5	2.360	**0.4**	2
22	1st	KANSO BW		24.59	1.50				60.01	3.65		
30	1st	KANSO BW	24.56	24.59	1.50	**0.0**	0		60.01	3.65		
50	1st	KANSO BW	26.41	24.59	1.50	**1.2**	7		60.01	3.65		
50	2nd	KANSO BW	24.45	24.59	1.50	**-0.1**	-1	60.30	60.01	3.65	**0.1**	0
60	1st	KANSO BW	Failed Calibration					60.05	60.01	3.65	**0.0**	0
60	2nd	KANSO BW	24.06	24.59	1.50	**-0.4**	-2	60.88	60.01	3.65	**0.2**	1

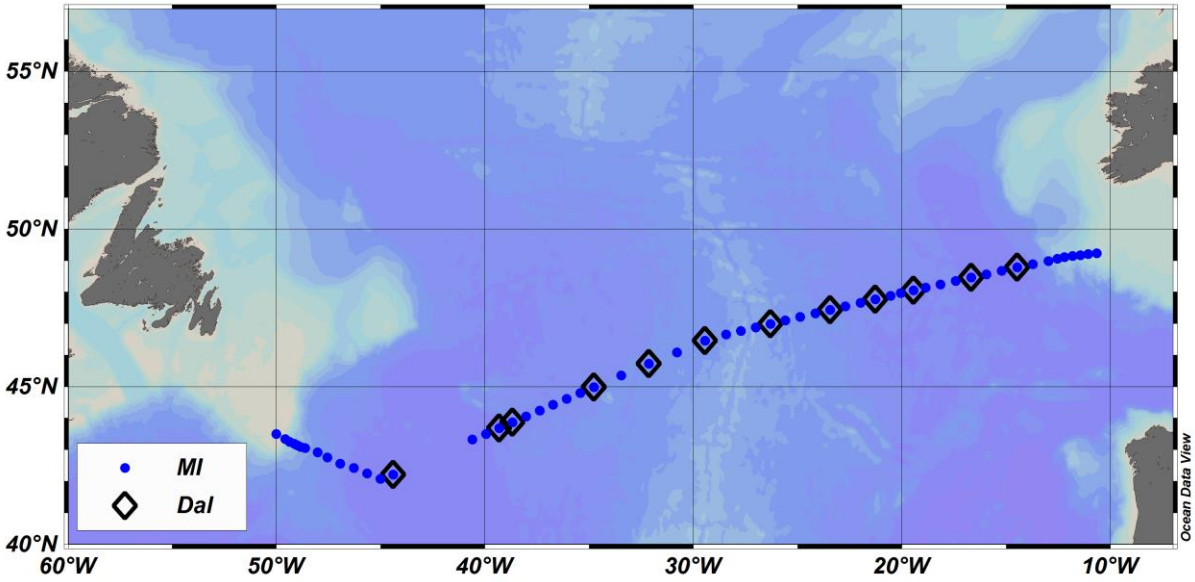


Figure 1. Station positions sampled along the GO-SHIP A02 trans-Atlantic survey completed in May 2017. The Marine Institute (MI) group sampled and analysed nutrient samples at every station along the transect, while the Dalhousie group (Dal) analysed nutrient samples from a selected number of sites, marked with a diamond. Both groups analysed samples over the full water column.

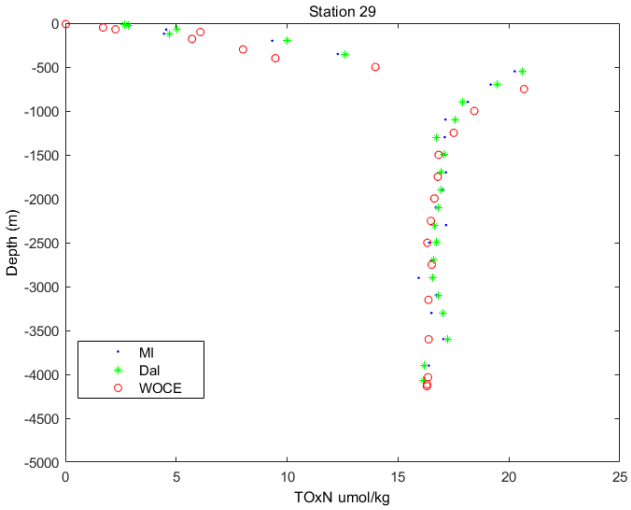


Figure 2a

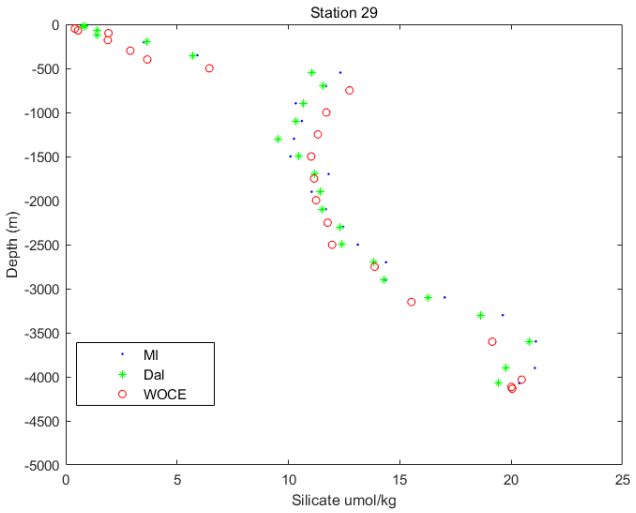


Figure 2b

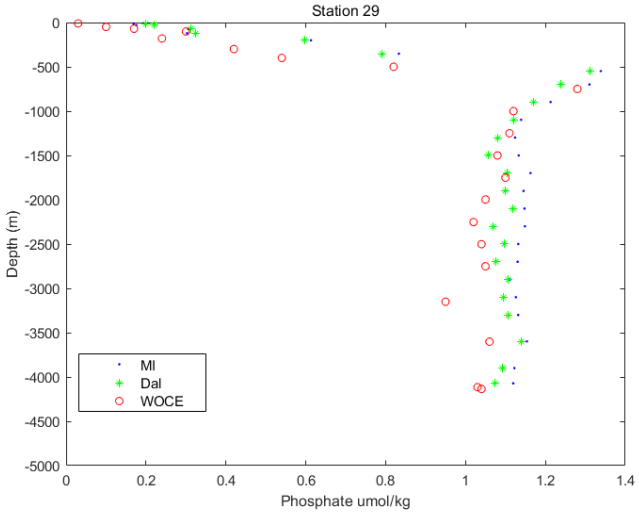


Figure 2c

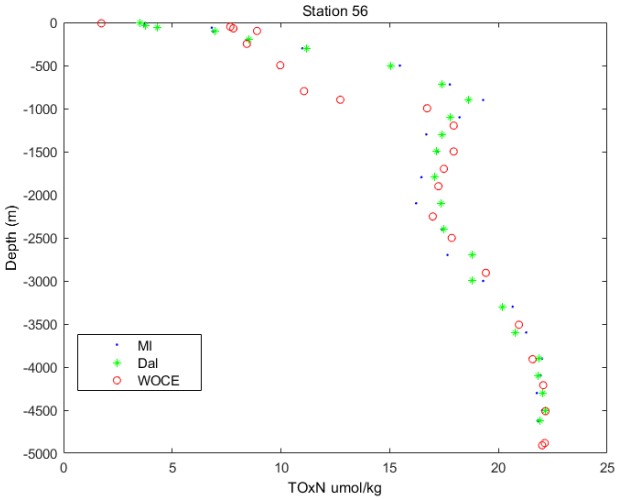


Figure 2d

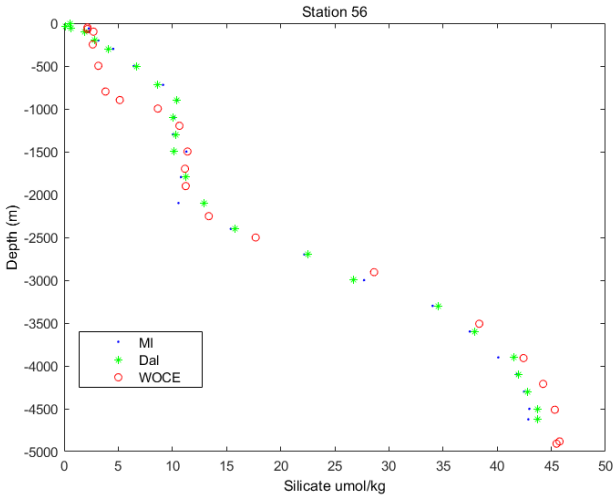


Figure 2e

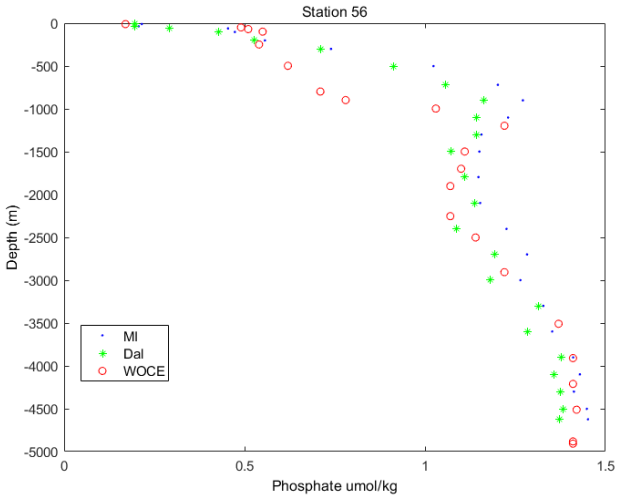
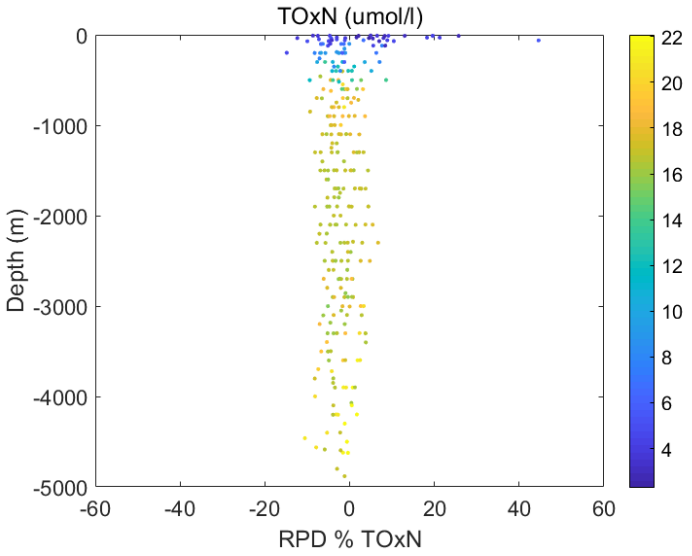


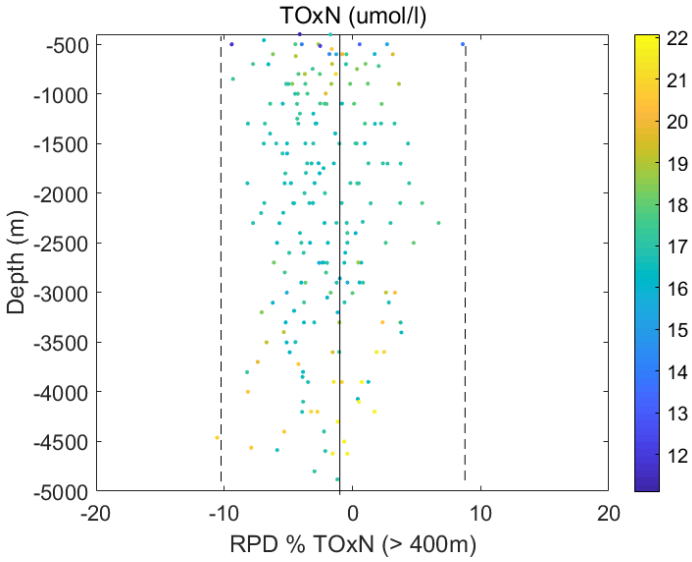
Figure 2f

Figure 2. Vertical profiles of TOxN, silicate and phosphate (in µmol/kg from the MI (Marine Institute), Dal (Dalhousie University) and WOCE (World Ocean Circulation Experiment) datasets. Only station 29 and 56 are included here, all other stations compared are in the Supplementary Material. Profiles are in µmol/kg since WOCE data was reported in µmol/kg rather than µmol/l.



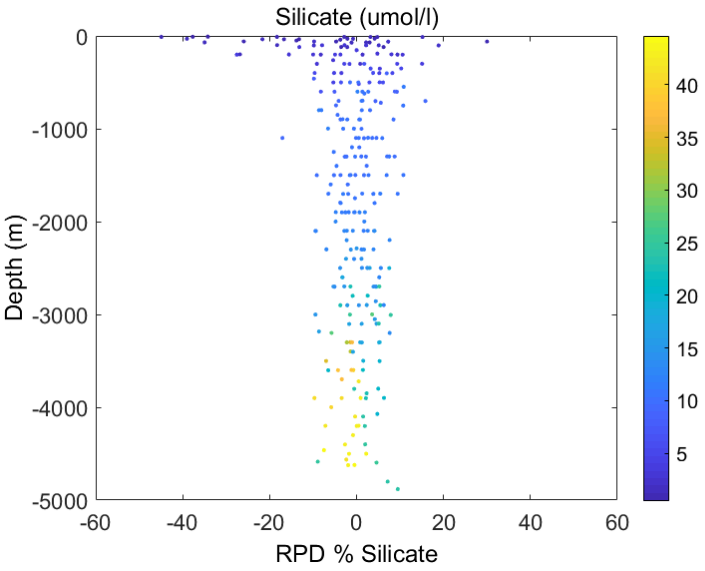
1030

1031 Figure 3a



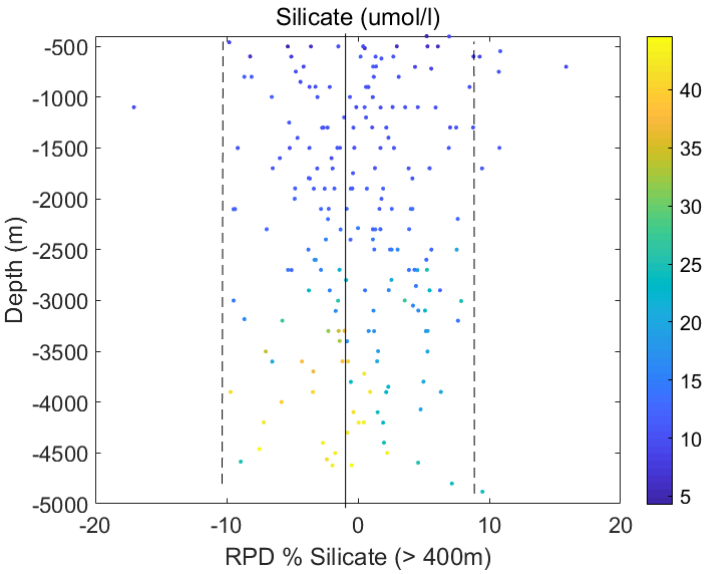
1032

1033 Figure 3b



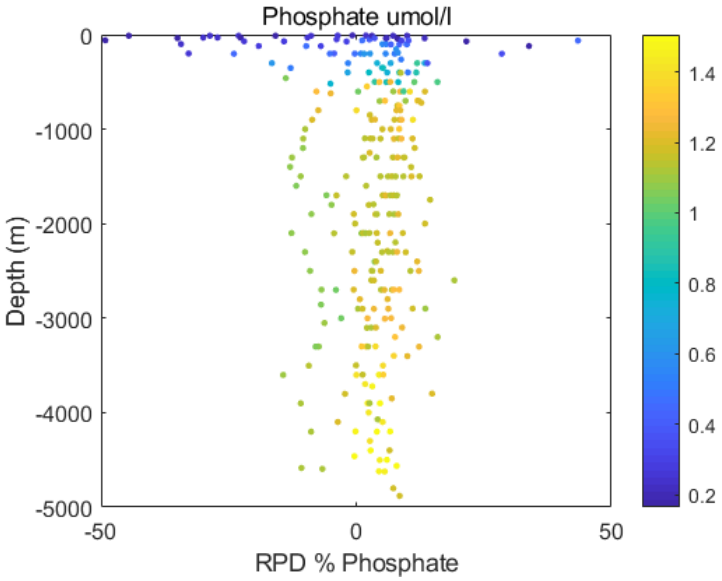
1034

1035 Figure 3c



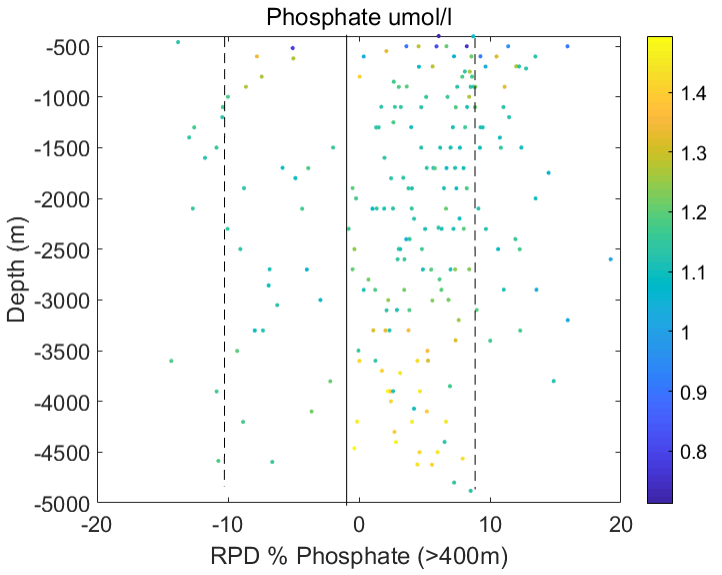
1036

1037 Figure 3d



1038

1039 Figure 3e

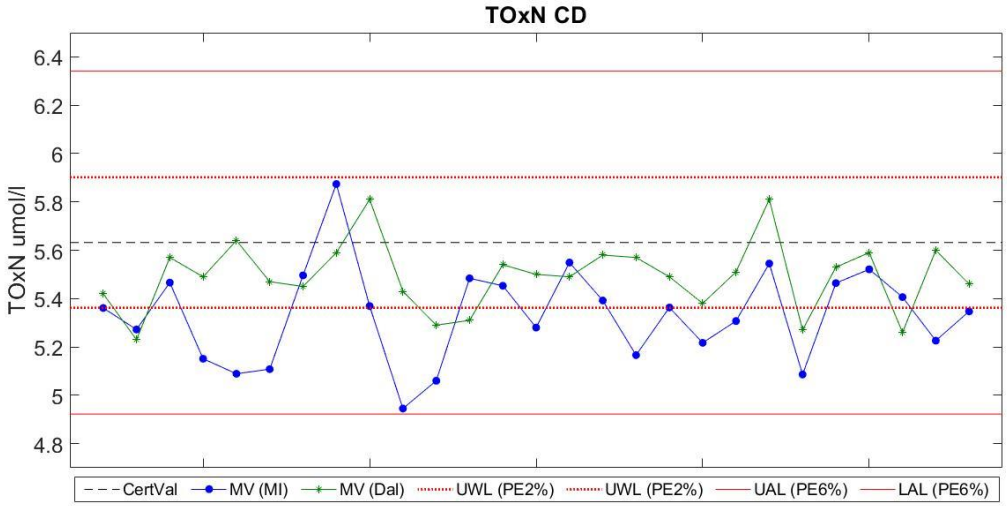


1040

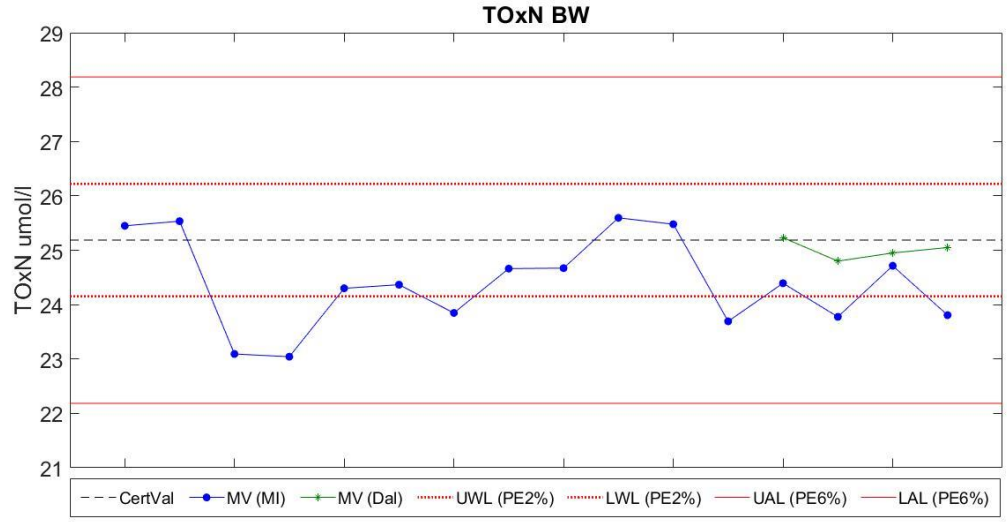
1041 Figure 3f

1042 Figure 3 (a-f): Relative percentage difference ($RPD_{MI\text{-}DAL}$) calculated as (MI conc - Dal conc)/average conc
1043 * 100% for each nutrient for the whole water column and for depths > 400m. The colour bar for each plot
1044 is the average concentration (µmol/l) of each nutrient (i.e. the average concentration from both systems)
1045 at that depth. Note the use of different Y-axis scales for the different subsets.

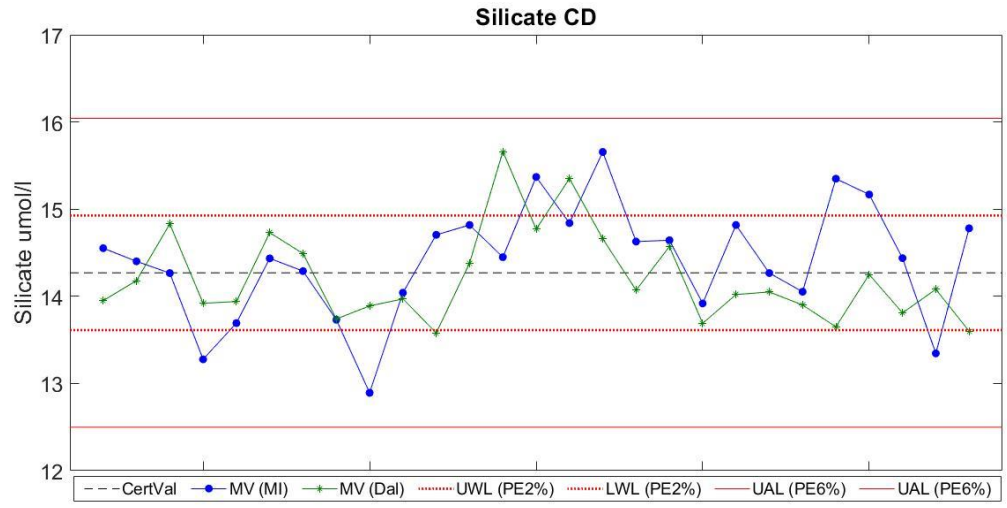
1046
1047



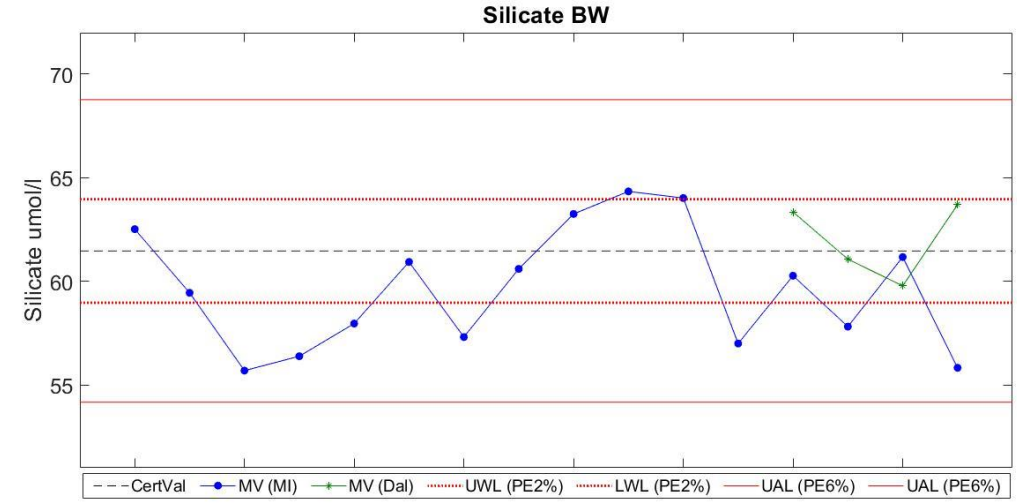
1048



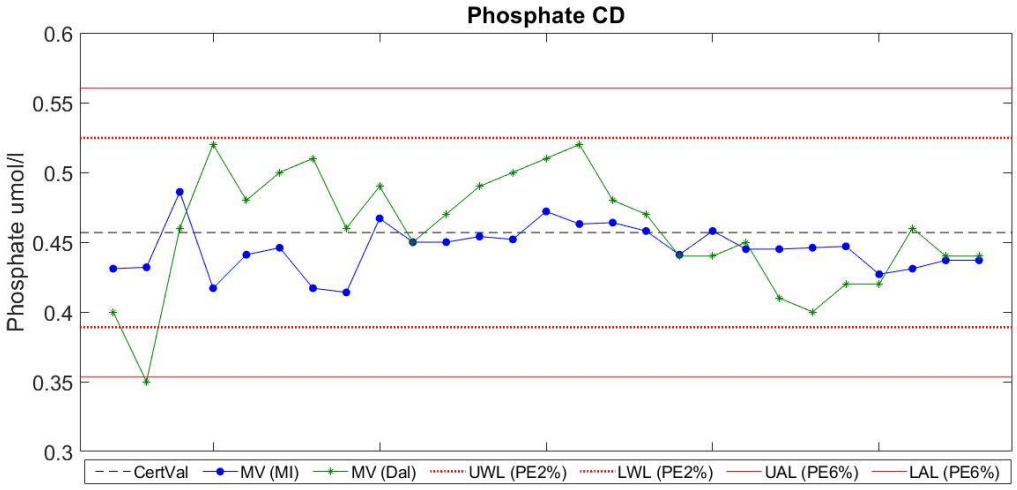
1049



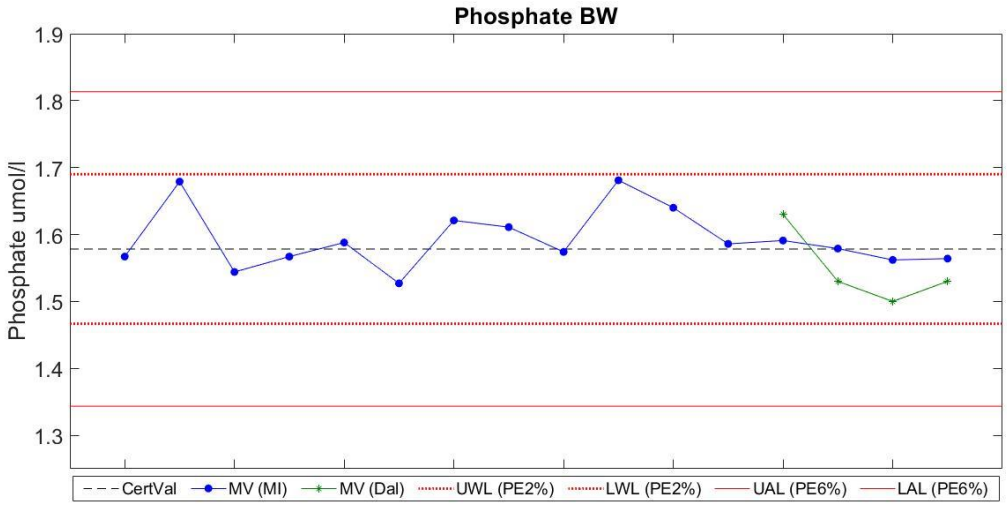
1050



1051



1052



1053

1054

1055 Figures 4 (a-f): Control charts of CRM concentrations from the MI and Dal systems. The dashed centre line
1056 represents the certified value for each CRM (CV), while the red upper (UAL, upper action limit) and lower
1057 (LAL, lower action limit) lines represent the z-score of 2 allowable limits criteria, where the z-scores were
1058 calculated with a proportional error of 6%. MV (MI) and MV (Dal) are the measured values from the MI and

Dal systems, respectively. The dash-dot and dotted lines represent the revised z-score limits with a proportional error of 2%. One CD CRM was run at the beginning and end of every run on both systems, and one BW CRM was analysed at the beginning of every run on the MI system. BW CRMs were run on only a selected number of runs of the Dal system for comparison.

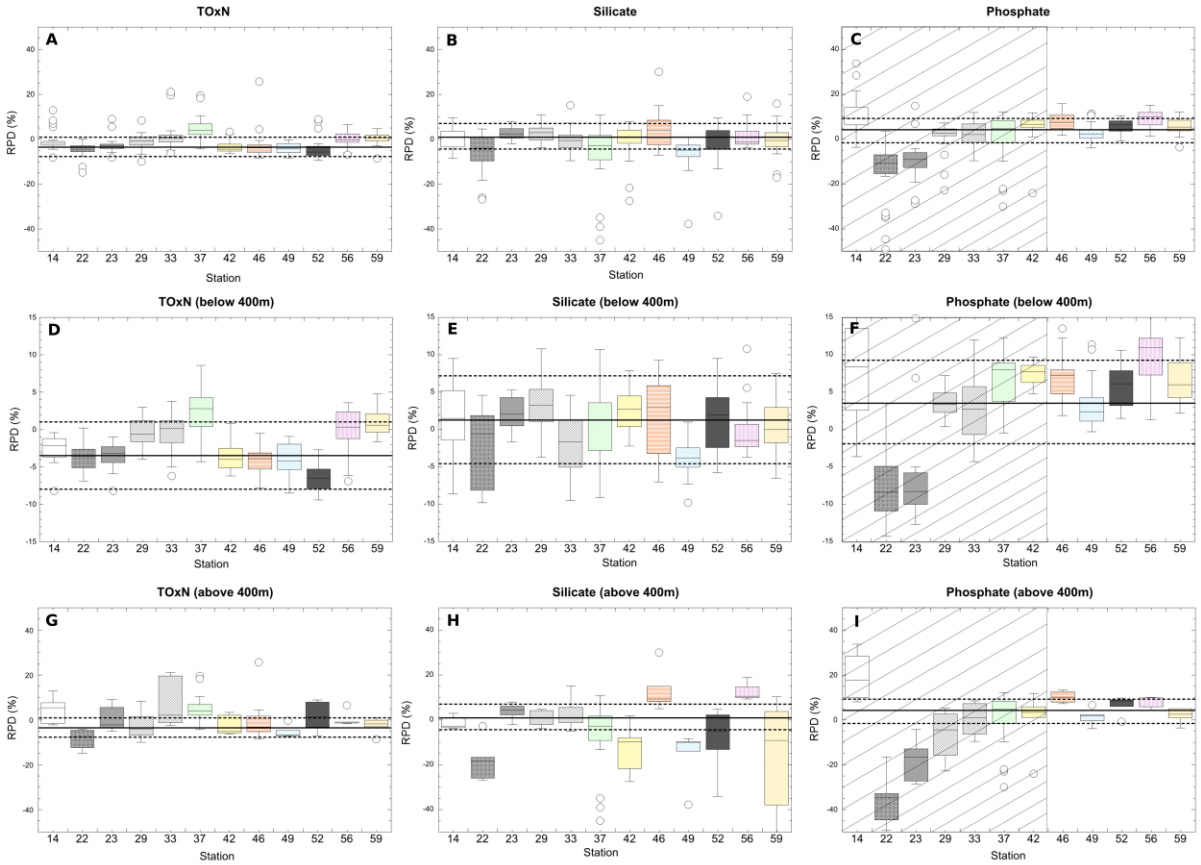


Figure 5. Boxplots of the cruise wide averages of the KANSO CD CRM during the cruise.

Box plots of relative percent differences (MI conc - Dal conc/average conc*100%) between MI and Dal results for stations used in the inter-comparison. The median RPD(%) defines the centre line of the box, and the entire box, representing the interquartile distance (IQD), is closed by the upper (UQ) and lower quartiles (LQ). Any points identify outliers, defined as LQ - 1.5*IQD and UQ + 1.5*IQD. The top row (a-c) represent the full depth profiles of TOxN (a), silicate (b), and phosphate (c), the middle row includes samples below 400 m depth (TOxN (d), Silicate (e), Phosphate (f)), and the bottom row includes samples in the surface waters, 400m to 0m (TOxN (g), Silicate (h), Phosphate (i)).

Note the use of a different y-axis for the "below 400m" plots, compared with "above 400m" and the full profile plots. Dashed lines represent -10, 0, and 10% RPDs on each plot. The solid horizontal line denotes the cruise-wide average % differences of MI CD CRMs - average % difference of Dal CD CRMs for each nutrient. The dashed lines represent ± 1 standard deviation, is calculated as the square root of the sum of the squared standard deviations of the differences from Certified Values measured on both the Dal and MI systems. TOxN and silicate were calculated using all CD CRMs measured during the cruise. For phosphate, the shaded area on the plot denotes the period prior to station 46, after which the Dal standard curve was altered. The lines for the average and standard deviation for phosphate relate only to this later portion of the cruise.