Newfoundland and Labrador Real-Time Water Quality Monitoring Network

Quality Assurance / Quality Control Procedures Assessment

Pilot Project Report

Government of Newfoundland & Labrador
Department of Environment and Conservation
Water Resources Management Division
St. John’s, NL, A1B 4J6 Canada

Table of Contents

Table of Figures

Important Terminology

ADRS – Automated Data Retrieval System: A database management system designed to collect data from a variety of telemetry sources such as GOES, Iridium, and cell network in order to populate a database of all water quality, hydrometric, and climate systems in the province. The system is also responsible for generating graphs for public consumption amongst other duties.

Field Sonde: The multi-parameter water quality monitoring device (typically a Hydrolab DS5X, in the Newfoundland and Labrador monitoring network) deployed on a 30 day schedule.

QAQC Sonde: The multi-parameter water quality monitoring device (typically a Hydrolab DS5X, in the Newfoundland and Labrador monitoring network) cleaned and recalibrated before each field visit, to be used as an unfouled data source for comparison to the Field Sonde.

Introduction

Response to environmental contamination typically occurs long after an event has passed – this is a natural consequence of periodic, scheduled field visits. Maintaining vigilance on any given location can be accomplished only through the use of unmanned water quality stations outfitted with remote telemetry. To this end, many legislative bodies and environmental groups around the world are involved in some form of continuous water quality monitoring. While the equipment used in such programs is well established, protocols for post-processing of data is hardly uniform across all regions of the globe.

Photo 1: Sensor Fouling

Post-processing of continuous water quality data largely revolves around two major factors: fouling and calibration drift. Fouling is a complex process that may involve both biotic and abiotic factors. Biotic fouling can be present in the form of algal growth or colonisation by various aquatic organisms. Abiotic fouling can present itself as silt deposition or, in serious cases, substrate burial. In either case, the process essentially reduces to a film of material covering the working surfaces of sensors making up a multi-parameter sonde. Such films reduce the accuracy of probes by reducing the contact between a probe and the sample water of interest. Calibration drift, on the other hand, can be described as ordinary wear and tear on sensors or dissolution of standards within the multi-parameter sonde; pH glass bulbs undergo abrasion from water-borne particles or dissolved oxygen cathodes, anodes and membranes age, becoming insensitive to oxygen concentrations.

Drift and fouling characteristics differ greatly from one region to the next because of variation in colonial organism assemblages. Additionally, geological differences from region to region may alter the characteristics of silt in terms of 'clinginess' or abrasiveness. To account for such differences, some jurisdictions may insist on more stringent post-processing methodology than others.

In theory, the effects of both processes – fouling and drift – can be corrected for through a host of post-processing methods. In a pilot project undertaken by the NL Department of Environment and Conservation, a set of rigorous post-processing protocols were instituted for a year to determine if the time and effort devoted to data correction resulted in increased data quality.

Methods

QAQC procedures have been an important part of the Real-Time network since its inception in 2001. However, movement towards the quantification of calibration and fouling drift and the application of correction factors to raw data represents a major paradigm shift in field processes.

Old QAQC Process

Historically, the QAQC component consisted of regimented 30 day deployment periods, thorough cleaning, and calibration of sensors in a controlled environment and the use of comparison instruments and grab samples (see Figure 1).

At deployment time, the Field and QAQC sondes were deployed side-by-side and allowed to equilibrate until all readings stabilized. At this point, a record was made of all parameters for both instruments and a grab sample was taken. A judgement of Field sonde functionality could be made on the spot based on the comparison of the *in situ* readings. Later, a further judgement on Field sonde reliability could be made by comparing *in situ* Field sonde readings to grab sample results.

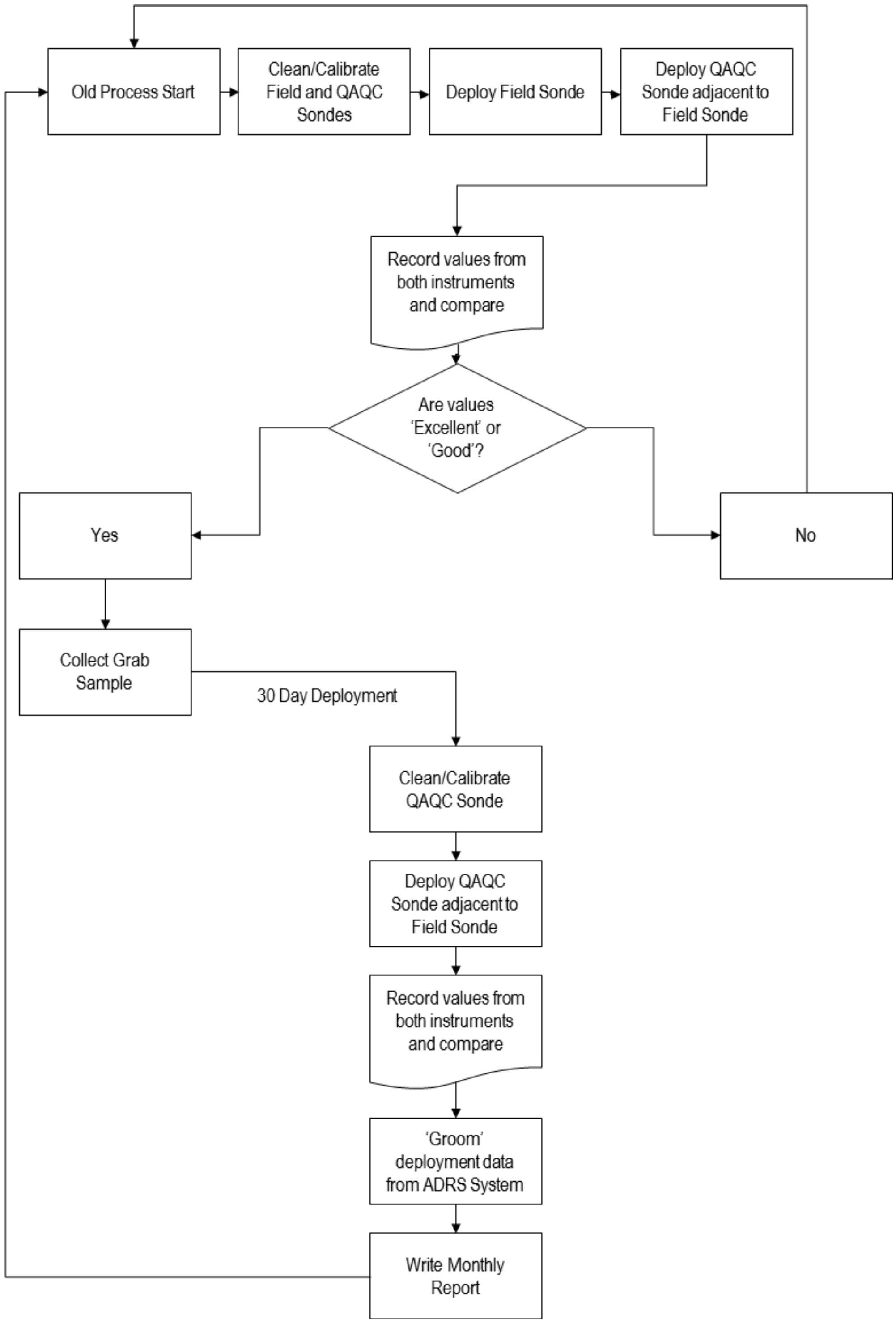
Comparisons between the Field and QAQC sonde were ascribed qualitative statements from “Poor” to “Excellent” for inclusion in monthly reports following the deployment period (see Table 1).

Table 1: Qualitative QAQC Rankings[1]

Parameter	Rank				
	Excellent	Good	Fair	Marginal	Poor
Temperature (^oC)	± 0.2	± 0.2 – 0.5	± 0.5 – 0.8	± 0.8 – 1.0	± 1.0
pH (units)	± 0.2	± 0.2 – 0.5	± 0.5 – 0.8	± 0.8 – 1.0	± 1.0
Specific Conductivity (µS/cm)	± 3	± 3 - 10	± 10 - 15	± 15 - 20	± 20
Dissolved Oxygen (mg/l)	± 0.3	± 0.3 – 0.5	± 0.5 – 0.8	± 0.8 – 1.0	± 1.0
Turbidity (NTU)	± 5	± 5 - 10	± 10 - 15	± 15 - 20	± 20

[1] Table adopted from Wagner, R.W., *et al* (2006).

Figure 1: Historical QAQC Process



Trial QAQC Process

In January 2010, at a Water Quality section meeting in Grand Falls – Windsor, the plan to implement a USGS-style correction process was explained to staff. The new processes are outlined below and in Figure 2.

In this trial QAQC Process, only Temperature, pH, conductivity, and turbidity were corrected. Dissolved oxygen was specifically omitted from the correction process due a confounding factor in the calibration process – the parameter of concern for DO is concentration, however, the probe is calibrated via the 100% saturation method in the lab. It was determined to be a point of

Field Cleaning

Photo 2: Field Cleaning

At the end of a 30 day deployment period, a freshly cleaned and calibrated QAQC multi-parameter sonde (QAQC Sonde) was deployed adjacent to the multi-parameter probe due for servicing (Field Sonde). Simultaneous readings were taken from both instruments and the Field Sonde was briefly removed from the water body. At this point, a thorough cleaning of the sensors was accomplished using a toothbrush, paper towels and cotton swabs to remove all traces of fouling debris. The Field Sonde was placed back in the river and readings were taken once again from both the QAQC and Field Sondes.

In Equation 1, the difference in value for each parameter is recorded from both the QAQC and Field Sondes before and after cleaning. The result of the equation accounts for the biofouling removed from the Field Sonde as well as accounting for variability in the water body occurring during the field cleaning process.

Equation 1

$$E_f = (F_{ac} - F_{bc}) - (Q_{ac} - Q_{bc})$$

Where:
Ef = Field Sonde Error due to fouling
Fbc = Field Sonde reading before cleaning
Fac = Field sonde reading after cleaning
Qbc = QAQC Sonde reading before cleaning
Qac = QAQC Sonde reading after cleaning

Calibration Drift

After removal, the sonde was returned to the laboratory and fully cleansed in a mild detergent. Next, calibration drift for each sensor was assessed by immersing the sonde in isothermic standards. Calibration drift was determined by Equation 2:

Photo 3: Lab Calibration

Equation 2

$$E_d = V_s - V_{FS}$$

Where:
Ed = Field Sonde Error due to drift
Vs = Value of known standard
Vfs = Field Sonde reading in known standard

Error for drift calculated in the equation above simply determines the difference between a known standard and the reading of that standard by the Field Sonde.

Total Error

In this study, fouling and drift were assumed to accrue linearly over the course of the 30 day deployment (Equation 3). For this reason, Total Error (the sum of fouling and calibration drift) was added – through interpolation – onto every reading over the deployment period.

Equation 3

$$V_c = V_{uc} + \left(E_t \times \frac{reading}{totalreadings} \right)$$

Where:
Vc = Corrected value
Vbc = Value before correction
Et = Total Error

Correction was only applied to records that fell within the Data Correction Criterion established by the USGS[2] (see Table 2).

---

[2] Reproduced from United States Geological Survey, Guidelines and Standard Procedures for Continuous Water-Quality Monitors: Station Operation, Record Computation, and Data Reporting, pp. 16 . 2006.

Table 2: Data Correction Criterion

Parameter	Data Correction Criterion
pH	± 0.2 units
Conductivity	± 5 µS/cm or 3%
Turbidity	± 0.5 NTU or 5%

Grab Sampling

Photo 4: Grab Sampling

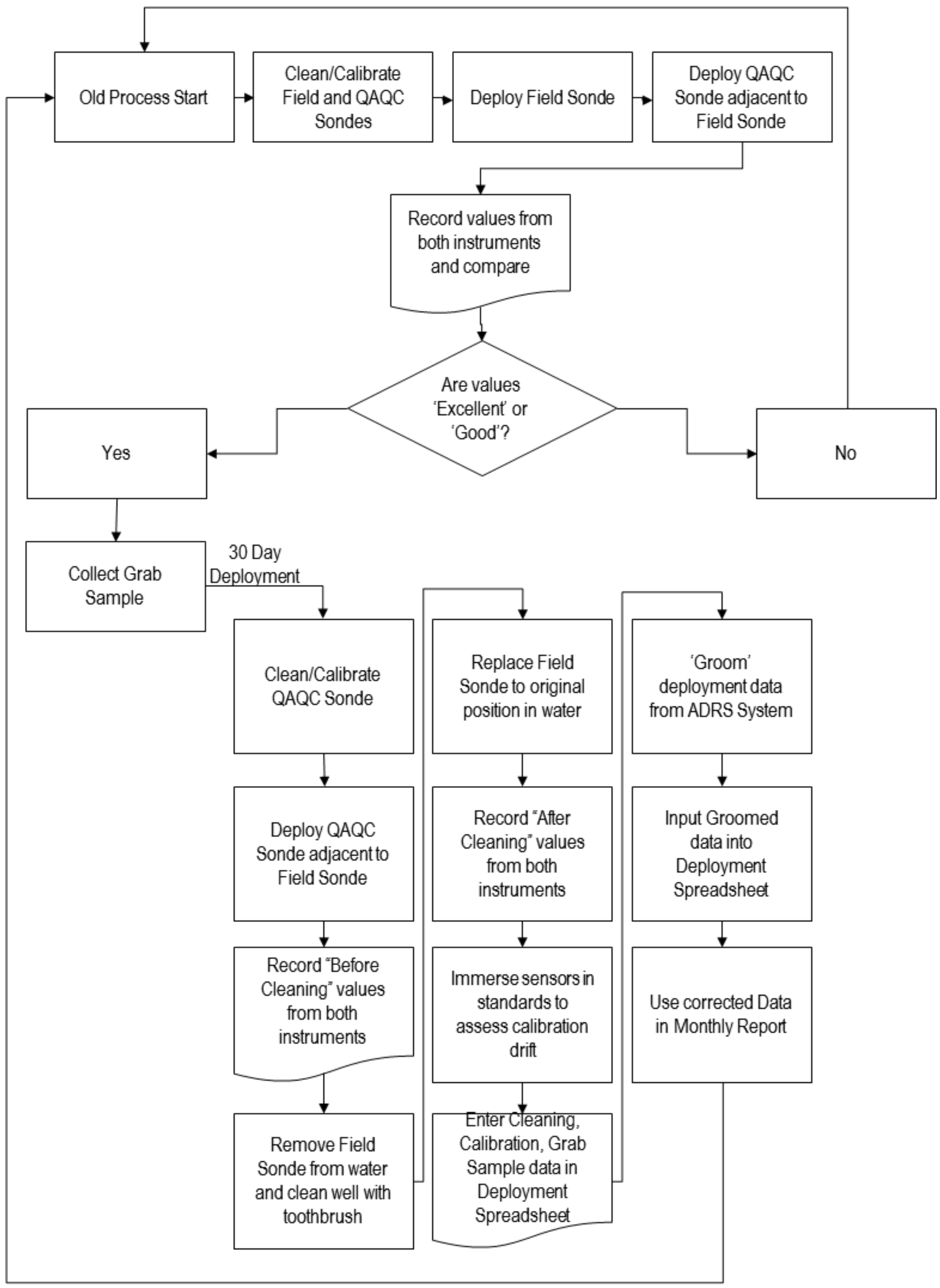
Routine grab sampling has always been an important supplement in the real-time water quality monitoring project. An alternative method to measuring the parameters recorded by *in situ* equipment is vital to ensure that the data is comparable to standardized laboratory practices. A further benefit is that grab samples report a thorough series of ion concentrations and physical parameters providing a useful dataset to generate predictive models based on recordings made by *in situ* equipment.

In the pilot project grab sample values were compared to simultaneous readings from *in situ* equipment. Based on the degree of agreement, a qualitative ranking factor (Excellent, Good, Fair, Marginal and Poor) is applied to assist in writing clear and concise monthly reports. Statistical comparisons are made between Field Sonde, QAQC Sonde, and Grab Sample values.

Automated Deployment Spreadsheet

To streamline the pilot project and ensure data was easy to retrieve for analysis, an automated deployment spreadsheet incorporating an easy to use Field Sheet was devised and put into action. The Field Sheet simplified the process of correcting data and generated the graphs necessary for monthly reports written by staff. Four parameters were corrected throughout the pilot project: temperature, pH, specific conductivity and turbidity.

Figure 2: Trial QAQC Process



Data

The range of data assessed began from October 15th, 2009 up to October 27th, 2010. Each deployment was assessed for presence/absence of correction to the parameters and then for the magnitude if correction was present.

In the first half of the results and discussion section, the magnitude and frequency of corrections are investigated in relation to Total Error. Next, a series of hypothesis tests are used to shed light on the validity of using qualitative comparisons between Field Sonde, QAQC Sonde and Grab Sample values for pH, conductivity, and Turbidity in monthly reports.

A total of 49 complete deployments out of an available 67 were used with some sites contributing as many as 11 deployments.

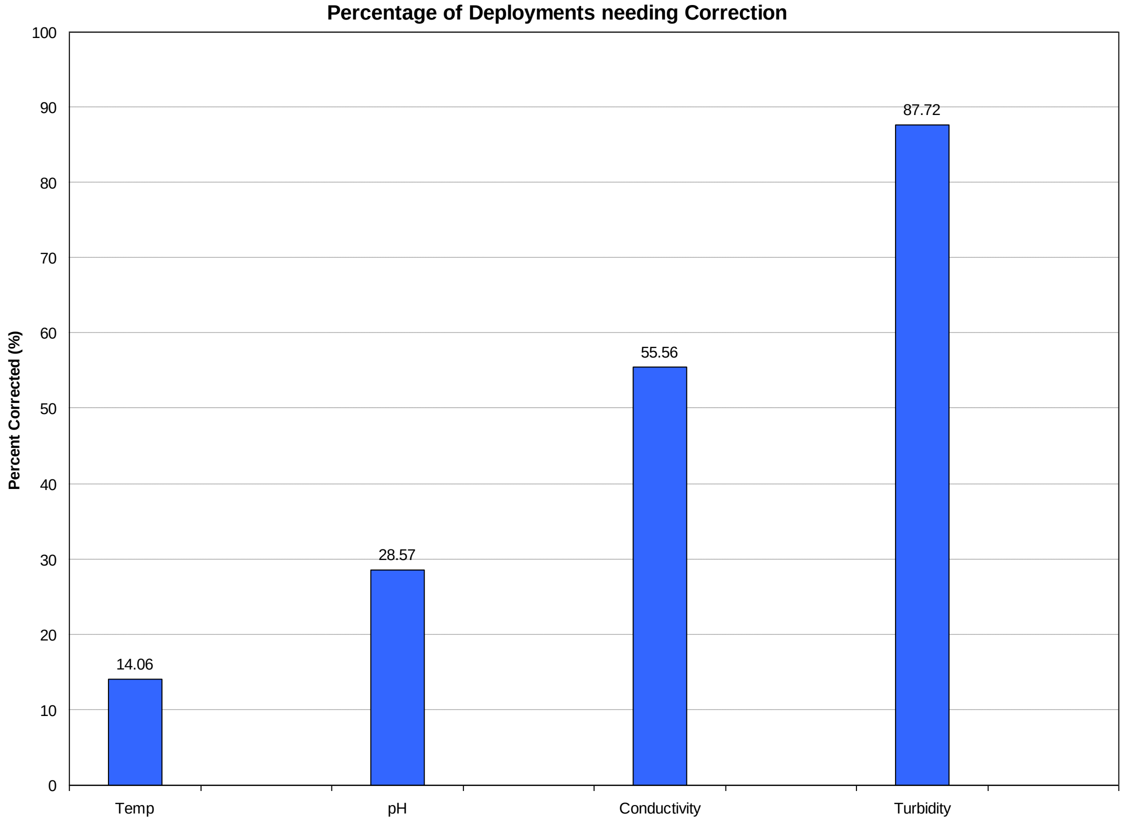
Results and Discussion

Data Corrections

Out of 247 potential corrections, 112 were made (correction 45% of the time). Not all parameters were corrected at the same rate, however. As shown in Figure 3, temperature values were rarely corrected at 14.06%, while turbidity values were frequently corrected at 87.72%.

In the figures below, the contributions of biofouling and calibration drift to total error are outlined using boxplots. Within the same figures, two red lines indicate the data correction criteria for each parameter.

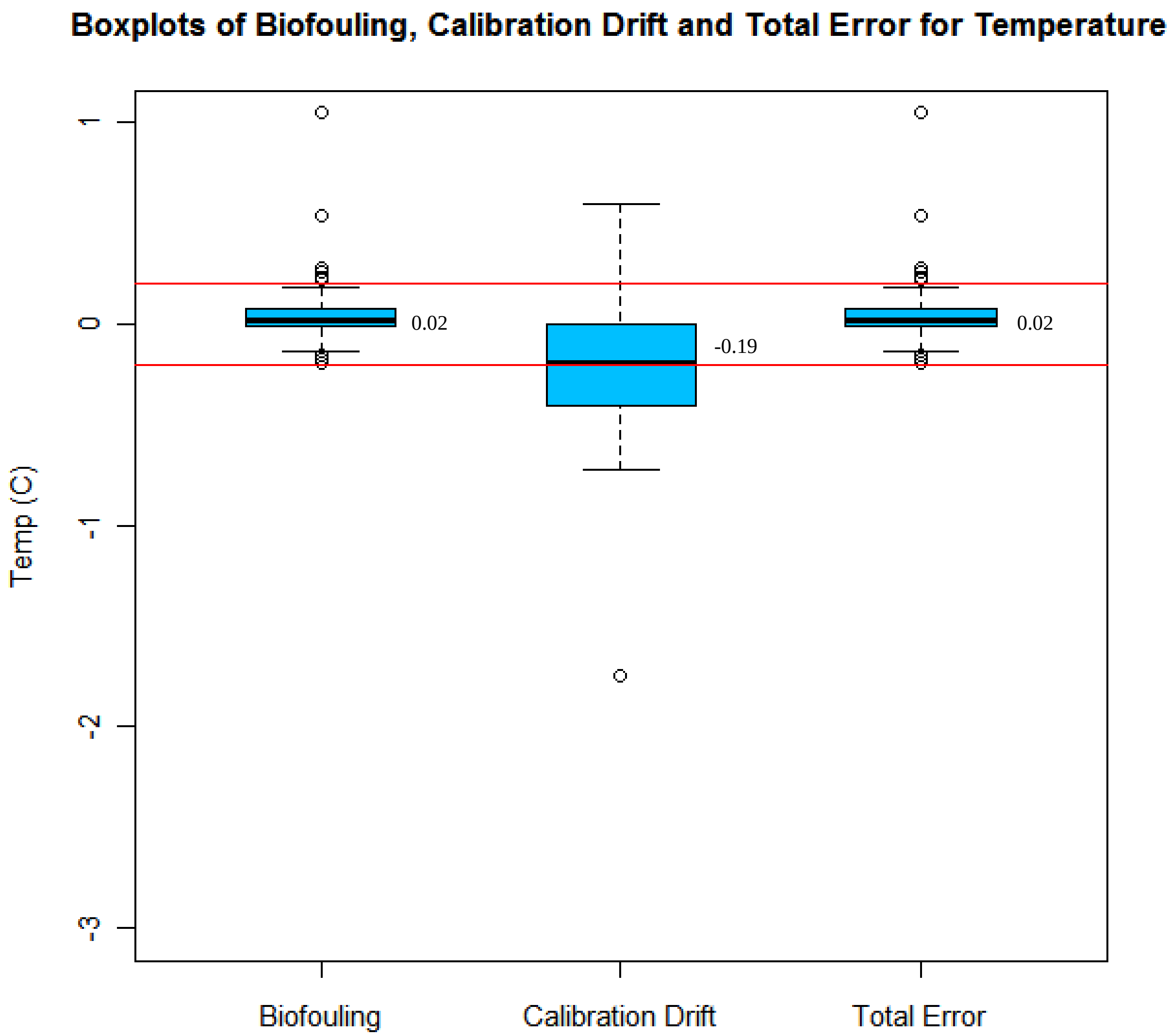
Figure 3: Percentage of Parameters Corrected



Temperature Correction

When correcting for water temperature error, only biofouling is included in the calculation of Total Error. Though calibration drift is sometimes estimated using a relatively low-accuracy thermometer, the thermistor used on Hydrolab sondes is very rugged and resistant to calibration drift due to the permanent factory calibration. Biofouling, "calibration drift", and Total Error values are included in Figure 4. The median biofouling value was found to be 0.02^{o}C, ten times less than the data correction criterion of 0.2^{o}C indicated by the red lines. Only 14% of deployments resulted in a correction being applied to temperature. Because it was felt that the magnitude of temperature correction was small when it did infrequently occur, specific conductivity was not recalculated based on the recalculated temperature.

Figure 4: Contribution of Biofouling and Calibration Drift to Total Error for Temperature

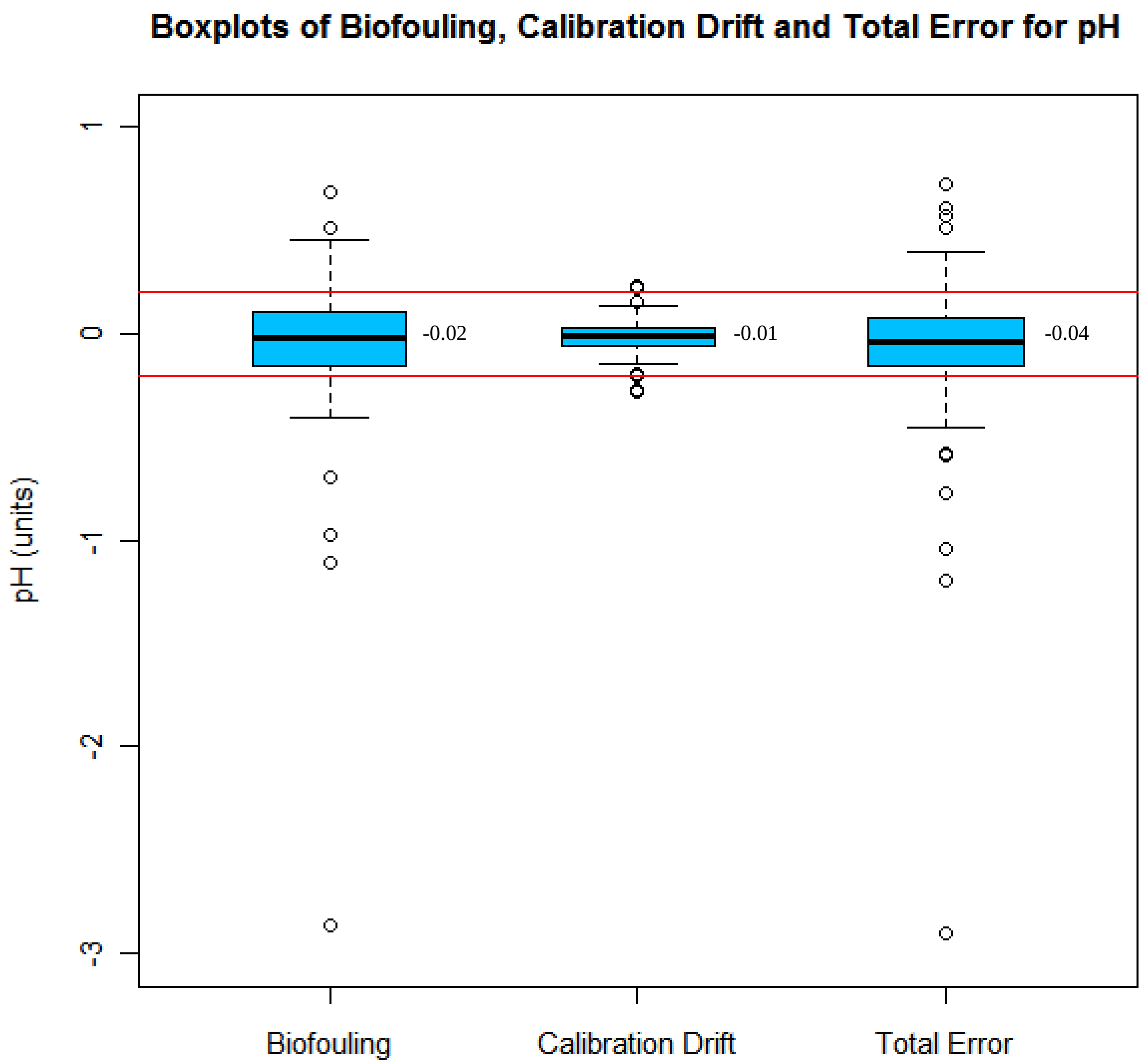


pH Correction

Boxplots in Figure 5 show that biofouling tends to have a more substantial impact on the Total Error for pH than calibration drift. Though the bulb is brushed prior to each reading by the self-cleaning apparatus on the Hydrolab DS5X, the bristles tend to wipe over the top portion of the bulb only and miss the sides. It is possible that the bristles are not cleaning the bulb thoroughly enough to reduce biofouling.

Regardless of the relative contribution by biofouling and drift, correction is rare and there is only a 29% chance that a correction of greater than ± 0.2 pH units will be applied; Total Error still falls within the range delineated by the red lines in Figure 5. During the pilot project, a maximum correction of 0.72 units was found.

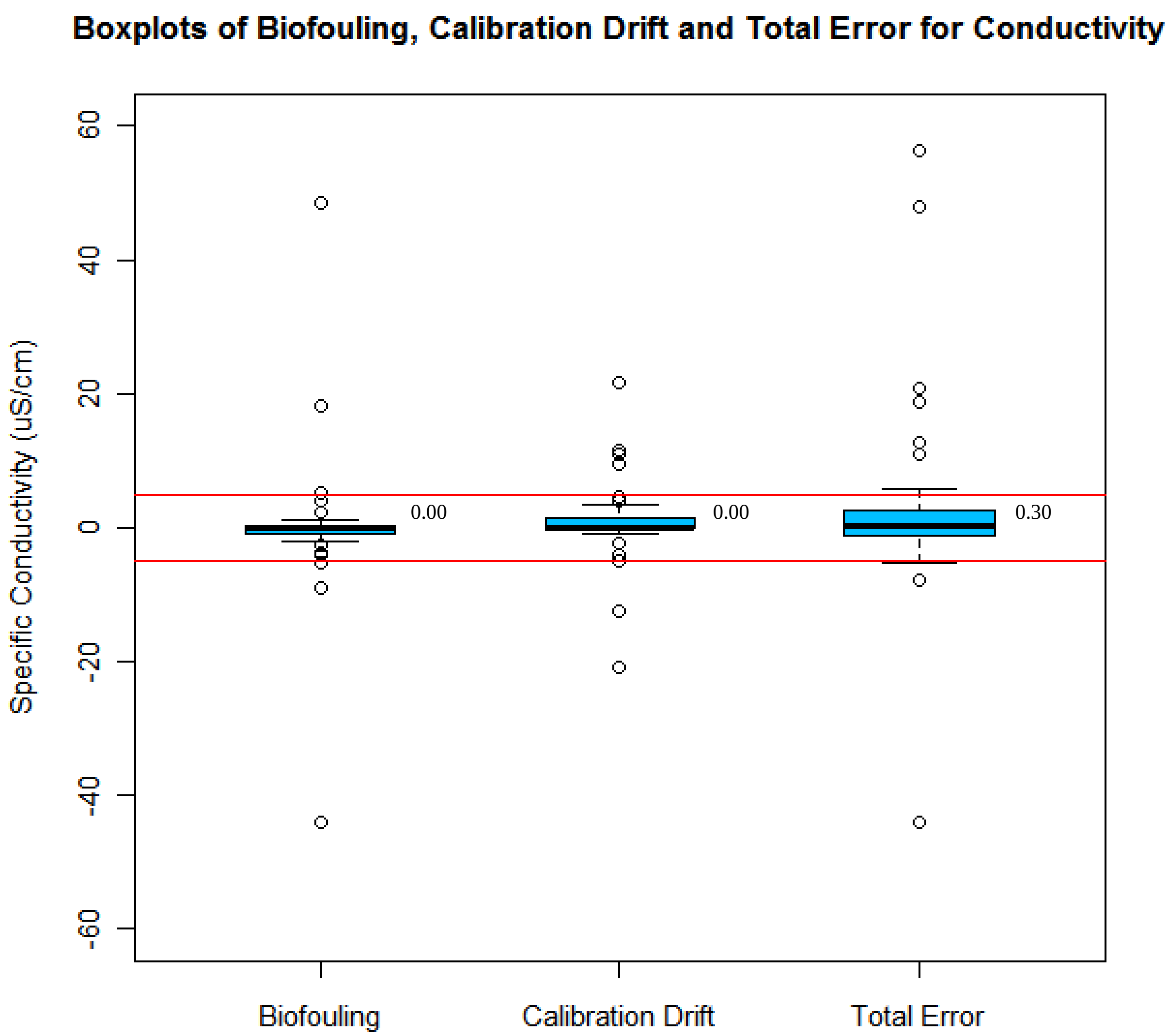
Figure 5: Contribution of Biofouling and Calibration Drift to Total Error for pH



Conductivity Correction

By the relative sizes of the boxplots in Figure 6, it appears that calibration drift contributes more to Total Error than biofouling. While the median Total Error value of 0.30 µS/cm falls well within the ±5 µS/cm data correction criterion, the criterion is actually a two-tired approach where correction may also be applied if Total Error is >3% of the measured value.

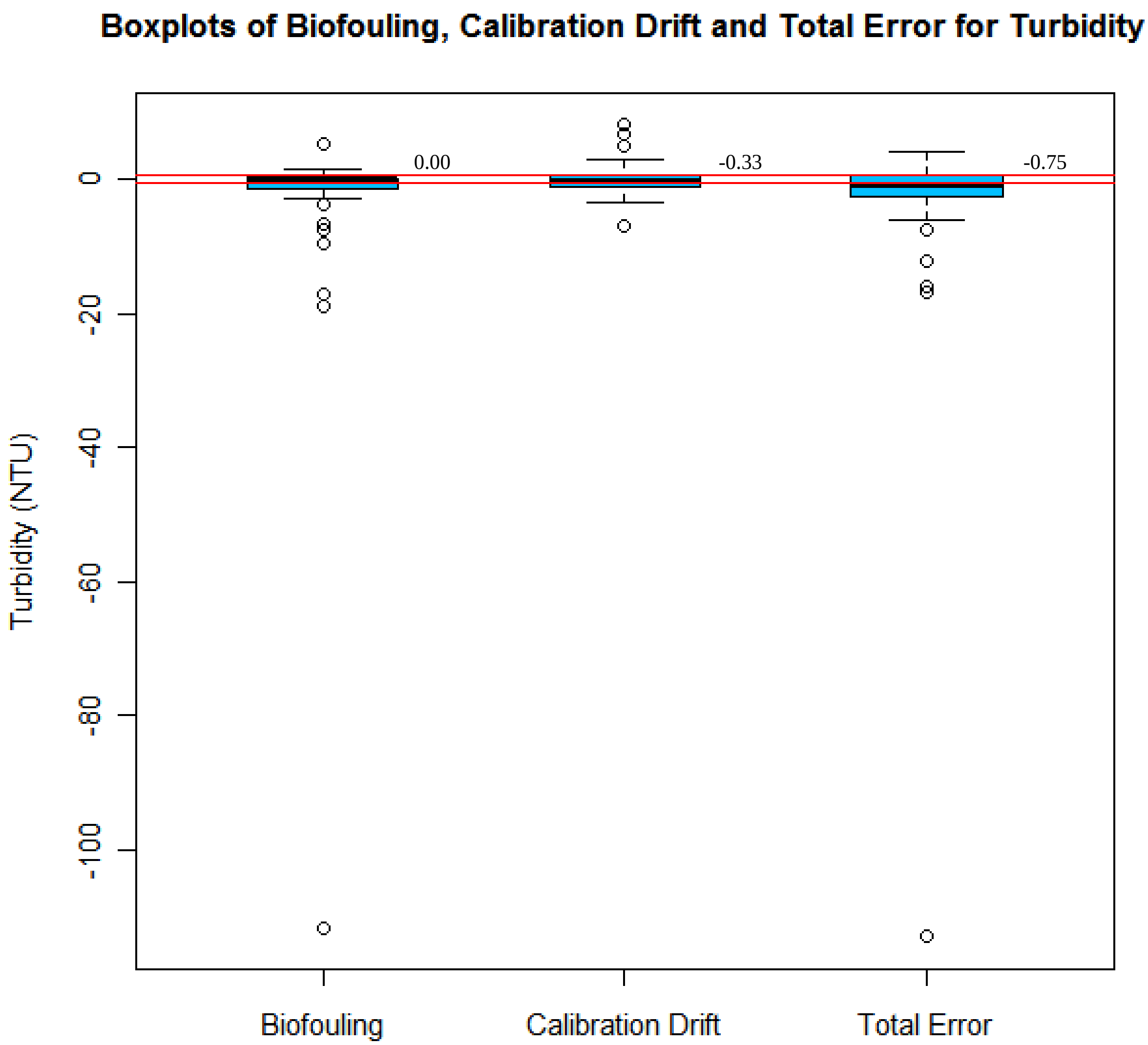
Figure 6: Contribution of Biofouling and Calibration Drift to Total Error for Conductivity



Turbidity Correction

Turbidity data undergoes frequent correction, mostly due to calibration drift, according to the relative size of boxplots in Figure 7. The magnitude of correction is generally very small with a median of only 0.75 NTU even though corrections are applied almost 88% of the time. It appears that the Correction Criterion for turbidity (± 0.5 NTU) may be unreasonably small.

Figure 7: Contribution of Biofouling and Calibration Drift to Total Error for Turbidity



Statistical Comparisons

There are three sources of water quality data in the Real-Time Water Quality network: Field Sonde readings, QAQC Sonde readings and Grab Sample results (abbreviated in this report as Field, QAQC and Grab, respectively). In the second half of the Results and Discussion Section, a series of hypothesis tests are carried out to assess the degree of similarity between the three data sources. Ideally, all three sources of data should be nearly indistinguishable from one another. The presence of a significant difference between any group indicates potential problem areas in quality control, sample analysis, inaccuracy of equipment or differences in analytical methods. A thorough presentation of all statistical procedures used in this report is presented in the Appendix.

Field Sonde vs. Grab Samples

As part of the QAQC process used in the real-time water quality monitoring network, qualitative statements on data quality are made based on the difference between Field Sonde and Grab Sample values. The rankings, given in Table 1 are intended to be easily understood but can be very misleading and unclear if the accuracy of data on which they are based is questionable.

In this section, differences in Field and Grab data sets will be assessed for pH, conductivity and turbidity. Other comparisons such as DO and temperature are based on Sonde to Sonde comparisons in the river.

pH

To determine if qualitative statements comparing pH values from Field and Grab should be included in monthly reports, a test must be made to determine if there is, in general, a difference between the two groups.

Figure 8 depicts the distributions of pH values from Field and Grab sources. Superficially, both appear to be approximately normal distributions and similar in range; however, descriptive statistics in Table 3 and a normality test in Table 15 identified Field pH values as a non-normal and skewed distribution.

Figure 8: Boxplots of Field Sonde and Grab Sample pH Values

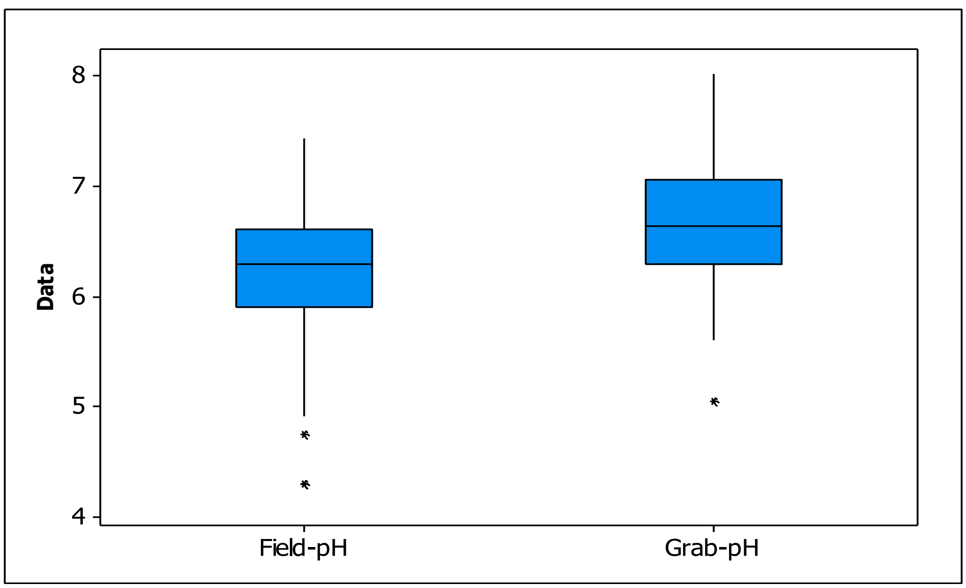


Table 3: Descriptive Statistics of Field Sonde and Grab Sample Conductivity Values

Variable	Mean	Minimum	Q1	Median	Q3	Maximum
Field-pH	6.1947	4.3000	5.9100	6.3000	6.6100	7.4300
Grab-pH	6.6457	5.0500	6.3000	6.6400	7.0600	8.0200

Three tests, a Mann-Whitney test and t-test on raw and log-transformed data, are presented in Table 4. The Mann-Whitney test is a non-parametric test of differing sample medians. New normalized distributions are formed for each sample by ranking individual. A comparison is made between the three new sample medians.

All three tests concur that a difference is evident between Field and Grab pH values as implied by the box plots above. pH values recorded by the field sonde are generally lower than those as measured by Grab Sample.

Table 4: Tests to Determine Significant Difference between Field and Grab Sample pH

Test	Result	p-value
Mann Whitney test of population medians Ho: ή(Field) = ή(Grab) Ha: ή(Field) ≠ ή(Grab)	Reject null	p = 0.0007
Student's t-test of population means (no transform) Ho: μ(Field) = μ(Grab) Ha: μ(Field) ≠ μ(Grab)	Reject null	p = 0.000
Student's t-test of population means (log transform) Ho: μ(Field) = μ(Grab) Ha: μ(Field) ≠ μ(Grab)	Reject null	p = 0.001

Specific Conductivity

Figure 9 illustrates the highly skewed distributions of specific conductivity from Field Sonde and Grab Sample distributions. Although a preliminary glance may suggest that the two distributions are dissimilar, descriptive statistics in

Anderson-Darling tests find that the turbidity distributions for Field, QAQC and Grab samples are not normal (Table 19).

Table 20: Normality Tests for Field, QAQC and Grab Turbidity

Population	Anderson-Darling Statistic	p-value
Field Sonde	10.566	< 0.005
QAQC Sonde	10.536	< 0.005
Grab Sample	10.227	< 0.005

Turbidity: Raw

Kruskal-Wallis Test: Rank versus Factor

```
Kruskal-Wallis Test on Rank

Factor         N  Median  Ave Rank      Z
Field-Turb    51   52.00      74.0  -0.60
Grab-Turb     51   82.00      86.2   1.82
QA-Turb       51   49.00      70.8  -1.22
Overall      153              77.0

H = 3.44  DF = 2  P = 0.179
```

The test above indicates a failure to reject the null hypothesis at a p-value of 0.179 – there is no significant difference between Field, QA and Grab sample turbidity.

Kruskal-Wallis Test: Field-Turb_1 versus QA-Turb_1

```
Kruskal-Wallis Test on Field-Turb_1

QA-Turb_1       N  Median  Ave Rank      Z
Field-Turb_1  23   8.000      41.2   1.81
Grab-Turb_1   23   3.600      24.3  -3.14
QA-Turb_1     23   7.700      39.5   1.32
Overall       69              35.0

H = 9.93  DF = 2  P = 0.007
H = 9.93  DF = 2  P = 0.007  (adjusted for ties)
```

The test above assesses uses natural (unranked) data from Field, QAQC, and Grab samples with all 0.0 NTU sonde values and concurrent Grab Sample values removed. The test observes a difference between the groups at a p-value of 0.007.

Turbidity: 0.0 NTU removed

Table 21: Normality Tests for Field, QAQC and Grab Turbidity (no zeros) shows that first, second and third quartiles (Q1, median and Q3) are close for both sources.

Normality tests indicate that the two groups do not fit the standard distribution; therefore, parametric analysis has very limited power compared to non-parametric tests (both types are presented out of interest, however).

Figure 9: Boxplots of Field Sonde and Grab Sample Conductivity Values

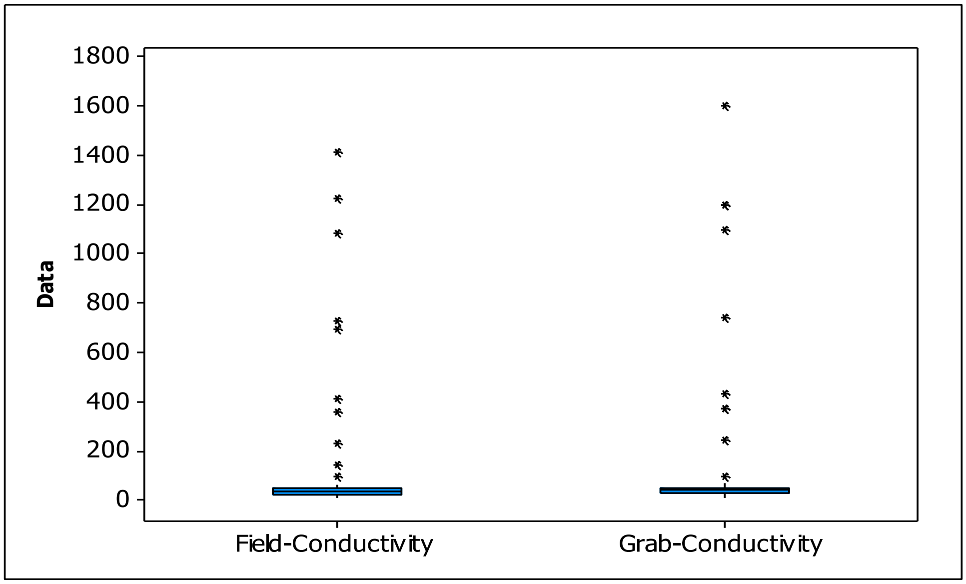


Figure 10: Descriptive Statistics of Field Sonde and Grab Sample Conductivity Values

Variable	Mean	Minimum	Q1	Median	Q3	Maximum
Field-Conductivi	152.0	11.0	25.3	38.0	50.0	1417.0
Grab-Conductivit	173.9	13.0	27.0	42.0	53.0	1600.0

All three tests presented in Table 5 indicate that the null hypothesis fails to be rejected. There is insufficient evidence to claim a significant difference in specific conductivity values between both Field and Grab sources.

Table 5: Tests to Determine Significant Difference between Field and Grab Sample Conductivity

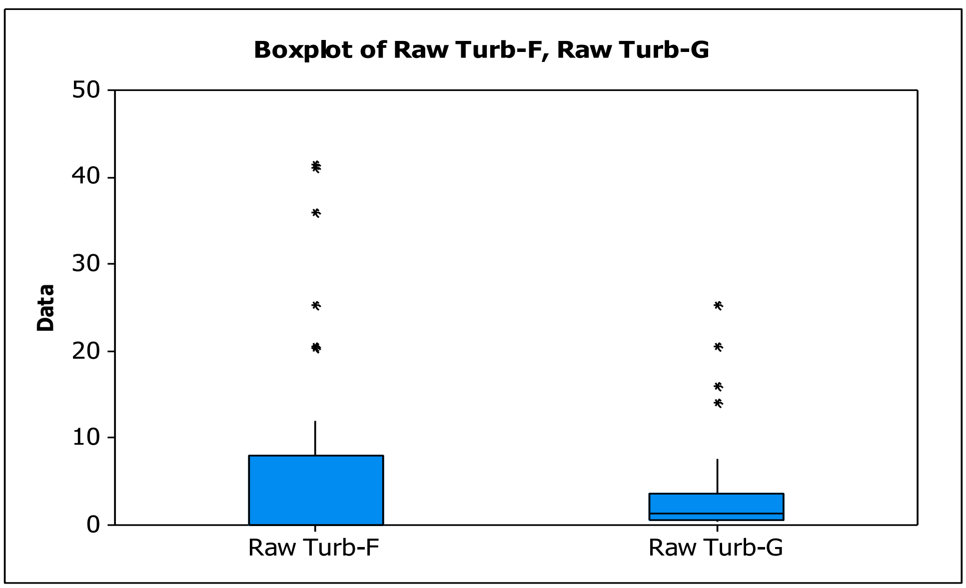
Test	Result	p-value
Mann Whitney test of population medians Ho: ή(Field) = ή(Grab) Ha: ή(Field) ≠ ή(Grab)	Fail to reject null	0.1873
Student's t-test of population means (no transform) Ho: μ(Field) = μ(Grab) Ha: μ(Field) ≠ μ(Grab)	Fail to reject null	0.743
Student's t-test of population means (log transform) Ho: μ(Field) = μ(Grab) Ha: μ(Field) ≠ μ(Grab)	Fail to reject null	0.562

Turbidity

Statistical analysis of turbidity is difficult due to the frequent occurrence of 0.0 NTU Field Sonde values. Frequent and repetitive zero values cause difficulty in attempting to attain reasonable distributions needed for hypothesis testing. Because zero values cannot be transformed except through ranking, the non-parametric Mann-Whitney test is used exclusively.

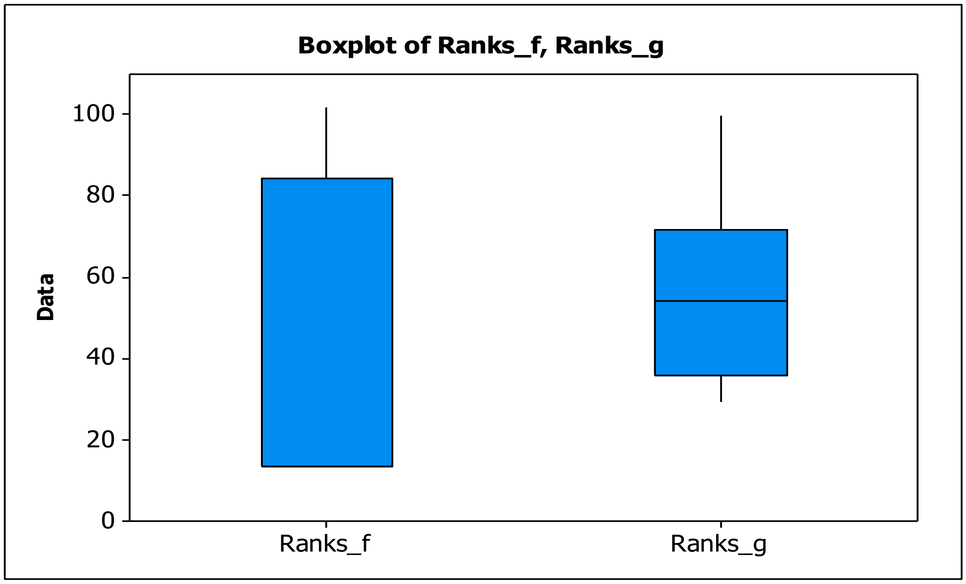
Figure 11 depicts the raw distributions of turbidity from Field and Grab. Due to the frequent and repetitive 0.0 NTU values recorded by the field sonde, the boxplots are very skewed.

Figure 11: Boxplots of Field Sonde and Grab Sample Turbidity Values



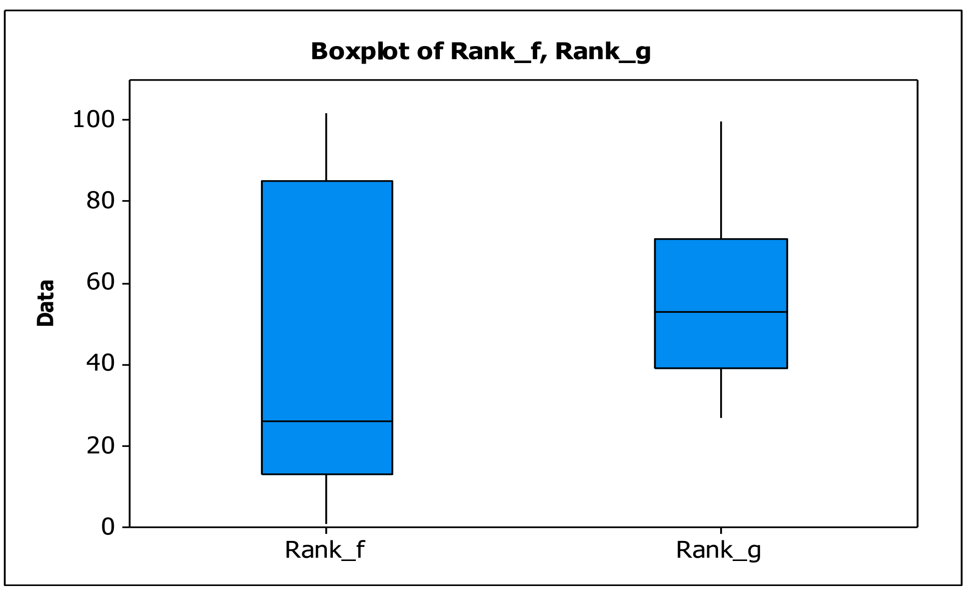
As part of the Minitab Mann-Whitney test method, ranks are tied in cases of equal values, leading to multiple instances of the same rank. Figure 12 details this effect in the distribution of ranked Field. Because of the frequent occurrence of 0.0 NTU values and multiple tied values, the median rank is pushed down, causing an artificially skewed distribution where the median is equal to the first quartile.

Figure 12: Boxplots of Ranked Field Sonde and Grab Sample Turbidity Values with Ties Unbroken



Breaking the ties by randomly assigning a rank to each equal value averts this problem. As a result, the median is elevated and the similarity between both distributions becomes even clearer in Figure 13.

Figure 13: Boxplots of Ranked Field Sonde and Grab Sample Turbidity Values with Ties Broken



Analysis of turbidity (Table 6) suggests that there is no significant difference between Field Sonde and Grab Sample turbidity values as a whole: there is a significant amount of overlap seen in Figure 11 between boxplots for Field and Grab values. Even ranking the data (with ties broken), a large degree of overlap is observed in Figure 13, making it difficult to distinguish the two distributions.

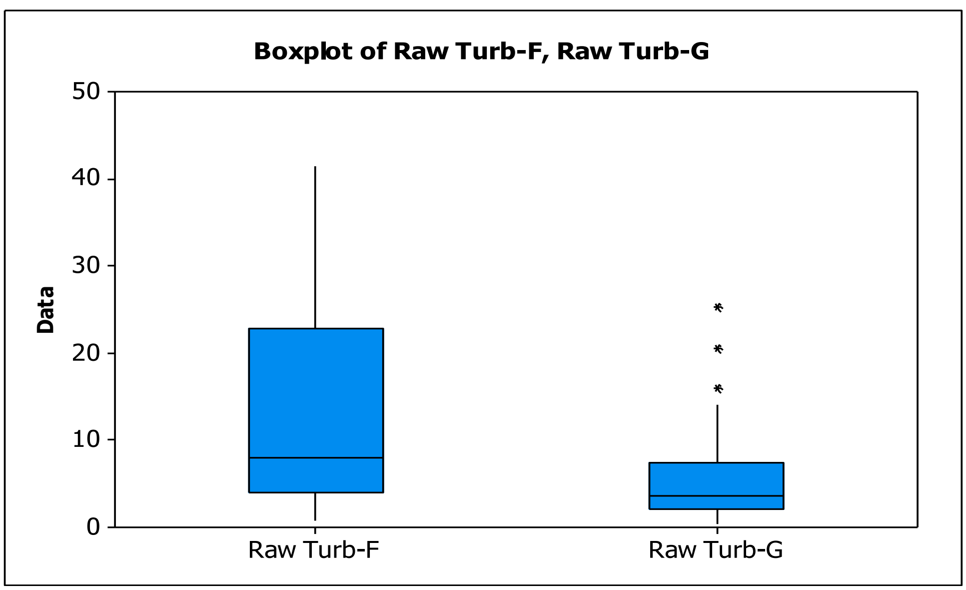
Table 6: Tests to Determine Significant Difference between Field and Grab Sample Conductivity

Test	Result	p-value
Mann Whitney test of population medians (Raw data) Ho: ή(Field) = ή(Grab) Ha: ή(Field) ≠ ή(Grab)	Fail to reject null hypothesis	0.1609
Mann Whitney test of population medians (Ranked with ties broken) Ho: ή(Field) = ή(Grab) Ha: ή(Field) ≠ ή(Grab)	Fail to reject null hypothesis	0.1619

A problem is revealed in examining the distributions for Field and Grab, however. Turbidity-free water is unusual in natural waters, which hinted towards a lack of low-end sensitivity in the Field Sonde values – Field Sonde values were often 0.0 NTU while concurrent Grab Sample values registered a detectable level between 0.1 and 0.9 NTU.

In an effort to identify if a lack of sensitivity in the low end inhibits the ability to detect differences in higher levels of turbidity, all 0.0 NTU values were removed from the Field Sonde dataset with concurrent Grab values also removed, resulting in Figure 14.

Figure 14: Boxplots of Field Sonde and Grab Sample Turbidity Values with 0.0 NTU Values Removed



The result, presented in Table 7, indicates that a significant difference is found between the two groups. Such manipulation in the dataset creates a bias towards moderate and high-level turbidity values which may be problematic to data interpretation. To avoid this, instead of removing all 0.0 NTU values, they were replaced with random values between 0.01 and 0.09 NTU – numbers below the level of resolution in the Hydrolab turbidity sensor. This test failed to detect a difference between both groups.

Table 7: Tests to Determine Significant Difference between Field and Grab Sample Turbidity (0.0 NTU values removed)

Test	Result	p-value
Mann Whitney test of population medians (0.0 NTU values removed) Ho: ή(Field) = ή(Grab) Ha: ή(Field) ≠ ή(Grab)	Reject null hypothesis	0.0083
Mann Whitney test of population medians (0.0 NTU values replaced randomly with 0.01 – 0.09 NTU) Ho: ή(Field) = ή(Grab) Ha: ή(Field) ≠ ή(Grab)	Fail to reject null hypothesis	0.1609

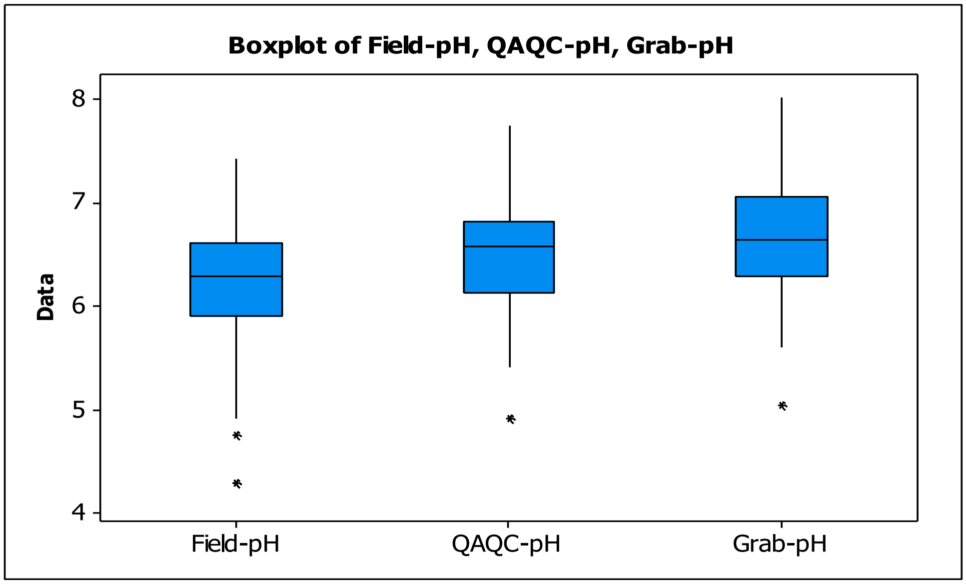
Field Sonde vs. QAQC Sonde vs. Grab Sample

The previous section dealt with comparisons between Field and Grab values with the aim of determining whether qualitative statements regarding data quality warrant inclusion in Monthly Reports. In this section comparisons between Field, QAQC and Grab values are made in the interest of determining if differences between the three groups are present. In the ideal world, all three groups should be very similar.

pH

Distributions of pH from each of the three sources in Figure 15 are very similar in shape. Both QAQC and Grab Sample resemble normal distributions, however, Field Sonde deviates from normality at a p-value of 0.033. Because only one of the three distributions just fails the test for normality, both non- and parametric tests are used to determine if there is a significant difference between the three groups.

Figure 15: Boxplots of pH from Field Sonde, QAQC Sonde and Grab Sample



In Table 8 the results from a Kruskal-Wallace and ANOVA concur that there is at least one significant difference in pH values between the three groups. These tests, however, do not indicate which difference is significant.

Table 8: Hypothesis Tests to Identify Differences between Field Sonde, QAQC Sonde and Grab Sample pH Values

Test	Method	p-value	Conclusion
Is there a difference in pH measurements between Field Sonde, QAQC Sonde, and Grab Sample?	Kruskal Wallace Test	0.002	There is a significant difference in population medians.
	ANOVA	0.002	There is a significant difference in population means.

In the previous section, it was determined that Field pH is less than Grab pH. However, to identify other differences, two methods were used: the Tukey family error rate (parametric) and pair-wise Mann-Whitney tests, shown in Table 9.

The results differ between the two methods, however, the Mann-Whitney tests should be considered more appropriate for discussion since the three sampling distributions do not quite fit the requirements for parametric testing. Significant differences were found between median Field and QAQC pH values (Field < QAQC) and the previously observed difference between median Field and Grab pH values (Field < Grab). No significant difference was detected between QAQC and Grab pH values.

Table 9: Hypothesis Tests to Identify Pair-wise Differences between Field, QAQC and Grab pH values

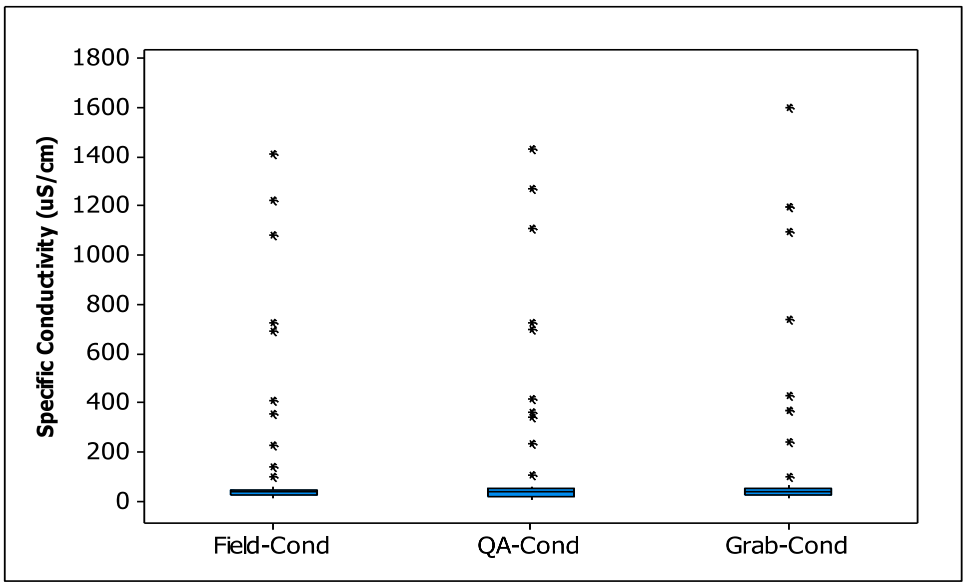
Test	Method	p-value	Conclusion
Are there significant pair-wise differences between Field, QAQC and Grab pH values?	ANOVA, with Tukey family error rate	-	Significant difference detected: - Mean Field pH is less than Grab pH - No difference between mean Field

Test	Method	p-value	Conclusion
			pH and QAQC pH values - No difference between mean QAQC and Grab pH values
	Mann-Whitney Test Ho: ή (Field) = ή (QAQC) Ha: ή (Field) ≠ ή (QAQC)	0.0270	Significant difference detected: median Field pH is less than median QAQC pH.
	Mann-Whitney Test Ho: ή (QAQC) = ή (Grab) Ha: ή (QAQC) ≠ ή (Grab)	0.1436	No significant difference detected between median QAQC pH and Grab Sample pH.
	Mann-Whitney Test Ho: ή (Field) = ή (Grab) Ha: ή (Field) ≠ ή (Grab)	0.0007	Significant difference detected: median Field pH is less than median Grab Sample pH.

Specific Conductivity

The distributions of raw Field, QAQC and Grab values are highly skewed due to the incidence of occasionally elevated conductivity values (Figure 16). Normality tests indicate that none of the groups are remotely normal in distribution.

Figure 16: Boxplots of Specific Conductivity from Field Sonde, QAQC Sonde and Grab Sample



Because of non-normal distributions, a non-parametric Kruskal Wallace test of population medians was applied to the dataset. The result is implied by the boxplots above: no significant difference was detected between the population medians for the three groups at a p-value of 0.284 (see Table 10).

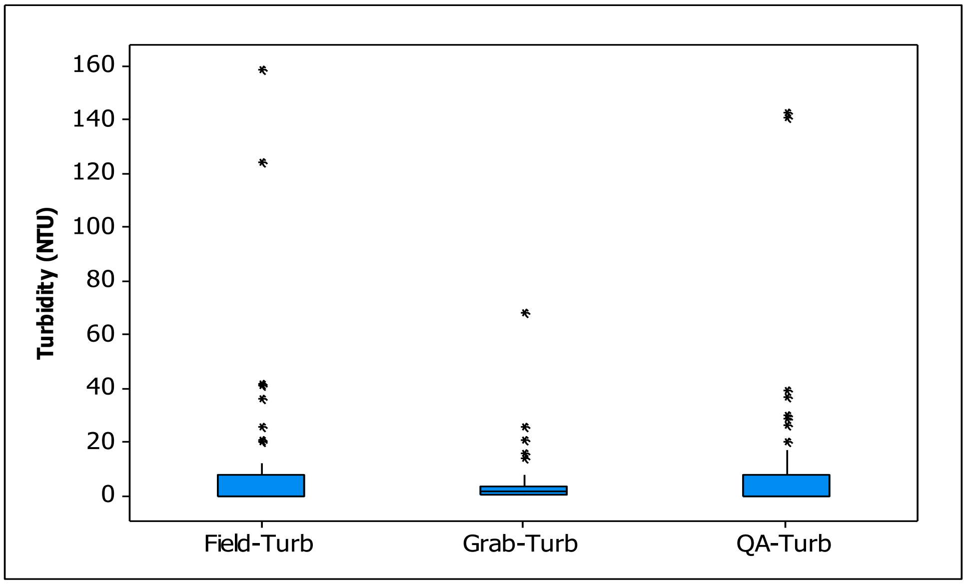
Table 10: Test to assess the difference between Field Sonde, QAQC Sonde and Grab Sample Specific Conductivity values

Test	Method	p-value	Conclusion
Is there a difference between the distributions for Field, QAQC and Grab Sample specific conductivity values?	Kruskal Wallace Test Ho: ή (Field) = ή (QAQC) = ή (Grab) Ho: ή (Field) ≠ ή (QAQC) ≠ ή (Grab)	0.284	Fail to reject null hypothesis. There is no significant difference between the three distributions

Turbidity

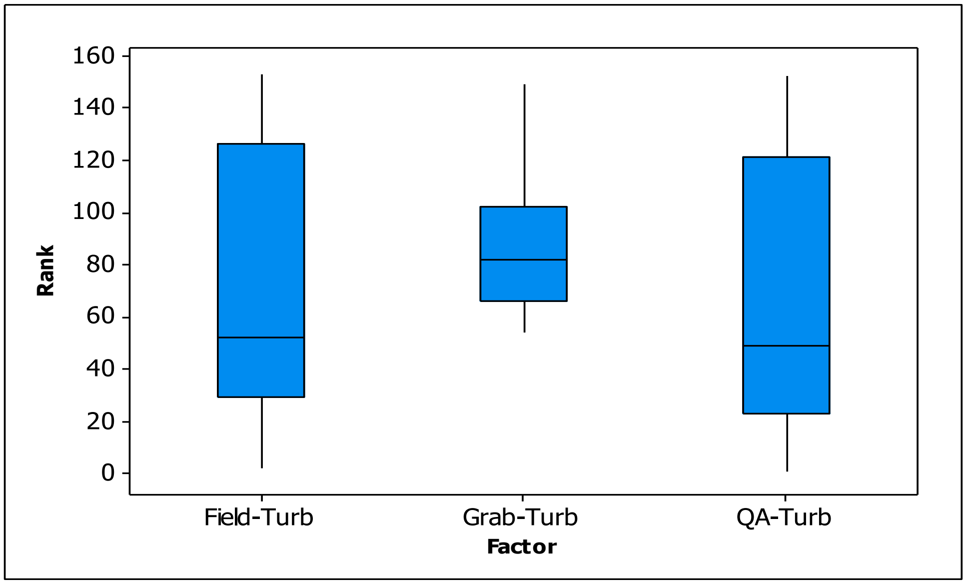
Turbidity values from Field, QAQC and Grab samples are not normally distributed. Typically turbidity values are very low with a low occurrence of very high values. With the addition of being bound by a lower limit of zero, turbidity distributions are almost always non-normal (Figure 17).

Figure 17: Boxplots of Turbidity from Field Sonde, QAQC Sonde and Grab Sample



To proceed with analysis, all values were pooled and ranked with ties broken randomly. From the ranked dataset, Figure 18 was constructed and illustrates a large degree of overlap between the three groups. A Kruskal-Wallace analysis of population medians was applied to this dataset.

Figure 18: Boxplots of Ranked Turbidity from Field Sonde, QAQC Sonde and Grab Sample



The result in Table 11 indicates that, with a p-value of 0.179, there is no significant difference between Field, QAQC and Grab Sample turbidity values.

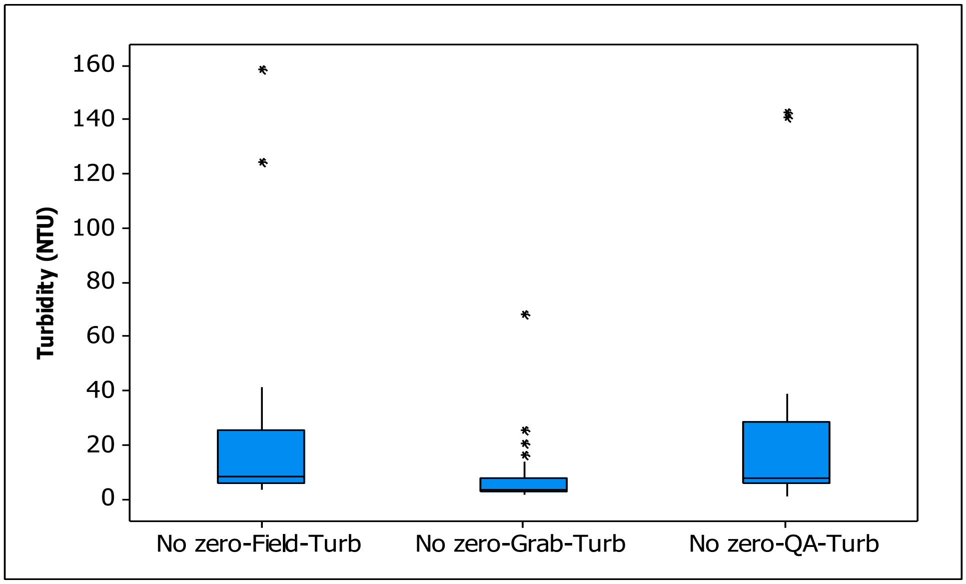
Table 11: Hypothesis Test to Identify Pair-wise Differences between Field, QAQC and Grab Turbidity Values

Test	Method	p-value	Conclusion
Is there a difference between the distributions for Field, QAQC and Grab Sample turbidity values?	Kruskal Wallace Test Ho: ή (Field) = ή (QAQC) = ή (Grab) Ho: ή (Field) ≠ ή (QAQC) ≠ ή (Grab)	0.179	Fail to reject null hypothesis. There is no significant difference between the three distributions

As discussed in the previous Field Sonde vs. Grab sample Turbidity section, there is evidence to suggest that a lack of sensitivity in the low end of turbidity measurements influences the sampling distribution. In the following set of tests, 0.0 NTU values are removed from the three sampling distributions and replaced with random values below the resolution limit of the Hydrolab sensor.

Figure 19 shows the distributions of all three groups with zero values removed. Superficially, it appears that Grab Sample turbidity values are lower than the other two groups with little difference between Field and QAQC values.

Figure 19: Boxplots of Turbidity Values from Field Sonde, QAQC Sonde and Grab Sample groups (0.0 NTU Removed)



The result of a Kruskal-Wallace test in Table 12 indicates that a difference between the three groups was detected.

Table 12: Test to assess the difference between Field, QAQC and Grab Sample turbidity values with 0.0 NTU values removed

Test	p-value	Conclusion
Kruskal-Wallace Ho: ή (Field) = ή (QAQC) = ή (Grab) Ho: ή (Field) ≠ ή (QAQC) ≠ ή (Grab)	P = 0.007	Reject null hypothesis: There is a difference between Field, QAQC and Grab Sample

To determine where the difference(s) reside, three pair-wise Mann-Whittney tests are presented in Table 13. These results support the superficial examination above: there is no difference between median Field and QAQC turbidity values, however, Grab turbidity values are less than Field and QAQC, as shown in the previous section.

Table 13: Hypothesis Tests to Identify Pair-wise Differences between Field, QAQC and Grab Turbidity values

Test	Method	p-value	Conclusion
Are there significant pair-wise differences between Field, QAQC and Grab Turbidity values?	Mann-Whitney Test Ho: ή (Field) = ή (QAQC) Ha: ή (Field) ≠ ή (QAQC)	0.8090	Fail to reject null: there is no difference between median Field and QAQC turbidity values
	Mann-Whitney Test Ho: ή (QAQC) = ή (Grab) Ha: ή (QAQC) ≠ ή (Grab)	0.0115	Reject null: there is a significant difference between median QAQC and Grab turbidity values
	Mann-Whitney Test Ho: ή (Field) = ή (Grab) Ha: ή (Field) ≠ ή (Grab)	0.0041	Reject null: there is a significant difference between median Field and Grab turbidity values

In the final test, 0.0 NTU values found within the Field and QAQC groups were replaced with turbidity values ranging from 0.01 – 0.09 NTU with the intention of avoiding the insertion of bias into the groups by deleting values, but introducing variability in the dataset through the addition of values below the resolution of the Hydrolab turbidity sensor.

A Kruskal-Wallace test of sample medians indicates that no difference was detected (Table 14).

Table 14: Test to assess the difference between Field, QAQC and Grab Sample turbidity values with 0.0 NTU values replaced

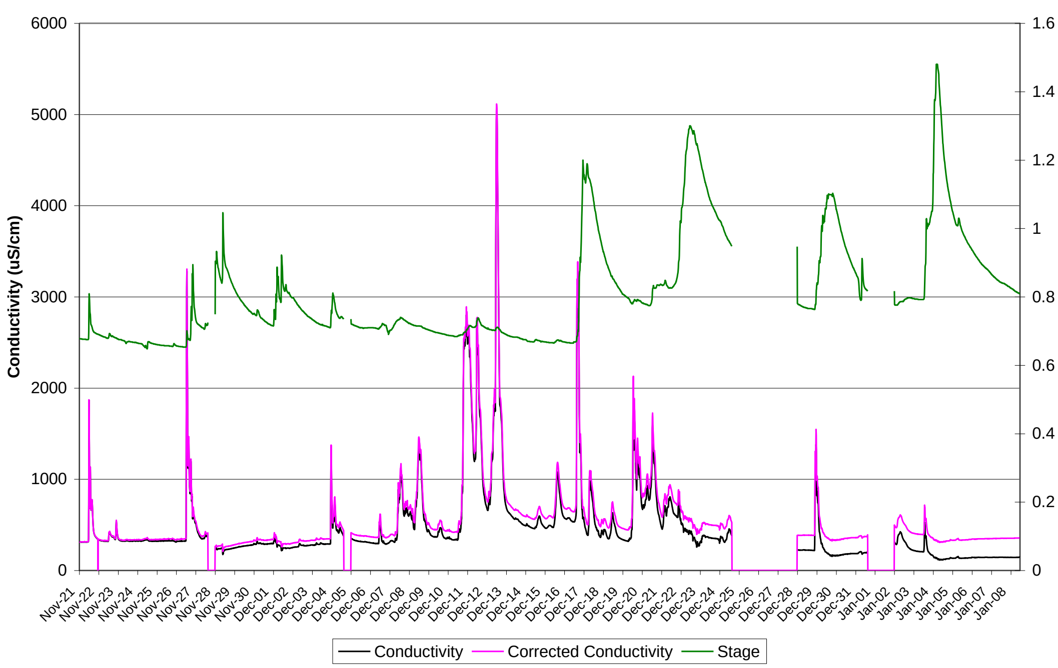
Test	p-value	Conclusion
Kruskal-Wallace Ho: ή (Field) = ή (QAQC) = ή (Grab) Ho: ή (Field) ≠ ή (QAQC) ≠ ή (Grab)	P = 0.184	Fail to reject null: there is no difference between median Field, QAQC and Grab turbidity values

Conclusions

Data Corrections

In theory, post-deployment data correction based on Total Error is a sound idea that accounts for fouling and calibration drift due to environmental and instrumental factors. In practice, however, the correction process was found to be time consuming and of minimal benefit. Some corrections were also found to be questionable and produced confusing results, such as in Figure 20. These corrections are difficult to verify without additional site visits. Corrections also failed to account for temporal fouling that may have been caused by silt deposition in a high flow event and subsequently removed by another.

Figure 20: Ramping behaviour



Boxplots indicate that the Biofouling, Calibration, and Total Error factors for temperature, pH, conductivity, and turbidity often fall within the data correction criterion. In the case of conductivity and turbidity correction, where more than half of deployments faced correction, most corrections were based on the two-tired data correction criterion. Corrections may be invoked by Total Error exceeding the data correction criterion or by Total Error Exceeding an individual reading by a certain percentage where the percent difference was usually much lower than the data correction criterion itself.

While a workaround exists for this problem (such as changing the criteria to a single-tired system), most effort in the correction method is directed at the pH probe where drift calculation involves taking readings from the sensor using standard buffers both before and after changing the standard KCL reference junction. This method, which is time consuming, resulted in correction of about 29% of deployments – even then, most corrections were small except for a few instances where corrections of < -1 were encountered[3].

Most importantly, given the objective of the Real-Time Water Quality Monitoring Network in Newfoundland and Labrador, corrections are not necessary at this time. Long term trend monitoring and event detection is not hampered by the small amount of biofouling encountered in the cool waters of Newfoundland and Labrador. Ensuring a 30 day deployment schedule ensures that fouling in all forms are minimal and that sensors are in their best condition. At this time, it is felt that data correction is unnecessary and will be discontinued with effort instead directed towards providing better data products to partners and the public, alike.

Statistical Comparisons

Describing data quality qualitatively eases the task of understanding how comparable real-time results are to 'true' values. It has been a long-standing QAQC process to make direct comparisons between the Field and QAQC sondes at the end of a deployment for incorporation into reports. With the new automated Deployment Spreadsheet, the scope of comparisons was broadened to include comparisons between Field and Grab samples as well. These comparisons hinge on the assumption that values from the QAQC and Grab Samples are accurate and true representations of water quality data and the Field Sonde deviates somewhat from these values in cases where biofouling and drift have occurred.

[3] Due to a lack of detailed information, it is possible that these extreme corrections may be related to instrument error or a failure to follow protocol instead of actual biofouling/drift.

For this reason, it could be assumed that QAQC and Grab values should be indistinguishable with differences between Field and QAQC/Grab values. The first part of the section dealing with the validity of qualitative statements between Field and Grab samples indicates that the only obvious deviation between the two is pH where Field pH is lower, in general, than concurrent Grab Sample values. There was no difference between the two in terms of specific conductivity and turbidity, however this is open to debate given the issues with 0.0 NTU values.

Comparisons between the three groups offers insight into whether drift and biofouling induces differences between Field and QAQC/Grab samples. In the case of pH, this tendency may be indicated: non-parametric tests find that Field pH values are less than QAQC and Grab values and there is no difference between Grab and QAQC values (which are not affected by biofouling/drift).

Specific Conductivity was found not to differ between the three groups and a difference in turbidity between Field, QAQC, and Grab was only detected when all 0.0 NTU values were removed, inducing a large degree of bias in the sample distributions.

Path Forward

Data corrections based on Total Error (as derived from Biofouling and Calibration drift) take a great deal of effort to implement and the application of these corrections has been questioned in many situations. Corrections for conductivity and turbidity are most often applied due to a small percentage error instead of a Total Error exceeding stated data correction criteria. In Figure 6 and Figure 7, Total Error barely exceeds the stated data correction criterion (red lines) in most cases. For this reason, it has been determined that the calculation of biofouling and calibration drift cease. Raw data as presented publicly on the internet, suits the goals and objectives of the Newfoundland and Labrador Real-Time Water Quality program with the reasoning that drift and biofouling is not a concern.

Data quality statements for Field Sonde/QAQC Sonde comparisons will be maintained, but Field Sonde/Grab Sample comparisons will be discontinued. Since the time between capture and analysis of Grab samples can be prolonged, the grab sample values could be questionable in many cases while Field and QAQC readings are taken in situ and with no time delay. Furthermore, Field/QAQC comparisons have been a long-standing process that all staff members appreciate and support.

Through this pilot project, a great deal of conformity was required to ensure that all staff members were following protocol across all regions of Newfoundland and Labrador. To assist in this process, a well-defined

Deployment Spreadsheet was constructed incorporating automated data corrections, graphing and qualitative statement generation. This spreadsheet is the product of a great deal of research and peer review and has streamlined field and laboratory procedures. With some changes, the Deployment Spreadsheet will continue to be updated as needed, especially with the aim of bringing consistency to the management of all real-time stations across the province.

Since turbidity was shown to be the most variable parameter of concern, further literature review and testing of various turbidity probes will be performed to better understand turbidity in provincial waters.

With the completion of the evaluation of the QA/QC protocols and procedures, a manual describing the existing protocols will be completed and posted to the departmental webpage.

Acknowledgements

The efforts of Keith Abbott, Ian Bell, Tara Clinton, Grace Gillis, Renée Paterson, Ryan Pugh, Joanne Sweeney, and Rob Wight are appreciated. This project introduced a great deal of change in long standing procedures and a lack of immediate help to clarify the details of new processes while in the field was undoubtedly frustrating at times.

Further appreciation goes to Tara Clinton, Grace Gillis and Renee Paterson for their input into the processes used in the Pilot Project and for Shibly Rahman and Ryan Pugh's efforts in assembling a final analysis and report on the pilot project.

References

Wagner, R.J., Boulger, R.W., Jr., Oblinger, C.J., and Smith, B.A., 2006, Guidelines and standard procedures for continuous water-quality monitors – Station operation, record computation, and data reporting: U.S. Geological Survey Techniques and Methods 1-D3, 51 p. + 8 attachments; accessed April 10, 2006, at http://pubs.water.usgs.gov/tm1d3

Prepared by:
Ryan Pugh
Department of Environment and Conservation
Water Resources Management Division
Phone: 709.729.1681
Fax: 709.729.3020

Appendix

Statistical Comparisons: Field Sonde vs. Grab Samples

pH

pH values from Field Sonde and Grab Sample sources show different degrees of normal distribution (according to Table 15). Grab sample pH values tend to b more normally distributed than Field Sonde values, according to Anderson-Darling tests (p=0.910 vs. p=0.033). Log transforms did not force a normal distribution. This limits the power of parametric tests of population means. Non parametric tests are more appropriate in this case and the Mann-Whitney test will be used primarily.

Table 15: Normality Tests for Field and Grab pH

	Anderson-Darling Statistic	p-value
Field Sonde pH	0.811	0.033
Grab Sample pH	0.181	0.910

Mann-Whitney Test and CI: pH_F, pH_G

```
       N  Median
pH_F  51  6.3000
pH_G  51  6.6400

Point estimate for ETA1-ETA2 is -0.4100
95.0 Percent CI for ETA1-ETA2 is (-0.6400,-0.1801)
W = 2119.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0007
The test is significant at 0.0007 (adjusted for ties)
```

In the above test result, since the p-value of 0.007 is less than α, we reject the null hypothesis and favour the alternate hypothesis that there is a difference between the field sonde and grab sample values for pH.

Two-Sample T-Test and CI: pH_F, pH_G

```
       N   Mean  StDev  SE Mean
pH_F  51  6.195  0.695    0.097
pH_G  51  6.646  0.552    0.077

Difference = mu (pH_F) - mu (pH_G)
Estimate for difference:  -0.451
95% CI for difference:  (-0.698, -0.204)
T-Test of difference= 0 (vs not =): T-Value= -3.63 P-Value= 0.000 DF=95
```

In the above test result, since the p-value of 0.000 is less than α, we reject the null hypothesis and favour the alternate hypothesis that there is a difference between the field sonde and grab sample values for pH.

Two-Sample T-Test and CI: pH_log_F, pH_log_G

```
           N    Mean   StDev  SE Mean
pH_log_F  51  0.7891  0.0520   0.0073
pH_log_G  51  0.8210  0.0368   0.0052
```

```
Difference = mu (pH_log_F) - mu (pH_log_G)
Estimate for difference:  -0.03193
95% CI for difference:  (-0.04965, -0.01421)
T-Test of difference = 0(vs not =): T-Value= -3.58 P-Value= 0.001 DF=90
```

In the above test, since the p-value of 0.001 is less than α, we reject the null hypothesis and favour the alternate hypothesis that there is a difference between the field sensor and grab sample values for pH.

Specific Conductivity

Conductivity values for Field Sonde and Grab Sample values are not normally distributed as seen in Table 16. Because log-normal transformation was not possible, non-parametric tests are preferred though parametric t-tests of log transformed and raw data are presented for interest.

Table 16: Normality Tests for Field and Grab Conductivity

	Anderson-Darling Statistic	p-value
Field Sonde Conductivity	11.704	< 0.005
Grab Sample Conductivity	11.697	< 0.005

Mann-Whitney Test and CI: Conductivity_F, Conductivity_G

```
                 N  Median
Conductivity_F  51    38.0
Conductivity_G  51    42.0
Point estimate for ETA1-ETA2 is -3.7
95.0 Percent CI for ETA1-ETA2 is (-11.0, 2.7)
W = 2429.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1873
The test is significant at 0.1871 (adjusted for ties)
```

Since the p-value of 0.1873 is greater than α, we fail to reject the null hypothesis that there is no difference between the specific conductance field sensor and grab sample values.

Two-Sample T-Test and CI: Field-Conductivity, Grab-Conductivity

```
                 N  Mean   StDev SE Mean
Field-Conductivi  51   152    315    44
Grab-Conductivit  51   174    354    50

Difference = mu (Field-Conductivity) - mu (Grab-Conductivity)
Estimate for difference:  -21.8275
95% CI for difference:  (-153.4147, 109.7598)
T-Test of difference = 0 (vs not =): T-Value = -0.33  P-Value = 0.743  DF = 98
```

Since the p-value of 0.743 is greater than α, we fail to reject the null hypothesis that there is no difference between the specific conductance field sensor and grab sample values.

Two-Sample T-Test and CI: Cond_log_F, Cond_log_G

```
             N   Mean  StDev  SE Mean
Cond_log_F  51  1.708  0.543    0.076
Cond_log_G  51  1.770  0.536    0.075
```

```
Difference = mu (Cond_log_F) - mu (Cond_log_G)
Estimate for difference:  -0.062
95% CI for difference:  (-0.274, 0.150)
T-Test of difference= 0 (vs not =):T-Value= -0.58 P-Value= 0.562  DF=99
```

In the above test result, since the p-value of 0.562 is greater than α, we fail to reject the null hypothesis that there is no difference between the specific conductance field sensor and grab sample values.

Turbidity

Due to the frequent presence of zero values for turbidity, transformation is difficult and alternative strategies were used to accomplish statistical analysis of the very non-normal distributions seen in Table 17. Mann Whitney tests are used to assess for difference between both sources of turbidity data.

Table 17: Normality Tests for Field and Grab Turbidity

	Anderson-Darling Statistic	p-Value
Field Sonde Turbidity	10.566	0.005
Grab Sample Turbidity	10.227	0.005

Mann-Whitney Test and CI: Turb_F, Turb_G (with the zero values)

```
         N  Median
Turb_F  51   0.000
Turb_G  51   1.300

Point estimate for ETA1-ETA2 is -0.500
95.0 Percent CI for ETA1-ETA2 is (-0.701,1.601)
W = 2416.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1609
The test is significant at 0.1572 (adjusted for ties)
```

In the above test result, since the p-value of 0.1609 is greater than α, we fail to reject the null hypothesis that there is no difference between the turbidity field sensor and grab sample values.

Mann-Whitney Test and CI: rank_f, rank_g

```
         N  Median
rank_f  51   26.00
rank_g  51   53.00

Point estimate for ETA1-ETA2 is -14.00
95.0 Percent CI for ETA1-ETA2 is (-26.00,7.00)
W = 2417.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1619
```

In the test above, the turbidity values from Field Sonde and Grab Sample sources were pooled and ranked accordingly. In the case of ties (many 0.0 values from Field Sonde were present), they were ranked randomly so as to avoid duplicate ranks. At a p-value of 0.1619, the null hypothesis fails to be rejected indicating that there is no significant difference between Field Sonde and Grab Samples.

Mann-Whitney Test and CI: Ranks_f, Ranks_g

```
          N  Median
Ranks_f  51   13.50
Ranks_g  51   54.00

Point estimate for ETA1-ETA2 is -16.00
95.0 Percent CI for ETA1-ETA2 is (-27.99,7.01)
W = 2416.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1609
The test is significant at 0.1572 (adjusted for ties)
```

In this test, turbidity ranks from both sources were pooled and ranked, however ties were not broken. The resulting p-value of 0.1609 indicates a failure to reject null: there is no significant difference between Field Sonde and Grab Sample. The boxplot in **Error! Reference source not found.** shows a similar setup as the previous statistical test, however the median value for Field Sonde has been driven down due to the repeated ties for zero.

Mann-Whitney Test and CI: Turb-F (no 0.0), Turb-G_1 (no 0.0)

```
                    N  Median
Turb-F (no 0.0)    25    7.90
Turb-G_1 (no 0.0)  25    3.60

Point estimate for ETA1-ETA2 is 4.00
95.2 Percent CI for ETA1-ETA2 is (0.90,7.49)
W = 774.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0083
The test is significant at 0.0083 (adjusted for ties)
```

In the test above, all 0.0 NTU Field Sonde and corresponding Grab Sample values were removed from the analysis. The result is a significant p-value of 0.0083, indicating a difference between Field and Grab.

Mann-Whitney Test and CI: Turb-F_1 (random), Turb-G_1 (random)

```
                    N  Median
Turb-F_1 (random)  51   0.100
Turb-G_1 (random)  51   1.300

Point estimate for ETA1-ETA2 is -0.420
95.0 Percent CI for ETA1-ETA2 is (-0.631,1.602)
W = 2416.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1609
The test is significant at 0.1606 (adjusted for ties)
```

This test replaces all 0.0 values with values ranging from 0.01 – 0.09 NTU to account for a lack of resolution in the very low turbidity range. A p-value of 0.1609 still indicates a lack of significance – no difference between Field and Grab Samples.

Statistical Comparisons: Field Sonde vs. QAQC Sonde vs. Grab Sample

pH

Values for pH from QAQC and Grab samples are found to be normally distributed while Field sonde fails the normality test by a small margin (Table 18). Because Field just fails this test, there is some power in parametric testing, however, more weight should be given to results from Non-parametric testing.

Table 18: Normality Tests for Field, QAQC and Grab pH

Population	Anderson-Darling Statistic	p-value
Field Sonde	0.811	0.033
QAQC Sonde	0.486	0.216
Grab Sample	0.181	0.910

Kruskal-Wallis Test on Rank pH

```
Factor       N  Median  Ave Rank      Z
Field-pH    51   58.00      60.6  -3.24
Grab-pH     51   96.00      91.3   2.82
QAQC-pH     51   84.00      79.1   0.41
Overall    153              77.0

H = 12.40  DF = 2  P = 0.002
```

A Kruskal Wallis test on population medians indicates that, at a p-value of 0.002, there is a significant difference in medians between the three groups tested. From a Mann-Whitney test earlier, it is known that Field and Grab values are different; a test of Field vs. QAQC is performed below.

One-way ANOVA: pH_Ranks versus Collection Method

```
Source              DF      SS     MS     F      P
Collection Method    2   24350  12175  6.66  0.002
Error              150  274055   1827
Total              152  298405

S = 42.74   R-Sq = 8.16%   R-Sq(adj) = 6.94%

                         Individual 95% CIs For Mean Based on
                         Pooled StDev
Level   N   Mean  StDev  --------+---------+---------+---------+-
Field  51  60.57  42.75  (-------*-------)
Grab   51  91.24  43.01                      (-------*-------)
QA     51  79.20  42.47              (-------*-------)
                         --------+---------+---------+---------+-
                                60        75        90       105
Pooled StDev = 42.74

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Collection Method

Individual confidence level = 98.09%

Collection Method = Field subtracted from:
```

```
Collection
Method      Lower  Center  Upper  ---+---------+---------+---------+------
Grab        10.62   30.67  50.72                  (-------*-------)
QA          -1.42   18.63  38.68             (-------*-------)
                                  ---+---------+---------+---------+------
                                   -25         0        25        50
Collection Method = Grab subtracted from:
Collection
Method       Lower  Center  Upper  ---+---------+---------+---------+------
QA          -32.09  -12.04   8.01  (-------*-------)
                                   ---+---------+---------+---------+------
                                    -25         0        25        50
```

The result of the Tukey's family comparison, above, indicates that Grab Sample pH > Field Sonde pH. No inference can be made between the relationship of Field pH and QAQC pH or Grab pH and QAQC pH based on the above result.

Mann-Whitney Test and CI: Field-pH, QAQC-pH

```
           N  Median
Field-pH  51  6.3000
QAQC-pH   51  6.5800

Point estimate for ETA1-ETA2 is -0.2500
95.0 Percent CI for ETA1-ETA2 is (-0.4699,-0.0201)
W = 2295.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0270
The test is significant at 0.0270 (adjusted for ties)
```

The test above indicates that there is a significant difference detected between Field and Grab pH values. QAQC pH values are found to be significantly higher than concurrently measured Field pH values.

Mann-Whitney Test and CI: QAQC-pH, Grab-pH

```
          N  Median
QAQC-pH  51  6.5800
Grab-pH  51  6.6400

Point estimate for ETA1-ETA2 is -0.1600
95.0 Percent CI for ETA1-ETA2 is (-0.3701,0.0599)
W = 2407.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.1436
The test is significant at 0.1436 (adjusted for ties)
```

In the test above, it is determined that there is no significant difference between QAQC Sonde and Grab Sample pH values.

Mann-Whitney Test and CI: Field-pH, Grab-pH

```
           N  Median
Field-pH  51  6.3000
Grab-pH   51  6.6400

Point estimate for ETA1-ETA2 is -0.4100
95.0 Percent CI for ETA1-ETA2 is (-0.6400,-0.1801)
```

```
W = 2119.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0007
The test is significant at 0.0007 (adjusted for ties)
```

The Mann Whitney test above rejects the null hypothesis at a p-value of 0.0007 – A significant difference exists between Field and Grab sample pH values.

Specific Conductivity

According to normality testing, Field, QAQC and Grab samples fail normality testing (Table 19). A non-parametric Kruskal-Wallis test is used.

Table 19: Normality Tests for Field, QAQC and Grab Conductivity

Population	Anderson-Darling Statistic	p-value
Field Sonde	11.704	< 0.005
QAQC Sonde	11.337	< 0.005
Grab Sample	11.697	< 0.005

Kruskal-Wallis Test: Rank versus Factor

```
Kruskal-Wallis Test on Rank

Factor         N  Median  Ave Rank      Z
Field-Cond    51   65.00      73.2  -0.75
Grab-Cond     51   94.00      85.0   1.59
QA-Cond       51   70.00      72.7  -0.84
Overall      153              77.0

H = 2.52  DF = 2  P = 0.284
```

The test above indicates that, at a p-value of 0.284, the null hypothesis fails to be rejected: there is no significant difference between Field, QAQC and Grab Sample conductivity.

Turbidity

Anderson-Darling tests find that the turbidity distributions for Field, QAQC and Grab samples are not normal (Table 19).

Table 20: Normality Tests for Field, QAQC and Grab Turbidity

Population	Anderson-Darling Statistic	p-value
Field Sonde	10.566	< 0.005
QAQC Sonde	10.536	< 0.005
Grab Sample	10.227	< 0.005

Turbidity: Raw

Kruskal-Wallis Test: Rank versus Factor

```
Kruskal-Wallis Test on Rank

Factor         N  Median  Ave Rank      Z
```

```
Field-Turb   51   52.00      74.0  -0.60
Grab-Turb    51   82.00      86.2   1.82
QA-Turb      51   49.00      70.8  -1.22
Overall     153              77.0

H = 3.44  DF = 2  P = 0.179
```

The test above indicates a failure to reject the null hypothesis at a p-value of 0.179 – there is no significant difference between Field, QA and Grab sample turbidity.

Kruskal-Wallis Test: Field-Turb_1 versus QA-Turb_1

```
Kruskal-Wallis Test on Field-Turb_1

QA-Turb_1       N  Median  Ave Rank       Z
Field-Turb_1  23   8.000      41.2   1.81
Grab-Turb_1   23   3.600      24.3  -3.14
QA-Turb_1     23   7.700      39.5   1.32
Overall       69              35.0

H = 9.93  DF = 2  P = 0.007
H = 9.93  DF = 2  P = 0.007  (adjusted for ties)
```

The test above assesses uses natural (unranked) data from Field, QAQC, and Grab samples with all 0.0 NTU sonde values and concurrent Grab Sample values removed. The test observes a difference between the groups at a p-value of 0.007.

Turbidity: 0.0 NTU removed

Table 21: Normality Tests for Field, QAQC and Grab Turbidity (no zeros)

Population	Anderson-Darling Statistic	p-value
Field Sonde	0.318	0.515
QAQC Sonde	0.321	0.510
Grab Sample	0.760	0.041

Kruskal-Wallis Test on No zero-rank

```
No zero-Factor       N  Median  Ave Rank       Z
No zero-Field-Turb  23   43.00      41.2   1.81
No zero-Grab-Turb   23   20.00      24.3  -3.12
No zero-QA-Turb     23   39.00      39.5   1.31
Overall             69              35.0

H = 9.81  DF = 2  P = 0.007
```

The test above indicates that, when zero values are removed from QAQC and Field Sonde groups (with corresponding values also removed from the Grab Sample group) a statistically significant difference is found. Below, pariwise Mann-Whitney tests determine which groups differ from one another.

Mann-Whitney Test and CI: No zero-Field-Turb, No zero-QA-Turb

```
                     N  Median
No zero-Field-Turb  23    8.00
No zero-QA-Turb     23    7.70
```

```
Point estimate for ETA1-ETA2 is 0.30
95.2 Percent CI for ETA1-ETA2 is (-4.50,4.92)
W = 552.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.8090
The test is significant at 0.8090 (adjusted for ties
```

Fail to reject null hypothesis: there is no difference between median Field and QAQC turbidity values.

Mann-Whitney Test and CI: No zero-Field-Turb, No zero-Grab-Turb

```
                     N  Median
No zero-Field-Turb  23    8.00
No zero-Grab-Turb   23    3.60

Point estimate for ETA1-ETA2 is 4.40
95.2 Percent CI for ETA1-ETA2 is (1.20,9.00)
W = 671.5
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0041
The test is significant at 0.0041 (adjusted for ties)
```

Reject null hypothesis: there is a significant difference between median Field and Grab sample turbidity values.

Mann-Whitney Test and CI: No zero-QA-Turb, No zero-Grab-Turb

```
                    N  Median
No zero-QA-Turb    23    7.70
No zero-Grab-Turb  23    3.60

Point estimate for ETA1-ETA2 is 4.10
95.2 Percent CI for ETA1-ETA2 is (1.09,13.51)
W = 656.0
Test of ETA1 = ETA2 vs ETA1 not = ETA2 is significant at 0.0115
The test is significant at 0.0115 (adjusted for ties)
```

Reject null hypothesis: there is a significant difference between median QAQC and Grab sample turbidity values

Turbidity: 0.0 NTU values replaced randomly with 0.01 – 0.09 NTU

Kruskal-Wallis Test: Rand-Turb versus Rand-Turb-Factor

```
Kruskal-Wallis Test on Rand-Turb

Rand-Turb-Factor    N   Median  Ave Rank      Z
Rand-Field-Turb    51  0.09000      73.7  -0.66
Rand-Grab-Turb     51  1.30000      86.2   1.82
Rand-QA-Turb       51  0.09000      71.1  -1.16
Overall           153               77.0

H = 3.39  DF = 2  P = 0.184
H = 3.39  DF = 2  P = 0.184  (adjusted for ties)
```

Fail to reject null: there is no difference between Field, QAQC and Grab turbidity values.