APPENDIX B2- CORE RECEIVING ENVIRONMENT MONITORING PROGRAM (CREMP), DESIGN DOCUMENT, VERSION 1 (DEC. 2012)

Version 1
Project No. AE-10-03
December 2012

Core Receiving Environment Monitoring Program (CREMP): Design Document 2012

Meadowbank Division, Nunavut

Prepared for:

Agnico-Eagle Mines Ltd.
Meadowbank Division

P.O. Box 540
Baker Lake,
Nunavut X0C 0A0

Prepared by:

AZIMUTH

Azimuth Consulting Group
218-2902 West Broadway Vancouver,BC V6K 2G8
Tel: 604-730-1220 • Fax: 604-739-8511

VERSION 1

Core Receiving Environment Monitoring Program (CREMP): Design Document 2012

Meadowbank Mine

Prepared for

Agnico-Eagle Mines Ltd.
Meadowbank Division
P.O. Box 540
Baker Lake, Nunavut X0C 0A0

December 2012

Azimuth Consulting Group Inc.
218-2902 West Broadway
Vancouver, BC
V6K 2G8

Project No. AE-10-03

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

DOCUMENT CONTROL

Version	Date (YMD)	Section	Page	Revision
1	2012-12	All	All	Core Receiving Environmental Monitoring Program (CREMP) Design; updates portions of document Core Receiving Environmental Monitoring Program (CREMP) 2010 Plan Update - Meadowbank Gold Project (Version 1)

Version 1:
Reviewed/approved by:

Ryan VanEngen
Environmental Biologist- Meadowbank

Approved by:

Kevin Buck
Environmental Superintendent

1. INTRODUCTION AND OBJECTIVE

This document presents a revised design for the Core Receiving Environment Monitoring Program (CREMP). The CREMP, which was formerly referred to as the Aquatic Effects Management Program (AEMP) for Agnico-Eagle Mines' (AEM) Meadowbank Gold Mine and was included as part of the Environmental Assessment (EA) for the project in 2005 (AEMP 2005), has been formally implemented since 2006. The Type A water license (2AM-MEA0815) for the project issued by the Nunavut Water Board (NWB) in 2008 required a revised AEMP, and specified some of the requirements for that revision. Most importantly, while the 2005 AEMP was focused only on receiving environment studies at the level of basins and lakes, the revised AEMP (Azimuth 2012) needs to be broader in scope to comply with the following licence requirements (stipulated in Part I-1):

- *A detailed monitoring protocol to verify that the Canadian Council of Ministers of Environment Fresh Water Aquatic Life Guidelines are met thirty (30) metres from the outfall diffusers*[1]*;*
- *Annual reporting for more immediate adaptive management*[2]*;*
- *Mechanisms to measure changes to productivity in the lake as a result of the mine adding nutrients*[3]*;*
- *Sampling and Analysis Plans*[4]*; and*
- *Monitoring under Fisheries Authorizations, NWB Licence Compliance Monitoring, Environmental Effects Monitoring, and Groundwater Monitoring.*

The last requirement diverges from traditional AEMPs (INAC, 2009) and required AEM to propose a new approach, which was presented in draft to the NWB (March 2-3, 2010 in Yellowknife) and necessitated the restructuring of the AEMP. As a result the AEMP was restructured to serve as an overarching 'umbrella' that conceptually provides an opportunity to integrate results of individual, but related, monitoring programs in accordance with the Type A water license requirements. The scope of the 2005 AEMP, which was essentially the core receiving environment monitoring, is now one of the monitoring programs that is integrated under the restructured AEMP and has been

[1] This component is included in quarterly environmental effects monitoring (EEM) receiving environment monitoring under the *Metal Mines Effluent Regulation*.

[2] This applies to most monitoring programs. Some programs, such as the Effects Assessment Studies (EAS) are conducted only when needed.

[3] This is conducted as part of the Core Receiving Environment Monitoring Program (CREMP).

[4] This is part of the CREMP and other programs.

renamed the Core Receiving Environment Monitoring Program (CREMP) to minimize confusion.

The CREMP is the core, broad scale program that is aimed at detecting potential impacts at the scale of lakes or basins. It is intended to monitor large-scale basin-wide changes in physical and biological variables to evaluate potential impacts from all mine related stressors to the receiving environment. It therefore serves as the most important monitoring program for evaluating short-term and long-term potential impacts, for which other programs provide additional support and verification. In 2011, AEM submitted an updated CREMP design (Azimuth, 2010), which superseded the original design (AEMP 2005). At the same time, AEM commissioned Azimuth to begin a thorough review of the historical data to ensure that the CREMP design would be able to detect potential mine-related impacts and would address most of the requirements in the Type A water license listed above.

This document presents the revised CREMP design, which aims to ensure that the data generated from the CREMP will be adequate for detecting potential mine-related impacts in a timely manner, and to determine whether management actions (i.e., further assessment or mitigation) are warranted. The document is organized as follows:

- **Section 2** briefly describes the components of the CREMP and the available data.
- **Section 3** describes the methodology for development of criteria for action (i.e., determining when action should be taken based on results of monitoring), and applies that methodology to the components of the CREMP.
- **Section 4** presents the experimental design for evaluating CREMP data, and applies that design to current data to determine optimal sampling intensities.
- **Section 5** summarizes the implications for CREMP sampling of the analyses in Section 4.
- **Section 6** discusses links to the broader AEMP including how CREMP findings are used in identifying whether management actions are warranted.

Note to reader: the majority of the statistical analyses conducted for this report were conducted on data collected through part or all of 2010. Some components (e.g., zooplankton) included data through 2011.

2. CREMP COMPONENTS AND DATA

2.1. CREMP Components and Monitoring Variables

The components (and associated parameters) of the CREMP are derived from conditions specified in the Nunavut Water Board A Licence # 2AM-MEA0815 and Nunavut Impact Review Board Project Certificate – NIRB- 004 for the Meadowbank Gold Mine, which is based on the following:

- Nuvavut Water Board (NWB; i.e., board decision after public hearings).
- Baseline AEMP approach (AEMP, 2005) (i.e., the approach for monitoring presented during the Nunavut Impact Review Board [NIRB] environmental impact assessment process).

The CREMP includes the following components and variables:

- Water chemistry – All parameters required by the Type A water license for the receiving environment water (Schedule I, Table 1, Group 4). This list includes total metals, dissolved metals, several anions and nutrients, and several conventional parameters. Water chemistry is the foundation of the CREMP as it is the first indicator of potential effects to the receiving environment. Furthermore, water chemistry results collected during the CREMP are very important as the results are assessed collectively in the AEMP (Azimuth 2012) together with other mine-related surface water sampling (i.e., NWB license A and MMER) and groundwater sampling data.
- Sediment chemistry – Sediment chemistry is an important CREMP component because many contaminants are more likely to be associated with sediment than with water. In particular, sediment chemistry trends can also be an important indicator of longer-term patterns. While the main focus is on metals for which federal (CCME) guidelines exist (i.e., arsenic, cadmium, chromium, copper, lead, mercury, and zinc), data for other metals are also reported. Phytoplankton – Phytoplankton data are collected as part of the CREMP as they are a key indicator of pelagic primary productivity. Specific metrics including total biomass and total number of species (as a measure of richness) are measured and used for monitoring purposes.
- Benthic invertebrate community – Benthic invertebrate community data are collected as an indicator of potential effects on the benthic environment. Specific metrics including total abundance and richness (as total taxa) are measured and used for monitoring purposes.
- Zooplankton – Zooplankton, an indicator of secondary productivity, were initially excluded from the program (AEMP 2005) due to high variability. They were then

added to the CREMP for 2010 and 2011 to formally assess their suitability as a monitoring endpoint. Specific metrics including total biomass and richness are measured and used for monitoring purposes.

- Periphyton – Periphyton, which represent primary productivity on hard substrates, were initially included in the program to characterize baseline conditions across the project lakes. The available data were included herein to formally assess their utility as a monitoring tool. As with phytoplankton, specific metrics including total biomass are measured and used for monitoring purposes.

2.2. CREMP Stations and Control/Impact Designations

The location of CREMP stations for Meadowbank and Baker Lake are shown in **Figure 1** and **Figure 2**, respectively. **Table 1** summarizes the years for which data are available for each station, and distinguishes general designations for 'control' and 'impact' data. Given that TSS has been the major driver distinguishing 'control' data from potential 'impact' data, more detail is provided for 2008 and 2009 in **Table 2**. Some key observations regarding the stations are as follows:

- Collection of data at Baker Lake began in 2008, and activities that could cause potential impacts were already present in that year. Therefore, there are no true 'before' data for Baker Lake stations. To reduce uncertainties in the analyses, an additional reference station was added in 2012.
- There are two reference stations for Meadowbank (INUG and PDL) that are removed from the mine area and could be expected to function as controls in perpetuity, but data collection for PDL began only in 2009. Two other stations are currently used as controls but could potentially be impacted in the future – TPS is currently an internal control in Third Portage Lake, while TEFF is currently a far-field station that has not been impacted.
- A reference station for Wally Lake has not been established. While the characteristics of Wally Lake are somewhat unique (it is much shallower than other lakes), further evaluation of the advantages and disadvantages of establishing a separate reference station (versus using existing reference stations) is needed in advance of the commencement of construction activities at Wally Lake.

The designation of a station as 'control' or 'impact' in a given time period is dependent on the variable[5] being analyzed and the type of analysis. For example, a given station

[5] In general, we refer to any measured entity (e.g., cadmium concentration or phytoplankton biomass) as a variable. In some cases, this term can be used interchangeably with others (e.g., "parameter" is commonly

may be considered as 'impact' for one time period, but as 'control' for subsequent time periods if recovery occurs. If recovery occurs quickly for one variable (e.g., surface water TSS) and more slowly for another variable (e.g., benthic community abundance) then a given station/month or station/year combination may be 'control' for some variables and 'impact' for other variables. Furthermore, even if a station/month or station/year combination is conceptually considered to be 'impact' for a particular variable, if the data shows no impacts then data for that station/month or station/year combination may be pooled with 'control' data to strengthen the data set. This approach has been used in this report occasionally (**Section 4**), but will be less important in future years when the data set is larger.

2.3. Data Summary

This section summarizes the data that were available and were used for analyses to support the CREMP design revision. In general, the data that were analyzed as part of the design process included data from 2006-2010 collected specifically for the CREMP re-design.

- Water chemistry data up to and including May 2010 are summarized in **Table 3**. Of the 352 total samples, 76 were depth replicates (i.e., bottom or integrated samples collected at the same time/location as the standard surface sample) and 30 were duplicate samples (a second sample collected at the same time, location and depth as the standard sample). The intensity of sampling was increased beginning in May 2009, in part to develop a better understanding of the sources of variability in these data. The water chemistry data are discussed in more detail in **Section 4.3**.
- Sediment chemistry data up to and including 2009 are summarized in **Table 4**. In this case, the limited number of duplicate samples are not included in the table, but are discussed in **Appendix B**. Importantly, the sediment data include both core sample data (i.e., sediment cores where the top 1 cm of a core is submitted for analysis) and grab sample data (i.e., Petite-Ponar grab samples with a penetration depth of 5 cm or more). While both sets of data are considered, the core results are more relevant for monitoring because it is the thin surface layer of sediment where potential mining-related impacts are most likely to be detected. Some of the core sample data reflected in the table were collected as part of the East Dike Effects Assessment Study (EAS), rather than for the CREMP

used for water quality concentrations and "endpoint" for biological values). In other contexts (e.g., statistical modelling where "variables" [e.g., dependent and independent] and "parameters" [e.g., slope and intercept] have different meanings) they are not interchangeable.

specifically – these data are included as they are comparable and help to augment the limited CREMP data set.

- The phytoplankton monitoring data set is similar to the water chemistry data set, with a total of 308 samples collected for the CREMP from 2006 through to September 2010 (see **Table 5**; not including field duplicate samples and depth replicates).
- Benthic invertebrate monitoring data have been collected as part of the CREMP once each year in August from 2006 to 2010. There have been 242 'standard' samples collected (see **Table 6**), which excludes a few samples because of differences in mesh size, sampling procedures, and inadequate sample preservation.
- The zooplankton monitoring data for the CREMP are limited to 45 samples collected in August 2010 (see **Appendix E**).
- The periphyton data for the CREMP are limited to 70 samples (5 spatial replicates at each of 7 Meadowbank stations in 2007 and 2008; see **Appendix F**).

3. DEVELOPMENT OF CRITERIA FOR ACTION

Implementation of the Management Response Plan for the AEMP requires the development of criteria for action for the CREMP and other AEMP programs.

3.1. Conceptual Early Warning Framework

For the CREMP, criteria for action have been developed with the assumption that action will be considered *before* certain monitored parameters reach levels that cause or have the potential to cause adverse effects to aquatic biota. The criteria for action should provide an early warning framework under which management responses may be considered, taking into account findings from other AEMP component programs. Two types of criteria are developed:

- *Thresholds* – license limits, regulatory guidelines or other discrete benchmarks, below which unacceptable adverse effects are not expected and above which unacceptable adverse effects may occur. If thresholds do not exist or are not used for a particular variable, then early warning triggers will be developed *without* thresholds.
- *Triggers*- site specific early warning criteria that lead to action. In cases where thresholds are established, the triggers are set at values that are more conservative than the thresholds. Triggers ensure that action is taken before a threshold has

been reached. . For variables where no thresholds exist, the triggers are set using statistical methods based on existing data.

The conceptual framework of thresholds and triggers is applied to each of the monitoring variables listed in the subsections below.

3.2. Water Chemistry

3.2.1. Methods for Development of Thresholds and Triggers

At a federal level, the CCME has established water quality guidelines for key water chemistry variables. The CCME guidelines are generally appropriate for use as thresholds because they have a toxicological basis and are relatively conservative. Where CCME guidelines do not exist, or where the underlying science is outdated or not applicable, published guidelines or standards in other jurisdictions can be considered.

Where CCME water quality guidelines have been developed for metals, the guidelines are assumed to apply to total metals (i.e., dissolved and particulate-bound metals) if not explicitly stated in CCME documents. This approach is conservative, since most toxicological studies use metal compounds that are more soluble and therefore more bioavailable than metals in typical waterbodies. The guidelines are not explicitly applied as thresholds for dissolved metals (which is unnecessary if they already apply to total metals). However, any evaluation of the potential effects of a metal would consider the metal in dissolved or particulate forms.

The method of developing thresholds and triggers for water chemistry parameters is shown in **Figure 3**. Application of the triggers is also shown in the figure but is not discussed in detail until **Section 4.3**. There are three methods of trigger development as follows:

1. When a threshold (e.g., CCME guideline) is established, the trigger was set as the maximum of either (a) the value halfway between the baseline median and the threshold ("Method A"), or (b) the 95th percentile of the baseline data ("Method B").
2. When a threshold is not established, the trigger was set equal to the 95th percentile of the baseline data ("Method B"), except in cases where less than 5% of the data exceeded the current detection limit (DL) – in the latter case, the trigger was set equal to two times the DL ("Method C").

There were exceptions to the above approach for a few special cases, specifically t-Al, t-Cd, t-Mn, t-Zn, d-Al, ammonia-N, t-P, pH and TSS. These exceptions are explained in detail in **Appendix A**.

The data set upon which triggers were developed excluded duplicates and depth replicates. The rationale for restricting the data set in these ways is two-fold. First, variation among duplicates and depth replicates is small compared to variation among samples and stations, therefore inclusion of duplicates and depth replicates results in little gain in our ability to detect potential impacts. More importantly, duplicates and depth replicates are considered as pseudoreplicates in the experimental design framework used for data analysis (see detailed discussion in **Appendix A**). In addition, individual data points for which values were below DL but for which the DL was higher than usual were excluded. Remaining data points with values below the usual DL were set equal to the DL.

3.2.2. Results for Thresholds and Triggers

Thresholds were established for 22 variables based on water quality guidelines – these are listed in **Table 7**. Triggers for total metals, dissolved metals and nutrients/conventionals are summarized in **Table 8**, **Table 9**, and **Table 10** respectively. Importantly, in some cases the 95th percentile of baseline data is above the threshold, in which case the trigger is developed using Method B and is greater than the threshold.

Separate thresholds and triggers are developed for Meadowbank and Baker Lake where appropriate, because Baker Lake is influenced by seawater and is therefore fundamentally different from the Meadowbank lakes. This is evidenced in many of the variables that reflect seawater such as Na, Cl, Ca, Mg, hardness, conductivity.

Further differentiation of triggers beyond the Baker/Meadowbank distinction is not pursued. As discussed in **Appendix A**, there are some notable differences among stations for some variables, with Wally Lake in particular having higher values for some variables. These differences could be used as rationale for development of station-specific triggers. We don't believe this is warranted, for two reasons. First, we would expect that a range for a variable that is normal for one of the project lakes is unlikely to cause effects in another nearby lake, even if the second lake has naturally lower levels of that variable. Ecological data for all of the project lakes suggests that they are all functioning with reasonably similar primary, secondary and fish species assemblages. Second, while the data set includes stations with higher natural concentrations of some variables (e.g. Wally Lake), those are balanced to some degree by stations with lower natural levels (e.g., INUG), thus we do not expect that the data are biased. Nevertheless, the derived triggers may be somewhat less or more conservative than planned for any given station. This is not an issue for variables with thresholds for which triggers are developed using Method A, because in those cases triggers will serve their early-warning purpose regardless. However, under Method B the 95th percentile of baseline data would vary depending on which stations are included in the baseline data. Since the use of the 95th percentile for Method B is somewhat arbitrary anyway, we don't think it is of critical importance that the triggers are somewhat less or more conservative for particular

stations, particularly since the CREMP water chemistry data are not evaluated in isolation – other components of the CREMP (e.g., biological effects variables) and other AEMP programs are also used to detect potential mine-related impacts on the receiving environment. In short, the complexity associated with developing, tracking and applying several hundred station-specific, variable-specific triggers does not seem warranted.

3.2.3. Time Frame for Trigger Application

The water license currently requires monthly sampling at CREMP stations including external reference stations during the open water season, and at a smaller number of representative stations near the mine site in the winter (through ice). Water chemistry is expected to demonstrate natural seasonal variation, which would be captured by monthly year-round sampling. However, the statistical power to detect effects using a BACI approach does not improve much beyond 6 months of water chemistry data, and therefore the value of monthly sampling year round is diminished. Furthermore, there are several seasonal constraints on sampling that include:

- The major mine-related contaminant inputs are expected to be from effluent discharges and dust (blasting and general), both of which are expected to occur during the open water season (i.e., mid-July through early October). While the generation and deposition of dust will occur year round, the two-meter ice cap will prevent contact with water for most of the year. Changes in water quality related to dust would be expected to be highest in the spring (at TPE and SP based on proximity to blasting and prevailing wind direction), but possibly offset by runoff-related dilution.
- Winter sampling of reference areas presents significantly higher risks to the sampling team due to darkness, lack of air (helicopter) support, and the need to cover long distances via tundra buggy. Furthermore, transport via tundra buggy is often impractical in the shoulder seasons when snowpacks are marginal (e.g., November, March).
- Lack of sampling of reference areas during winter months diminishes the value of collecting any winter data. Not having paired data (i.e., collection of data at both reference and exposure areas) reduces the ability to distinguish between potential naturally-occurring seasonal effects and mine-related effects (i.e., any observed effects must be assumed due to the mine, when in fact they may be reflective of natural variation. . Furthermore, un-paired winter data provide no statistical improvement for the BACI model. Currently, the available winter data for reference areas are extremely limited, and would continue to be limited due to the safety concerns.
- Sampling in June, October and early November is highly dangerous due to thin ice conditions.

In light of the above, we propose to collect and analyze 6 months (April, May, July, Aug, Sept, plus November or December) of full water chemistry data for the annual period of paired sampling to support BACI analyses. This will maintain adequate statistical power to detect potential impacts at the scale of lakes or basins in the receiving environment while ensuring safe sampling conditions. In any given year the actual number of samples that would be collected may range from 4 to 6 events.

Nevertheless, there is a current licence requirement for monthly sampling throughout the year (including in winter through the ice) at a subset of the CREMP areas nearest to the mine-site. This will continue as it is recognized that failing to sample for a period of several months during the winter could be problematic if there are serious changes (e.g., low oxygen, changes to stratification, etc.) that might be slow to manifest but be related to mine activities. Consequently, basic field water quality data will be collected at near-field areas (i.e., TPN, TPE, SP and eventually Wally) at least once mid-winter. These data would not add to the statistical power of the CREMP analyses, but would reduce uncertainty regarding potential changes during the winter period. Qualified technicians or biologists will collect all field data and if changes are observed, they may collect the full suite of water chemistry parameters for comparison to the triggers at that time.

The implications of monitoring water quality six times per year is explored in **Appendix A** and **Section 4.3**. In **Figure 3**, the application of trigger values to data for one month and annual (based on six months) is depicted, whereby an exceedance for one month is a flag and an exceedance for an annual (six-month) mean is the early warning trigger for action.

3.3. Sediment Chemistry

3.3.1. Methods for Development of Thresholds and Triggers

At federal level, the CCME has established sediment quality guidelines for the seven metals which are formally evaluated in the CREMP. The CCME guidelines are generally appropriate for use as thresholds because they have a toxicological basis and are relatively conservative.

The method of developing thresholds and triggers for water quality variables is shown in **Figure 3**. Application of the triggers is also shown in the figure but is not discussed in detail until **Section 4.3**. Since all sediment chemistry parameters had thresholds, trigger development is more straight forward than for water chemistry. For each sediment chemistry parameter, the trigger was set as the maximum of either (a) the value halfway between the baseline median and the threshold ("Method A"), or (b) the 95th percentile of the baseline data ("Method B").

The data set upon which triggers were developed focused on core samples[6], with duplicates excluded. In addition, individual data points for which values were below DL but for which the DL was higher than usual were excluded. Remaining data points with values below the usual DL were set equal to the DL. **Appendix B** summarizes the data.

3.3.2. Results for Thresholds and Triggers

Thresholds were established for 22 variables based on water quality guidelines – these are listed in **Table 11**. Triggers are summarized in **Table 12**. Importantly, in some cases the 95th percentile of baseline data is above the threshold, in which case the trigger is developed using Method B and is greater than the threshold. Separate thresholds and triggers are developed for Meadowbank and Baker Lake where appropriate.

Further differentiation of triggers beyond the Baker/Meadowbank distinction is not pursued. As discussed in **Appendix B** (and similar to analysis for water chemistry in the previous section), there are some notable differences among stations for some parameters; these differences could be used as rationale for development of station-specific triggers. We don't believe this is warranted, for the same reasons stated in **Section 3.2.2**

3.3.3. Time Frame for Trigger Application

Sediment chemistry data are not expected to vary in the manner that would be expected for water chemistry. Changes to sediment chemistry may occur as a result of limited duration events, but those changes would not be expected to dissipate quickly as they might with water chemistry. Consequently, sampling frequency for sediment chemistry as part of the CREMP should be less frequent. We propose that sampling take place approximately every three years. To maximize the potential utility of the data, sampling could be aligned with the sampling times for benthic invertebrates required for the EEM program. Thus, sediment chemistry cores were collected in 2012 (because there have been no core data collected since 2009), and would in future be matched to the envisioned EEM sampling years which are 2014 and then every three years (depending on results in 2014). During each sampling event, there should be multiple subsamples in each monitoring area (see **Section 4.4**).

[6] Sediment chemistry is also measured in grab samples to support benthic invertebrate community monitoring (i.e., benthos findings from grab samples are matched to chemistry grab samples).

3.4. Biological Effects Variables

3.4.1. Thresholds and Triggers

There are no established guidelines for biological variables that can be used directly as absolute thresholds and triggers for the Meadowbank Gold Mine. Unlike water or sediment, where environmental quality guidelines can be used to develop thresholds or triggers, there are no universal benchmarks for biological variables such as abundance, biomass or diversity. Rather, the magnitude of change or difference relative to expected conditions must be used to establish "critical effect sizes" (CES) for biological variables.

In a review of methods for deriving CES, Munkittrick et al. (2009) noted that "the ideal method for setting CES would be to define the level of protection that prevents ecologically relevant impacts". The authors concluded that there are very few examples of CES being developed on the basis of ecological relevance. They summarize alternative methods, which include identifying values that would represent the extreme of natural variability among or within reference areas. Recent guidance for metal mines under EEM (Environment Canada 2012a) attempts to define CES for fish and for benthic invertebrates that would be associated with effects. For fish endpoints, the CES are expressed as a percentage deviation from the reference mean (e.g., 10% for condition factor; 25% for other variables). For benthic invertebrate community endpoints (e.g., density, richness), the CES are defined as +/- two standard deviations where the standard deviation is derived from within-reference-area data.

The CREMP should aim to define CES that are ecologically relevant. There is precedent for defining ecologically relevant CES in related applications in Canada:

- CCME water quality guideline derivation process – CCME (2007) defined thresholds for "no effect" (i.e., an effect on 10% or less of exposed individuals) and "negative effect" (i.e., an effect level on more than 15–20% of the exposed individuals of a species).
- General risk assessment practice in Canada – ecological risk assessments often involve the quantification of adverse effects related to environmental contamination in aquatic environments. Common effect measures are *in vitro* toxicity tests or *in situ* biological studies (e.g., field-based metrics related to benthic invertebrates or fish). Recent draft guidance for the Federal Contaminated Sites Action Plan (Environment Canada 2012b) does not specify acceptable CES for any endpoints, largely to provide flexibility and to allow for site-specific considerations (e.g., local biology, input from regulators or stakeholders). Effects to non-listed (i.e., not threatened or endangered) species in Ontario (2005) are assessed relative to a lowest observable effects level (LOEL; usually defined as the lowest effect statistically significant from the control group); by definition this may actually represent a range of effect magnitudes, but could be well above

20%. In British Columbia (SAB, 2008), site-specific toxicity tests are typically interpreted based on a 20% effect size (e.g., an inhibition concentration [IC_{20}] that corresponds to a 20% reduction in a continuous endpoint relative to the control or to reference samples). Measured or predicted effects are generally reported relative to the 20% effect threshold (i.e., below which risks would be considered "acceptable") and possibly to higher levels (e.g., 50% effects) if warranted to distinguish degrees of unacceptability (e.g., between marginally and clearly unacceptable, which may be communicated as moderate to high risk). Confidence in the effect sizes is usually addressed through a discussion of uncertainty on a site specific basis.

For purposes of exploring the statistical power of the CREMP, a 50% effect size was defined as the threshold of interest and a 20% effect size as the trigger. However, the terms "threshold" and "trigger" for biological variables are not used as strictly as for water and sediment chemistry parameters , because:

1) *Statistical Power* - For most biological variables, natural variability can make it difficult (if not impossible) to statistically detect effect sizes as low as 20%. It is more realistic to detect larger effect sizes such as 50%.

2) *Causality* – Even if statistically-significant changes are documented (at whatever effect size), the cause of the change needs to be understood in order to effectively manage the situation. For the Meadowbank biological data, effect sizes exceeding 50% have been observed due to natural causes in the baseline data.

Nevertheless, although effect sizes of 20% may be easily observed by chance, it is worth evaluating biological data in more detail whenever small effect sizes such as 20% are observed. The method of developing thresholds and triggers for biological variables is shown in **Figure 3**. Application of the triggers is also shown in the figure but is not discussed in detail until **Section 4.3**. **Appendices C to F** summarize the data and present analyses of four sets of biological variables: phytoplankton, benthic invertebrate community, zooplankton, and periphyton. Results for the analyses of statistical design are provided in **Section 4.5**.

3.4.1.1. Comparison of CREMP and EEM Benthic Invertebrate Effect Sizes

Given their different purposes and application, critical effect sizes may differ between an EEM program and a CREMP (INAC 2009). Nevertheless, it is important to understand the differences so that the findings from CREMP and EEM studies can be interpreted together..The MMER defines CES for fish on the basis of percentage change (10% or 25% depending on the variable), and for benthos on the basis of +/- two (within-reference area) standard deviations from the reference mean. The CREMP, meanwhile, defines CES for all biological variables on the basis of percentage change (50%, with 20% as an

early warning trigger). Furthermore, the CREMP does not include fish, and EEM does not include other biological components of the CREMP such as phytoplankton .Thus, the only component of overlap for Meadowbank EEM and CREMP monitoring is benthic invertebrate monitoring.

The CREMP uses percent change rather than standard deviations for benthos, in an effort to maintain a transparent (fixed) effect size that is more likely to be ecologically relevant. Where natural variability is high, use of two standard deviations for benthos could potentially mean that large and ecologically-relevant effects could occur to some endpoints without being higher than the CES. On the other hand, the limitation of using percentage change to define the CES for benthos when variability is high is reduced statistical power to detect change. However, this disadvantage is not nearly as important when the experimental designs for EEM and CREMP are compared. The experimental design for EEM is fundamentally a control-impact (CI) design, with no before-after (BA) temporal component. The intent of EEM is to have a design that will detect effects in a given study (year), and if effects are detected in two consecutive monitoring events (36 months apart) then the cause must be investigated. The fundamental weakness of the CI design is the underlying assumption that any differences between C and I areas are considered an effect. Based on data evaluation for Meadowbank and our experience at other sites, this assumption is weak and can lead to erroneous conclusions regarding natural differences between C and I areas (i.e., falsely concluding that a difference between areas is mine related). Before-after-control-impact (BACI) designs are more robust in that they track changes in C and I areas over time; the target effect is differential change at I relative to C over time. Many EEM programs do not collect 'before' data as they only become subject to MMER after mine-related discharge occurs. In the BACI framework, the power to detect differences is greater when there are more monitoring events in the B and A periods included in the analysis[7].

In summary, EEM data are evaluated for a particular period (year), using a CI design that assumes that all differences between the two areas are mine-related (unless proven otherwise, such as by incorporating a physical covariate such as grain size). EEM's CES (+/- 2SD) typically results in reasonable statistical power, but may do so at the cost of allowing large and ecologically-relevant effects to occur In contrast, the CREMP started

[7] From a statistical standpoint, the CI design treats replicate samples in an area as true replicates. In a BACI design, where natural differences among areas or over time are acknowledged, the replicate samples collected in an area in a given time period are pseudoreplicates because each area in each time period is a replicate. The high power achieved by CI designs is therefore false if we admit that areas may differ naturally. For statistical tests, in the CI design the degrees of freedom will be based at least in part on the number of samples in an area, whereas for the BACI design the degrees of freedom will be based on the number of areas and time periods without consideration for the number of samples taken in a given area in a given period.

prior to mining, thus allowing the incorporation of baseline data into a BACI design that effectively characterizes natural spatial and temporal variability. Overall, the CREMP is capable of detecting large impacts in a short time period, but requires longer time periods to detect more subtle effects.

3.4.2. Time Frame for Trigger and Threshold Application

The time frame for trigger and threshold application must vary according to the type of biological variables that are monitored. For phytoplankton, which we expect to respond to stresses on a relatively short-term basis, sampling will be similar to that for water chemistry parameters (see **Section 3.2.3**) – up to 6 months of sampling per year. For the benthic invertebrate community analysis, for which responses are more likely to integrate effects over reasonably long time periods, we propose continued sampling once per year in August. For zooplankton and periphyton, analyses below in **Section 4.5** show that data are extremely variable and monitoring would be almost incapable of detecting effects in a given year. Consequently, no further routine sampling of zooplankton or periphyton as part of the CREMP is recommended. Further rationale for the time frame for each of the biological variables is found in **Section 4.5**.

4. EXPERIMENTAL DESIGN AND STATISTICAL ANALYSES

4.1. Sources of Variability in CREMP Data

The receiving environment program involves measurement of numerous variables representing environmental chemistry and productivity. There are several potential sources of variability that contribute to overall variability in the data, including:

- Lake to lake or basin to basin variability – in the case of the CREMP, the sampling stations represent entire lakes or basins.
- Spatial variability within a waterbody – the receiving environment program targets waterbodies or large basins within waterbodies. There may be variability within a waterbody associated with spatial structure (e.g., edges may be more prone to influence of spring run-off), depth or other factors.
- Subsample variability – this reflects small-scale variability for 2 samples taken at the same place at the same time at the same depth.
- Measurement error – that is, variability associated with imprecision in lab techniques and/or variability associated with field collection methods (e.g, homogenization of sediment samples).

- Variability with depth – we might expect variability in water chemistry or measures of productivity associated with lake stratification, and we can expect differences in other variables as well (e.g., benthic community structure will differ in deep waters). Variability within the water column has been found to be less important in the project lakes as lake stratification is limited (e.g., during open water season).
- Inter-annual variability – we can expect natural variability due for example to annual differences in rainfall or other climatic variables.
- Seasonal variability – seasonal differences can be expected over the year, associated with events such as spring run-off or conditions such as ice cover.
- Day to day variability – There may be differences day to day, for example associated with storm events or other short-term conditions. This component of variation is unlikely to be measured directly via sampling and will contribute to residual error in any analyses.

Both inter-lake and inter-annual variability are expected and are taken into account through the general BACI design (i.e., use of 'before' data and 'control' stations). In some cases, spatial variability within a lake/basin, sub-sample variability and measurement error can be simultaneously characterized through the use of field replicates – that is, true field replicates taken in difference places within a lake/basin during the same sampling event. In contrast, field duplicates (i.e., taken at the same location at the same time and depth) will reflect only sub-sample variability and measurement error. Finally, lab replicates taken from the same water sample (i.e., split samples) can be used to characterize only lab measurement error.

Regarding temporal variability and variability associated with depth, the 2009 receiving environment program (and 2008 in the case of sediment chemistry) included significant replication in time (multiple sampling events in one year) and at depth. Those data will be analyzed in later subsections of this document to evaluate the importance of depth.

The implications of seasonal variation in monitoring variables were discussed earlier in **Section 3**.

4.2. Impact Hypotheses and Statistical Design

Two general classes of impacts are hypothesized for the Meadowbank mine:

1. Pulse events for which potential impacts would be high for a short time and they may dissipate relatively quickly (e.g., in the case of water chemistry) or not (e.g., in the case of sediment chemistry). Pulse events could be associated with any phase of the mine, but are more likely to be associated with particular activities such as dike construction.

2. Long-term cumulative impacts that may be associated with ongoing activities. Long-term cumulative impacts are more likely to be associated with ongoing activities of mine operations.

Since operations began in 2010, the focus of monitoring to date has been on detecting pulse events associated with construction. The appropriate framework for analysis is a before-after-control-impact (BACI) that is aimed at detecting a potential impact in a particular lake or basin in a particular time period. The BACI framework can also be used to evaluate long-term impacts, but other tools such as time series regression analysis may also be appropriate for evaluating long-term trends. For this design document, we focus on the use of the BACI framework, recognizing that other tools such as time series regressions may be useful at a future date once sufficient time series data are available[8].

The classic BACI (paired) design has before/after periods α_i (i = B, A; I = 2), control/impact sites β_j (j = C, I; J = 2), and a total of K paired sampling times τ_k that are nested within period. A statistical model for this design is given by (Smith, 2002; equation 2):

$$(1) \qquad X_{ijk} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

The key term is the interaction ($\alpha\beta$), which can be tested using an F test with F = MS[($\alpha\beta$)]/MS[Resid] and degrees of freedom = 1, K – 2. As discussed by Smith (2002), this is equivalent to simply taking the differences between the control and impact values across times and using a two-sample (before-after) t test (Stewart-Oaten et al., 1986).

Model (1) can be extended to include additional control sites (e.g., "asymmetric" designs; Underwood, 1994) and/or additional impact sites. To be valid, the additional sites must be replicates rather than subsamples (i.e., as controls, they should be spatially independent of each other but representative of the impact sites, while replicates for impacts need to be spatially independent and (ideally) affected by independent disturbances). So whereas j = (C, I) in the classic BACIP, j may compose any combination of J total sites, for example J = 4 where j = (C_1, C_2, C_3, I). The general test of ($\alpha\beta$) still applies, but with degrees of freedom = (J – 1), (K – 2)(J – 1) (e.g., see Table 1 of Underwood (1994) and Table 9 of Smith (2002)).

[8] In theory, a BACI analysis that is appropriately framed should be capable of detecting changes associated with long-term trends.

In addition, there may be *n* replicate subsamples *s* at each site/time combination (*jk*), as assumed in Table 1 of Underwood (1994). In this case, we modify equation (1) as:

$$(2) \qquad X_{ijks} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + (\tau\beta)_{k(i)j} + \varepsilon_{ijks} \ ,$$

where subsamples now permit estimation of times-by-site interactions ($\tau\beta$). The appropriate *F* ratio for ($\alpha\beta$) is now $F = \text{MS}[(\alpha\beta)]/\text{MS}[(\tau\beta)]$ with df = $(J - 1), (K - 2)(J - 1)$. As Underwood demonstrates, specific comparisons (interaction terms, such as the impact site versus either "period" or a specific "time" unit) can be examined by partitioning variation accordingly (e.g., Underwood Table 2[9]).

Importantly, the BACI design is necessary only when there is reason to suspect natural inter-annual variability. For some variables, in particular sediment chemistry, over a time span of several years we may not have reason to suspect any significant inter-annual variability across years. In such a case, the comparison of interest is simply between baseline (before) and impact (after) data for a given station. There is no need or use for a "control station" under this assumption – the value of a control is to specifically "control" for potential temporal variation. Thus, the model would become a simple **"BA"** comparison (a t-test) between baseline data (**B**efore data) and a new year of impact data (**A**fter data). This test will generally have much greater power to detect a difference (real or not) than a full BACI model (which inherently assumes that temporal variation may be present, both across stations and within stations).

The BA model has a simple design:

$$(3) \qquad X_{is} = \mu + \alpha_i + \varepsilon_{is}$$

where *X* is the variable measured for period *i* and subsample *s*, and the estimate of interest is the difference in means for the before and after periods, represented by α_i, which is tested via a t-test.

4.3. Analyses for Water Chemistry

4.3.1. Data Review

The water chemistry data were summarized in **Table 3**. Further description and all analyses of the water chemistry data are provided in **Appendix A**. Some key conclusions regarding the data are as follows:

- Duplicate samples are derived by collecting a separate water sample at the same location, station, depth and time as a standard sample. Differences between a

[9] We note that there are mistakes in the tables presented by Underwood (1994).

duplicate sample and its associated standard sample reflect the combined variability from subsample variation and measurement error associated with field collection methods and lab methods. Analysis of duplicate samples in water shows that the variability due to these factors is generally much lower than general sample variation. From a QA/QC perspective, this confirms that precision in field collection and lab methods is reasonable relative to the natural variability among the samples and stations. Duplicate samples were not included in the power analyses of water chemistry below because they are not independent of the standard samples and are therefore pseudoreplicates of the standard samples.

- Depth samples were collected in 2009-2010, to determine the potential importance of differences in measurements associated with depth. Analysis showed that bottom samples (and depth-integrated samples to a lesser degree) tended to have lower values than surface samples for many variables as there were statistically significant differences in values by depth. . However, the magnitude of differences associated with depth is minimal compared to the magnitude of differences that occurs naturally between samples and stations[10]. We conclude that future sampling should focus on surface samples[11], in order to take advantage of the most baseline data and to incorporate a more conservative approach for most water chemistry parameters which demonstrate lower concentrations at depth. Depth samples were not included in power analyses below – as with duplicate samples, the depth and integrated samples are pseudoreplicates of the surface samples.
- Baker / Meadowbank data differences, as discussed in **Section 3.2.2**, are significant and meaningful. and are mostly attributable to the influence of seawater on Baker Lake.

4.3.2. Power Analysis and Implications

Using the BACI framework presented in **Section 4.2**, the power to detect changes in water chemistry variables from current baseline levels to trigger levels was estimated under various scenarios for the 'after' period ranging from one to six months of data and from one to three subsamples per month. As discussed in **Section 3.2.3**, it is proposed to

[10] The Meadowbank study lakes typically lack stratification, which results in a well-mixed water column with consistent physical and chemical characteristics across depth zones.
[11] Samples at depth could be added in cases where there is reason to suspect potential for large depth-specific effects. Given that the CREMP is intended to monitor large-scale basin-wide changes, depth-specific sampling is less likely to be warranted for the CREMP than for other programs that have a finer scale of resolution.

use data for a single month as a flag, whereas triggers for action would be based on mean values over several months.

The BACI analysis focused on Meadowbank data. Baker Lake data have only been collected since 2008, thus inferences would be limited based on the small data set. Results from Meadowbank should be generalizable to Baker Lake, given that the BACI analyses for Meadowbank compare a single control to each potential impact station individually (the same scenario as for Baker, which has one control and two impact stations).

The BACI analysis for Meadowbank focused on estimation of power for stations SP, TPE and TPN compared to INUG (as control). Power was estimated for a limited number of variables, using simulation – values used in the simulations were based on fits of mixed-effects models to the 'before' period data (control data only, no impact data). All details are provided in **Appendix A**.

Results of power analysis for t-Al, t-Mn, ammonia-N and t-P are shown in **Figure 6**, while results for pH are shown in **Figure 7**. More details are provided in **Appendix A**. In general, with an exception for t-P, power is high (around 80% or higher) with six month means. In many cases power is almost 100%. Results show that the benefits of subsampling are limited, with some gains in power associated with two subsamples but little improvement associated with three subsamples. In terms of duration of sampling, the results show significant increases in power associated with 3 months of sampling, and smaller increases associated with moving from 3 to 6 months. Consequently, we would not expect significant limitations if fewer than six samples are collected due to sampling constraints discussed in **Section 3.2.3**. Furthermore, it may be reasonable to intentionally reduce sampling from 6 samples in some cases, for example as follows:

1. Sampling for PDL could be limited to open water seasons, to reduce the safety hazard associated with sampling PDL during spring. PDL data would still be used as control, but the sample size in a given year would be slightly lower. As long as INUG continues to be sampled whenever the impact stations are sampled, it will be possible to detect natural anomalous month-effects if they occur. Since trigger criteria for water are based on means over several months, and since natural anomalous month-effects are not generally expected, the PDL data should be just as useful as the INUG data even when under ice sampling is not included.

2. Baker Lake sampling could also be reduced to open water seasons. Given that impacts at Baker Lake would be expected primarily in open water months when barges are moving and cargo is being loaded/unloaded, there is little chance of missing major potential impacts by not sampling through the ice.

4.4. Analyses for Sediment Chemistry

4.4.1. Data Review

The sediment chemistry data were summarized in **Table 4**. Further description and all analyses of the sediment chemistry data are provided in **Appendix B**. Some key conclusions regarding the data are as follows:

- Data collected by coring are more useful than grab sample data for monitoring changes in sediment chemistry. This is because the former targets a thin surface layer, whereas the grab sample collection targets a thicker layer that may be less reflective of potential mine-related impacts and more reflective of historic (pre-mining) conditions.
- Duplicate samples – Field duplicates are different for core and grab samples. In the case of core samples, the duplicates are separate samples collected at the same location. In the case of grab samples, the duplicates are split samples – that is, a different aliquot of sample is taken from the composite sample. Thus, differences among core sample duplicates will reflect fine-scale spatial variation and sampling/lab precision (measurement error), whereas for grab sample duplicates, differences should reflect homogenization of sediment and sampling/lab precision. Analysis of duplicate samples in water shows that the variability due to these factors is generally much lower than general sample variation. From a QA/QC perspective, this confirms that precision in field collection and lab methods is reasonable relative to the natural variability among the samples and stations. Duplicate samples were not included in the power analyses of sediment chemistry below because they are not independent of the standard samples and are therefore pseudoreplicates of the standard samples.
- Baker / Meadowbank differences – as discussed in **Section 3.3.2**, there are significant and meaningful differences in baseline data between Baker Lake and the Meadowbank lakes.

4.4.2. Power Analysis and Implications

Power analysis for sediment chemistry variables is based on the assumption that we would not expect any significant natural inter-annual variability in sediment chemistry over a limited number of years. This is reasonable given the low rates of natural sediment deposition in the project lakes. Thus, a before-after (BA) modeling framework is appropriate for sediment chemistry.

Using the BA framework presented in **Section 4.2**, and the data for sediment core samples, the power to detect changes in sediment chemistry parameters from current baseline levels to trigger levels was estimated under various scenarios for the number of subsamples ranging from 5 to 25. The objective was to detect a difference from baseline

in a given (single) year. Analysis considered five of the seven sediment chemistry variables (As, Cr, Cu, Hg, Zn) at the near-field Meadowbank stations SP, TPE and TPN. Results from these variables/stations give a reasonable cross-section and should be generalizable to other variables and stations. Power was estimated using one-tailed t-tests assuming an effect size equal to a shift from baseline mean to a trigger. All details are provided in **Appendix B**.

Statistical power was basically 100% except for two cases (Cr for TPN and Zn for TPE) in which the mean of baseline data was close to the trigger value (see **Figure 8**).

Importantly, there was potential evidence that differences in "random selection" of sampling location resulted in significant differences between years (see **Appendix B**). This is explained by reviewing the field sampling methods between the two years for INUG (in one year, the field crew clearly followed a linear path and did not randomly sample). Given this evidence of spatial heterogeneity within lakes or basins, it seems reasonable to have 10 subsamples, even though 5 would probably be sufficient in most cases. In addition, it seems appropriate to continue to sample INUG as a control to maintain the option of applying BACI models over time rather than the simplified BA model. Consequently, we recommend that for every sediment core chemistry sampling event, 10 subsamples be collected at each impact station, with 5 submitted for analysis and the additional 5 to be archived. If there is evidence of potential impacts, or if uncertainty is high, the additional 5 samples could be analyzed. For INUG, only 5 subsamples should be collected and analyzed, because the value of control data lies only in potential use of BACI models, not BA models.

Although sediment core samples are the most appropriate for evaluating sediment chemistry, sediment grab samples should be continued (to be collected at the same time as benthic invertebrate samples) because some important monitoring variables (e.g., particle size) are not or may not be measured with the core samples due to volume limitations.

4.5. Analyses for Biological Effects Variables

4.5.1. Phytoplankton

4.5.1.1. Data Review

The phytoplankton data are similar to the water chemistry data, with a total of 308 samples collected for the CREMP from 2006 through to September 2010 (see **Table 5**; not including field duplicate samples and depth replicates). Some key conclusions regarding the data are as follows:

- Duplicate samples are derived by collecting a separate sample at the same location, station, depth and time as a standard sample. Differences between a duplicate sample and its associated standard sample reflect the combined variability from subsample variation and measurement error associated with field collection methods and lab methods. Analysis of duplicate samples for phytoplankton shows that the variability due to these factors is generally lower than general sample variation. From a QA/QC perspective, this confirms that precision in field collection and lab methods is reasonable relative to the natural variability among the samples and stations. Duplicate samples were not included in the power analyses of phytoplankton below because they are not independent of the standard samples and are therefore pseudoreplicates of the standard samples.
- Depth samples were collected starting in 2009, to determine the potential importance of differences in measurements associated with depth. Analysis shows that for almost all variables there is no difference associated with depth, and for the few exceptions the magnitude of differences associated with depth are minimal compared to the magnitude of differences that occur naturally between samples and stations[12]. We conclude that future sampling should focus on surface (3 m below surface) samples[13], in order to take advantage of the most baseline data. Depth samples were not included in power analyses below – as with duplicate samples, the depth samples are pseudoreplicates of the surface samples.

4.5.1.2. Power Analysis and Implications

Using the BACI framework presented in **Section 4.2**, the power to detect 20% and 50% reductions in phytoplankton variables from current baseline levels was estimated under various scenarios for the 'after' period ranging from one to six months of data and from one to three subsamples per month. As discussed in **Section 3.4.2**, it is proposed to use data for a single month as a flag, whereas triggers for action would be based on mean values over several months.

The BACI analysis focused on Meadowbank data. Baker Lake data have only been collected since 2008, thus inferences would be limited based on the small data set.

[12] The Meadowbank study lakes typically lack stratification, which results in a well-mixed water column with consistent physical and chemical characteristics across depth zones.
[13] Samples at depth could be added in cases where there is reason to suspect potential for large depth-specific effects. Given that the CREMP is intended to monitor large-scale basin-wide changes, depth-specific sampling is less likely to be warranted for the CREMP than for other programs that have a finer scale of resolution.

Results from Meadowbank should be generalizable to Baker Lake, given that the BACI analyses for Meadowbank compare a single control to each potential impact station individually (the same scenario as for Baker, which has one control and two impact stations).

The BACI analysis for Meadowbank focused on estimation of power for stations SP, TPE and TPN compared to INUG (as control). Power was estimated for two variables (total biomass and species counts), using simulation – values used in the simulations were based on fits of mixed-effects models to the 'before' period data. All details are provided in **Appendix C**.

Results of power analysis are shown in **Figure 9**. More details are provided in **Appendix C**. Results show that the power to detect 20% reductions in phytoplankton variables is much lower than power to detect 50% reductions. This is particularly true for total biomass, and we would expect power to be even lower for all other biomass variables (except Chrysophytes) given their higher variability. Results also show that the benefits of subsampling are limited, with very little gain in power associated with more than two subsamples. In terms of duration of sampling, the results show notable increases in power associated with 3 months of sampling, and smaller increases associated with moving from 3 to 6 months. Consequently, there may not be a large benefit associated with sampling phytoplankton under ice, since 3 samples collected during open water (July, August, September) would provide as much power. We propose to sample phytoplankton whenever water samples are collected (i.e., up to 6 months per year), but to only analyze the data for July to September while archiving the samples for other months.

4.5.2. Benthic Community

4.5.2.1. Data Review

Benthic invertebrates have been collected as part of the CREMP once each year in August from 2006 to 2010. Each sample is a composite of two Petite Ponar grabs, sieved using a 500 um mesh screen. If we exclude a few samples that were collected with a different mesh size (250 um) or that were based on single grabs rather than composites, and if we also exclude some 2006 data where samples were poorly preserved and denoted as not useful, there have been 242 samples collected (see **Table 6**). Of these, 162 were designated as control and 80 samples were designated as impact (shaded cells in the table). There were no duplicate samples collected for benthos.

4.5.2.2. Power Analysis and Implications

Using the BACI framework presented in **Section 4.2**, the power to detect 20% and 50% reductions in benthic invertebrate community variables from current baseline levels was

estimated under various scenarios for the 'after' period ranging from one to three years data and from 3 to 20 subsamples per month.

The BACI analysis focused on Meadowbank data. Baker Lake data have only been collected since 2008, thus inferences would be limited based on the small data set. Results from Meadowbank should be generalizable to Baker Lake, given that the BACI analyses for Meadowbank compare a single control to each potential impact station individually (the same scenario as for Baker, which has one control and two impact stations).

The BACI analysis for Meadowbank focused on estimation of power for stations SP, TPE and TPN compared to INUG (as control). Power was estimated for two variables (total abundance and total taxa), using simulation – values used in the simulations were based on fits of mixed-effects models to the 'before' period data. All details are provided in **Appendix D**.

Results of power analysis are shown in **Figure 10**. More details are provided in **Appendix D**. The presence of year-by-station effects (i.e., natural station-specific inter-annual variation) will make it very difficult to distinguish between impacts and natural variations given only one year of "impact" data (or even after several years for relatively minor effects). Results show that the power to detect 20% reductions in total abundance or total taxa for benthic invertebrates is generally not high even after 3 years and with many subsamples per year. Power is higher for detecting a 50% reduction, particularly for total taxa. Increasing the number of replicates (subsamples) from the current 5 replicates would not result in significant improvements to power or precision in most cases.

4.5.3. Zooplankton

The zooplankton data for the CREMP include 45 samples collected in August 2010 (5 spatial replicates at each of the 9 Meadowbank stations) and 90 samples collected in August 2011 (10 spatial replicates at each of the 9 stations; see **Appendix E**). Using the BACI framework presented in **Section 4.2**, the power to detect 20% and 50% reductions in zooplankton variables from current baseline levels was estimated under various scenarios for the 'after' period ranging from one to three years data and from 5 to 20 subsamples per sampling event. In addition, because there were only 2 years of 'before' data, additional 'before' data were simulated so that three scenarios could be considered (before years = 2, 3, or 4). The power analyses were conducted using a single control (INUG) as well as two controls (INUG and PDL), paired with one impact station (SP, TPE or TPN). Power was estimated for three variables (wet biomass, dry biomass and abundance) using simulation – all details are provided in **Appendix E**.

Results of power analysis are provided in **Appendix E**. Power is always low when trying to detect a 20% change, and when trying to detect a 50% change in one year (except a

couple of cases when there are a high number of subsamples). Power is only reasonable (around 0.7 or higher) for abundance and especially for dry biomass when detecting a 50% reduction with 2 or 3 years of 'after' period data. The number of spatial replicates (5, 10 or 20) appears somewhat less important than the number of years of 'after' period data. Overall, zooplankton variables are not realistically capable of detecting effects in a given year. Consequently, we do not recommend that zooplankton be included as a monitoring component of the CREMP. Zooplankton may still be useful (particularly dry biomass, which appears to be more sensitive than abundance) where focused EAS studies are warranted, for example using intensive spatial-gradient designs.

4.5.4. Periphyton

The periphyton data for the CREMP include 105 samples over the period 2007 to 2011 (see **Appendix F**). Most of the samples were collected in 2007 and 2008 (5 spatial replicates at each of 7 Meadowbank stations for a total of 35 samples each year). For 2009 to 2011 there were 5 spatial replicates collected in SP and also at reference (SP-DT)

Using the BACI framework presented in **Section 4.2**, the power to detect 20% and 50% reductions in periphyton biomass variables from current baseline levels was estimated under various scenarios for the 'after' period ranging from one to three years data and from 5 to 20 subsamples per sampling event. In addition, because there were only 2 years of 'before' data, additional 'before' data were simulated so that three scenarios could be considered (before years = 2, 3, or 4). The power analyses were conducted using a single control (INUG), paired with one impact station (SP, TPE or TPN). Power was estimated for three variables (Cyanobacteria biomass, diatom biomass and total biomass) using simulation – all details are provided in **Appendix F**.

Results of power analysis are provided in **Appendix F**. Power is generally low, except for detecting a 50% decrease in total biomass when there are multiple years of before and after data. The number of spatial replicates (5, 10 or 20) appears relatively unimportant, as power is driven mainly by the number of years of data in the 'before' and 'after' periods. Overall, periphyton variables are not realistically capable of detecting effects in a given year, and are even less powerful than zooplankton variables for detecting medium term trends (e.g., 2-3 years each of 'before' and 'after' data). The most useful variable, total biomass, does not achieve high (0.8) power with less than 4 years of 'after' data. Furthermore, collection of periphyton data is subject to bias because sampling crews must subjectively choose the rocks that are sampled – while efforts can be made to standardize this process, there is nevertheless a risk of year-to-year variability in field sample collection methods that is greater than for other monitoring variables. For all of these reasons, we do not recommend that periphyton be included as a monitoring component of the CREMP. Periphyton may still be useful where focused EAS studies are warranted, for example using intensive spatial-gradient designs.

5. SUMMARY OF IMPLICATIONS FOR THE CREMP

The previous sections of this design document have derived thresholds and triggers for action for CREMP water chemistry parameters and biological monitoring variables, and have evaluated the ability of alternative monitoring program designs to detect changes relative to those thresholds and triggers. The key implications of the analyses for CREMP sampling are as follows, and the program is summarized in **Table 13**:

- Water chemistry data will be collected up to 6 months per year (April, May, July, August, September and November/December) for the annual period of paired sampling to support BACI statistical analyses, recognizing that in any given year the actual number of samples that would be collected may range from four to six depending on logistical constraints (e.g., snow and ice conditions). Sampling will be limited to open water months only (July, August and September) for PDL and for the Baker Lake stations. It is recommended that two randomly located subsamples be collected at each station in each month. All samples should be surface samples (3 m from the surface). In addition to the core water chemistry program, basic water quality data will be collected at key near-field areas (i.e., TPN, TPE, SP and eventually Wally) at least once mid-winter to reduce uncertainty regarding the potential occurrence of changes over winter.
- Sediment chemistry core sampling for the CREMP is intended to detect long term trends, therefore a sampling frequency of approximately every three years is recommended. To maximize the potential utility of the data, sampling will be aligned with the sampling times for benthic invertebrates required for the EEM program. Thus, sediment chemistry cores were collected in 2012 (because there have been no core data collected since 2009), and then matched to the envisioned EEM sampling years which are likely to be 2014 and then every three years (depending on results in 2014). Samples will be collected during open water (August) at potentially impacted stations as well as INUG and BAP. There is no need for sampling at PDL or other reference stations other than INUG and BAP. For each sampling event, 10 randomly[14] located subsamples should be collected at each impact station, with 5 submitted for analysis and the additional 5 archived. If there is evidence of potential impacts, or if uncertainty is high, the additional 5 samples could be analyzed. At INUG, only 5 subsamples should be collected and analyzed – there is no need to archive an additional 5 samples for INUG.
- Sediment grab sampling synoptically collected with benthic invertebrate samples (i.e., once per year) should also be continued to ensure collection of basic physical

[14] Importantly, the random locations should be selected in advance using random number generators and GIS software, and should not be selected by the sampling team in the field.

variables (e.g., particle size and organic carbon) not covered by sediment core sampling (due to volume limitations or lack of direct spatial compatibility with benthic sample replicates) but which may nevertheless affect benthic invertebrates.

- Phytoplankton data will be collected whenever the water chemistry data are collected, but only the open water samples (July to September) will be analyzed while it is recommended that the other samples be archived. It is recommended that two randomly located subsamples be collected at each station for each sampling event. All samples should be surface samples (3m from the surface).
- Benthic invertebrates should be collected once per year in August at all stations, with 5 subsamples per station.
- Zooplankton sampling (5 subsamples per station) should be discontinued for the CREMP, due to low power to detect effects.[15]
- Periphyton sampling should be discontinued for the CREMP, due to its low power to detect effects even with several years of data[16].

The above recommendations should be re-visited periodically based on accumulated data for the CREMP and updated power analyses for each group of variables.

6. LINKS TO AEMP

A management response plan (MRP) has been developed for the AEMP (Azimuth, 2012), of which the CREMP is one component. The general MRP for the Meadowbank Mine AEMP is shown in **Figure 11**. Following the integration of the results from each independent program, the response actions are based on the cumulative results of all programs. Therefore, while we expect management actions to be taken in cases where criteria for action are exceeded, the specific actions are not linked to outcomes of the CREMP alone because the CREMP is only one of the monitoring programs under the AEMP. In other words, it is not possible or appropriate to describe the specific management actions that will be taken when CREMP triggers or thresholds are exceeded.

Nevertheless, as shown in **Figure 3**, **Figure 4**, **Figure 5**, and **Figure 11**, there are two general classes of management actions – those aimed at further assessment and those aimed at mitigation. In general, exceedance of early warning triggers will trigger further

[15] Zooplankton and periphyton may still be useful where focused EAS studies are warranted, for example using intensive spatial-gradient designs.

[16] See previous footnote.

assessment, which may then lead to mitigation, whereas exceedances of thresholds could possibly lead directly to mitigation. It is expected that CREMP triggers will be exceeded occasionally due to chance (given the large number of variables that are monitored, particularly for water chemistry), therefore further assessment will almost always be important.

The specific management action that would be appropriate in a given case depends on the underlying cause. For example, if a metal becomes elevated in receiving water, the identification of options for further assessment and/or mitigation options would be different if the source of the metal is groundwater versus effluent versus dust.

The timing of management actions is also case-specific. In cases where further assessment is warranted, that assessment should begin as soon as practically possible. In cases where mitigation is considered, mitigation should begin as soon as the weight of evidence indicates that mitigation is warranted, and the benefits of commencing mitigation immediately outweigh the disadvantages of waiting for further information. Consultation with regulators and stakeholders is important for determining management actions (see Azimuth, 2012).

7. REFERENCES

AEMP (Aquatic Effects Management Program). 2005. A report prepared by Azimuth Consulting Group, Vancouver for Cumberland Resources Ltd. October, 2005.

Azimuth. 2010. Core Receiving Environment Monitoring Program (CREMP) – 2010 Plan Update. Report prepared by Azimuth Consulting Group Inc., Vancouver, BC for Agnico-Eagle Mines Ltd., Baker Lake, NU. June 2010.

Azimuth. 2012. Aquatic Effects Monitoring Program (AEMP) Version 2. Report prepared by Azimuth Consulting Group Inc., Vancouver, BC for Agnico-Eagle Mines Ltd., Baker Lake, NU. December 2012.

CCME (Canadian Council of Ministers of the Environment) 2007. Canadian Water Quality Guidelines for the Protection of Freshwater Aquatic Life, 1999, updated December 2007.

Cohen, J. 1988. Statistical power analysis for the behavioural sciences. 2nd ed. L. Erlbaum, Hillsdale, NY, USA.

Environment Canada. 2012a. Metal mining technical guidance for Environmental Effects Monitoring. ISBN 978-1-100-20496-3.

Environment Canada. 2012b. FCSAP Ecological Risk Assessment Guidance. Environmental Stewardship Branch, Vancouver, BC. March 2012.

INAC (Indian and Northern Affairs Canada). 2009. Guidelines for designing and implementing aquatic effects monitoring programs for development projects in the Northwest Territories. Yellowknife. June 2009.

Munkittrick, K., C. Arens, R. Lowell, G. Kaminski. 2009. A review of potential methods of determining critical effect size for designing environmental monitoring programs. Environmental Toxicology and Chemistry 28:1361-1371.

Science Advisory Board for Contaminated Sites in BC (SAB). 2008. Detailed Ecological Risk Assessment (DERA) in British Columbia: Technical Guidance. Submitted to BC Ministry of Environment.

Smith, E. 2002. BACI design. Pages 141-148, in: A.H. El-Shaarawi and W.W. Piegorsch [Eds], Encyclopedia of Environmetrics, John Wiley & Sons, Chicherster.

Stewart-Oaten, A., W. Murdoch, K. Parker. 1986. Environmental impact assessment: "Pseudorelication" in time? Ecology 67:929-940.

Underwood, A. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications 4:3-15.

Table 1. CREMP Station Summary.

	Baker Lake			Meadowbank Lakes							Wally	
Year	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	
2006				0		0	0		0	0	0	0	Known
2007				0		0	0		0	0	0	0	
2008	0	1	1	0		Aug	Aug		0	0	0	0	
2009	0	1	1	0	0	1	1	0	Aug	March	March	0	
2010	0	1	1	0	0	1	1	0	1	1	0	0	
2011	0	1	1	0	0	1	1	0	1	1	0	0	
2012	0	1	1	0	0	1	1	?	1	1	?	0	Guess
2013	0	1	1	0	0	1	1	?	1	1	?	?	
2014	0	1	1	0	0	1	1	?	1	1	?	1	

Notes:

	Always "control"
	Far-field: possible "impact"; considered "control" until potential exposure occurs.
	Near-field: Impact (as of start of activities that could result in exposure; "control" prior to that).
Blank	denotes that the station was not part of the monitoring program that year.
0, 1	This denotes whether the station-year combination was considered "control" (0) or "impact" (1).
	Baker Lake stations to be compared to BAP control station.
	Meadowbank stations to be compared to INUG, PDL and TEFF (while "control").
	Wally Lake to be compared to new control lake.

Table 2. General TSS Conditions at Meadowbank CREMP Stations for 2008, 2009 and 2010.

	2008				2009					2010				
	July	**Aug**	**Sept**	**Oct - Dec**	**Jan - Jun**	**July**	**Aug**	**Sept**	**Oct - Dec**	**Jan - Jun**	**July**	**Aug**	**Sept**	**Oct - Dec**
Third Portage Lake														
North Basin (TPN)	Background	Background	Background	Background	Dewatering - no apparent TSS increase	Dewatering stopped	Background	Background	Background?	Dewatering, TSS < 1 mg/L	Dewatering, TSS < 1 mg/L	Dewatering, TSS < 1 mg/L	Dewatering, TSS < 1 mg/L	Dewatering, TSS < 1 mg/L
South Basin (TPS)	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background
East Basin (TPE)	Background	Background	Background	Background	Background	BG Dike construction starts	TSS concentrations gradually increase, but mostly at depth	Strong winds mix TSS throughout water column and basin; TSS source finished	TSS drops over time; still above HVH trigger by end of October	TSS increases again during causeway construction	TSS generally < 2 mg/L	TSS generally < 5 mg/L	TSS generally < 5 mg/L	TSS drops?
Second Portage Lake														
Main Basin (SP)	Background	ED construction starts; TSS rises dramatically in third week	TSS drops over time; meets targets by end of month	TSS drops?	TSS drops initially, then dewatering starts in March	Background	Some TSS input from BG dike construction, but low concentrations	TSS increases slightly, but limited spatially	TSS drops over time	TSS generally < 1 mg/L; 2 mg/L in Apr/May	TSS generally < 1 mg/L	TSS generally < 1 mg/L	TSS generally < 1 mg/L	TSS drops?
NW Arm	Background	TSS rises in parallel to main basin, but lower concentrations than main basin	TSS reduced, but rises again during WC Dike construction	TSS drops	TSS drops initially, then increases due to dewatering	TSS elevated due to dewatering activities; dewatering stops in early July	Isolated	Isolated	Isolated	Isolated	Isolated	Isolated	Isolated	Isolated
Drilltrail Arm	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background
Tehek Lake														
Near-field Basin (TE)	Background	TSS rises during third week	TSS drops over time; meets targets by end of month	TSS drops?	TSS drops?	Background	Very slight increase in TSS	TSS slightly above background?	TSS drops over time?	TSS drops over time?	Background?	Background?	Background?	Background?
Far-field Basin (TEFF)	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background	Background

Notes: Shaded months indicate ice cover on lakes.
Acronyms: ED = East Dike; BG = Bay-Goose Dike; TSS = total suspended solids; DCM = dike construction monitoring.
"?" indicates uncertain condition.
CREMP stations shown in (), where applicable

Table 3. Water chemistry samples collected for the CREMP up to May 2010.

Notes: Shading denotes "impact" periods. Some of the data include field duplicates and depth replicates. See text for details.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul				1		1	1		1	1	1	1	7
	Aug				1		1	1		1	2	1	1	8
2007	Jul				1		1	2		1	1	1	1	8
	Aug				1		1	1		2	1	1	1	8
2008	Jul	1	1	1	1		1	1		1	1	1	2	11
	Aug	1	1	1	2		1	1		1	1	1	1	11
	Sep	2	1	1			1	1		1	4	1		12
2009	May				14		7	13		7	7			48
	Jul	3	4	3	7	3	3	8	2	3	4	3	4	47
	Aug	1	1	1	3	1	1	4	1	1	2	1	1	18
	Sep	3	3	4	7	3	4	7	3	4	3	3	4	48
	Nov				7		3	8		3	4			25
	Dec						1	3		2	1	3		10
2010	Jan						5	3		3	3	7		21
	Feb						2	3		1	1	3		10
	Mar						3	7		4	4	7		25
	Apr						1	3		2	1	3		10
	May						4	7		4	3	7		25
Total		11	11	11	45	7	41	74	6	42	44	44	16	352

Table 4. Core and grab sediment samples (excluding duplicates) collected for the CREMP program.

Notes: Core samples for July 2009 were collected for East Dike Effects Assessment Study, not for the CREMP, but have been included in the CREMP analysis. Shading denotes "impact" periods.

CORE		Baker			Meadowbank									
Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul													0
2007	Aug													0
2008	Jul				15		15	15		15	15	15	15	105
	Aug	15	6	15										36
2009	Jul				15		15	15		15				60
	Aug													0
Total		**15**	**6**	**15**	**30**	**0**	**30**	**30**	**0**	**30**	**15**	**15**	**15**	**201**

GRAB		Baker			Meadowbank									
Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul				1		1	1		1	1	1	1	7
2007	Aug				1		1	1		1	1	1	1	7
2008	Jul													0
	Aug	1	1	1	1		1	1		1	1	1	1	10
2009	Jul													0
	Aug	3	3	3	3	3	3	3	3	3	3	3	3	36
Total		**4**	**4**	**4**	**6**	**3**	**6**	**6**	**3**	**6**	**6**	**6**	**6**	**60**

Table 5. Phytoplankton samples collected for the CREMP.

Notes: Duplicates and depth replicates removed. Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul				1		1	1		1	1	1	1	7
	Aug				2		2	2		2	2	2	2	14
	Sep				1		1	1		1	1	1	1	7
2007	Jul				1		1	1		1	1	1	1	7
	Aug				1		1	1		1	1	1	1	7
2008	Jul	1	1	1	1		1	1		1	1	1	1	10
	Aug	1	1	1	1		1	1		1	1	1	1	10
	Sep	1	1	1			1	1		1	1	1		8
2009	Jul	3	3	3	3	3	3	3	2	3	3	3	3	35
	Aug	1	1	1	1	1	1	1	1	1	1	1	1	12
	Sep	3	3	3	3	3	3	3	3	3	3	3	3	36
	Nov				3		3	3		3	3			15
	Dec						1	1		1	1	1		5
2010	Jan						3	2		3		3		11
	Feb						1	1		1	1	1		5
	Mar						3	3		3	3	3		15
	Apr						1	1		1	1	1		5
	May						3	3		3	3	3		15
	Jul	3	3	3	3	3	3	3	3	3	3	3	3	36
	Aug	1	1	1	1	1	1	1	1	1	1	1	1	12
	Sep	3	3	3	3	3	3	3	3	3	3	3	3	36
Total		17	17	17	25	14	38	37	13	38	35	35	22	308

Table 6. Summary of August benthic invertebrate samples for the CREMP (mesh size = 500 um).

Note: Shading denotes "Impact" periods.

Year	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006				6		6	4		6	3	6	6	37
2007				5		5	5		5	5	5	5	35
2008	5	5	5	5		5	5		5	5	5	5	50
2009	5	5	5	5	5	5	5	5	5	5	5	5	60
2010	5	5	5	5	5	5	5	5	5	5	5	5	60
Total	15	15	15	26	10	26	24	10	26	23	26	26	242

Table 7. Summary of thresholds for water variables.

Note: See text for derivation of triggers based on thresholds.

Variable	Source	Description of guidelines
t-Aluminum (Al)	CCME	The CCME guideline for t-Al in water is 0.005 mg/L when pH < 6.5, and 0.1 mg/L when pH $\geq$ 6.5. See **Appendix A** for details.
t-Arsenic (As)	CCME	The CCME water quality guideline (aquatic life) for t-As is 0.005 mg/L.
t-Boron (B)	BC MOE	There is no CCME guideline for t-B. However, the BC Ministry of Environment (BC MOE, www.env.gov.bc.ca) guideline for freshwater aquatic life is 1.2 mg/L.
t-Cadmium (Cd)	CCME, EPA	The hardness-dependent CCME guideline for t-Cd (mg/L) is $0.001*10^{0.86*\log(H)-3.2}$ where H = hardness (mg/L $CaCO_3$). This guideline was deemed inappropriate. The EPA guideline for t-Cd of 0.0001 mg/L was used instead. See **Appendix A** for details.
t-Copper (Cu)	CCME	The CCME guideline for t-Cu is 0.002 mg/L for hardness < 120 mg/L $CaCO_3$.
t-Chromium (Cr)	CCME	The CCME guideline for hexavalent chromium (the most common form in surface waters) is 0.001 mg/L.
t-Iron (Fe)	CCME	The CCME guideline for t-Fe is 0.3 mg/L.
t-Lead (Pb)	CCME	The CCME guideline for t-Pb is 0.001 mg/L for hardness < 60 mg/L $CaCO_3$.
t-Manganese (Mn)	BC MOE	There is no CCME guideline for t-Mn in water. The hardness-dependent BC MOE guideline for t-Mn in mg/L is 0.0044*H + 0.605, where H = hardness (mg/L $CaCO_3$.See **Appendix A** for details.
t-Mercury (Hg)	CCME	The CCME guideline for total inorganic mercury is 26 ng/L (0.00026 mg/L).
t-Molybdenum (Mo)	CCME	The CCME guideline for t-Mo in water is 0.073 mg/L.
t-Nickel (Ni)	CCME	The CCME guideline for t-Ni is 0.025 mg/L for hardness < 60 mg/L $CaCO_3$.
t-Selenium (Se)	CCME	The CCME guideline for t-Se in water is 0.001 mg/L.
t-Thallium (Tl)	CCME	The CCME water quality guideline for t-Tl is 0.0008 mg/L.
t-Zinc (Zn)	CCME, EVS (2004)	The CCME water quality guideline for t-Zn is 0.030 mg/L. However, this guideline does not take into account hardness, and zinc toxicity is known to be hardness-dependent. An assessment for Ekati by EVS (2004) compiled data on species applicable to oligotrophic systems with low hardness, and developed a chronic benchmark for t-Zn that was hardness dependent. See **Appendix A** for details.
d-Aluminum (Al)	BC MOE	A pH-dependent water quality guideline for d-Al (mg/L) has been developed by BC MOE for protection of freshwater aquatic life as follows: $d\text{-}Al = e^{(1.6-3.327*pH + 0.402*K)}$ where $K = pH^2$. See **Appendix A** for details.
Ammonia-N	CCME	The CCME guidelines for total ammonia in freshwater are pH and temperature dependent, with more stringent guidelines applying at higher pH and higher temperature. See **Appendix A** for details.
Nitrate-N	CCME	The CCME guideline for NO3-N is 2.9 mg/L.
Nitrite-N	CCME	The CCME guideline for nitrite-N is 0.06 mg/L.
t-Phosphorous (P)	CCME	The CCME does not specify a particular guideline for t-P, but instead establishes a guidance framework for site-specific application. See **Appendix A** for details.

pH	CCME	The CCME guideline for pH is a range from 6.5 to 9.0. See **Appendix A** for details.
TSS	CCME	For water bodies with low natural TSS, the CCME guideline is a maximum increase of 25 mg/L over background for short periods (e.g., 24h) and 5 mg/L for longer periods (e.g., 24h to 30 days). See **Appendix A** for details.

Table 8. Total metals in water: summary of trigger values for Meadowbank and Baker stations.

Notes: For each variable, thresholds are shown if applicable (see text for discussion); DL = current detection limit; N = sample measurements; >DL = number of measurements above DL; medians and 95th percentiles are shown for variables with at least one value >DL (with all values <DL set equal to DL); Method = method used to determine the trigger, where A = halfway from median (or DL if all values <DL) to threshold, B = 95th percentile, and C = 2*DL.

Variable	Threshold	DL	N	>DL	Median	95th	Trigger	Method
Aluminum (T)	0.1	0.005	120	85	0.0068	0.0127	0.053	A
Antimony (T)		0.0005	120	0			0.0010	C
Arsenic (T)	0.005	0.0005	120	0			0.0028	A
Barium (T)		0.02	120	0			0.04	C
Beryllium (T)		0.001	120	0			0.002	C
Boron (T)	1.2	0.1	120	0			0.65	A
Cadmium (T)	0.0001	0.000017	120	7	0.000017	0.000017	0.000059	A
Chromium (T)	0.001	0.001	120	0			0.001	A
Copper (T)	0.002	0.001	120	1	0.001	0.001	0.0015	A
Iron (T)	0.3	0.03	120	2	0.03	0.03	0.165	A
Lead (T)	0.001	0.0005	120	3	0.0005	0.0005	0.00075	A
Lithium (T)		0.005	120	0			0.010	C
Manganese (T)	See text	0.0003	120	114	0.0010	0.0025	0.32	See text
Mercury (T)	0.000026	0.00001	120	0			0.000018	A
Molybdenum (T)	0.073	0.001	120	0			0.037	A
Nickel (T)	0.025	0.001	120	0			0.013	A
Selenium (T)	0.001	0.001	120	0			0.001	A
Strontium (T)			6	6	0.0081	0.0089	0.0089	B
Thallium (T)	0.0008	0.0002	120	0			0.0005	A
Tin (T)		0.0005	120	0			0.0010	C
Titanium (T)		0.01	120	0			0.02	C
Uranium (T)		0.0002	120	0			0.0004	C
Vanadium (T)		0.001	120	0			0.002	C
Zinc (T)	See text	0.005	120	3	0.005	0.005	0.007	See text

Table 9. Dissolved metals in water: summary of trigger values for Meadowbank and Baker stations.

Notes: For each variable, thresholds (guidelines) are shown if applicable (see text for discussion); DL = current detection limit; N = sample measurements; >DL = number of measurements above DL; medians and 95th percentiles for variables with at least one value >DL (with all values <DL set equal to DL); Method = method used to determine the trigger, where A = halfway from median (or DL if all values <DL) to threshold, B = 95th percentile, and C = 2*DL.

Variable	Threshold	DL	N	>DL	Median	95th	Trigger	Method
Aluminum (D)	See text	0.005	71	2	0.005	0.005	0.016	See text
Antimony (D)		0.0005	71	0			0.0010	C
Arsenic (D)		0.0005	71	0			0.0010	C
Barium (D)		0.02	71	0			0.04	C
Beryllium (D)		0.001	71	0			0.002	C
Boron (D)		0.1	71	0			0.2	C
Cadmium (D)		0.000017	71	0			0.000034	C
Chromium (D)		0.001	71	0			0.002	C
Copper (D)		0.001	71	0			0.002	C
Iron (D)		0.03	71	0			0.06	C
Lead (D)		0.0005	71	0			0.0010	C
Lithium (D)		0.005	71	0			0.010	C
Manganese (D)		0.0003	71	25	0.0003	0.0014	0.0014	B
Mercury (D)		0.00001	71	0			0.00002	C
Molybdenum (D)		0.001	71	0			0.002	C
Nickel (D)		0.001	71	0			0.002	C
Selenium (D)		0.001	71	0			0.002	C
Strontium (D)			6	6	0.0082	0.0092	0.0092	B
Thallium (D)		0.0002	71	0			0.0004	C
Tin (D)		0.0005	71	0			0.0010	C
Titanium (D)		0.01	71	0			0.02	C
Uranium (D)		0.0002	71	0			0.0004	C
Vanadium (D)		0.001	71	0			0.002	C
Zinc (D)		0.005	71	0			0.010	C

Table 10. Nutrients and conventional parameters in water: summary of trigger values for Meadowbank and Baker stations.

Notes: For each variable, thresholds (guidelines) are shown if applicable (see text for discussion); DL = current detection limit; N = sample measurements; >DL = number of measurements above DL; medians and 95th percentiles for variables with at least one value >DL (with all values <DL set equal to DL); Method = method used to determine the trigger, where A = halfway from median (or DL if all values <DL) to threshold, B = 95th percentile, and C = 2*DL.

			Meadowbank (or both)						Baker (if different)					
Variable	Threshold	DL	T	N	Med	95th	Trigger	Method	T	N	Med	95th	Trigger	Method
Ammonia-N	0.126	0.02	120	47	0.020	0.052	0.073	A						
TKN		0.05	110	106	0.097	0.181	0.181	B	10	10	0.180	0.194	0.194	B
Nitrate-N	2.9	0.005	110	3	0.005	0.005	1.45	A	10	10	0.014	0.027	1.46	A
Nitrite-N	0.06	0.001	120	1	0.001	0.001	0.031	A						
Ortho-phosphate		0.001	108	8	0.0010	0.0011	0.002	C						
Total phosphorous	0.004	0.002	98	44	0.0020	0.0047	0.0047	B	10	10	0.0032	0.0061	0.0061	B
TOC			110	110	1.63	2.51	2.51	B	10	10	3.06	3.17	3.17	B
DOC			110	110	1.59	2.31	2.31	B	10	10	3.02	3.42	3.42	B
Reactive silica		1.0	99	0			2.0	C						
Bicarbonate alkalinity			98	98	4.60	9.42	9.42	B	10	10	9.00	9.46	9.46	B
Chloride		0.5	110	52	0.5	0.79	0.79	B	10	10	67.8	119.7	119.7	B
Carbonate alkalinity		2.0	108	0			4.0	C						
Conductivity			110	110	15.9	27.4	27.4	B	10	10	273.5	464.9	464.9	B
Hardness			110	110	5.7	12.3	12.3	B	10	10	32.6	49.5	49.5	B
Calcium			110	110	1.3	3.3	3.3	B	10	10	3.8	5.2	5.2	B
Potassium		2.0	120	2	2.0	2.0	4.0	C						
Magnesium			110	110	0.64	0.96	0.96	B	10	10	5.59	8.90	8.90	B
Sodium		2.0	110	0			4.0	C	10	10	36.2	65.1	65.1	B
Sulphate			110	110	1.43	2.61	2.61	B	10	10	9.53	17.31	17.31	B
pH (Upper)	9.0		110	110	6.80	8.00	8.00	B	10	10	7.17	7.59	8.09	A
pH (Lower)	6.5		110	110	6.80	6.47[a]	6.47	B	10	10	7.17	6.97[a]	6.84	A
Total alkalinity			98	98	4.60	9.42	9.42	B	10	10	9.00	9.46	9.46	B
TDS		10.0	110	62	11.0	19.0	19.0	B	10	10	138.5	245.8	245.8	B
TSS	5.0	1.0	120	8	1.0	1.6	3.0	A						

[a] For pH (Lower), the 5th percentile is reported.

Table 11. Summary of CCME guidelines for sediment variables.

Variable	Source	Description of guidelines
t-Arsenic (As)	CCME	The CCME ISQG for As is 5.9 mg/kg.
t-Cadmium (Cd)	CCME	The CCME ISQG for Cd is 0.6 mg/kg.
t- Chromium (Cr)	CCME	The CCME ISQG for Cr is 37.3 mg/kg.
t-Copper (Cu)	CCME	The CCME ISQG for Cu is 35.7 mg/kg
t-Lead (Pb)	CCME	The CCME ISQG for Pb is 35 mg/kg.
t-Mercury (Hg)	CCME	The CCME ISQG for Hg is 0.17 mg/kg.
t-Zinc (Zn)	CCME	The CCME ISQG for Zn is 123 mg/kg.

Table 12. Summary of trigger values for sediment variables for Meadowbank and Baker stations.

Notes: For each variable, thresholds (guidelines) are shown (details for thresholds are in **Table 7**); DL = current detection limit; N = sample measurements; >DL = number of measurements above DL; medians and 95th percentiles for variables with at least one value >DL (with all values <DL set equal to DL); Method = method used to determine the trigger, where A = halfway from median (or DL if all values <DL) to threshold, and B = 95th percentile.

			Meadowbank (or both for Lead)						Baker					
Variable	Threshold	DL	N	>DL	Med	95th	Trigger	Method	N	>DL	Med	95th	Trigger	Method
Arsenic (As)	5.9	5.0	135	133	26.5	120.0	120.0	B	15	3	5.0	8.3	8.3	B
Cadmium (Cd)	0.6	0.50	135	64	0.50	1.10	1.10	B	15	0			0.55	A
Chromium (Cr)	37.3		135	135	76.5	114.3	114.3	B	15	15	16.7	20.3	27.0	A
Copper (Cu)	35.7		135	135	61.8	126.0	126.0	B	15	15	5.3	7.8	20.5	A
Lead (Pb)	35	30	150	11	30	32	32.5	A						
Mercury (Hg)	0.17	0.005	135	135	0.038	0.064	0.104	A	15	13	0.006	0.011	0.088	A
Zinc (Zn)	123		135	135	91.0	121.3	121.3	B	15	15	24.1	32.1	73.6	A

Table 13. Sampling frequency and intensity for CREMP components

Note: Wally is shown together with Meadowbank for simplicity, although intensive sampling for Wally has not yet begun.

	Baker Lake	Meadowbank / Wally
Water Chemistry for Annual Period Statistical Analysis (all samples at 3 m from surface)	• Open water months (July, Aug, Sept) at all stations • 2 randomly located subsamples per station per event	• Open water months (July, Aug, Sept) for PDL • All other stations: Up to 6 months per year (April, May, July, Aug, Sept, Oct/Nov) depending on logistical & safety constraints • 2 randomly located subsamples per station per event
Water Chemistry Mid-Winter (all samples at 3 m from surface)	• None	• Minimum one sampling event at stations TPN, TPE, SP, WAL
Sediment Chemistry Cores	• Sampling in 2012, 2014, and approximately every three years thereafter, always in open water (August) • All impact stations; for control stations only BAP and INUG • Analyze 5 subsamples per station per event, and (except for INUG and BAP) archive an additional 5 samples (to be used only if the initial 5 subsamples suggest differences)	
Phytoplankton	• Sample collection same stations and frequency and depth as water chemistry, but only open water (July, Aug, Sept) samples should be analyzed – all other samples to be archived • 2 randomly located subsamples per station per event	
Benthic Community	• Sample collection once per year in August at all stations • 5 subsamples per station per event	

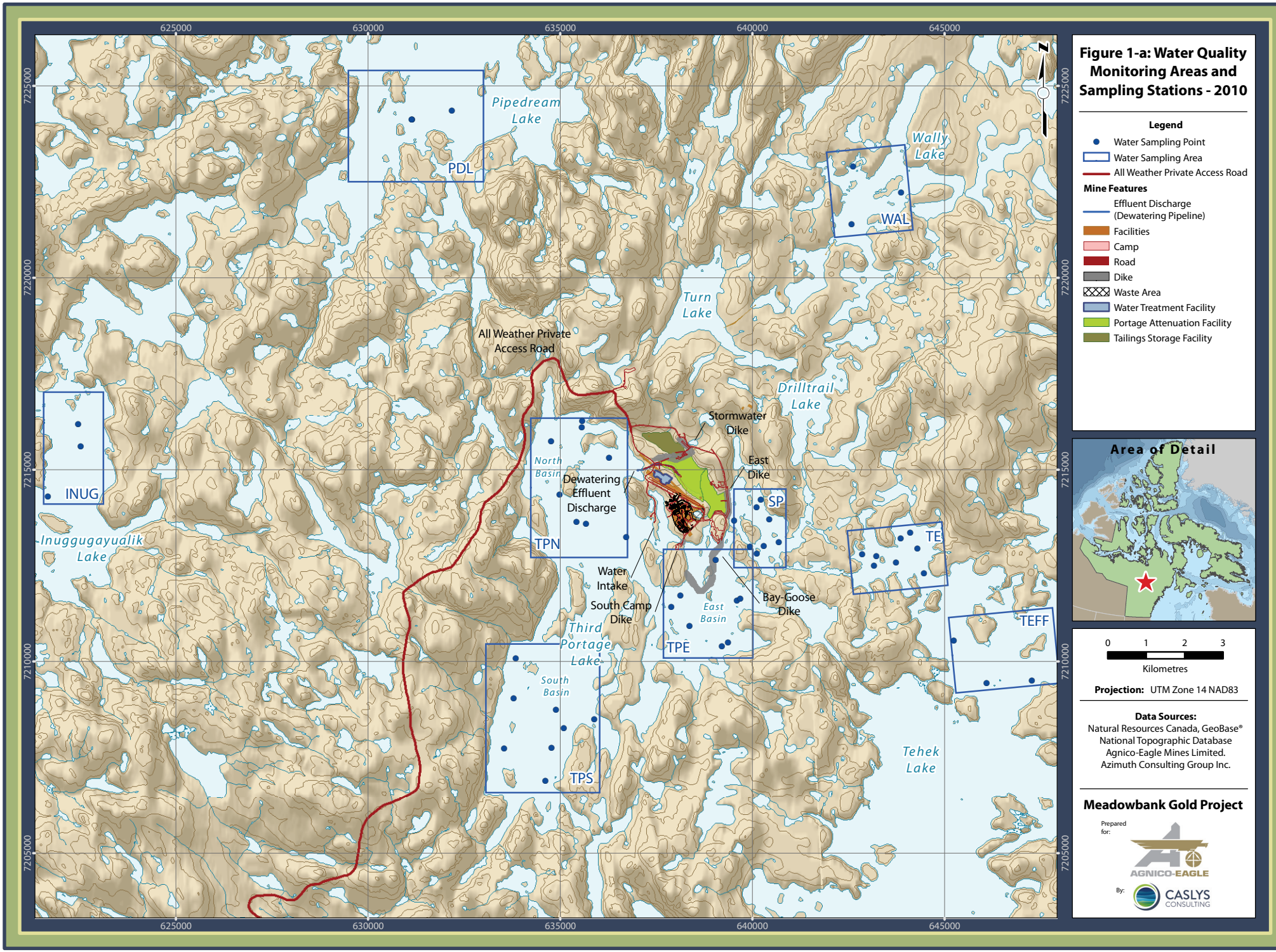


Figure 1-a: Water Quality Monitoring Areas and Sampling Stations - 2010

Legend

- Water Sampling Point
- Water Sampling Area
- All Weather Private Access Road

Mine Features

- Effluent Discharge (Dewatering Pipeline)
- Facilities
- Camp
- Road
- Dike
- Waste Area
- Water Treatment Facility
- Portage Attenuation Facility
- Tailings Storage Facility

Area of Detail

0 1 2 3

Kilometres

Projection: UTM Zone 14 NAD83

Data Sources:
Natural Resources Canada, GeoBase®
National Topographic Database
Agnico-Eagle Mines Limited.
Azimuth Consulting Group Inc.

Meadowbank Gold Project

Prepared for: AGNICO-EAGLE

By: CASLYS CONSULTING

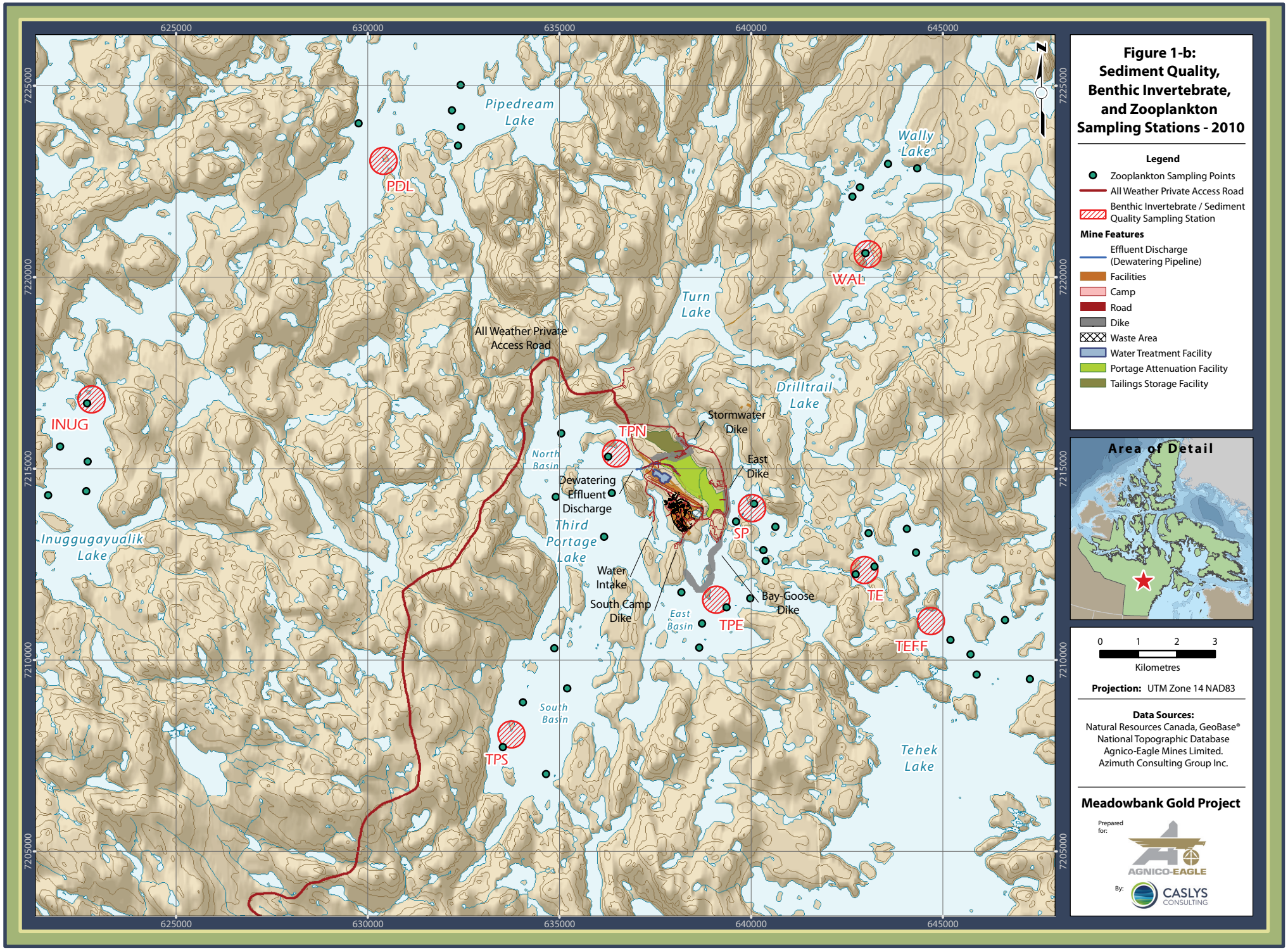


Figure 1-b: Sediment Quality, Benthic Invertebrate, and Zooplankton Sampling Stations - 2010

Legend

- Zooplankton Sampling Points
- All Weather Private Access Road
- Benthic Invertebrate / Sediment Quality Sampling Station

Mine Features

- Effluent Discharge (Dewatering Pipeline)
- Facilities
- Camp
- Road
- Dike
- Waste Area
- Water Treatment Facility
- Portage Attenuation Facility
- Tailings Storage Facility

Area of Detail

0 1 2 3

Kilometres

Projection: UTM Zone 14 NAD83

Data Sources:
Natural Resources Canada, GeoBase®
National Topographic Database
Agnico-Eagle Mines Limited.
Azimuth Consulting Group Inc.

Meadowbank Gold Project

Prepared for: AGNICO-EAGLE

By: CASLYS CONSULTING

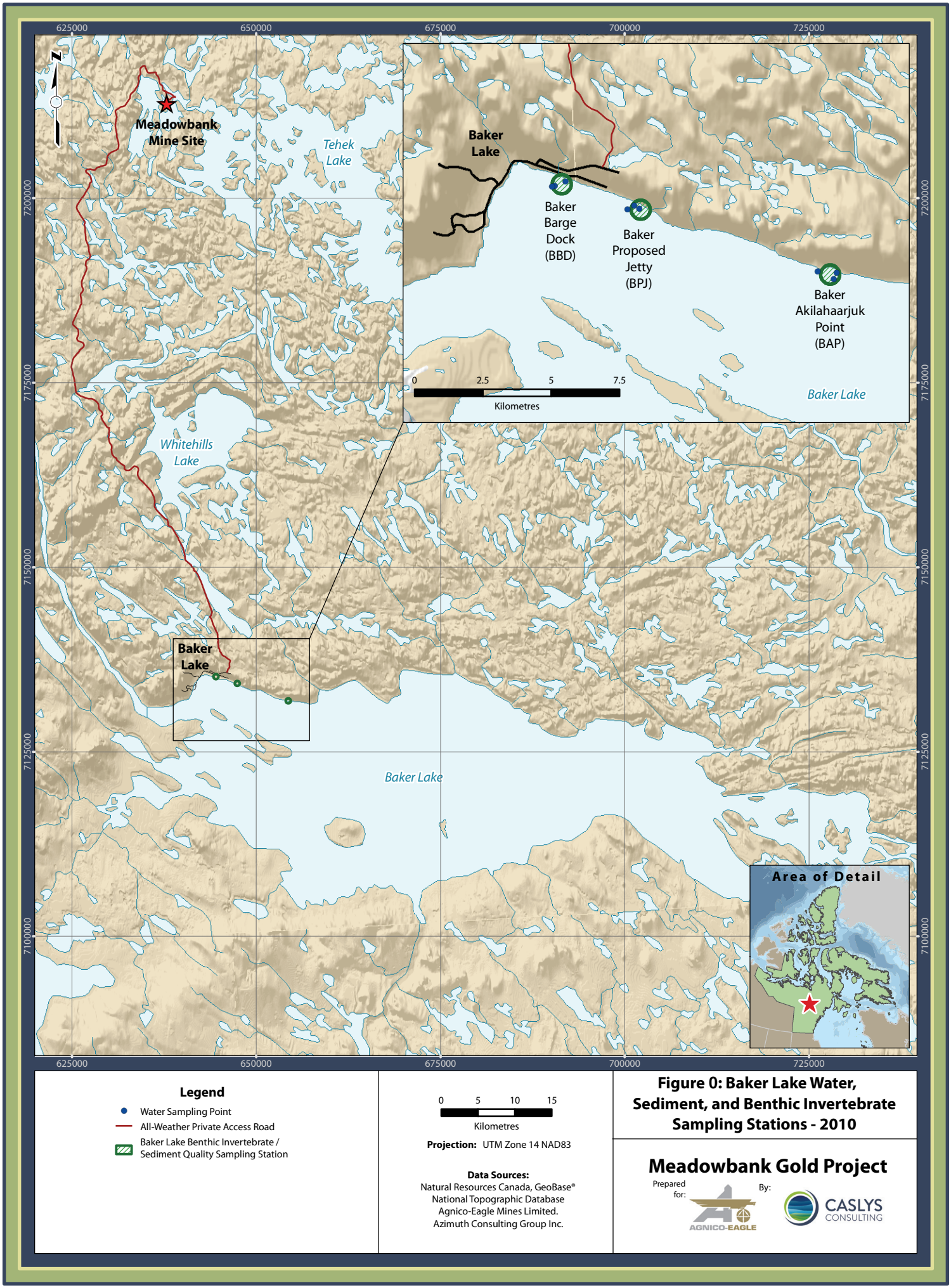


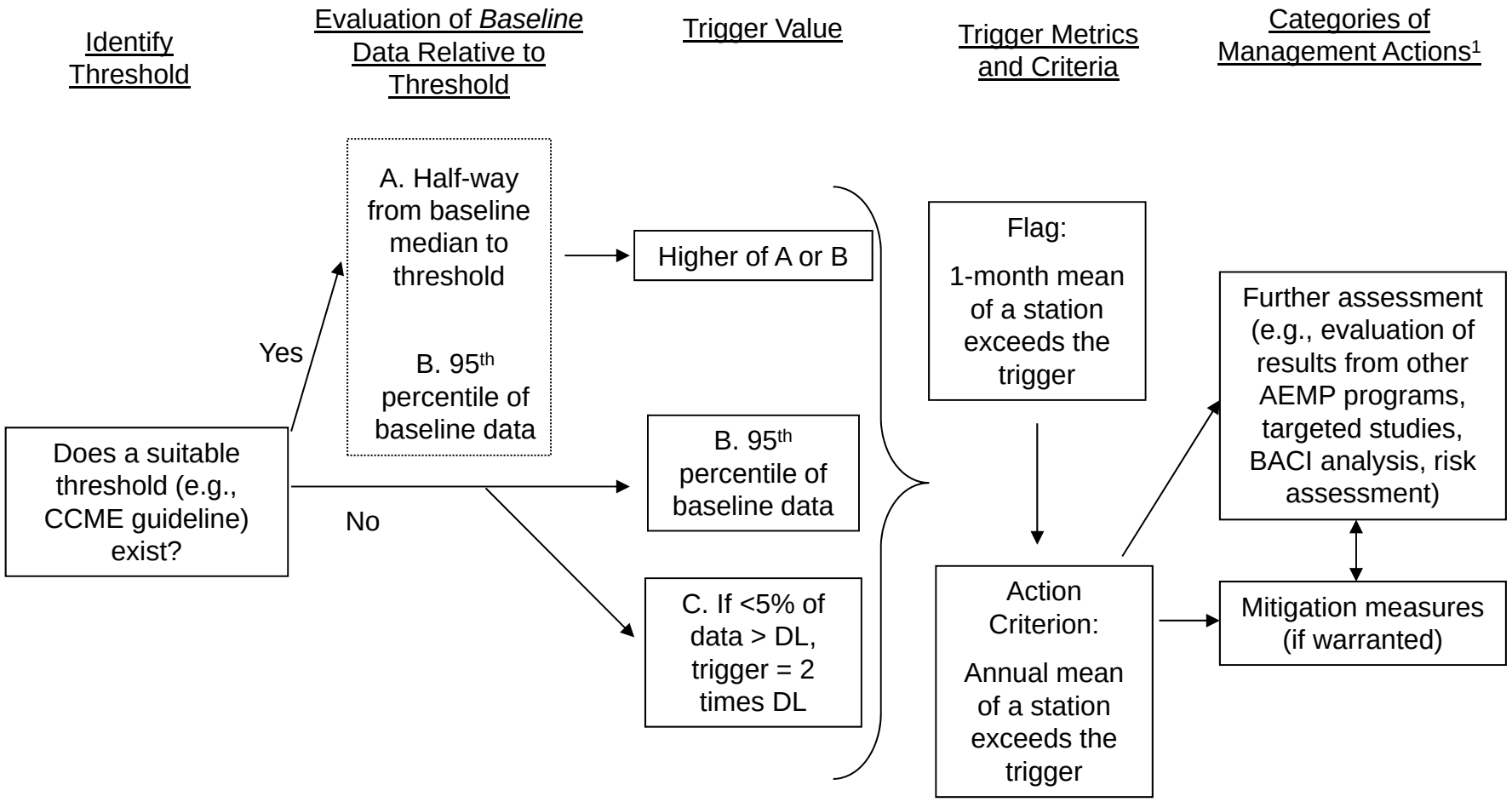
Figure 0: Baker Lake Water, Sediment, and Benthic Invertebrate Sampling Stations - 2010

Meadowbank Gold Project

Prepared for: AGNICO-EAGLE

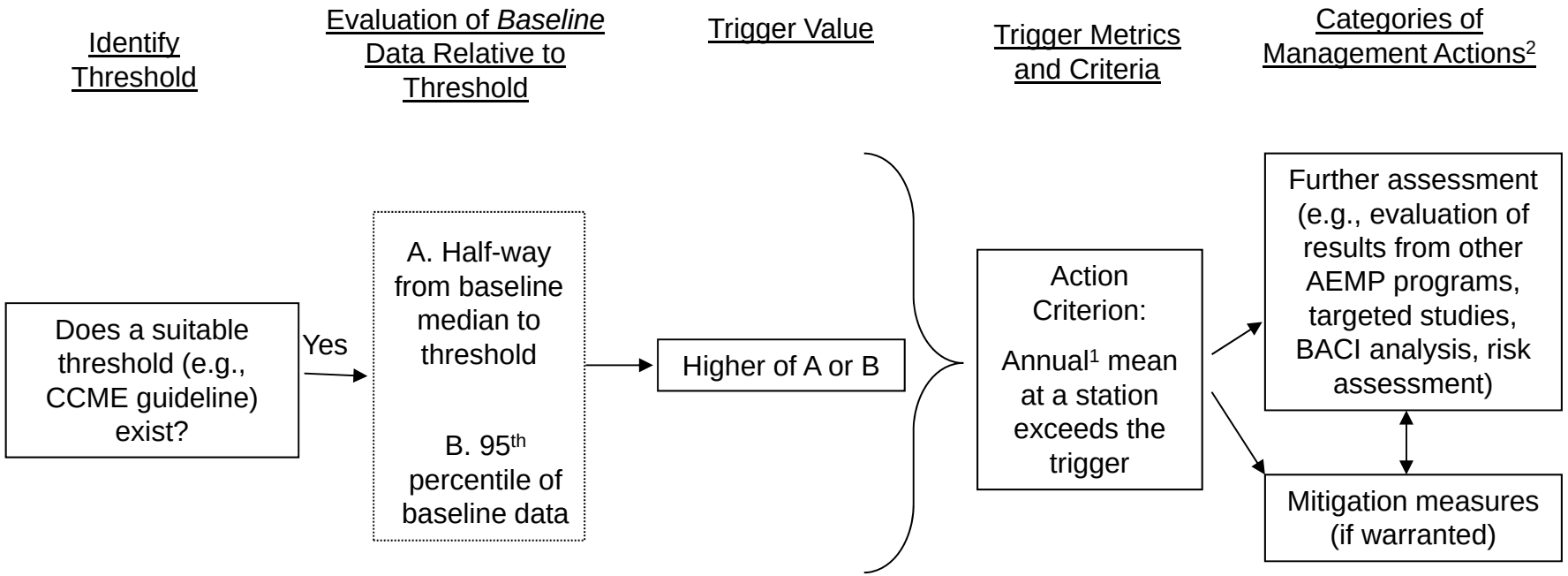
By: CASLYS CONSULTING

Figure 3. Development and Application of Monitoring Thresholds and Triggers for Water Chemistry Variables



[1] Potential management actions are considered as part of the broader AEMP, where results of the CREMP and other programs are considered together.

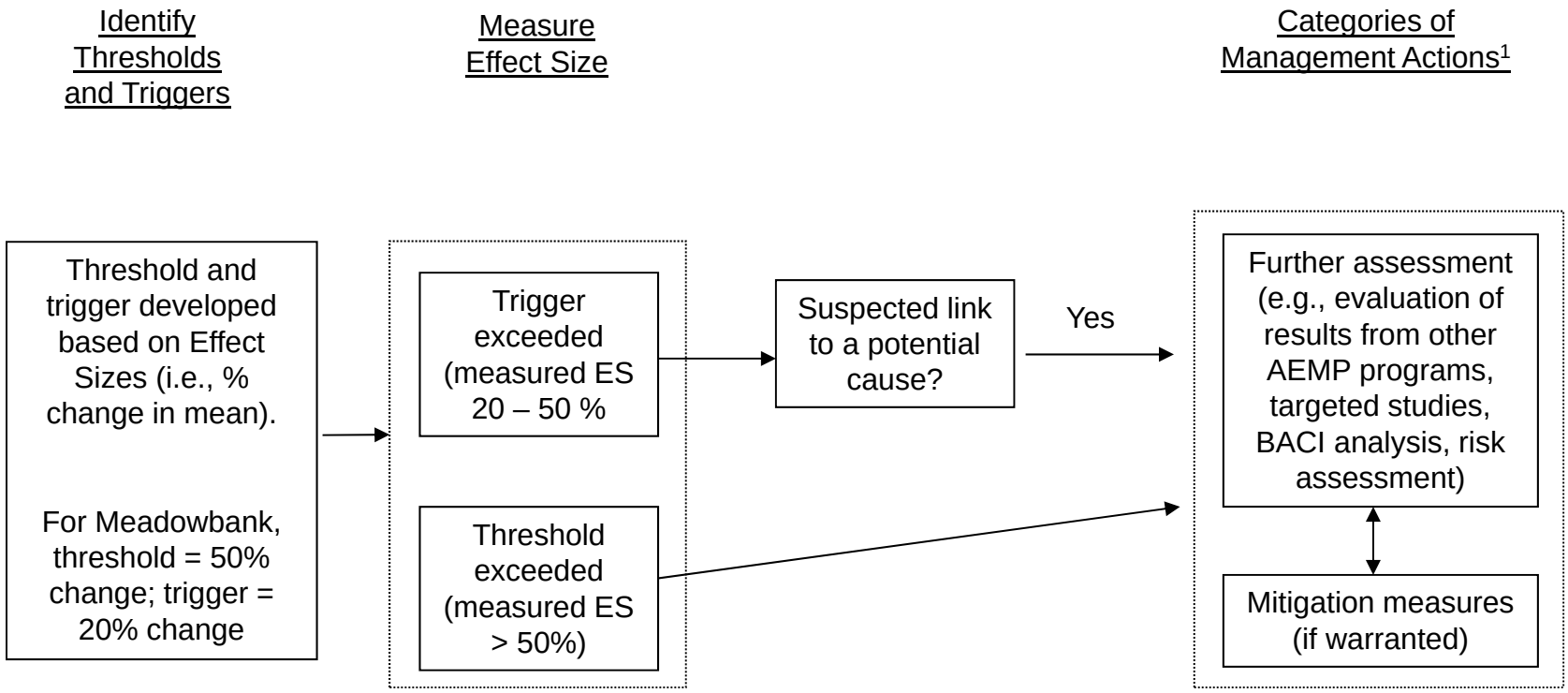
Figure 4. Approach for Development and Application of Monitoring Thresholds and Triggers for Sediment Chemistry Variables



[1] Annual means for sediment chemistry variables are calculated only in the years when sediment chemistry data are collected.

[2] Potential management actions are considered as part of the broader AEMP, where results of the CREMP and other programs are considered together.

Figure 5. Approach for Development and Application of Monitoring Thresholds and Triggers for Biological Effects Variables



[1] Potential management actions are considered as part of the broader AEMP, where results of the CREMP and other programs are considered together.

Figure 6. Estimates of BACI statistical power for detecting a significant increase (one-tailed test, alpha = 0.05) in a given water chemistry variable (rows) by station (columns) as a function of sampling months (after period) and the number of sub-samples per month.

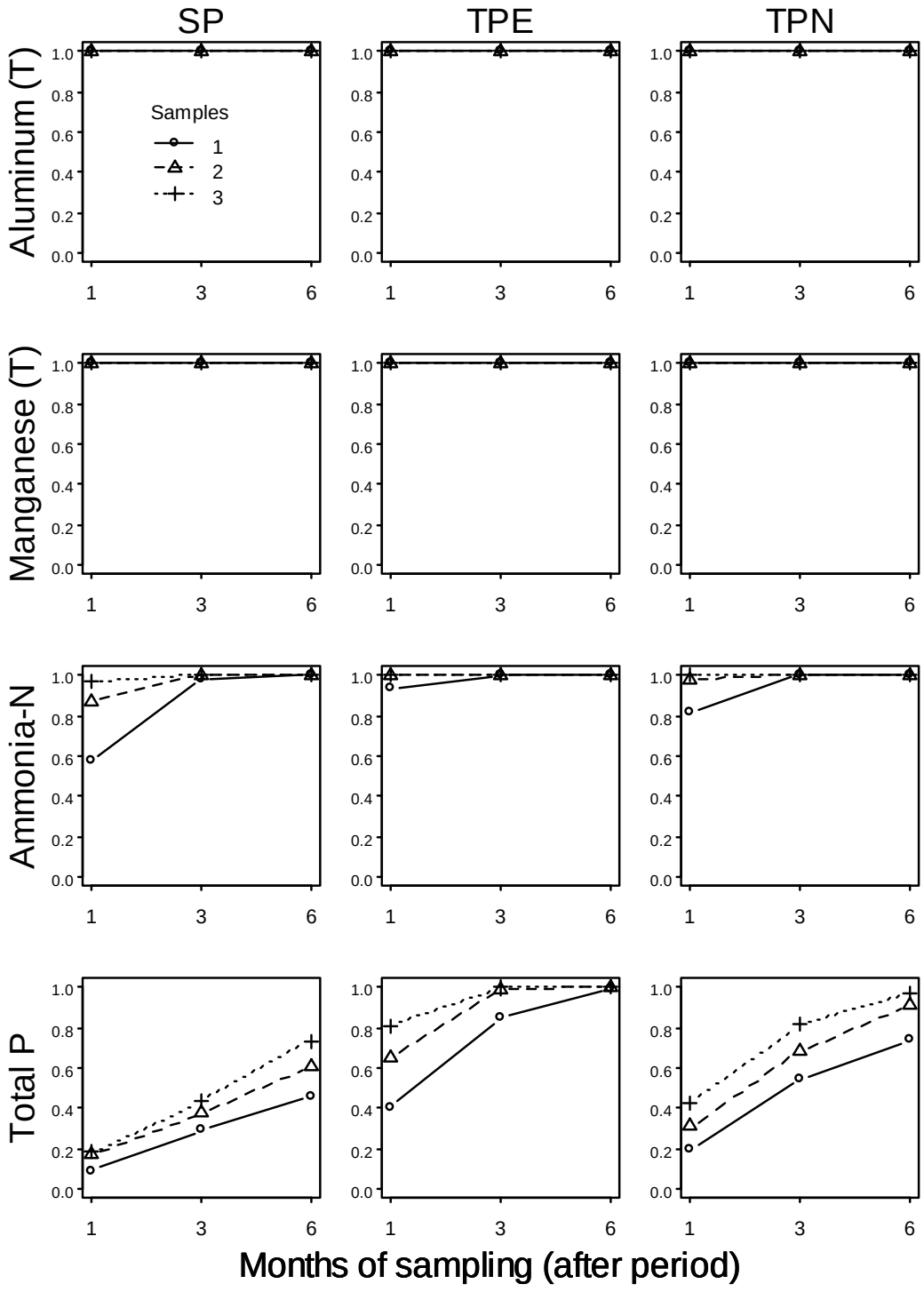


Figure 7. Estimates of BACI statistical power for detecting a significant increase in pH (top row) or decrease in pH (bottom row) by station (columns) for two-tailed tests (alpha = 0.05) as a function of sampling months (after period) and the number of sub-samples per month.

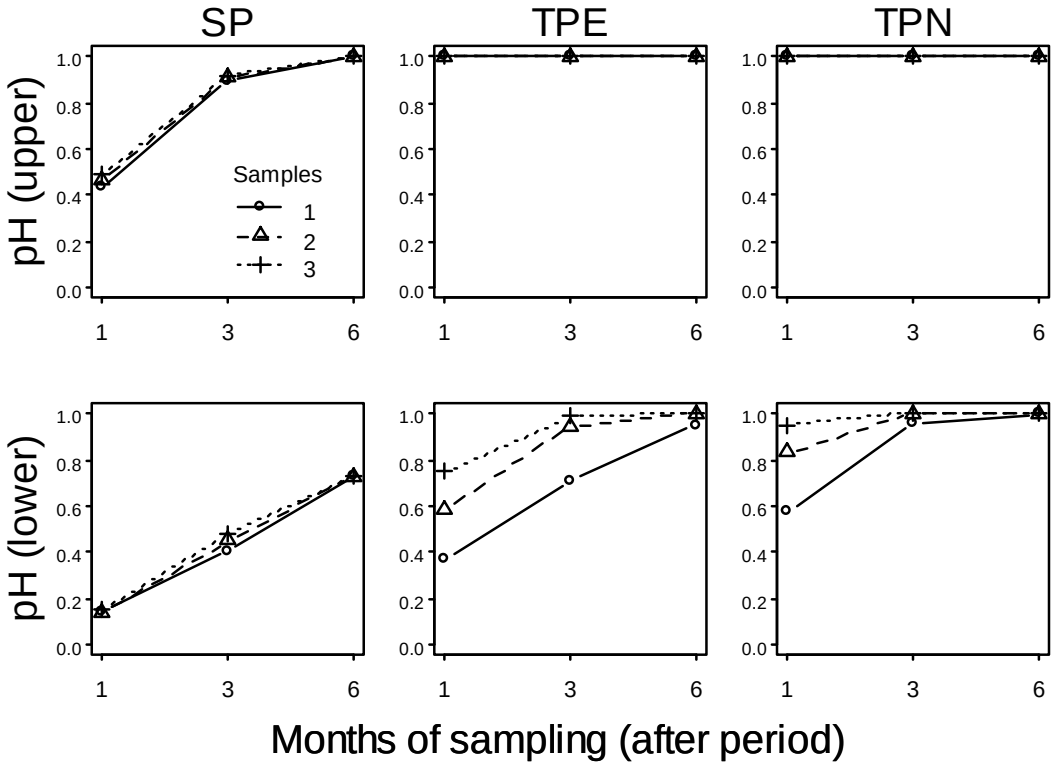


Figure 8. Estimates of BA statistical power for detecting a significant increase (one-tailed test, alpha = 0.05) in a given sediment chemistry variable (rows) by station (columns) as a function of the number of sub-samples in the impact year (after period).

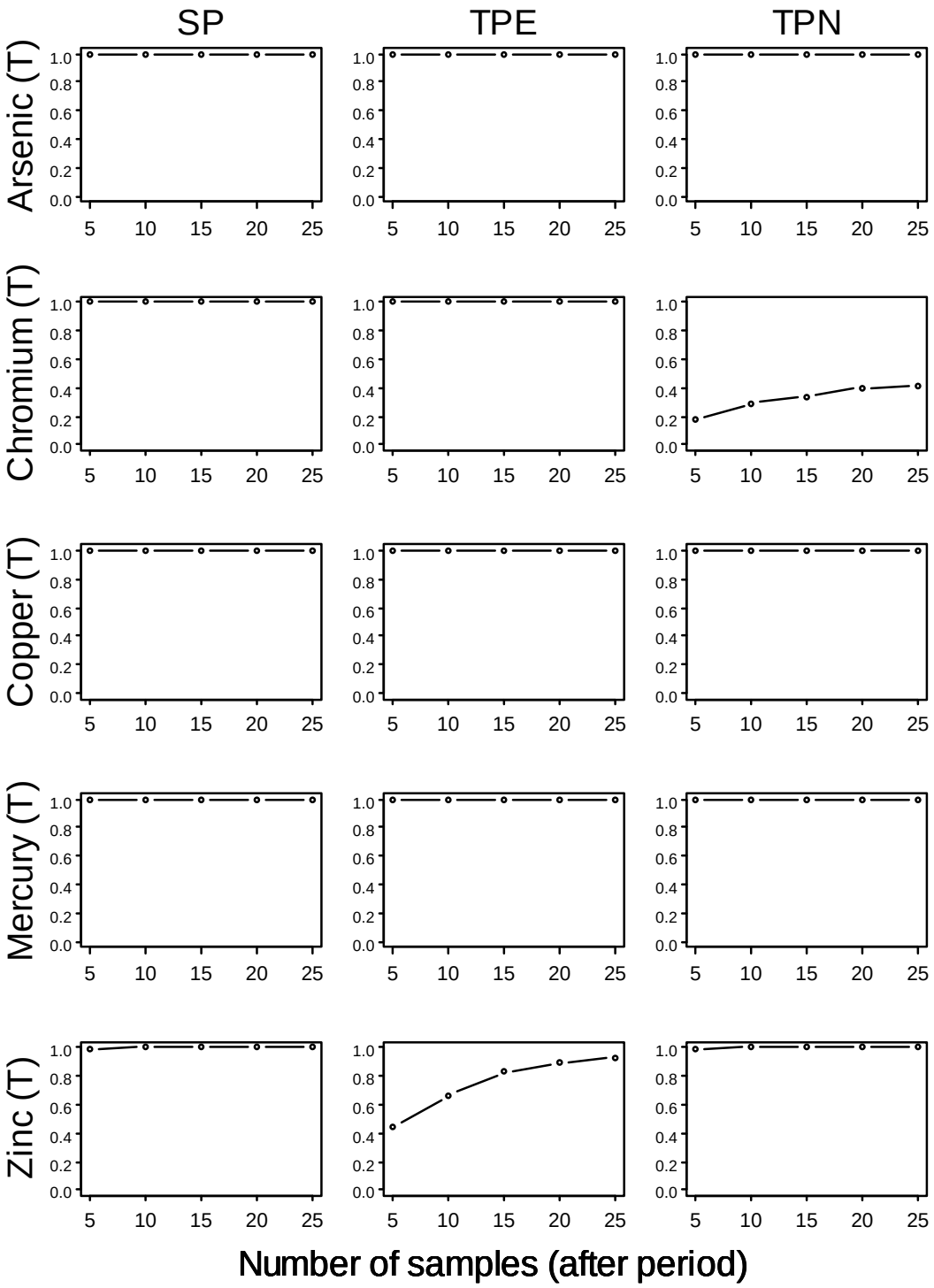


Figure 9. Estimates of BACI power for detecting a significant decrease (one-tailed, alpha = 0.05) in a given phytoplankton variable (Total biomass or Simpsons diversity) and effect size (rows) by station (columns) as a function of sampling months (after period) and the number of sub-samples per month (see legend).

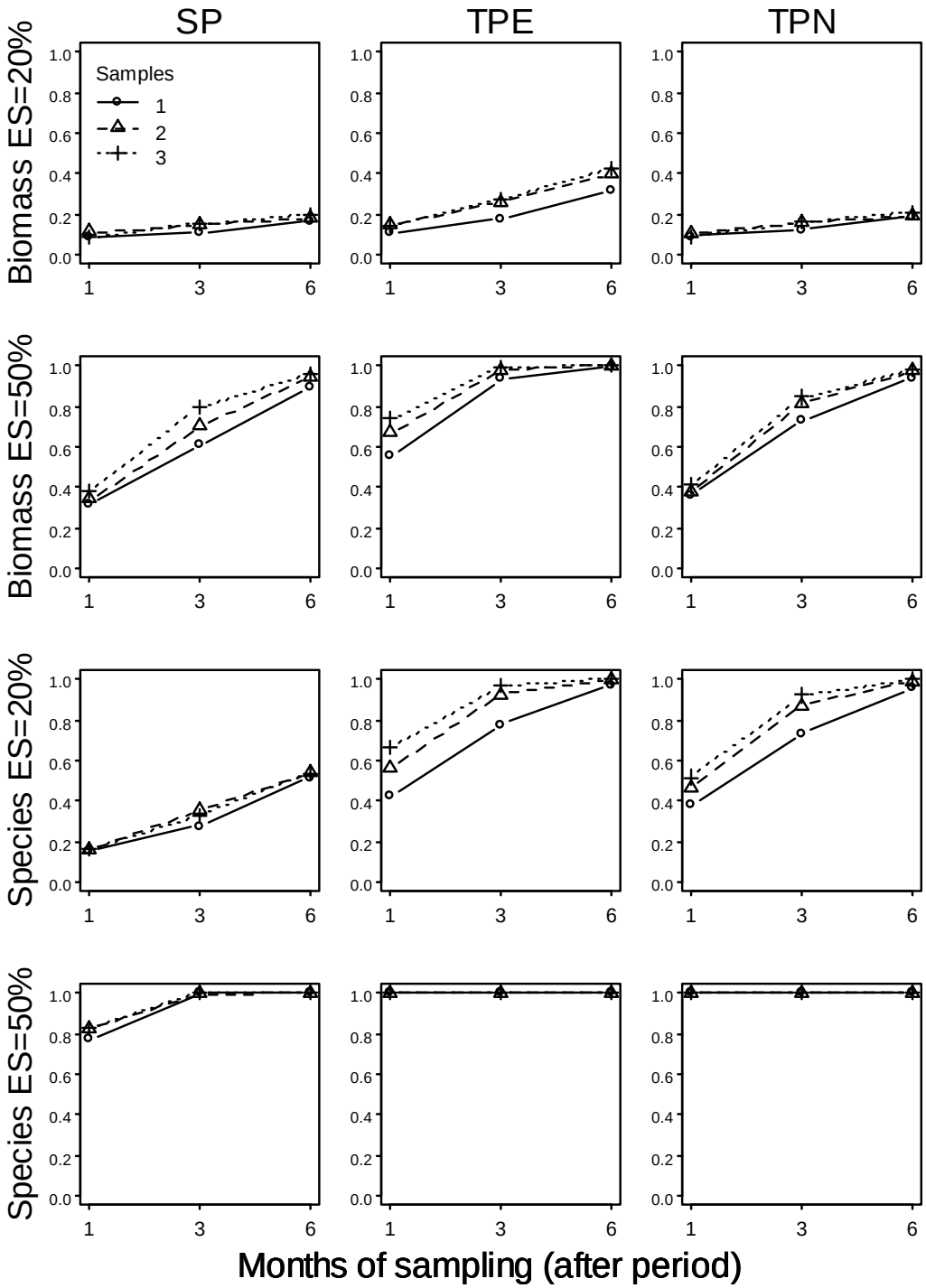


Figure 10. Estimates of BACI power for detecting a significant decrease (one-tailed, alpha = 0.05) in a given benthic invertebrate community variable (Total N or Total taxa) and effect size (rows) by station (columns) as a function of sampling years (after period) and the number of sub-samples per year (see legend).

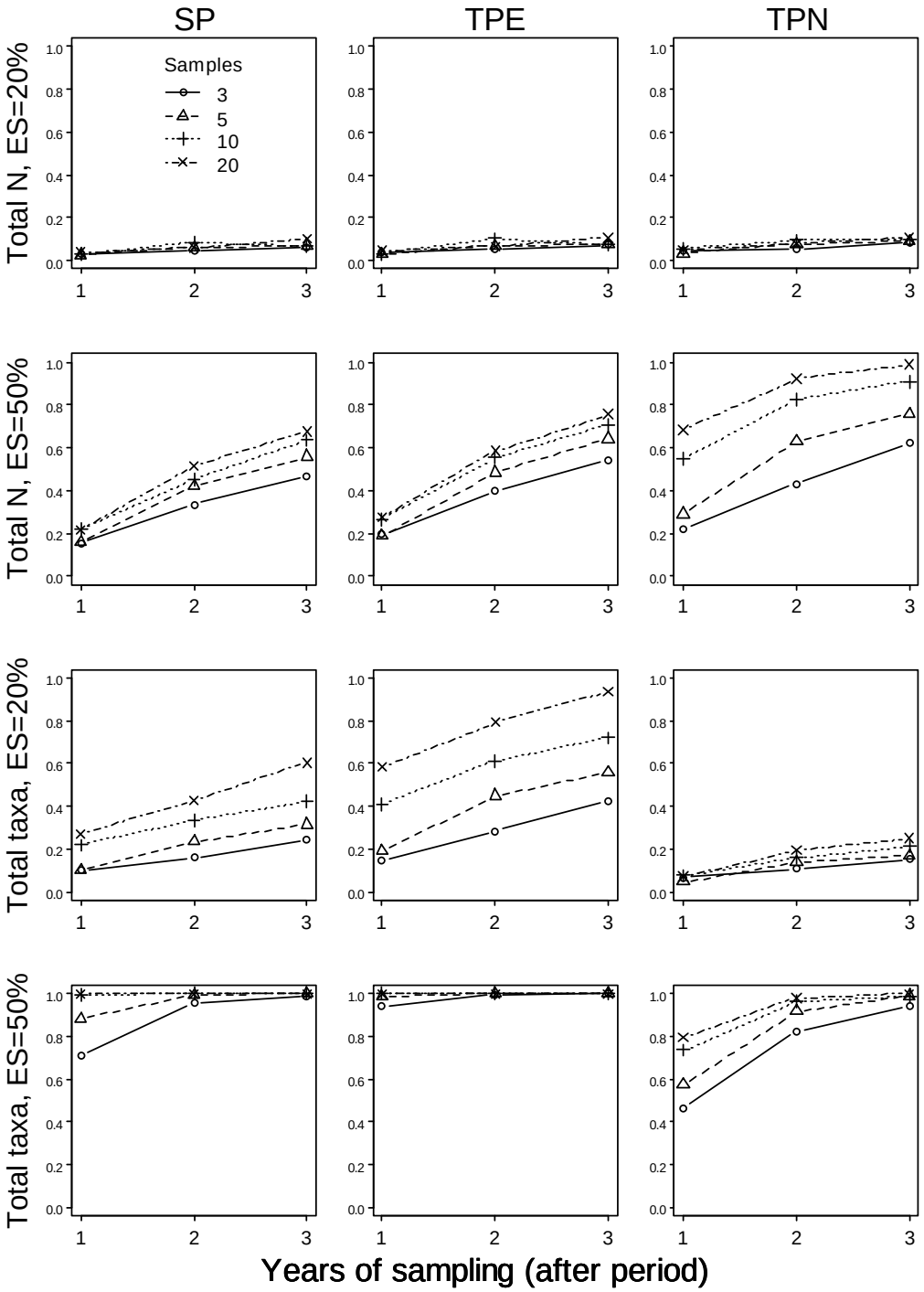
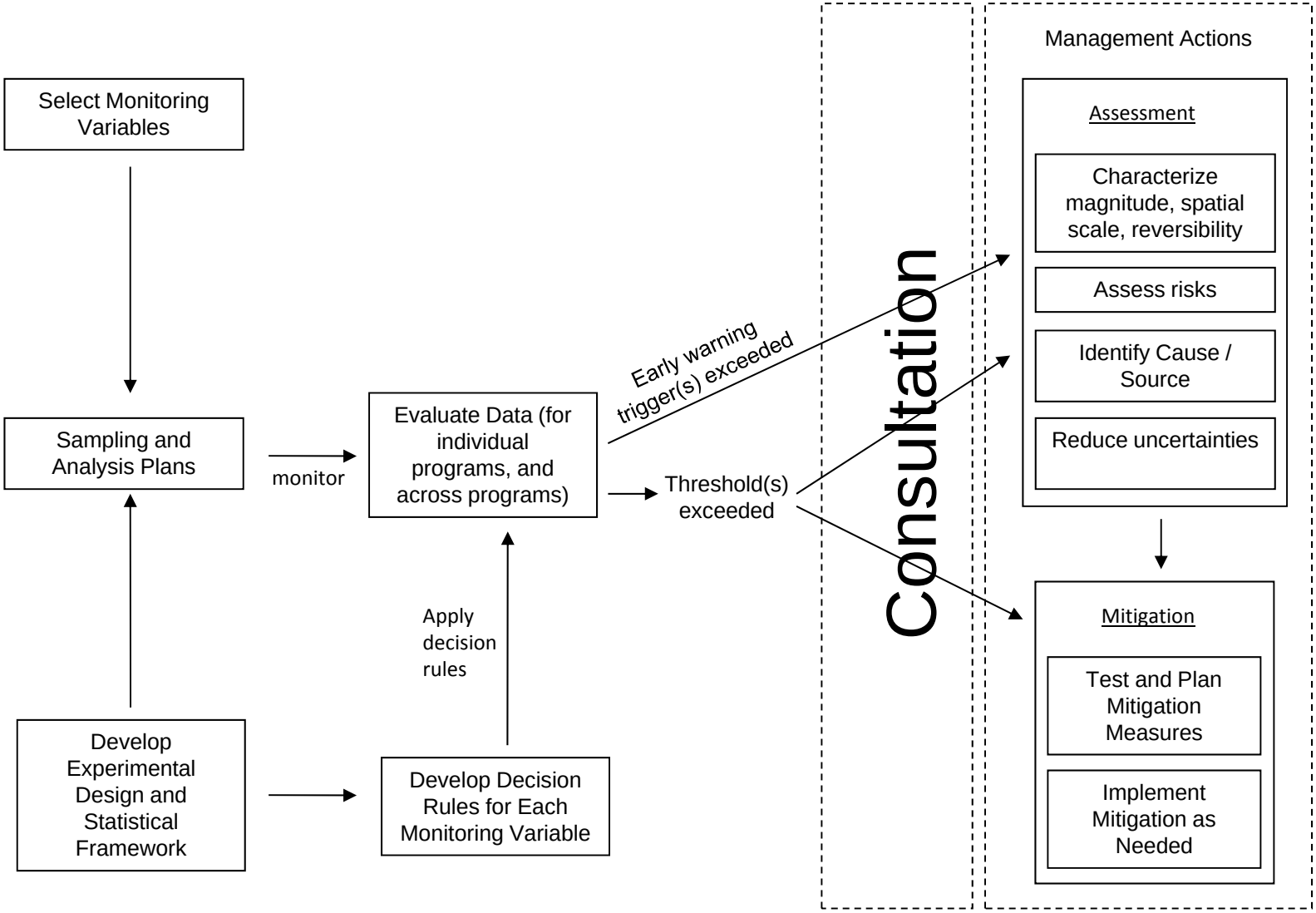


Figure 11. Management Response Plan for the Meadowbank Mine AEMP[1]



[1] Management responses under the AEMP are based on the combined results of individual monitoring programs including the CREMP.

APPENDIX A – STATISTICAL ANALYSES FOR WATER CHEMISTRY

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This appendix contains the following analyses.

- Summary of CREMP water samples
- Analysis of field duplicates
- Analysis of depth
- Analysis of Baker/Meadowbank differences
- Determination of triggers
- Analysis of sampling design

Some text, tables and figures are repeated from the main document so that the analyses contained in this appendix can be read and understood without reference to the main document. The material in this appendix assumes understanding of basic statistical methods (Venables and Ripley 2002), mixed-effects models (Pinheiro and Bates 2000), before-after-control-impact (BACI) experimental design (Stewart-Oaten et al. 1986; Underwood 1994; Smith 2002), and use of simulation in statistical analysis (Gelman and Hill 2006).

2. SUMMARY OF CREMP WATER SAMPLES

There are three basic types of samples:

1. Standard samples – specific to a given location, month (a single sampling day), and depth (typically surface samples).
2. Depth replicates – collected at a given location and time (typically one surface, one bottom, and one integrated sample).
3. Duplicate samples – occasionally collected; these consisted of a second water sample specific to a given location, time, and depth.

For many station-month combinations, more than one "standard" sample was collected. These "sub-samples" represent spatial replication (a different location within the designated station). In addition, samples were designated as "control" or "impact" depending on operation activities near a given station (see main text for differentiation of control and impact).

For the CREMP program (current dataset), a total of 352 samples have been collected across 5 years (2006 up to and including May 2010) and 12 stations (Baker Lake and Meadowbank programs). The full data set is summarized in Table 1. Of the total of 352

samples, 242 were considered standard samples (i.e., unique location/day samples; **Table 2**); 76 were depth replicates of surface samples (i.e., either bottom or integrated samples; **Table 3**), and 30 were duplicate samples (**Table 4**). The distribution of samples by month and station are shown in **Tables 1-4** for total samples, standard samples, depth replicates, and duplicates; shaded areas indicate designated impact periods. Of the 242 standard samples, all were surface samples except for 15 integrated samples (collected mostly in July/August of 2007); 120 samples were designated control and 126 samples designated impact.

3. ANALYSIS OF FIELD DUPLICATES

Field duplicates are water samples collected at the same location, time and depth as a standard sample. If there is little difference between duplicates and the standard samples with which they are associated, then duplicates should be omitted when evaluating data because they are pseudo replicates of the standard samples. On the other hand, if the differences are large (i.e., similar to what is observed among standard samples), there may be rationale for including the duplicates when evaluating data. Presumably, differences among duplicates largely reflect field sampling error and lab precision (measurement error). This variation will be a component of the variation observed among station/month sub-samples. Thus, comparing the two provides insight into the extent of real sub-sample variation (due to differing locations within a station).

Data – Duplicates have been collected for about 10% of samples (n=30; see **Table 4**). We would not expect control and impact samples to differ in terms of duplicate measures, so all samples were examined (control = 9; impact = 21). Analysis was limited to variables with at least 30 sample measurements above detection limits (with 30 pairs of sample/duplicate values, the total number of possible measurements was 60). **Tables 5-7** summarize the number of measurements and values above detection limits for total metals (**Table 5**), dissolved metals (**Table 6**), and nutrients/conventional parameters (**Table 7**). For some samples, a given variable was not measured and hence $n < 60$. In total, there were 15 variables with at least 30 measurements above the detection limits: t-Al, t-Mn, TKN, t-P, TOC, DOC, bicarbonate alkalinity[1], Cl, conductivity, hardness, Ca, Mg, sulphate, pH and TDS.

Methods – For each variable, xy plots were produced using all data, with values below DL set equal to DL. For remaining analyses, data pairs (standard sample and its duplicate) were limited to those pairs with both values above DL. For each variable, we computed the median across all samples, the median absolute difference (MAD) between

[1] Total alkalinity was excluded as the values were identical to bicarbonate alkalinity

samples and duplicates, and Spearman rank correlation between samples and duplicates. In addition, we fit a mixed-effects model (Pinheiro and Bates 2000) of the form:

Log(y) ~ Location + Error

where y is the water quality variable, Location denotes random effects for sample pairs (variability among pairs due to sample location/month) and Error denotes the residual error (variability between samples and duplicates).

Results – Results of the analysis of duplicates are shown in **Figure 1** (xy plots) and **Table 8** (medians, MADs, and Spearman r). For most variables, the duplicates are very similar to the associated standard samples, with a notable exception for TKN, and occasional anomalous values for some variables. The standard deviations (SDs) resulting from the mixed-effect model show the relative variability due to sample location/month (SD.Location) versus duplicate variability (SD.Error). In most cases SD.Error is much smaller than SD.Location, which just reaffirms that duplicate measures are very precise compared to general sample variation (i.e., minimal measurement error or fine-scale spatial variation). In the case of TKN, however, the pattern is reversed, suggesting that measurement precision for TNK is quite poor (obvious from plot in **Figure 1**). Results for t-P and pH were also reversed, but not after removing the single anomalous data pair for each of those variables. These results suggests that while sample/duplicate measures are reasonably consistent for t-P and pH, there is potential for occasional large differences due to laboratory measurement error or fine-scale spatial variation.

4. ANALYSIS OF DEPTH

Most CREMP water samples have been collected near the surface. However, some data have been collected near the bottom or as integrated samples whereby the water is collected across a depth range. Evaluation of the importance of depth was needed in order to determine if surface samples are likely to be representative of the water column, and if not, to determine whether sampling at different depths is warranted in future.

The primary null hypothesis is that a given variable is the same across sample types for depth (i.e., surface, bottom, integrated); however, depth differences may be related to station and especially season. Potential stratification of the water column is expected to be much greater during "winter" months (November through May) than for "summer" months (July, August, September), though July may be more similar to winter conditions.

Data – In the there were 39 cases or "depth tests" where water samples were simultaneously collected at differing depths, with samples classified as surface (n = 39), bottom (n = 37), or integrated (n = 39). The total sample size was therefore 115. The

depth replicates (i.e., bottom and/or integrated samples) were summarized in **Table 3**, and can be further delineated as follows:

Type	INUG	TE	TPS	Total
Control	11		9	20
Impact		19		19
Total	11	19	9	39

Month/Year	INUG	TE	TPS	Total
May.09	4	4		8
Jul.09	2	2		4
Aug.09	1	1		2
Sep.09	2	2		4
Nov.09	2	2		4
Dec.09		1	1	2
Jan.10		1	2	3
Feb.10		1	1	2
Mar.10		2	2	4
Apr.10		1	1	2
May.10		2	2	4
Total	11	19	9	39

A more detailed summary of the 115 depth samples is presented in **Table 9.** Including impact samples doubles the sample size and provides contrast (via station TE) across months, consequently TE is included in the analyses below.

Methods – Analysis was limited to variables that had at least 40 of the 115 sample measurements above detection limits (for some samples, a given variable was not measured and hence n < 115). **Tables 5-7** summarize the number of measurements and values above detection limits for total metals (**Table 5**), dissolved metals (**Table 6**), and nutrients/conventional parameters (**Table 7**). In total, there were 14 variables with at least 40 measurements above the detection limits: t-Al, t-Mn, TKN, TOC, DOC, bicarbonate alkalinity[2], Cl, conductivity, hardness, Ca, Mg, sulphate, pH and TDS. For each variable, xy plots were produced using all data, with values below DL set equal to DL. In addition, we fit three mixed-effects models (Pinheiro and Bates 2000) of the following forms:

Model 1: Log(y) ~ Type + Location + Error

Model 2: Log(y) ~ Type + Station + Type*Station + Location + Error

Model 3: Log(y) ~ Type + Season + Type*Season + Location + Error

where "Type" (with fixed effects) is the key factor variable of interest denoting depth types (surface, bottom, integrated), "Station" is the CREMP sampling station (e.g., TPE)

[2] Total alkalinity was excluded as the values were identical to bicarbonate alkalinity

that is actually a broad area, and "Location" denotes unique sample locations (with random effects) that were sampled within a given station at a given time. In Models 2 and 3, the terms of interest are the interactions (fixed effects) between Type and Station (i.e., potential station-specific effects of depth) and Type and Season (potential season-specific effects of depth). Two options were examined for season models. First," Season" was defined as a factor variable with two levels separating Summer months (July, August, September) and Winter months (all others). In the second analysis, July was considered a Winter month. For all models, data were limited to those cases (locations) in which all three sample values >DL (i.e., each location had to include valid measures for surface, bottom, and integrated samples). This was done to ensure that consistent data were used to compare the three sample types. Thus, cases 24 and 25 (**Table 9**) were always omitted because they lacked "bottom" samples, providing a maximum sample N = 111. For Model 3 (Season) analysis, data for station TPS were excluded because no "summer" months were sampled (see Table 3; thus maximum n=84).

Results – The statistical significance for the key depth terms of models 1 to 3 is presented in **Table 10** (p-values, based on F tests). There were significant differences ($P < 0.05$) among depth samples (Type) for 8 of the 14 variables, but relatively few cases of significant interactions between depth type and station or season. Consequently, we focus initially on results for Model 1. **Table 11** shows depth type coefficients for Model 1. The coefficients were defined as "Bottom – Surface" and "Integrated – Surface" to make explicit comparisons with surface samples (the standard sample type). Since all data were log-transformed, the coefficient (X) has to be translated into a proportional effect size (ES) relative to (untransformed) surface samples[3]. The percentage ES for each variable is shown in **Table 11**. For example, for Aluminum (T), the estimated ES = -15.9% for bottom samples (i.e., Aluminum (T) in bottom samples was 15.9% less than that observed in surface samples, on average) and ES = 12.9% for integrated samples (i.e., 12.9% greater than for surface samples, on average). In general, the predominant pattern was toward lower values in bottom samples relative to surface samples (negative ES for 12 of 14 variables, $P < 0.05$ for 10 of these[4]). The trend was less clear for integrated samples, with only four significant coefficients, but these were also negative. **Figure 2** shows bottom v. surface and integrated v. surface measures for each variable (values < DL set equal to DL). Also shown are the implied differences (ES in **Table 11**) for which $P < 0.1$ (dashed lines, which in log-log scale are parallel lines to the 1:1 line).

[3] The following steps are needed for calculating ES: (1) X = log(Bottom) – log(Surface); (2) X = log(Bottom/Surface); (3) Exp(X) = Bottom/Surface; (4) ES = (Bottom – Surface)/Surface = Bottom/Surface – 1; (5) ES = Exp(X) – 1; (6) ES (%) = (Exp(X) – 1)*100.

[4] Removing one outlier for pH for the bottom-surface comparison reduces the p-value for pH from 0.562 to 0.018, though the ES is still marginal (-0.6%).

The key question of interest is how large are the significant depth differences compared to variation among locations and among depth replicates (sample types). As shown in **Figure 2**, differences between solid and dashed lines are minimal compared to the full data ranges. As an example, the variable with the largest depth effect was t-Al – its percent ES was -15.9%, yet the variation among surface samples / areas is an order of magnitude[5]. Importantly, most variables have effect sizes much lower than that of t-Al.

For model 2, there were two variables with significant depth interactions (**Table 10**): t-Al and TOC. Box-plots capture the marginal station-depth differences found in these data (**Figure 3**). For both variables, there are strong differences among stations but less so among depth types. For t-Al, the significant interaction for Type*Station was largely due to station TPS, for which integrated samples were clearly higher than surface/bottom samples, whereas for INUG and TE the general pattern was surface concentrations were higher than bottom concentrations, and surface and integrated samples were roughly similar. For TOC, the differences are subtle, but model coefficients reveal a significance pattern of surface > integrated > bottom, except for station TE, for which all depth types are similar.

For model 3, two variables had significant depth interactions (**Table 10**): t-Al when July was treated as a summer month, and bicarbonate alkalinity when July was a winter month. Again, box-plots capture the season-depth differences found in these data (**Figure 4**). There are strong differences among stations for both variables, and among seasons for bicarbonate alkalinity at station TE, but relatively minor differences among depth types. For t-Al, depth measures were similar in summer months but for winter, surface > bottom in winter, particularly driven by TE. For bicarbonate alkalinity, depth measures were similar in winter months but for summer, surface > bottom and integrated.

Conclusion – Analysis shows that for many variables there are statistically significant differences in values by depth, with bottom samples (and depth-integrated samples to a lesser degree) tending to have lower values than surface samples. However, the magnitude of differences associated with depth is minimal compared to the magnitude of differences that occur naturally between samples and stations. We conclude that future sampling should focus on surface samples, in order to take advantage of the most baseline data. This is somewhat conservative, as most variables demonstrate lower concentrations at depth. However, in cases where there is reason to suspect potential for large depth-specific effects, a bottom sample could be added occasionally to test for a hypothesized depth effect. Given that the CREMP is intended to monitor large-scale basin-wide changes, depth-specific sampling is less likely to be warranted for the

[5] The top panel of Figure 3 is actually more informative in this regard, as it shows variation in t-Al among surface samples at a given station. The ES of -15.9% is not large compared to variation among surface samples.

CREMP than for other programs that have a finer scale of resolution. Depth samples were not included in power analyses later in this appendix – as with duplicate samples, the depth samples are pseudo-replicates of the surface samples with which they are associated.

5. ANALYSIS OF BAKER/MEADOWBANK DIFFERENCES

Baker Lake and the Meadowbank area lakes are fundamentally different, because Baker Lake is connected to and influenced by salt water. Consequently, development of separate thresholds and triggers may be appropriate for the two areas for any variables that exhibit large differences.

Data – Potential differences between Baker/Meadowbank were examined by considering the control samples (i.e., potential impact stations removed), not including duplicate samples and depth replicates. This provided n=10 samples for Baker (station BAP) and n=110 across all Meadowbank stations (see **Table 2**).

Methods – Analysis was limited to the 18 variables for which at least 5 of the Baker measurements were above detection limits (see **Tables 5-7**). The variables included t-Al, t-Mn, ammonia-N, TKN, nitrate-N, t-P, TOC, DOC, bicarbonate alkalinity[6], Cl, conductivity, hardness, Ca, Mg, Na, sulphate, pH and TDS. In general, values below DLs were set equal to DLs, with a couple of exceptions[7]. Box-plots were produced for each variable to compare the Baker and Meadowbank samples. In addition, the Baker and Meadowbank data were compared using Mann-Whitney tests (on the raw data) and t-tests (using log data). The Mann-Whitney tests is more robust in this case as it is appropriate for non-normal data (which occurs when there are numerous values = DL).

Results – There are large differences between Baker and Meadowbank samples for several variables (**Figure 5** and **Table 12**). The percentage ES is calculated the same way as above for analysis of depth. There are relatively small differences for t-Al and ammonia-N. The difference for pH appears small (difference in medians = 0.4) but for pH that could be biologically quite meaningful. For all other variables there are large or very large differences between Baker and Meadowbank (ratio of medians ≥ 1.6 or ES ≥ 66%).

---

[6] total alkalinity was excluded because values were identical to bicarbonate alkalinity.

[7] Across the 18 test variables, Baker measures were <DL for a few cases of t-Al and ammonia-N These were set = DL. Meadowbank data for these two variables had different DLs in some cases. For t-Al, there were DLs greater than 0.005; these measurements were removed to be consistent with Baker. For Ammonia-N, there were DLs and valid measurements less than 0.02; these were all set = 0.02 to be consistent with Baker.

When considering "means" there is clearly a strong rationale for developing separate Baker/Meadowbank triggers for numerous variables. However, it is unclear how triggers based on maximums or 95th percentiles will reflect obvious differences in medians/means among Baker/Meadowbank measures. This is explored further in the next section.

6. THRESHOLDS AND TRIGGERS

Data – The data set upon which thresholds and triggers were developed was all standard control samples – duplicates and depth replicates were excluded as they are pseudo-replicates of "standard" samples. In addition, individual data points for which values were below DL but for which the DL was higher than usual were excluded. Remaining data points with values below the usual DL were set equal to the DL. For variables with very few values above DL, the reported medians will equal the DL, and so may 95th percentiles.

Methods – The main text has described the rationale and approach for development of thresholds and triggers. There are three methods of trigger development as follows:

1. When a threshold (e.g., CCME guideline) is established, the trigger was set as the maximum of either (a) the value halfway between the baseline median and the threshold ("Method A"), or (b) the 95th percentile of the baseline data ("Method B").
2. When a threshold is not established, the trigger was set equal to the 95th percentile of the baseline data ("Method B"), except in cases where less than 5% of the data exceeded the current detection limit (DL) – in the latter case, the trigger was set equal to two times the DL ("Method C").

Medians and 95th percentiles were chosen as metrics rather than means, standard deviations, or maximums, because the former are generally robust to skewed distributions and potential outliers.

There were exceptions to the above approach for a few special cases, specifically t-Al, t-Cd, t-Mn, t-Zn, d-Al, ammonia-N, t-P, pH and TSS. These exceptions are explained in detail below.

Different triggers were set for Baker and Meadowbank for many conventional parameters, for which there obvious differences between Meadowbank and Baker data (see earlier analysis above). However, for total metals and dissolved metals, the same triggers were set for Meadowbank and Baker stations (as discussed below, in cases where the data suggested different triggers could be appropriate, the most conservative or lowest value was used for both.

Results – Thresholds and triggers are summarized in **Tables 13-15** for total metals, dissolved metals, and nutrients/conventionals. Thresholds were established for 22 variables based on water-quality guidelines (**Table 16**). In most cases, the threshold was equal to a given guideline, but there were exceptions for a few variables as discussed below. It should be noted that in cases where a water quality guideline exists but Method B was used for trigger development (i.e., cases where baseline data already exceed the guideline for > 5% of cases), it is possible for the trigger to be more extreme than the guideline (e.g., this occurs for total phosphorus and for the lower bound for pH) – in such cases the guideline is reported as the threshold but is not used as a criterion for action; rather, the trigger is the only criterion for action as is the case for variables lacking water quality guidelines.

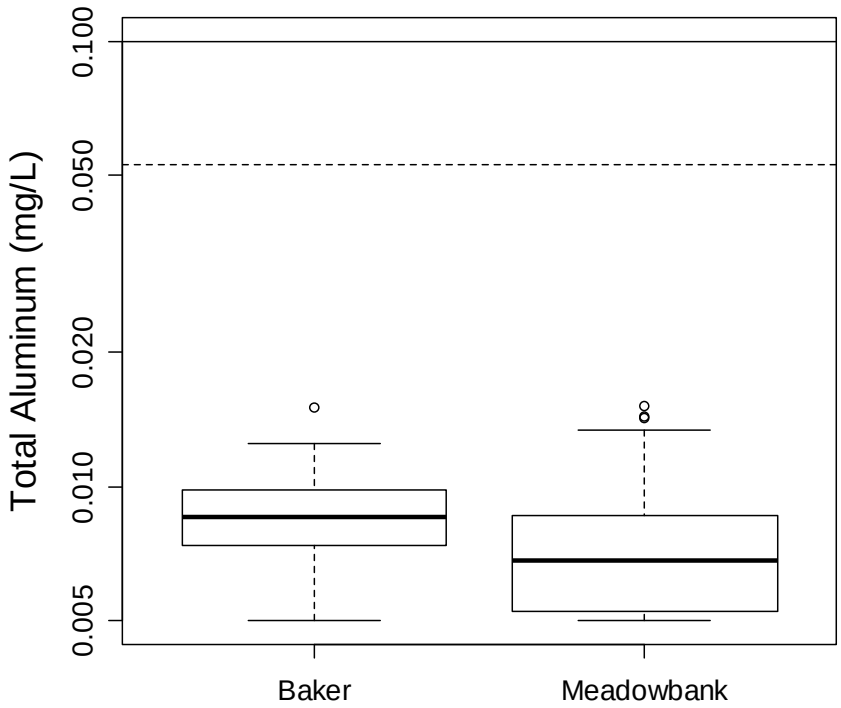
There were two cases where the water quality guideline was equal to the current DL. For these variables (t-Cr and t-Se), the trigger was set equal to the DL.

There are three variables (t-Cu, t-Pb, t-Ni) for which the water quality guidelines are specific to water hardness ranges below 120, 60, and 60 mg/L $CaCO_3$, respectively. Hardness levels for both Meadowbank and Baker samples were consistently below 60 mg/L $CaCO_3$. For example, as reported in **Table 15**, the 95th percentiles for hardness were 12.3 and 49.5 for Meadowbank and Baker samples, respectively. Thus for these three variables, the guidelines associated with low hardness ranges were used as thresholds.

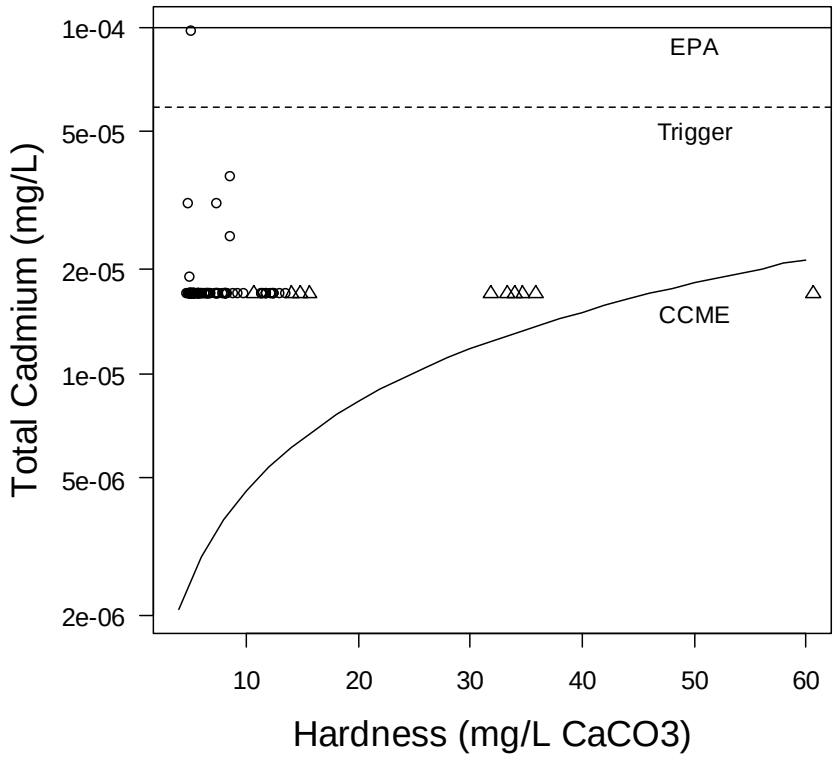
Special Cases – There were several special cases for which exceptions to the default approach to threshold and trigger development was warranted. These are discussed in the following pages.

Total aluminum -- The CCME guideline for t-Al in water is 0.005 mg/L when pH < 6.5, and 0.1 mg/L when pH ≥ 6.5. Across baseline samples for Meadowbank/Baker (n = 120), there were only 8 cases of pH < 6.5 (2 for station TPE and 6 for TPS), of which only three were below pH = 6.47 (values = 6.18, 6.20, and 6.34). For these 8 samples, only three t-Al measurements were above the current DL (0.005 mg/L), with a maximum value of 0.007 mg/L.

Given the strong tendency for pH to equal or exceed 6.5 across baseline samples, the CCME guideline of 0.1 mg/L was adopted as the threshold for t-Al (**Table 13**). Across the 120 samples, the median t-Al was 0.0068 mg/L and the 95th percentile was 0.0127 mg/L (**Table 13**). Based on Method A, the value halfway between the median t-Al and the threshold is 0.053 mg/L (i.e., [0.1 – 0.0068]/2 = 0.053), which is larger than the 95% percentile (Method B), and thus the proposed trigger for t-Al is 0.053 mg/L. As an example, the following figure shows box-plots of t-Al (in log scale) for each of Baker/Meadowbank, as well as the guideline (solid lines) and proposed trigger (dashed line).

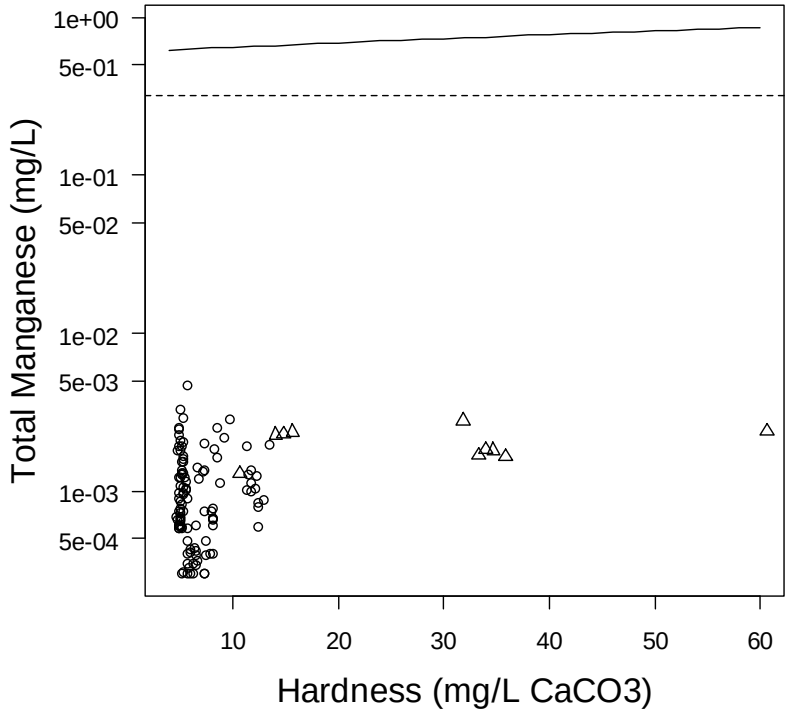


Total cadmium -- The hardness-dependent CCME guideline for t-Cd (mg/L) is $0.001*10^{0.86*\log10(H)-3.2}$ where H = hardness (mg/L $CaCO_3$). This relationship is illustrated in the figure below (solid curve) across the range of baseline observations of hardness for Baker (triangles) and Meadowbank (circles). Note that the CCME guideline is below the current DL for t-Cd (0.000017 mg/L) except at high hardness values > 46 mg/L $CaCO_3$. Moreover, several baseline t-Cd measurements exceeded the CCME guideline at low hardness. Consequently, the CCME guideline was deemed inappropriate for determining a threshold for Meadowbank/Baker data. The EPA water quality guideline for t-Cd (0.0001 mg/L, irrespective of hardness) was used instead (solid line in figure below). The corresponding trigger (based on Method A; **Table 13**) is 0.000059 mg/L (dashed line). One t-Cd baseline sample exceeded the proposed trigger (station TPE, hardness = 5.05, t-Cd = 0.000098 mg/L).



Total manganese -- There is no CCME water quality guideline for t-Mn. The hardness-dependent BC MOE guideline for t-Mn (mg/L) is 0.0044*H + 0.605, where H = hardness (mg/L $CaCO_3$). This guideline is based on numerous studies for fish, invertebrates and plants. The relationship is illustrated in the figure below (solid curve) across the range of baseline observations of hardness for Baker (triangles) and Meadowbank (circles). The guideline greatly exceeds observed t-Mn values for either Meadowbank or Baker samples.

For simplicity, we propose a single, conservative t-Mn trigger for Meadowbank/Baker data regardless of sample hardness or lake. To compute the t-Mn trigger, we first computed the guidelines corresponding to the median values of hardness observed for Meadowbank samples (median hardness = 5.7, t-Mn guideline = 0.63) and Baker samples (median hardness = 32.6, t-Mn guideline = 0.75). (At this stage, separate values were computed for each lake in recognition of the significant differences between their hardness values; e.g., see **Table 12**.) The corresponding lake-specific triggers for t-Mn (using Method A) are 0.32 for Meadowbank and 0.38 for Baker. Thus, the proposed trigger for t-Mn is 0.32 mg/L, which is the lower (more conservative) of the two lake-specific values. This trigger is depicted as a dashed line in the figure below (note that although the trigger is halfway between the median and guideline for Meadowbank data, it appears much closer to the guideline because of the log-scale used for t-Mn).



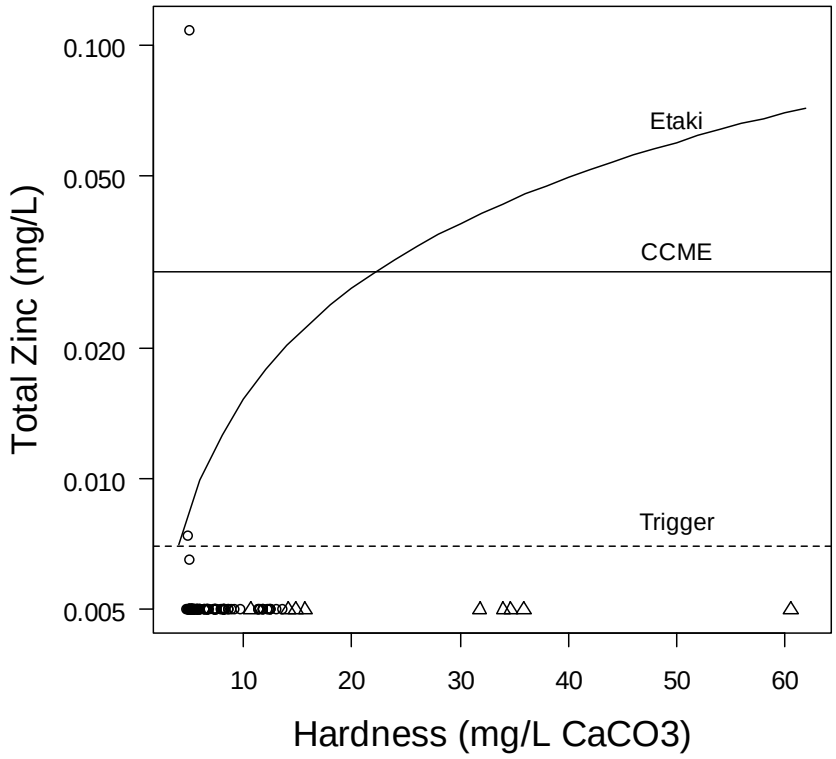
Total zinc -- The CCME water quality guideline for Zn is 0.030 mg/L. However, this guideline does not take into account hardness, and zinc toxicity is known to be hardness-dependent. An assessment for Ekati by EVS (2004) compiled data on species applicable to oligotrophic systems with low hardness, and developed a chronic benchmark for that was hardness dependent. The Ekati benchmark, denoted HC_5, represents the concentration of t-Zn at which 95% of species are likely to be protected against chronic effects (EVS 2004):

Hardness (mg/L CaCO3)	t-Zn level (HC_5) (mg/L)
5	0.0085
10	0.0153
15	0.0216
20	0.0276
25	0.0334
30	0.0389
35	0.0444
40	0.0497
45	0.0549
50	0.0600
55	0.0651
60	0.0700
65	0.0750
70	0.0798
75	0.0846
80	0.0894

To express benchmarks as a continuous relationship, we fit a power function to the hardness-HC_5 data, which provided a near-perfect fit: t-Zn (HC_5) in mg/L = $0.00217*H^{0.8486}$, where H = hardness (mg/L $CaCO_3$). This relationship is illustrated in the figure below (solid curve) across the range of baseline observations of hardness for Baker (triangles) and Meadowbank (circles). Note that the Ekati t-Zn benchmark is lower than the CCME guideline of 0.03 mg/L for hardness less than about 23. Only three baseline measurements of t-Zn have exceeded the current DL of 0.005 mg/L (the anomalous high value of t-Zn = 0.109 mg/L occurred for station TPS in July 2009).

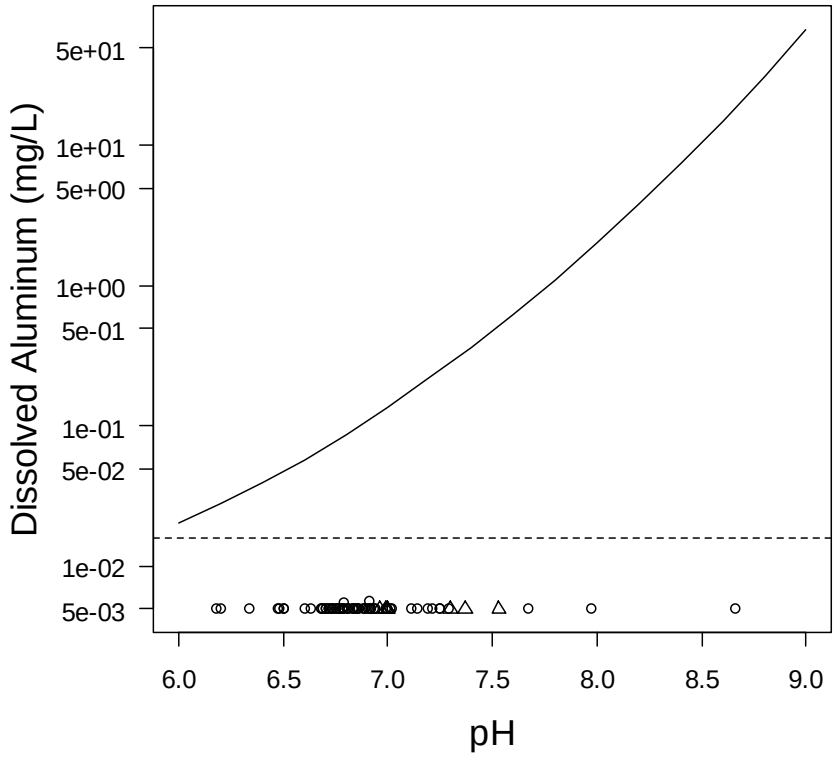
As for t-Mn, we propose a single t-Zn trigger for Meadowbank/Baker data regardless of sample hardness or lake. First, we computed the t-Zn benchmark (HC_5) corresponding to the median values of hardness observed for Meadowbank samples (median hardness = 5.7, t-Zn benchmark = 0.010) and Baker samples (median hardness = 32.6, t-Zn

benchmark = 0.041). However, for Baker, the lower CCME guideline of 0.03 mg/L was used instead as the Baker threshold. The corresponding lake-specific triggers for t-Zn (using Method A) are 0.007 for Meadowbank and 0.018 for Baker. Thus, the proposed trigger for t-Zn is 0.007 mg/L (dashed line in figure), which is the lower of the two lake-specific values.



Dissolved aluminum – There is no CCME guideline for d-Al in water. However, a pH-dependent water quality guideline for d-Al (mg/L) has been developed by BC MOE for protection of freshwater aquatic life as follows: $d\text{-}Al = e^{(1.6-3.327*pH + 0.402*K)}$ where $K = pH^2$. This relationship is illustrated in the figure below (solid curve) across the range of baseline observations of pH for Baker (triangles) and Meadowbank (circles). The BC MOE guideline greatly exceeds the observed values of d-Al. Only two baseline measurements of d-Al exceeded the current DL of 0.005 mg/L (maximum = 0.0057).

Again, we propose a single d-Al trigger for Meadowbank/Baker data regardless of sample pH or lake. Based on the minimum pH observed (6.02), the corresponding BC MOE guideline of d-Al is 0.027 mg/L. Via Method A, with median = 0.005 and threshold = 0.027, the proposed trigger for d-Al is computed as 0.016 mg/L (dashed line in figure).



Ammonia-N -- The CCME guideline for total ammonia in freshwater is pH and temperature dependent, with more stringent guidelines applying at higher pH and higher temperature. The proposed threshold for Ammonia-N (Meadowbank and Baker) was conservatively derived using two discrete CCME guidelines corresponding to specific pH and temperature values. Note that the maximum pH among baseline data for Meadowbank/Baker is 8.66, while maximum temperatures in the lakes are around 16 to 18 degrees. The two CCME guidelines that span these maximum (i.e., worst-case) conditions are as follows: (1) total ammonia = 0.239 mg/L for pH = 8.5 and temperature = 15 degrees; and (2) total ammonia = 0.067 mg/L for pH = 9.0 and temperature = 20 degrees. The mid-point of these two values is 0.153 mg/L, which when converted from total ammonia to total ammonia as N is 0.126 mg/L.

Thus, the proposed threshold for ammonia-N is 0.126 mg/L. Application of this threshold provided a proposed trigger value of 0.073 mg/L (**Table 15**). Only at extreme pH and temperature would this trigger potentially exceed the CCME guideline. Whenever the trigger is exceeded, the concentrations of ammonia-N should be compared to the CCME guideline based on the specific pH and field temperature of each sample.

Total P -- The CCME does not specify a particular guideline for total phosphorus, but instead establishes a guidance framework for site-specific application. Under that framework, the specification for ultra-oligotrophic lakes is for total-P of <0.004 mg/L. The framework notes that up to a 50% increase in total-P over baseline is generally considered acceptable. Regardless, the 95th percentiles for Total-P exceeded 0.004 mg/L for both Meadowbank samples and Baker samples (**Table 15**). Consequently, the proposed lake-specific triggers were set equal to these 95th percentiles (Method B, **Table 15**).

pH -- The CCME guideline for pH in freshwater is a range from 6.5 to 9.0. Thus, for pH, there is both an upper threshold (9.0) and a lower threshold (6.5), with associated upper and lower triggers (**Table 15**). The corresponding proposed triggers for pH were lake specific (given the significant differences in pH observed between lakes – see **Table 12**), with Meadowbank triggers based on Method B (the 95th percentile as the upper trigger and 5th percentile as the lower trigger) and Baker triggers based on Method A (the distance halfway from median to threshold). In the case of Meadowbank data, baseline pH levels are often below the lower threshold of 6.5, and as such, the lower pH trigger (6.47) is slightly lower as well (**Table 15**).

TSS -- For water bodies with low natural TSS, the CCME guideline is a maximum increase of 25 mg/L over background for short periods (e.g., 24h) and a maximum increase of 5 mg/L over background for longer periods (e.g., 24h to 30 days). If we conservatively assume a background TSS of 0 mg/L, then thresholds of 25 mg/L and 5 mg/L would apply for short-term and long-term exposures, respectively. However, because sampling occurs only at most once per month, it will be unknown whether a

given TSS measure is a short-term (< 24 h) or longer term (> 24 h) phenomenon. We therefore propose a TSS trigger based on the lower threshold of 5 mg/L, which thereby addresses both short and long durations. The resulting trigger, based on Method A, is 3.0 mg/L for both lakes (**Table 15**).

Evaluation of Triggers – The method of trigger development above was simplified in some ways. First, the same trigger was applied to Meadowbank and Baker programs for total and dissolved metals, although lake-specific triggers could be justified for total manganese (significant differences between lakes for t-Mn and hardness) and total zinc (due to hardness). Second, and more importantly, no allowance was made for potential differences in variables by season or (for Meadowbank only) by station. Therefore, it is important to examine the data more closely in relation to the proposed triggers.

To address this issue, we examined plots of month-specific and station-specific data relative to triggers.

Specific box plots are shown for Baker Lake data by station (**Figure 6**) and by month (**Figure 7**) and similarly for Meadowbank by station (**Figure 8**) and by month (**Figure 9**). Plots are limited to those variables with at least 40 of the 120 control observations above detection limits, and exclude duplicates and depth replicates. In addition, plots also include an additional 20 impact samples for Baker (versus 10 control) and 106 impact samples for Meadowbank (versus 110 control). In the plots, if a station median exceeds a trigger, it is likely that the trigger was exceeded in several months, and similarly if a monthly median exceeds a trigger, it is likely that the trigger was exceeded for several stations.

For Baker Lake, results show reasonably consistent measures across the three stations (one control, two impact). There is no case in which a station median exceeds a given trigger, and only a few cases where the median is quite close to the trigger (e.g., TKN and Bicarbonate alkalinity). However, for Baker months, there are numerous examples where the median (and often all three stations) exceeds the trigger. Clearly, September 2008 was unique – samples at all three stations exceed the trigger for eight variables (t-P, Cl, conductivity, Ca, hardness, Mg, sulphate, and TDS). This could represent higher sea water influence compared to other months. Obviously, there is high month-specific variation in these data, and so presumably the 95th percentile of monthly means (were there a lot of data) might suggest much higher triggers are justified to depict natural variation. Ultimately, however, it will be easy to identify such month-specific patterns as "natural" (given coherence across control/impact stations and across variables), whereas a 3-month or 6-month shift specific to a given impact station would indeed appear anomalous. Overall, then, the Baker triggers appear to be reasonable (based on the limited data).

For Meadowbank, triggers also appear reasonable overall. However, several variables appear much higher for Wally Lake (medians at or above triggers) compared to other control/impact stations, including bicarbonate alkalinity, conductivity, hardness, Ca, Mg, and sulphate. Consequently, it could be argued that some or all Wally data should have been omitted, and some triggers may be too high as a result. For some variables (e.g., t-Al, hardness, Mg), there is clearly a lot of station-specific variability, which may have

implications for triggers (but likely providing conservative triggers, that is, to the extent that control stations like INUG were heavily represented in the baseline data but had naturally lower concentrations than impact stations like SP and TPE). For months, there is a lot of variability and apparent trends as well (e.g., TOC and hardness), but very few instances of month-specific medians exceeding triggers (one case each for t-Al, t-P and upper pH). In the case of t-Al, all 120 baseline samples were < 0.02 mg/L, but many or most 'impact' samples for stations SP and TPE (across months for 2010) were in the range of 0.5 mg/L (the trigger is 0.053 mg/L) – this is not surprising because we expect t-Al to be high when TSS is high (associated with dike construction).

7. ANALYSIS OF SAMPLING DESIGN

Impact hypotheses and statistical design – Two general classes of impacts are hypothesized for the Meadowbank mine:

1. Pulse events for which potential impacts would be high for a short time but would then (in the case of water chemistry) dissipate relatively quickly. Pulse events could be associated with any phase of the mine, but are more likely to be associated with particular activities such as dike construction.
2. Long-term cumulative impacts that may be associated with ongoing activities. Long-term cumulative impacts are more likely to be associated with ongoing activities of mine operations.

As operations have just begun in 2010, the focus of monitoring to date has been on detecting pulse events associated with construction. The appropriate framework for analysis is a before-after-control-impact (BACI) that is aimed at detecting a potential impact in a particular lake or basin in a particular time period. The BACI framework can also be used to evaluate long-term impacts, but other tools such as time series regression analysis may also be appropriate for evaluating long-term trends. For this design document, we focus on the use of the BACI framework, recognizing that other tools such as time series regressions may be useful at a future date once sufficient time series data are available[8].

The classic BACI (paired) design has before/after periods α_i (i = B, A; I = 2), control/impact sites β_j (j = C, I; J = 2), and a total of K paired sampling times τ_k that are nested within period. A statistical model for this design is given by (Smith 2002, equation 2):

[8] In theory, a BACI analysis that is appropriately framed should be capable of detecting changes associated with long-term trends.

(1) $$X_{ijk} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

The key term is the interaction ($\alpha\beta$), which can be tested using an F test with F = MS[($\alpha\beta$)]/MS[Resid] and degrees of freedom = 1, $K - 2$. As discussed by Smith (2002), this is equivalent to simply taking the differences between the control and impact values across times and using a two-sample (before-after) t test (Stewart-Oaten et al. 1986).

Model (1) can be extended to include additional control sites (e.g., "asymmetric" designs; Underwood 1994) and/or additional impact sites. To be valid, the additional sites must be replicates rather than subsamples (i.e., as controls, they should be spatially independent of each other but representative of the impact sites, while replicates for impacts need to be spatially independent and (ideally) affected by independent disturbances). So whereas j = (C, I) in the classic BACIP, j may compose any combination of J total sites, for example $J = 4$ where $j = (C_1, C_2, C_3, I)$. The general test of ($\alpha\beta$) still applies, but with degrees of freedom = $(J - 1)$, $(K - 2)(J - 1)$ (e.g., see Table 1 of Underwood (1994) and Table 9 of Smith (2002)).

In addition, there may be n replicate subsamples s at each site/time combination (jk), as assumed in Table 1 of Underwood (1994). In this case, we modify equation (1) as:

(2) $$X_{ijks} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + (\tau\beta)_{k(i)j} + \varepsilon_{ijks} \ ,$$

where subsamples now permit estimation of times-by-site interactions ($\tau\beta$). The appropriate F ratio for ($\alpha\beta$) is now F = MS[($\alpha\beta$)]/MS[($\tau\beta$)] with df = $(J - 1)$, $(K - 2)(J - 1)$. As Underwood demonstrates, specific comparisons (interaction terms, such as the impact site versus either "period" or a specific "time" unit) can be examined by partitioning variation accordingly (e.g., Underwood Table 2[9]).

Methods – The analysis here assessed the expected precision and power of BACI estimates for different after-period (impact) durations and different numbers of sub-samples (random spatial samples collected each month). Formal analysis of the sampling design uses Meadowbank data. Baker Lake data have only been collected since 2008, thus inferences would be limited based on the small data set. Results from Meadowbank should be generalizable to Baker Lake given that the BACI analyses below for Meadowbank compare a single control to each potential impact station individually (the same scenario as for Baker, which has one control and two impact stations). Separate analyses were conducted for the three primary impact stations SP, TPE, and TPN. In each case, INUG was used as the control station. Separate analyses were conducted for **five** variables -- total aluminum, total manganese, total ammonia, total phosphorous, and pH. These variables were selected as they have definable thresholds and most data above

[9] We note that there are mistakes in the tables presented by Underwood (1994).

detection limits, therefore the results would be particularly meaningful. We assume that for the most part if the experimental design is adequate for five variables it is likely to be adequate for most variables. For a given variable, the effect size or ES (fixed across months) was equal to the change in mean from baseline (i.e., during the before period) to the trigger value. For example, if the before-period mean = 10 and trigger = 15, then ES = 5 (=15-10). In the case of pH, both upper and lower triggers were assessed (for the latter, this implies a reduction in mean). Thus, for all reported results, we are assessing the power and precision of BACI estimates under the assumption that the true effect size resulted in a new mean equal to the trigger value during the entire after period (of course, the observed after-period mean will differ somewhat from the trigger value due to simulated random variation).

For purposes of the BACI analyses, the before-period data included all available data except "impact" samples (summarized below). This assumes that these data, regardless of station or month, are reasonably representative of "natural" conditions. This assumption should be reasonable as all "impact" data are excluded.

After-period data were simulated using variances consistent with observed data. Specifically, the following mixed-effects model was fit to the "before data" for a given control-impact station pair:

$$X_{jks} = \beta_j + \tau_k + (\tau\beta)_{kj} + \varepsilon_{jks}$$

The fit provided estimates of before-period station means β_j, random-effects variances for month ($\sigma^2[\tau]$) and month-by-station ($\sigma^2[\tau\beta]$), and the residual variance for sub-samples ($\sigma^2[\varepsilon]$).

After-period data were simulated using after-period means (β_control, β _impact + **ES**) and the above variance estimates for three durations (1, 3, and 6 months) and three sub-sample scenarios (1, 2, and 3 subsamples per month). **Thus, a total of 9 scenarios of after-period duration and sub-sampling were examined.** In all cases, log-transformed data were used. For each scenario, 500 simulations were used.

Data summary – Sample numbers by station for the "before period" are shown in **Table 17**. Sample sizes (N) of valid measurements are shown in **Table 18** for each variable and station; a few total aluminum measurements were excluded (high detection limits), and six measures of total phosphorous were missing for each station. Among valid measurements, most were above the detection limit (DL) for t-Al and Mn (**Table 18**), whereas only two-thirds (roughly) were > DL for t-P, and only a third (roughly) for ammonia-N. All values < DL were set = DL for the remaining analysis. As noted above, values used in BACI simulations were obtained from fits of mixed-effects models to data

used to represent the "before period". These estimates are summarized in **Table 19**. In each case, data were paired with INUG samples (means for INUG, which were used as the after-period means, are of minor importance and are not shown). The metrics of interest are the effect sizes for log data, ES(log), and estimates of standard deviations (SD, log units). Power will be a function of ES(log) relative to the standard error of ES(log), which will depend firstly on SD(M x S) and secondarily on SD(Error). In relative terms, power will be high when ES(log) >> SD(M x S) and SD(Error) and low when the opposite is true.

Results – Estimates of statistical power are shown in **Tables 20 and 21**, and repeated in **Figures 10 and 11**. For all variables except pH **(Table 20; Figure 10)**, the *a priori* hypothesis is that impacts will result in increases, so power is based on one-tailed tests (alpha = 0.05 and 0.10 are both shown in **Table 20**; alpha = 0.05 for **Figure 10**). For pH, we are concerned about either increases or decreases, so power is for two-tailed tests **(Table 21; Figure 11)**. For example, for t-Al, power is 1.0 (100%) for all stations even when the after period data is a single sample (month = 1, sample = 1). This is because the specified effect size, ES(log), is very large compared to the key source of variation (SD for M x S and Error, see **Table 19**). In contrast, for t- P, power was generally low for small sample sizes in the after period because the specified effect size was low, especially for SP (**Table 19**). Results for pH (lower) illustrate another important factor affecting power. For station TPE, we observe useful increases in power for both an increase in months and samples. In this case, SD(M x S) is essentially zero (**Table 19**), so the driver is SD(Error). When SD(Error) is dominant, increasing sub-samples improves precision and power, especially when months = 1 or 3. In contrast, for SP, increasing months of sampling improves power somewhat, but increasing the number of sub-samples does very little. This is because SD(M x S) > SD(Error) in **Table 19**. In summary, for 6-month means, power is > 80% when alpha = 0.05 in all cases except total P and pH (lower) for station SP (for total P, high power is achieved at alpha = 0.10 for 6-month means with two subsamples per month; **Table 20**). In some cases, use of two subsamples provides important increases in power (e.g., **Figures 10-11**); however, relative improvements for three subsamples are typically minor.

A measure of precision for BACI estimates is shown in **Figures 12 and 13**. For log data, a useful measure of precision is the coefficient of variation (CV = SE[BACI estimate]/ [BACI estimate]). Results are averages across 500 trials, and do not depend on tails or alpha. These results mirror those for power. As a rough guide, power > 80% when CV < 35%.

Summary – Results show that the benefits of subsampling are limited, with some gains in power associated with two subsamples but little improvement associated with three subsamples. In terms of duration of sampling, the results show significant increases in

power associated with 3 months of sampling, and smaller increases associated with moving from 3 to 6 months. Implications of these findings are discussed in the main text.

8. REFERENCES

CCME (Canadian Council of Ministers of the Environment) 2007. Canadian Water Quality Guidelines for the Protection of Freshwater Aquatic Life, 1999, updated December 2007.

EVS Environmental Consultants (EVS). 2004. Tier I Ecological Rish Assessment for Zinc, Ekati Diamond Mine, NWT. A report prepared by EVS Environmental Consultants, North Vancouver, BC, for BHB Billiton Diamonds, Yellowknife, NWT. January, 2004.

Gelman, A, and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models – Analytical Methods for Social Research. Cambridge University Press, New York, 625 pp.

Pinheiro, J., and D. Bates. 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York, 528 pp.

Smith, E. 2002. BACI design. Pages 141-148, in: A.H. El-Shaarawi and W.W. Piegorsch [Eds], Encyclopedia of Environmetrics, John Wiley & Sons, Chicherster.

Stewart-Oaten, A., W. Murdoch, and K. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67: 929-940.

Venables, W., and B. Ripley. 2002. Modern Applied Statistics. Springer, New York, 495 pp.

Underwood, A. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications 4(1): 3-15.

Table 1. Total water samples collected for the CREMP.

Note: Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul				1		1	1		1	1	1	1	7
	Aug				1		1	1		1	2	1	1	8
2007	Jul				1		1	2		1	1	1	1	8
	Aug				1		1	1		2	1	1	1	8
2008	Jul	1	1	1	1		1	1		1	1	1	2	11
	Aug	1	1	1	2		1	1		1	1	1	1	11
	Sep	2	1	1			1	1		1	4	1		12
2009	May				14		7	13		7	7			48
	Jul	3	4	3	7	3	3	8	2	3	4	3	4	47
	Aug	1	1	1	3	1	1	4	1	1	2	1	1	18
	Sep	3	3	4	7	3	4	7	3	4	3	3	4	48
	Nov				7		3	8		3	4			25
	Dec						1	3		2	1	3		10
2010	Jan						5	3		3	3	7		21
	Feb						2	3		1	1	3		10
	Mar						3	7		4	4	7		25
	Apr						1	3		2	1	3		10
	May						4	7		4	3	7		25
Total		11	11	11	45	7	41	74	6	42	44	44	16	352

Table 2. Standard water samples collected for the CREMP.

Note: Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul				1		1	1		1	1	1	1	7
	Aug				1		1	1		1	1	1	1	7
2007	Jul				1		1	1		1	1	1	1	7
	Aug				1		1	1		1	1	1	1	7
2008	Jul	1	1	1	1		1	1		1	1	1	1	10
	Aug	1	1	1	1		1	1		1	1	1	1	10
	Sep	1	1	1			1	1		1	4	1		11
2009	May				6		6	6		6	6			30
	Jul	3	3	3	3	3	3	3	2	3	3	3	3	35
	Aug	1	1	1	1	1	1	1	1	1	1	1	1	12
	Sep	3	3	3	3	3	3	3	3	3	3	3	3	36
	Nov				3		3	3		3	3			15
	Dec						1	1		1	1	1		5
2010	Jan						3	2		3	3	3		14
	Feb						1	1		1	1	1		5
	Mar						3	3		3	3	3		15
	Apr						1	1		1	1	1		5
	May						3	3		3	3	3		15
Total		10	10	10	22	7	35	34	6	35	38	26	13	246

Table 3. Depth replicates (Bottom and Integrated samples) collected for the CREMP.

Note: Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul													0
	Aug													0
2007	Jul													0
	Aug													0
2008	Jul													0
	Aug													0
	Sep													0
2009	May				8			7						15
	Jul				4			4						8
	Aug				2			2						4
	Sep				4			4						8
	Nov				4			4						8
	Dec							2				2		4
2010	Jan							1				4		5
	Feb							2				2		4
	Mar							4				4		8
	Apr							2				2		4
	May							4				4		8
Total		0	0	0	22	0	0	36	0	0	0	18	0	76

Table 4. Duplicate samples collected for the CREMP.

Note: Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul													0
	Aug										1			1
2007	Jul							1						1
	Aug									1				1
2008	Jul												1	1
	Aug				1									1
	Sep	1												1
2009	May						1			1	1			3
	Jul		1					1			1		1	4
	Aug							1			1			2
	Sep			1			1			1			1	4
	Nov							1			1			2
	Dec									1				1
2010	Jan						2							2
	Feb						1							1
	Mar									1	1			2
	Apr									1				1
	May						1			1				2
Total		1	1	1	1	0	6	4	0	7	6	0	3	30

Table 5. Total metals: summary of sample measurements (N) and values above detection limits (>DL) for duplicate samples (control/impact), depth replicates (control/impact), standard control samples (total for Baker and Meadowbank, excluding duplicates and depth replicates), and Baker Lake control samples.

	Duplicate		Depth		Control (total)		Baker (control)	
Variable	N	>DL	N	>DL	N	>DL	N	>DL
Aluminum (T)	60	51	115	102	120	85	10	9
Antimony (T)	60	0	115	0	120	0	10	0
Arsenic (T)	60	4	115	0	120	0	10	0
Barium (T)	60	1	115	0	120	0	10	0
Beryllium (T)	60	0	115	0	120	0	10	0
Boron (T)	60	0	115	0	120	0	10	0
Cadmium (T)	60	3	115	6	120	7	10	0
Chromium (T)	60	2	115	1	120	0	10	0
Copper (T)	60	2	115	5	120	1	10	0
Iron (T)	60	23	115	14	120	2	10	0
Lead (T)	60	0	115	8	120	3	10	0
Lithium (T)	60	0	115	0	120	0	10	0
Manganese (T)	60	58	115	111	120	114	10	10
Mercury (T)	60	0	115	0	120	0	10	0
Molybdenum (T)	60	0	115	0	120	0	10	0
Nickel (T)	60	2	115	0	120	0	10	0
Selenium (T)	60	0	115	0	120	0	10	0
Strontium (T)	2	2	11	11	6	6	0	0
Thallium (T)	60	0	115	0	120	0	10	0
Tin (T)	60	0	115	0	120	0	10	0
Titanium (T)	60	2	115	0	120	0	10	0
Uranium (T)	60	2	115	0	120	0	10	0
Vanadium (T)	60	0	115	0	120	0	10	0
Zinc (T)	60	0	115	5	120	3	10	0

Table 6. Dissolved metals: summary of sample measurements (N) and values above detection limits (>DL) for duplicate samples (control/impact), depth replicates (control/impact), standard control samples (total for Baker and Meadowbank, excluding duplicates and depth replicates), and Baker Lake control samples.

	Duplicate		Depth		Control (total)		Baker (control)	
Variable	N	>DL	N	>DL	N	>DL	N	>DL
Aluminum (D)	48	10	115	6	71	2	7	0
Antimony (D)	48	0	115	0	71	0	7	0
Arsenic (D)	48	2	115	0	71	0	7	0
Barium (D)	48	0	115	0	71	0	7	0
Beryllium (D)	48	0	115	0	71	0	7	0
Boron (D)	48	0	115	0	71	0	7	0
Cadmium (D)	48	0	115	4	71	0	7	0
Chromium (D)	48	0	115	0	71	0	7	0
Copper (D)	48	0	115	2	71	0	7	0
Iron (D)	48	1	115	0	71	0	7	0
Lead (D)	48	0	115	1	71	0	7	0
Lithium (D)	48	0	115	0	71	0	7	0
Manganese (D)	48	26	115	34	71	25	7	3
Mercury (D)	48	0	115	1	71	0	7	0
Molybdenum (D)	48	0	115	0	71	0	7	0
Nickel (D)	48	0	115	0	71	0	7	0
Selenium (D)	48	0	115	0	71	0	7	0
Strontium (D)	2	2	11	11	6	6	0	0
Thallium (D)	48	0	115	0	71	0	7	0
Tin (D)	48	0	115	0	71	0	7	0
Titanium (D)	48	0	115	0	71	0	7	0
Uranium (D)	48	0	115	0	71	0	7	0
Vanadium (D)	48	0	115	0	71	0	7	0
Zinc (D)	48	0	115	4	71	0	7	0

Table 7. Nutrients and conventional parameters: summary of sample measurements (N) and values above detection limits (>DL) for duplicate samples (control/impact), depth replicates (control/impact), standard control samples (total for Baker and Meadowbank, excluding duplicates and depth replicates), and Baker Lake control samples.

	Duplicate		Depth		Control (total)		Baker (control)	
Variable	N	>DL	N	>DL	N	>DL	N	>DL
Ammonia-N	60	17	115	31	120	47	10	6
TKN	60	56	115	101	120	116	10	10
Nitrate-N	60	27	115	37	120	13	10	10
Nitrite-N	60	1	115	1	120	1	10	1
Ortho-phosphate	54	5	92	4	108	8	10	0
Total phosphorous	54	31	92	33	108	54	10	10
TOC	60	60	115	115	120	120	10	10
DOC	60	60	115	115	120	120	10	10
Reactive silica	54	0	115	0	99	0	7	0
Bicarbonate alkalinity	54	54	92	92	108	108	10	10
Chloride	60	32	115	82	120	62	10	10
Carbonate alkalinity	54	0	92	0	108	0	10	0
Conductivity	60	60	115	115	120	120	10	10
Hardness	60	60	115	115	120	120	10	10
Calcium	60	60	115	115	120	120	10	10
Potassium	60	2	115	0	120	2	10	2
Magnesium	60	60	115	115	120	120	10	10
Sodium	60	4	115	0	120	10	10	10
Sulphate	60	60	115	115	120	120	10	10
pH	60	60	115	115	120	120	10	10
Total alkalinity	54	54	92	92	108	108	10	10
TDS	59	39	115	81	120	72	10	10
TSS	59	15	115	13	120	8	10	0

Table 8. Summary of statistical comparisons between sample/duplicate pairs.

Notes: Limited to pairs with both values above detection limits). N = total samples (pairs = N/2); Median = median of all samples; MAD = median absolute difference among sample/duplicate pairs; r = Spearman rank correlation among sample/duplicate pairs; SD = standard deviation of random effects for Location (variability among sample pairs) and residual errors (variability between samples and duplicates) fit to log-transformed data.

					Model SD (log data)	
Variable	N	Median	MAD	r	Location	Res. Error
Aluminum (T)	50	0.016	0.002	0.97	1.272	0.088
Manganese (T)	58	0.00121	0.00005	0.97	0.669	0.074
TKN	54	0.10	0.03	0.23	0.127	0.280
Total phosphorous	26	0.004	0.001	0.47	0.147	0.574
TOC	60	1.65	0.07	0.88	0.287	0.098
DOC	60	1.55	0.08	0.90	0.291	0.062
Bicarbonate alkalinity	54	6.35	0.20	0.93	0.309	0.122
Chloride	30	0.68	0.01	0.97	1.753	0.013
Conductivity	60	20.10	0.20	0.99	0.773	0.016
Hardness	60	7.90	0.10	0.97	0.514	0.033
Calcium	60	1.83	0.03	0.99	0.414	0.044
Magnesium	60	0.80	0.01	1.00	0.620	0.011
Sulphate	60	1.84	0.02	0.99	0.582	0.021
pH	60	6.87	0.04	0.74	0.000	0.065
TDS	28	17.0	2.0	0.55	0.862	0.148

Table 9. Summary of depth samples (N = 115) collected at 39 locations.

Notes: Depth (m) is shown for "Surface" and "Bottom" samples; depth range is shown for "Integrated" samples.

Case	Station/Sample	Date	Surface	Bottom	Integrated
1	INUG-1	3-May-09	2	12	2 to 8
2	INUG-2	3-May-09	2	5.5	2 to 4.5
3	INUG-1	28-May-09	3	7	3 to 7
4	INUG-2	28-May-09	2.5	5	2.5 to 4.5
5	INUG-4	24-Jul-09	3	9	0.5 to 9
6	INUG-5	24-Jul-09	3	9	0.5 to 9
7	INUG-7	19-Aug-09	3	6.7	0.5 to 6.7
8	INUG-8	16-Sep-09	3	7	0.5 to 7.0
9	INUG-9	16-Sep-09	3	10	0.5 to 10
10	INUG-11	28-Nov-09	3	5.5	2 to 5
11	INUG-12	28-Nov-09	3	4.5	2 to 4.5
12	TE-1	4-May-09	3	5	3 to 6
13	TE-2	4-May-09	3	10	3 to 8
14	TE-1	30-May-09	3	8	3 to 8
15	TE-2	30-May-09	3	4.5	2.5 to 4.5
16	TE-4	19-Jul-09	3	10	0.5 to 8
17	TE-5	19-Jul-09	3	10.5	0.5 to 8
18	TE-7	15-Aug-09	3	8	0.5 to 8
19	TE-8	12-Sep-09	3	7.5	0.5 to 7.5
20	TE-9	12-Sep-09	3	7.5	0.5 to 8.0
21	TE-11	26-Nov-09	3	7	3 to 7
22	TE-12	26-Nov-09	3	8	3 to 8
23	TE-14	21-Dec-09	3	12	3 to 12
24	TE-17	16-Jan-10	2.8		2.2 to 3.8
25	TE-18	26-Feb-10	3		3 to 5
26	TE-19	21-Mar-10	3	10	3 to 10
27	TE-20	21-Mar-10	2.5	4.5	2.5 to 4.5
28	TE-22	9-Apr-10	3	5	3 to 5
29	TE-23	21-May-10	3	5	2.5 to 5
30	TE-24	21-May-10	3	5	2.5 to 5
31	TPS-8	22-Dec-09	3	6.5	3 to 6.5
32	TPS-10	19-Jan-10	2.5	5	2.5 to 5
33	TPS-9	19-Jan-10	3	12	3 to 12
34	TPS-12	24-Feb-10	3	12	3 to 12
35	TPS-13	24-Mar-10	3	12	3 to 12
36	TPS-14	24-Mar-10	3	12	3 to 12
37	TPS-16	10-Apr-10	3	12	3 to 12
38	TPS-17	22-May-10	3	6	3 to 6
39	TPS-18	22-May-10	3	14	3 to 14

Table 10. P-values for F tests of the key fixed-effect terms for depth.

Notes: Models include Type (Model 1), Type*Station (Model 2), and Type*Season (Model 3; Seas1). Shaded cells denote $0.05 \leq P < 0.01$ (light shade) and $P \leq 0.01$ (dark shade). N = number of samples (locations = N/3).

		Type	Type*Station	Type*Season (Model 3)		
Variable	N	(Model 1)	(Model 2)	N	July summer	July winter
Aluminum (T)	90	0.000	0.001	72	0.003	0.081
Manganese (T)	99	0.204	0.062	75	0.940	0.889
TKN	87	0.303	0.451	72	0.212	0.067
TOC	111	0.000	0.007	84	0.335	0.601
DOC	111	0.003	0.346	84	0.372	0.437
Bicarbonate alk.	90	0.003	0.761	63	0.107	0.019
Chloride	72	0.052	0.995	69	0.569	0.836
Conductivity	111	0.001	0.479	84	0.123	0.296
Hardness	111	0.002	0.146	84	0.141	0.202
Calcium	111	0.060	0.917	84	0.656	0.896
Magnesium	111	0.000	0.361	84	0.169	0.313
Sulphate	111	0.001	0.553	84	0.638	0.367
pH	111	0.500	0.105	84	0.166	0.136
TDS	54	0.351	0.302	45	0.623	NA

Table 11. Summary of coefficients for Model 1 (log[y] ~ Type + Location).

Notes: Est = estimate; SE = standard error; P = P-value based on t-test; ES (%) = proportional effect size in untransformed units relative to surface samples, where ES (%) = (exp[Est] – 1)*100. Shaded cells denote $0.05 \leq P < 0.01$ (light shade) and $P \leq 0.01$ (dark shade).

	Bottom – Surface				Integrated - Surface			
Variable	Est	SE	P	ES (%)	Est	SE	P	ES (%)
Aluminum (T)	-0.174	0.061	0.006	-15.9%	0.122	0.061	0.051	12.9%
Manganese (T)	0.108	0.097	0.268	11.5%	0.174	0.097	0.079	18.9%
TKN	0.109	0.086	0.213	11.5%	0.123	0.086	0.158	13.1%
TOC	-0.071	0.018	0.000	-6.8%	-0.008	0.018	0.638	-0.8%
DOC	-0.060	0.022	0.008	-5.9%	0.014	0.022	0.516	1.4%
Bicarbonate alk.	-0.068	0.019	0.001	-6.6%	-0.040	0.019	0.040	-4.0%
Chloride	-0.047	0.019	0.016	-4.6%	-0.021	0.019	0.270	-2.1%
Conductivity	-0.048	0.012	0.000	-4.6%	-0.017	0.012	0.166	-1.7%
Hardness	-0.045	0.013	0.001	-4.4%	-0.033	0.013	0.013	-3.2%
Calcium	-0.036	0.016	0.029	-3.5%	-0.031	0.016	0.056	-3.1%
Magnesium	-0.054	0.011	0.000	-5.3%	-0.039	0.011	0.001	-3.8%
Sulphate	-0.061	0.015	0.000	-5.9%	-0.034	0.015	0.028	-3.4%
pH	-0.004	0.006	0.562	-0.4%	-0.007	0.006	0.241	-0.7%
TDS	-0.048	0.047	0.322	-4.6%	0.020	0.047	0.673	2.0%

Table 12. Summary of comparisons between Baker and Meadowbank control samples.

Notes: Ratio = median(Baker)/median(Meadow); MW test P = P-value for the Mann-Whitney test; Est = difference in means (Baker – Meadow) for log-transformed data; P = P-value based on t-test; ES (%) = proportional effect size in untransformed units relative to Meadowbank samples, where ES (%) = (exp[Est] – 1)*100.

	Sample size (N)		Medians			MW	t-test (log data)		
Variable	Baker	Meadow	Baker	Meadow	Ratio	test P	Est	P	ES (%)
Aluminum (T)	10	103	0.009	0.007	1.3	0.055	0.21	0.041	23%
Manganese (T)	10	110	0.002	0.001	2.4	0.000	0.81	0.000	126%
Ammonia-N	10	110	0.022	0.020	1.1	0.103	0.11	0.281	12%
TKN	10	110	0.180	0.097	1.9	0.000	0.51	0.000	66%
Nitrate-N	10	110	0.014	0.005	2.8	0.000	1.07	0.000	192%
Total phosphor.	10	98	0.003	0.002	1.6	0.000	0.59	0.000	80%
TOC	10	110	3.1	1.6	1.9	0.000	0.62	0.000	86%
DOC	10	110	3.0	1.6	1.9	0.000	0.65	0.000	92%
Bicarbonate alk.	10	98	9.0	4.6	2.0	0.000	0.52	0.000	69%
Chloride	10	110	67.8	0.5	135.5	0.000	4.44	0.000	8409%
Conductivity	10	110	273.5	15.9	17.3	0.000	2.47	0.000	1082%
Hardness	10	110	32.6	5.7	5.7	0.000	1.36	0.000	289%
Calcium	10	110	3.8	1.3	3.0	0.000	0.82	0.000	128%
Magnesium	10	110	5.6	0.6	8.8	0.000	1.87	0.000	549%
Sodium	10	110	36.2	2.0	18.1	0.000	2.52	0.000	1139%
Sulphate	10	110	9.5	1.4	6.7	0.000	1.57	0.000	379%
pH	10	110	7.2	6.8	1.1	0.000	0.04	0.017	5%
TDS	10	110	138.5	11.0	12.6	0.000	2.19	0.000	791%

Table 13. Total metals: summary of trigger values for Meadowbank and Baker stations.

Notes: For each variable, thresholds (guidelines) are shown if applicable (see text for discussion); DL = current detection limit; N = sample measurements; >DL = number of measurements above DL; medians and 95th percentiles are shown for variables with at least one value >DL (with all values <DL set equal to DL); Method = method used to determine the trigger, where A = halfway from median (or DL if all values <DL) to threshold, B = 95th percentile, and C = 2*DL.

Variable	Threshold	DL	N	>DL	Median	95th	Trigger	Method
Aluminum (T)	0.1	0.005	120	85	0.0068	0.0127	0.053	A
Antimony (T)		0.0005	120	0			0.0010	C
Arsenic (T)	0.005	0.0005	120	0			0.0028	A
Barium (T)		0.02	120	0			0.04	C
Beryllium (T)		0.001	120	0			0.002	C
Boron (T)	1.2	0.1	120	0			0.65	A
Cadmium (T)	0.0001	0.000017	120	7	0.000017	0.000017	0.000059	A
Chromium (T)	0.001	0.001	120	0			0.001	A
Copper (T)	0.002	0.001	120	1	0.001	0.001	0.0015	A
Iron (T)	0.3	0.03	120	2	0.03	0.03	0.165	A
Lead (T)	0.001	0.0005	120	3	0.0005	0.0005	0.00075	A
Lithium (T)		0.005	120	0			0.010	C
Manganese (T)	See text	0.0003	120	114	0.0010	0.0025	0.32	See text
Mercury (T)	0.000026	0.00001	120	0			0.000018	A
Molybdenum (T)	0.073	0.001	120	0			0.037	A
Nickel (T)	0.025	0.001	120	0			0.013	A
Selenium (T)	0.001	0.001	120	0			0.001	A
Strontium (T)			6	6	0.0081	0.0089	0.0089	B
Thallium (T)	0.0008	0.0002	120	0			0.0005	A
Tin (T)		0.0005	120	0			0.0010	C
Titanium (T)		0.01	120	0			0.02	C
Uranium (T)		0.0002	120	0			0.0004	C
Vanadium (T)		0.001	120	0			0.002	C
Zinc (T)	See text	0.005	120	3	0.005	0.005	0.007	See text

Table 14. Dissolved metals: summary of trigger values for Meadowbank and Baker stations.

Notes: For each variable, thresholds (guidelines) are shown if applicable (see text for discussion); DL = current detection limit; N = sample measurements; >DL = number of measurements above DL; medians and 95th percentiles for variables with at least one value >DL (with all values <DL set equal to DL); Method = method used to determine the trigger, where A = halfway from median (or DL if all values <DL) to threshold, B = 95th percentile, and C = 2*DL.

Variable	Threshold	DL	N	>DL	Median	95th	Trigger	Method
Aluminum (D)	See text	0.005	71	2	0.005	0.005	0.016	See text
Antimony (D)		0.0005	71	0			0.0010	C
Arsenic (D)		0.0005	71	0			0.0010	C
Barium (D)		0.02	71	0			0.04	C
Beryllium (D)		0.001	71	0			0.002	C
Boron (D)		0.1	71	0			0.2	C
Cadmium (D)		0.000017	71	0			0.000034	C
Chromium (D)		0.001	71	0			0.002	C
Copper (D)		0.001	71	0			0.002	C
Iron (D)		0.03	71	0			0.06	C
Lead (D)		0.0005	71	0			0.0010	C
Lithium (D)		0.005	71	0			0.010	C
Manganese (D)		0.0003	71	25	0.0003	0.0014	0.0014	B
Mercury (D)		0.00001	71	0			0.00002	C
Molybdenum (D)		0.001	71	0			0.002	C
Nickel (D)		0.001	71	0			0.002	C
Selenium (D)		0.001	71	0			0.002	C
Strontium (D)			6	6	0.0082	0.0092	0.0092	B
Thallium (D)		0.0002	71	0			0.0004	C
Tin (D)		0.0005	71	0			0.0010	C
Titanium (D)		0.01	71	0			0.02	C
Uranium (D)		0.0002	71	0			0.0004	C
Vanadium (D)		0.001	71	0			0.002	C
Zinc (D)		0.005	71	0			0.010	C

Table 15. Nutrients and conventional parameters: summary of trigger values for Meadowbank and Baker stations.

Notes: For each variable, thresholds (guidelines) are shown if applicable (see text for discussion); DL = current detection limit; N = sample measurements; >DL = number of measurements above DL; medians and 95th percentiles for variables with at least one value >DL (with all values <DL set equal to DL); Method = method used to determine the trigger, where A = halfway from median (or DL if all values <DL) to threshold, B = 95th percentile, and C = 2*DL.

			Meadowbank (or both)						Baker (if different)					
Variable	Threshold	DL	T	N	Med	95th	Trigger	Method	T	N	Med	95th	Trigger	Method
Ammonia-N	0.126	0.02	120	47	0.020	0.052	0.073	A						
TKN		0.05	110	106	0.097	0.181	0.181	B	10	10	0.180	0.194	0.194	B
Nitrate-N	2.9	0.005	110	3	0.005	0.005	1.45	A	10	10	0.014	0.027	1.46	A
Nitrite-N	0.06	0.001	120	1	0.001	0.001	0.031	A						
Ortho-phosphate		0.001	108	8	0.0010	0.0011	0.002	C						
Total phosphorous	0.004	0.002	98	44	0.0020	0.0047	0.0047	B	10	10	0.0032	0.0061	0.0061	B
TOC			110	110	1.63	2.51	2.51	B	10	10	3.06	3.17	3.17	B
DOC			110	110	1.59	2.31	2.31	B	10	10	3.02	3.42	3.42	B
Reactive silica		1.0	99	0			2.0	C						
Bicarbonate alkalinity			98	98	4.60	9.42	9.42	B	10	10	9.00	9.46	9.46	B
Chloride		0.5	110	52	0.5	0.79	0.79	B	10	10	67.8	119.7	119.7	B
Carbonate alkalinity		2.0	108	0			4.0	C						
Conductivity			110	110	15.9	27.4	27.4	B	10	10	273.5	464.9	464.9	B
Hardness			110	110	5.7	12.3	12.3	B	10	10	32.6	49.5	49.5	B
Calcium			110	110	1.3	3.3	3.3	B	10	10	3.8	5.2	5.2	B
Potassium		2.0	120	2	2.0	2.0	4.0	C						
Magnesium			110	110	0.64	0.96	0.96	B	10	10	5.59	8.90	8.90	B
Sodium		2.0	110	0			4.0	C	10	10	36.2	65.1	65.1	B
Sulphate			110	110	1.43	2.61	2.61	B	10	10	9.53	17.31	17.31	B
pH (Upper)	9.0		110	110	6.80	8.00	8.00	B	10	10	7.17	7.59	8.09	A
pH (Lower)	6.5		110	110	6.80	6.47[a]	6.47	B	10	10	7.17	6.97[a]	6.84	A
Total alkalinity			98	98	4.60	9.42	9.42	B	10	10	9.00	9.46	9.46	B
TDS		10.0	110	62	11.0	19.0	19.0	B	10	10	138.5	245.8	245.8	B
TSS	5.0	1.0	120	8	1.0	1.6	3.0	A						

[a] For pH (Lower), the 5th percentile is reported.

Table 16. Summary of thresholds for water variables; (CCME, 2007).

Variable	Source	Description of guidelines
t-Aluminum (Al)	CCME	The CCME guideline for t-Al in water is 0.005 mg/L when pH < 6.5, and 0.1 mg/L when pH $\geq$ 6.5. See text for details.
t-Arsenic (As)	CCME	The CCME water quality guideline (aquatic life) for t-As is 0.005 mg/L.
t-Boron (B)	BC MOE	There is no CCME guideline for t-B. However, the BC Ministry of Environment (BC MOE, www.env.gov.bc.ca) guideline for freshwater aquatic life is 1.2 mg/L.
t-Cadmium (Cd)	CCME EPA	The hardness-dependent CCME guideline for t-Cd (mg/L) is $0.001*10^{0.86*\log(H)-3.2}$ where H = hardness (mg/L $CaCO_3$). This guideline was deemed inappropriate. The EPA guideline for t-Cd of 0.0001 mg/L was used instead. See text for details.
t-Copper (Cu)	CCME	The CCME guideline for t-Cu is 0.002 mg/L for hardness < 120 mg/L $CaCO_3$.
t-Chromium (Cr)	CCME	The CCME guideline for hexavalent chromium (the most common form in surface waters) is 0.001 mg/L.
t-Iron (Fe)	CCME	The CCME guideline for t-Fe is 0.3 mg/L.
t-Lead (Pb)	CCME	The CCME guideline for t-Pb is 0.001 mg/L for hardness < 60 mg/L $CaCO_3$.
t-Manganese (Mn)	BC MOE	There is no CCME guideline for t-Mn in water. The hardness-dependent BC MOE guideline for t-Mn in mg/L is 0.0044*H + 0.605, where H = hardness (mg/L $CaCO_3$). See text for details.
t-Mercury (Hg)	CCME	The CCME guideline for total inorganic mercury is 26 ng/L (0.00026 mg/L).
t-Molybdenum (Mo)	CCME	The CCME guideline for t-Mo in water is 0.073 mg/L.
t-Nickel (Ni)	CCME	The CCME guideline for t-Ni is 0.025 mg/L for hardness < 60 mg/L $CaCO_3$.
t-Selenium (Se)	CCME	The CCME guideline for t-Se in water is 0.001 mg/L.
t-Thallium (Tl)	CCME	The CCME water quality guideline for t-Tl is 0.0008 mg/L.
t-Zinc (Zn)	CCME Ekati	The CCME water quality guideline for t-Zn is 0.030 mg/L. However, this guideline does not take into account hardness, and zinc toxicity is known to be hardness-dependent. An assessment for Ekati by EVS (2004) compiled data on species applicable to oligotrophic systems with low hardness, and developed a chronic benchmark for t-Zn that was hardness dependent. See text for details.
d-Aluminum (Al)	BC MOE	A pH-dependent water quality guideline for d-Al (mg/L) has been developed by BC MOE for protection of freshwater aquatic life as follows: $d\text{-}Al = e^{(1.6-3.327*pH + 0.402*K)}$ where $K = pH^2$. See text for details.

Ammonia-N	CCME	The CCME guidelines for total ammonia in freshwater are pH and temperature dependent, with more stringent guidelines applying at higher pH and higher temperature. See text for details.
Nitrate-N	CCME	The CCME guideline for NO3-N is 2.9 mg/L.
Nitrite-N	CCME	The CCME guideline for nitrite-N is 0.06 mg/L.
t-Phosphorous (P)	CCME	The CCME does not specify a particular guideline for t-P, but instead establishes a guidance framework for site-specific application. See text for details.
pH	CCME	The CCME guideline for pH is a range from 6.5 to 9.0. See text for details.
TSS	CCME	For water bodies with low natural TSS, the CCME guideline is a maximum increase of 25 mg/L over background for short periods (e.g., 24h) and 5 mg/L for longer periods (e.g., 24h to 30 days). See text for details.

Table 17. Samples by station used as the “before period” in BACI simulations. Shaded months denote periods designated as “impact” months by station (samples excluded).

Month	INUG	SP	TPE	TPN
Jul.06	1	1	1	1
Aug.06	1	1	1	1
Jul.07	1	1	1	1
Aug.07	1	1	1	1
Jul.08	1	1	1	1
Aug.08	1		1	1
May.09	6		6	
Jul.09	3		3	
Aug.09	1			
Sep.09	3			
Nov.09	3			
Total	**22**	**5**	**15**	**6**

Table 18. Number of valid measurements (N) and the percentage of N above detection limits (DL) by variable and station for data used as the “before-period” in BACI simulations.

Number of valid measurements (N)

	INUG	SP	TPE	TPN
Aluminum (T)	19	5	14	6
Manganese (T)	22	5	15	6
Ammonia-N	22	5	15	6
Total Phosphorous	16	5	9	6
pH	22	5	15	6

Percentage of N measurements > detection limit (DL)

	INUG	SP	TPE	TPN
Aluminum (T)	100%	100%	71%	67%
Manganese (T)	91%	100%	87%	100%
Ammonia-N	50%	60%	13%	33%
Total Phosphorous	69%	100%	33%	83%
pH	100%	100%	100%	100%

Table 19. Summary of "before-period" mixed-effects model estimates by variable and impact station (paired with control INUG).

Notes: The station mean, trigger, and effect size (ES = Trigger – Mean) are in raw units (e.g., mg/L). ES(log) denotes the effect size for log data (ES(log) = log[Trigger/Mean]). Estimates of the standard deviations (SD) for random-effects terms (Month and Month-by-Station, M x S) and residuals errors are for log-transformed data.

						SD (log data)		
Variable	Station	Mean	Trigger	ES	ES(log)	Month	M x S	Error
Aluminum (T)	SP	0.009	0.053	0.044	1.796	0.196	0.103	0.102
	TPE	0.008	0.053	0.045	1.857	0.164	0.167	0.092
	TPN	0.006	0.053	0.047	2.151	0.228	0.000	0.086
Manganese (T)	SP	0.001	0.320	0.319	5.513	0.636	0.000	0.133
	TPE	0.001	0.320	0.319	5.643	0.597	0.140	0.137
	TPN	0.001	0.320	0.319	6.142	0.594	0.215	0.133
Ammonia-N	SP	0.023	0.073	0.050	1.176	0.410	0.000	0.322
	TPE	0.020	0.073	0.053	1.286	0.466	0.000	0.261
	TPN	0.018	0.073	0.055	1.417	0.401	0.000	0.322
Total P	SP	0.003	0.005	0.002	0.448	0.132	0.000	0.330
	TPE	0.002	0.005	0.002	0.749	0.136	0.000	0.294
	TPN	0.003	0.005	0.002	0.586	0.139	0.000	0.331
pH (upper)	SP	7.00	8.00	1.00	0.134	0.018	0.041	0.018
	TPE	6.82	8.00	1.18	0.159	0.055	0.000	0.024
	TPN	6.83	8.00	1.17	0.159	0.058	0.000	0.016
pH (lower)	SP	7.00	6.47	-0.53	-0.078	0.018	0.041	0.018
	TPE	6.82	6.47	-0.35	-0.053	0.055	0.000	0.024
	TPN	6.83	6.47	-0.36	-0.054	0.058	0.000	0.016

Table 20. Estimates of BACI statistical power for detecting a significant increase (one-tailed test) in a given variable by station (SP, TPE, TPN) as a function of sampling months (after period) and the number of sub-samples per month.

Note: Power is shown for two levels of alpha (0.05 and 0.10).

			alpha = 0.05			alpha = 0.10		
Variable	Months	Samples	SP	TPE	TPN	SP	TPE	TPN
Aluminum (T)	1	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
	3	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
	6	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
Manganese (T)	1	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
	3	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
	6	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
Ammonia-N	1	1	0.58	0.93	0.82	0.81	0.98	0.94
		2	0.87	1.00	0.98	0.96	1.00	0.99
		3	0.96	1.00	1.00	0.99	1.00	1.00
	3	1	0.98	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
	6	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
Total P	1	1	0.09	0.41	0.19	0.26	0.59	0.37
		2	0.17	0.65	0.31	0.33	0.82	0.50
		3	0.18	0.81	0.43	0.43	0.93	0.67
	3	1	0.29	0.85	0.55	0.54	0.92	0.73
		2	0.38	0.99	0.68	0.64	1.00	0.88
		3	0.44	1.00	0.81	0.74	1.00	0.93
	6	1	0.46	0.99	0.74	0.67	1.00	0.89
		2	0.61	1.00	0.91	0.83	1.00	0.97
		3	0.73	1.00	0.97	0.93	1.00	1.00

Table 21. Estimates of BACI statistical power for detecting a significant increase in pH (upper trigger) or a significant decrease in pH (lower trigger) by station (SP, TPE, TPN) for two-tailed tests as a function of sampling months (after period) and the number of sub-samples per month.

Note: Power is shown for two levels of alpha (0.05 and 0.10).

			alpha = 0.05			alpha = 0.10		
pH Trigger	Months	Samples	SP	TPE	TPN	SP	TPE	TPN
Upper	1	1	0.18	0.99	1.00	0.43	1.00	1.00
		2	0.20	1.00	1.00	0.47	1.00	1.00
		3	0.21	1.00	1.00	0.49	1.00	1.00
	3	1	0.78	1.00	1.00	0.90	1.00	1.00
		2	0.78	1.00	1.00	0.91	1.00	1.00
		3	0.79	1.00	1.00	0.92	1.00	1.00
	6	1	0.97	1.00	1.00	1.00	1.00	1.00
		2	0.98	1.00	1.00	1.00	1.00	1.00
		3	0.99	1.00	1.00	1.00	1.00	1.00
Lower	1	1	0.04	0.20	0.38	0.15	0.37	0.58
		2	0.04	0.42	0.64	0.14	0.58	0.83
		3	0.04	0.54	0.86	0.15	0.75	0.95
	3	1	0.20	0.52	0.87	0.41	0.71	0.96
		2	0.23	0.88	0.98	0.45	0.95	1.00
		3	0.26	0.96	1.00	0.48	0.99	1.00
	6	1	0.49	0.90	0.99	0.73	0.95	1.00
		2	0.50	1.00	1.00	0.73	1.00	1.00
		3	0.51	1.00	1.00	0.73	1.00	1.00

Figure 1. Sample versus duplicate values of selected water variables (log-log scale).

Notes: Values below detection limits were set equal to the given detection limit. Solid line is the 1:1 line.

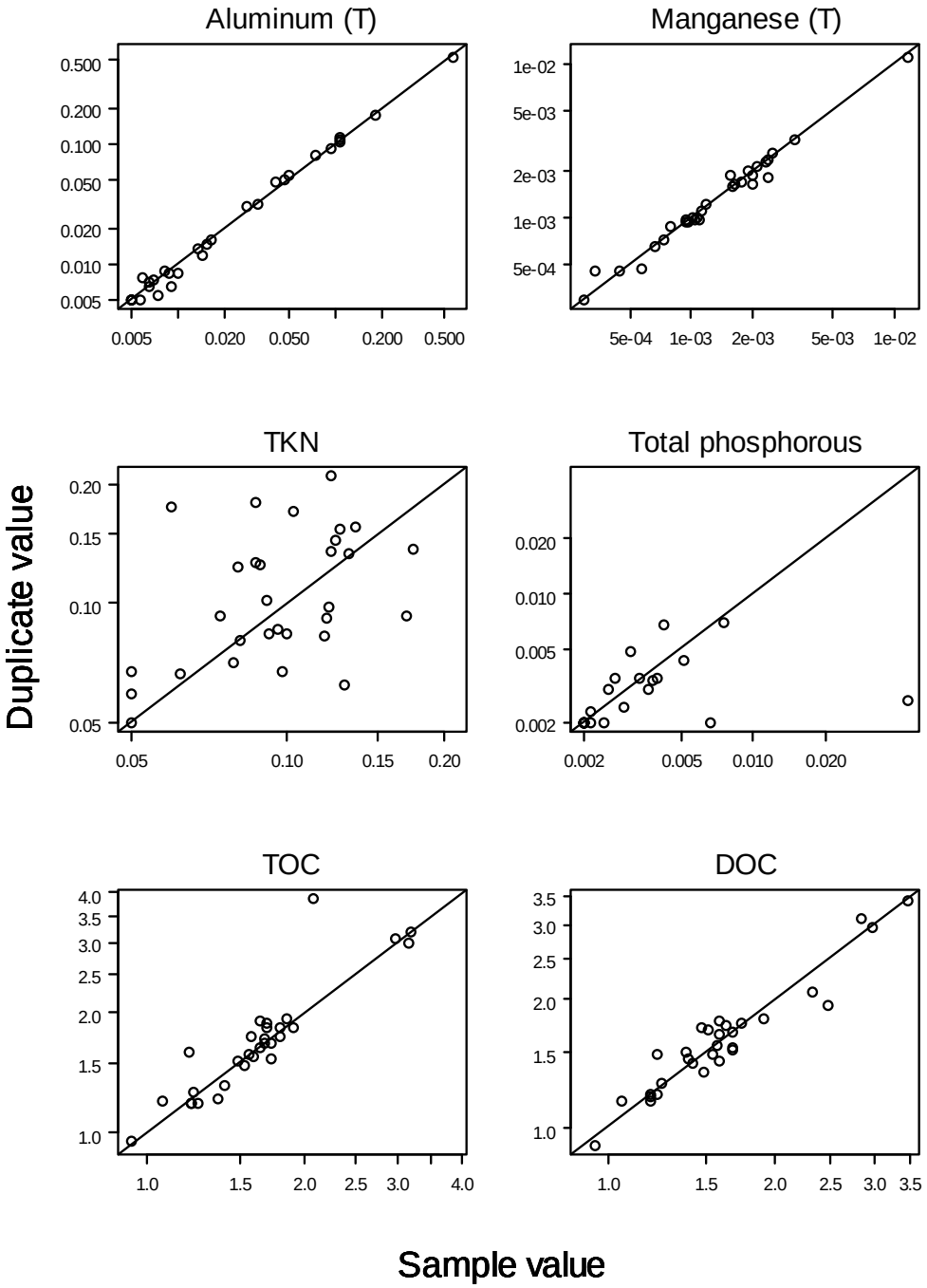


Figure 1 (page 2 of 3)

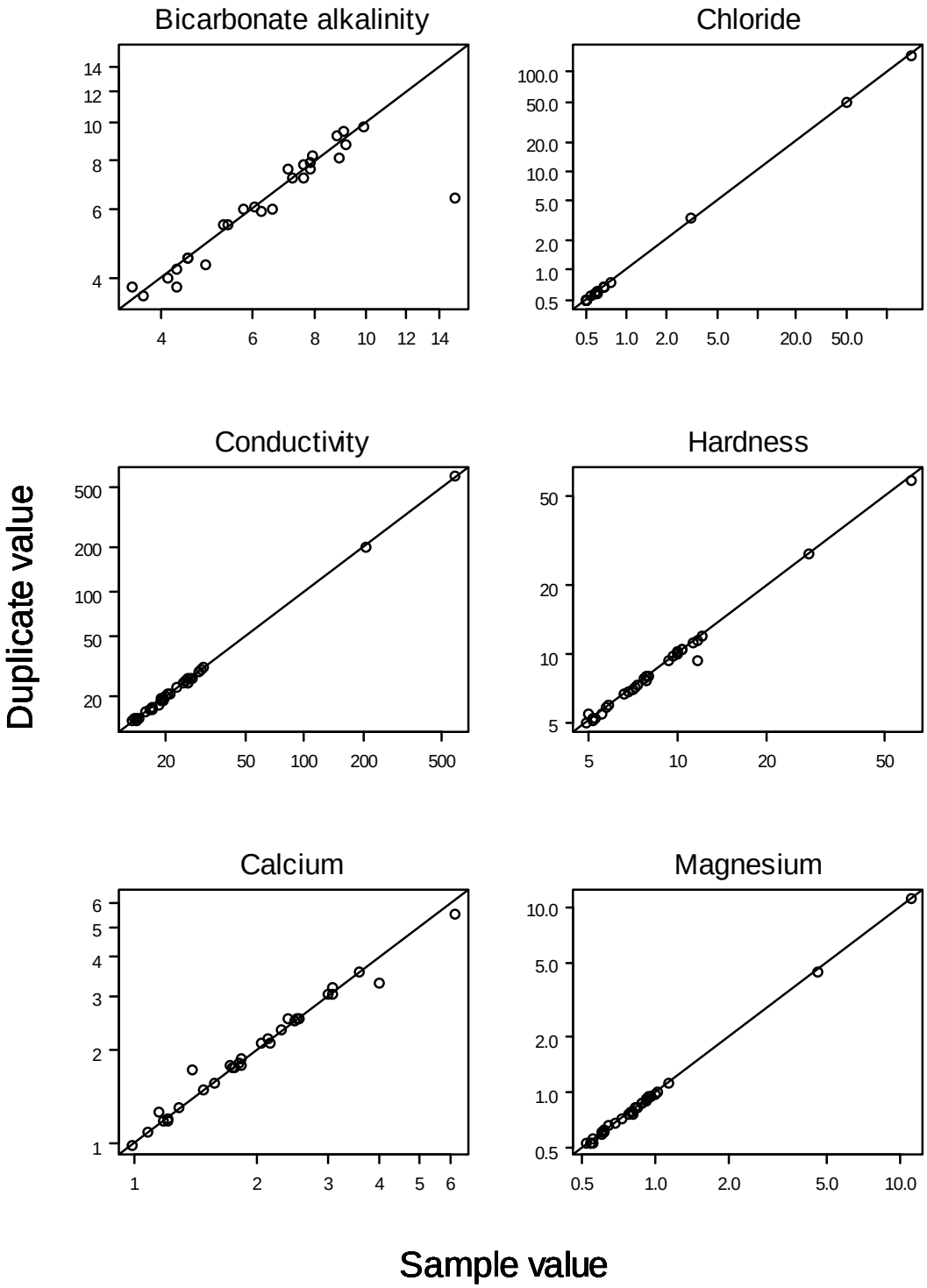


Figure 1 (page 3 of 3)

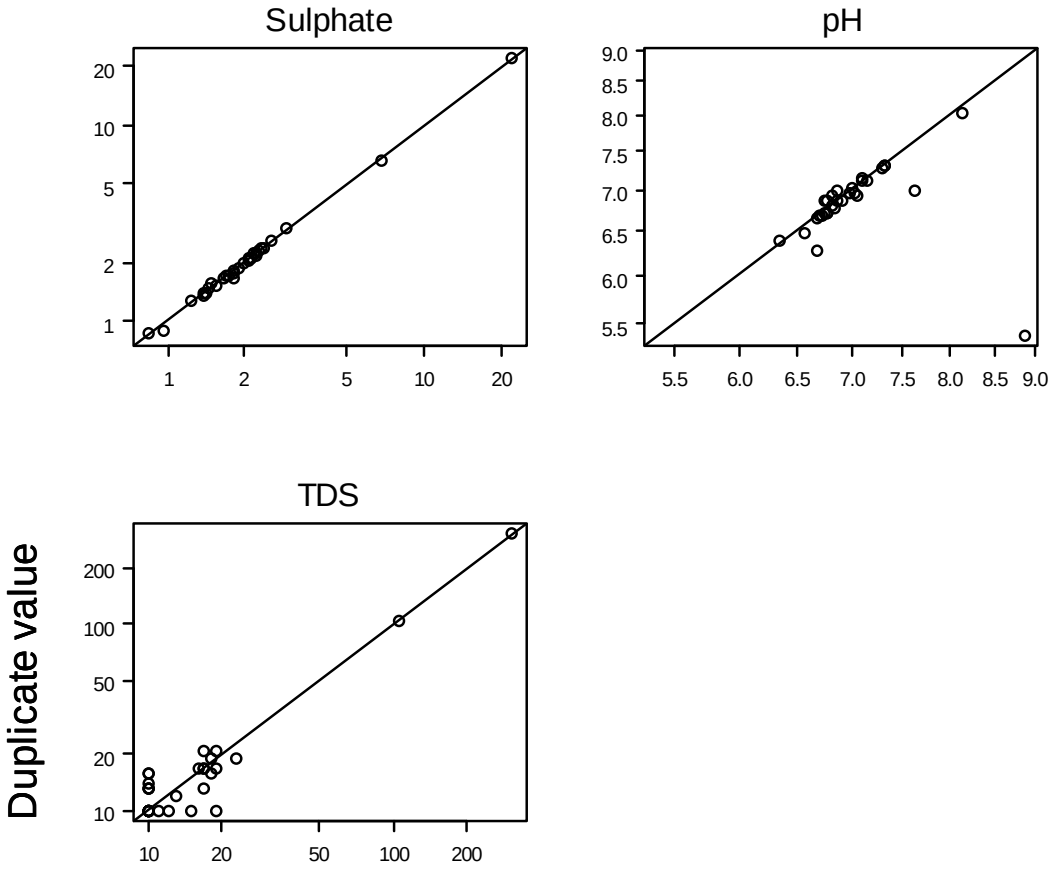


Figure 2. Surface versus Bottom samples (left side; circles) and Integrated samples (right side; triangles) for selected water variables (log-log scale).

Notes: Values < DL were set = DL. Dashed lines denote estimated differences (if P < 0.1) relative to Surface samples (see text). Solid line is the 1:1 line.

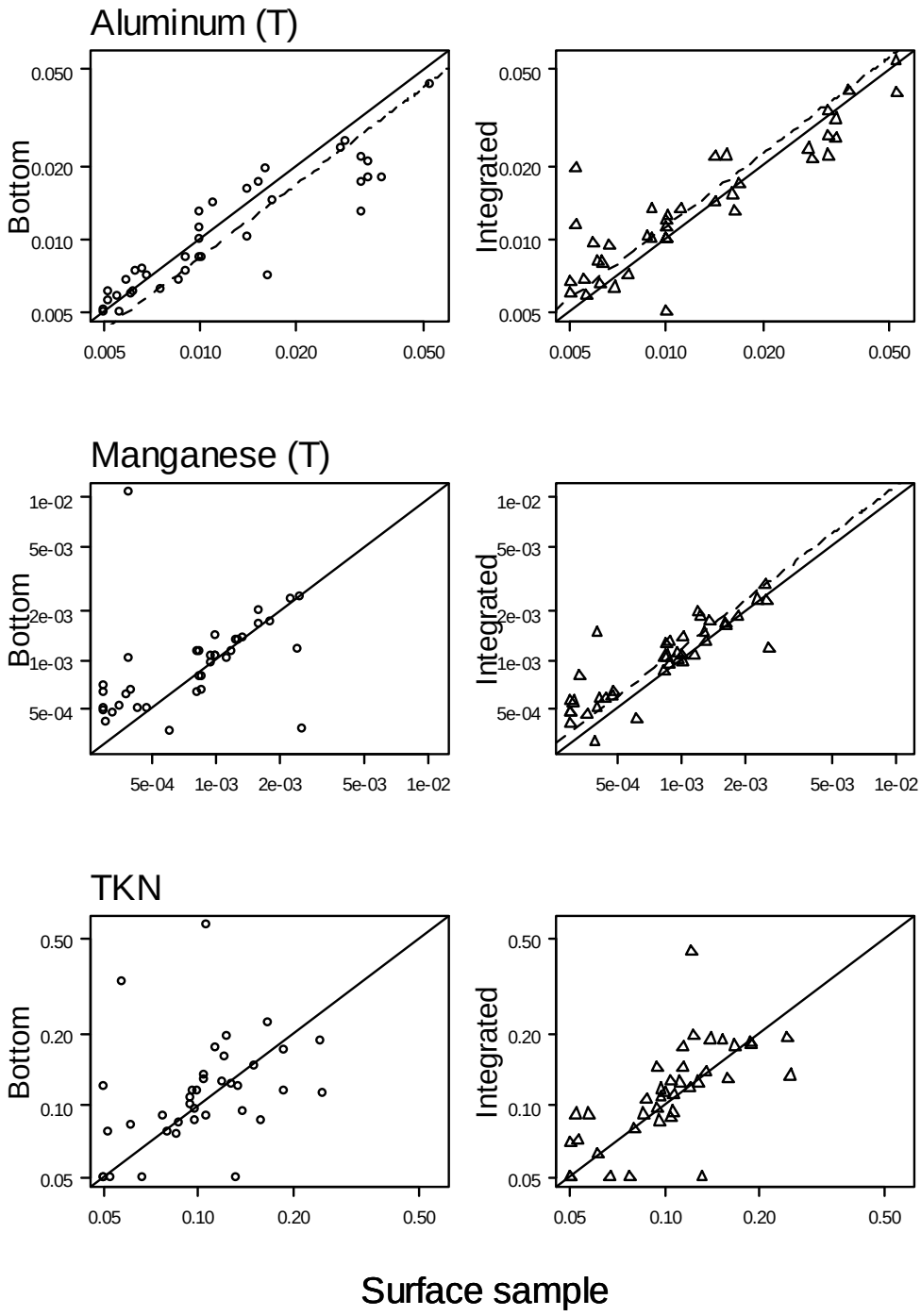


Figure 2 (page 2 of 5)

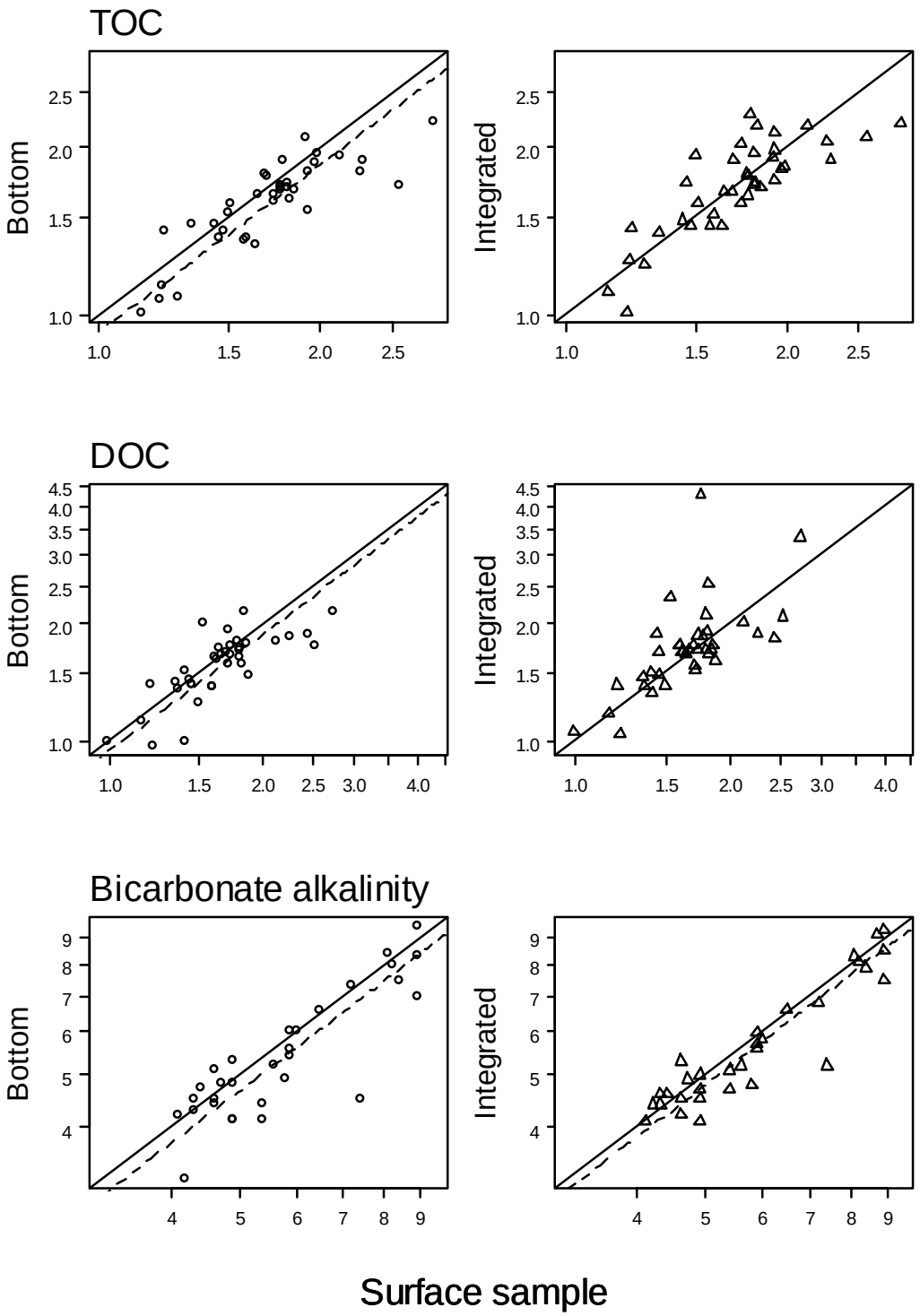


Figure 2 (page 3 of 5)

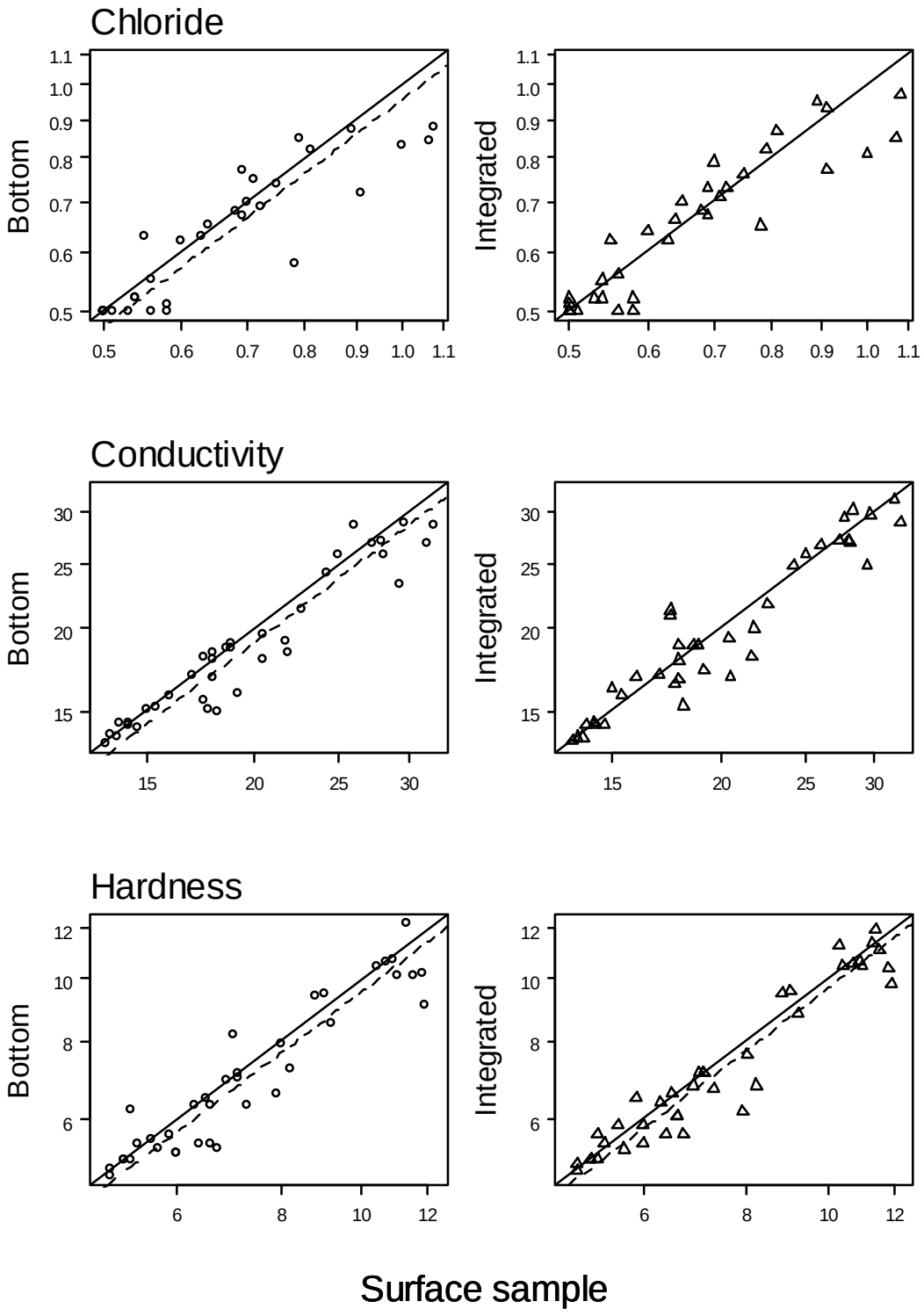


Figure 2 (page 4 of 5)

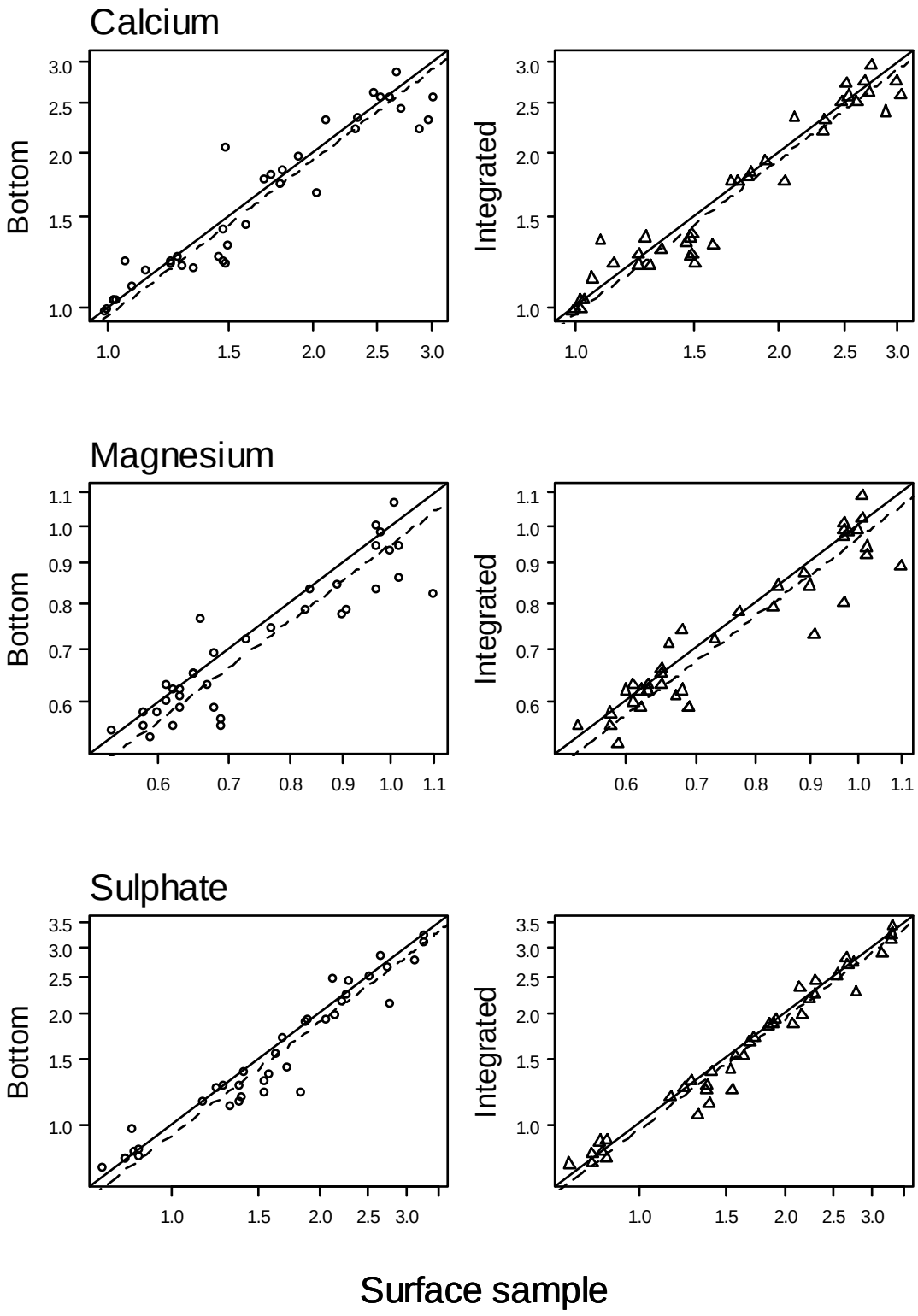
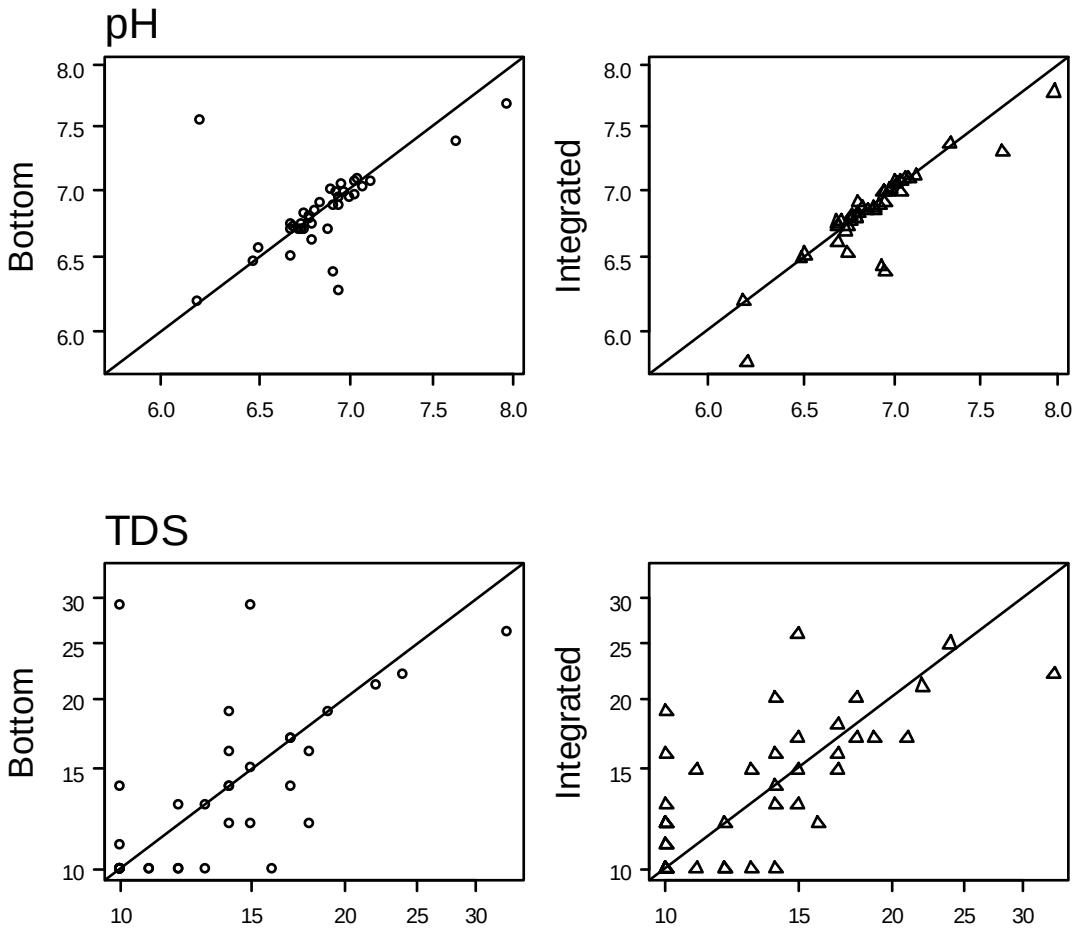


Figure 2 (page 5 of 5)



Surface sample

Figure 3. Box-plots of sample measurements by station (INUG, TE, and TPS) and depth type for total aluminum (top panel) and TOC (bottom panel).

Notes: S = surface, .B = bottom, and .Int = integrated.

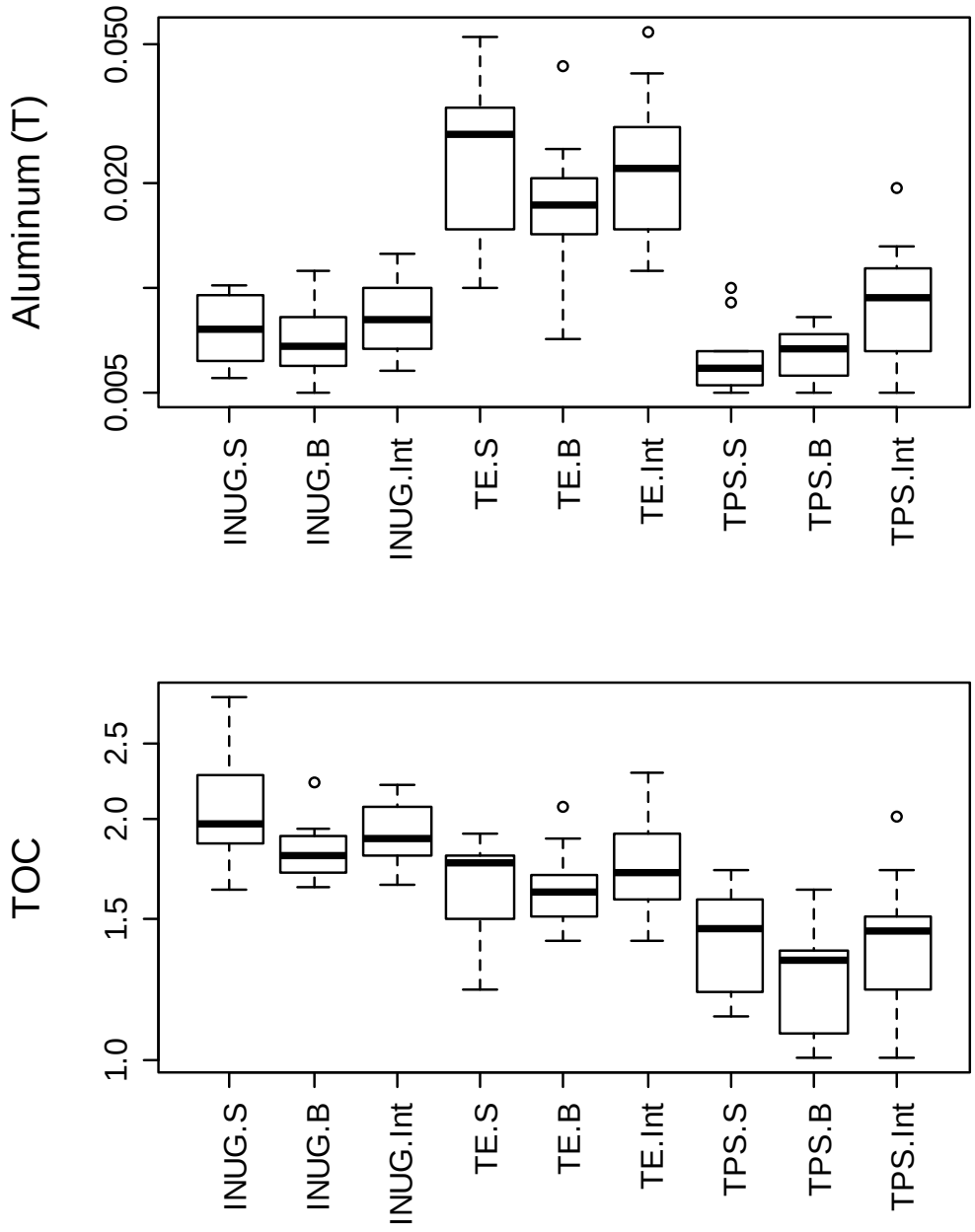


Figure 4. Box-plots of sample measurements by season (Summer or Winter) and depth type (.S = surface, .B = bottom, and .Int = integrated) for total aluminum (top panels) and Bicarbonate alkalinity (bottom panels), for each of stations INUG (left side) and TE (right side).

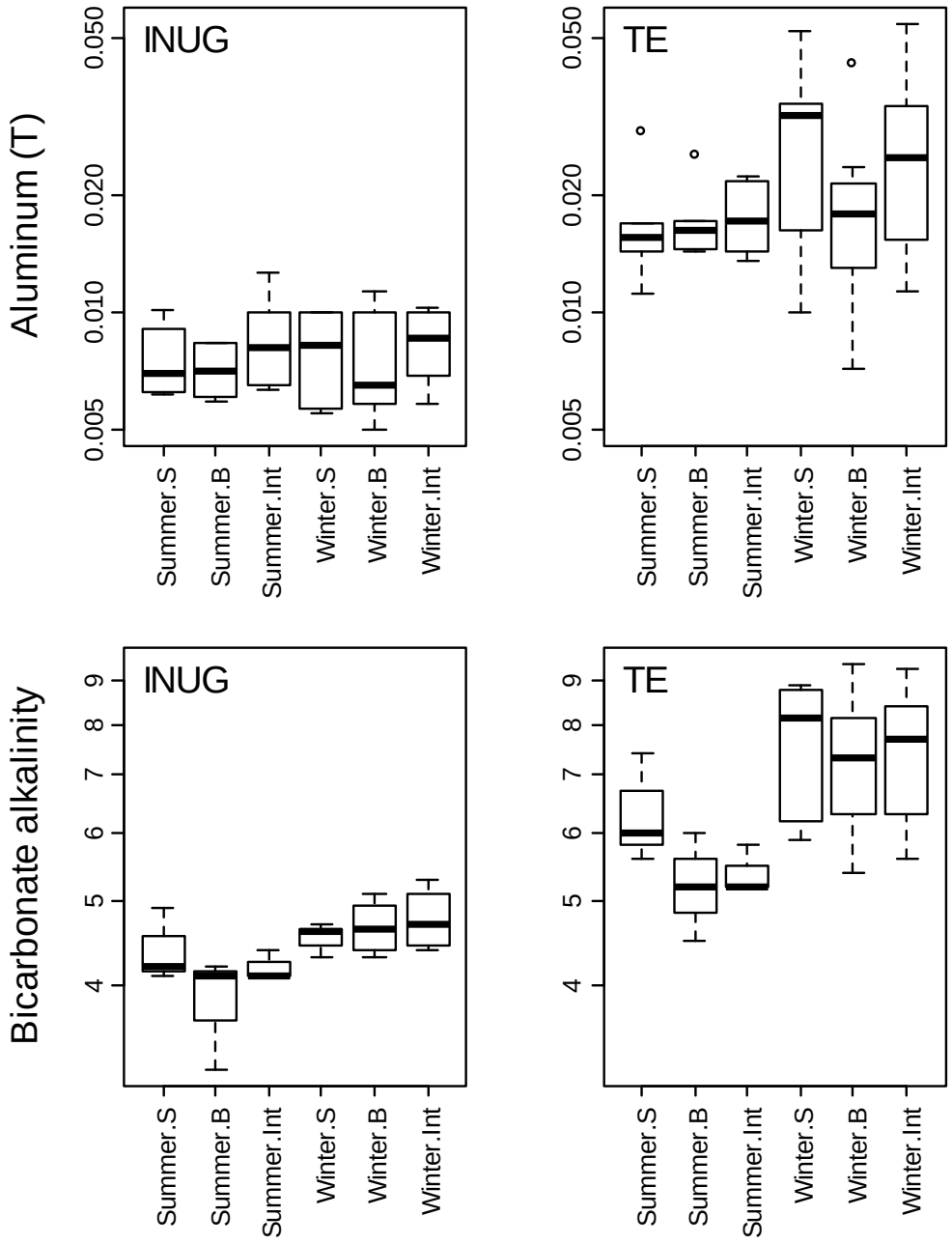


Figure 5. Box-plots of sample measurements (standard control samples) by project lake for selected water variables (log-log scale).

Notes: Values < DL were set = DL (see text for details).

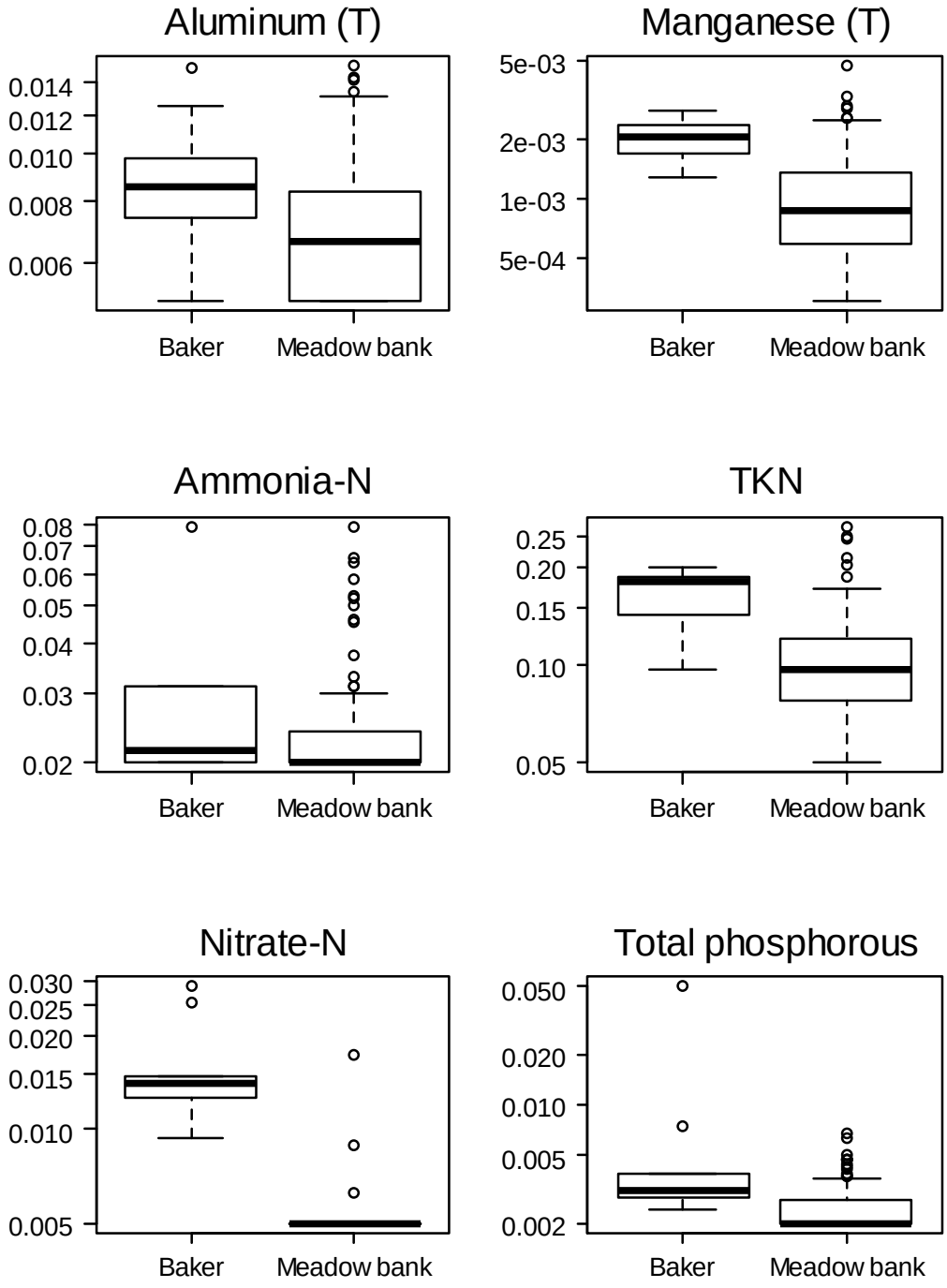


Figure 5 (page 2 of 3)

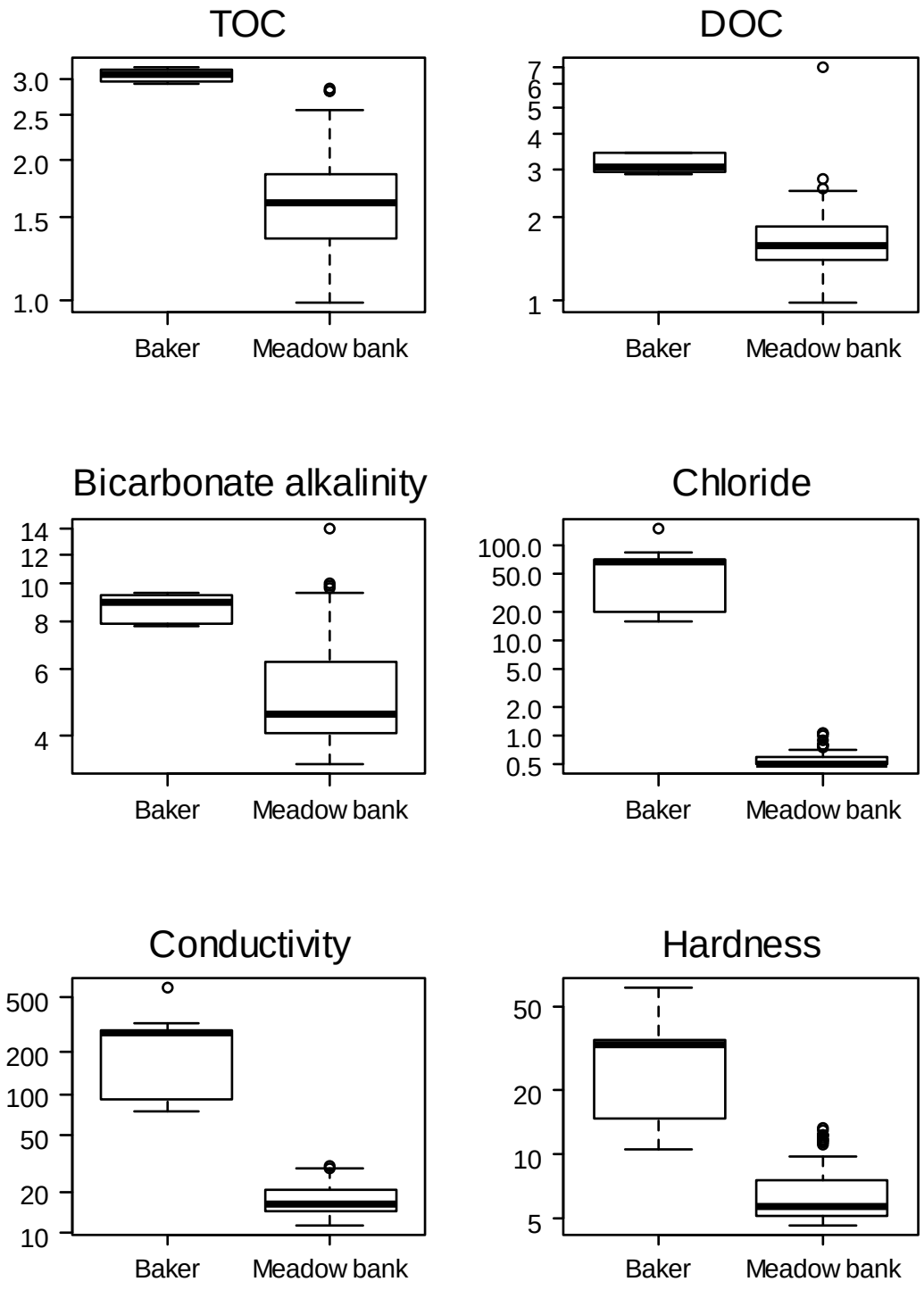


Figure 5 (page 3 of 3)

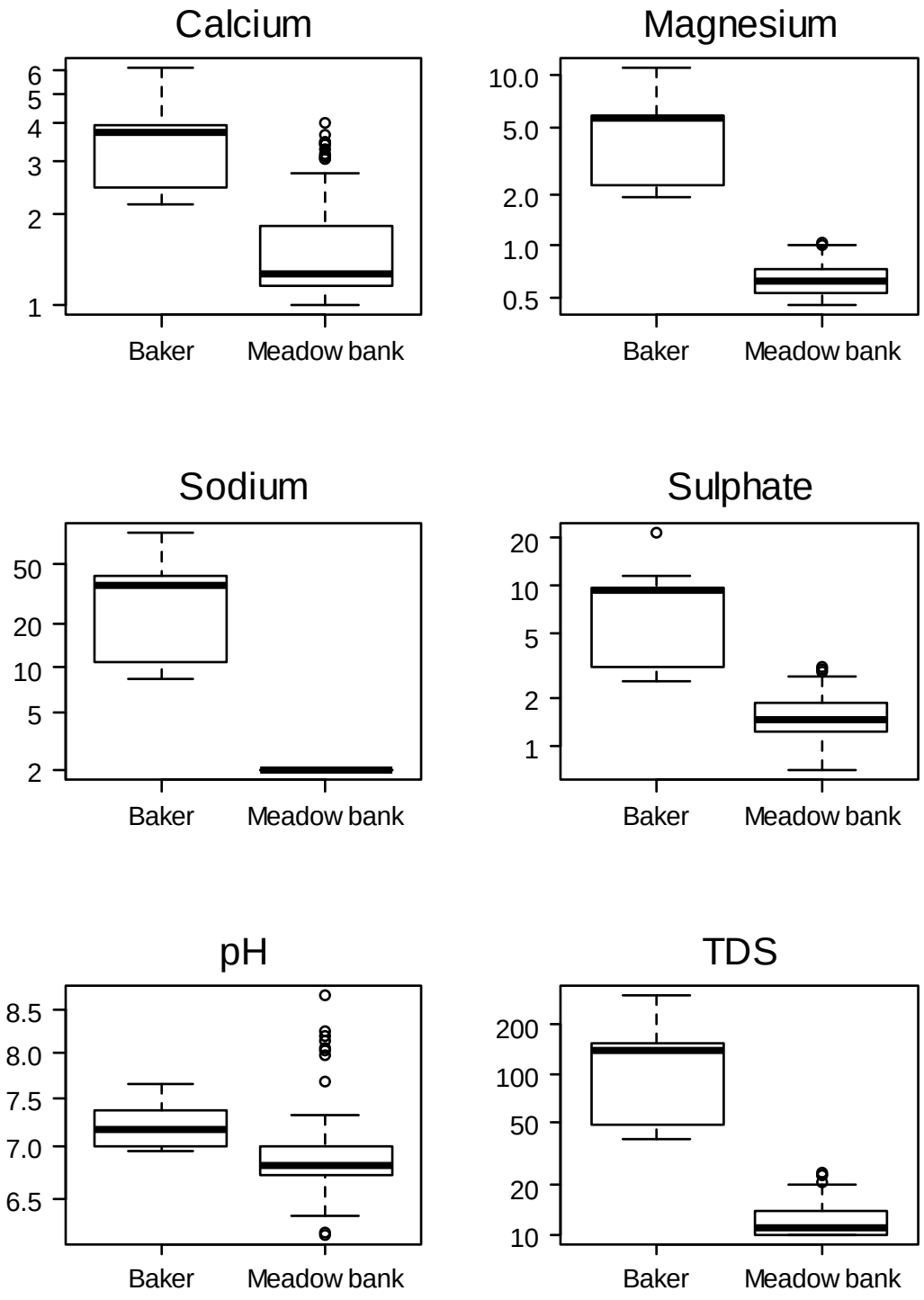


Figure 6. Box-plots of sample measurements for select variables for Baker Lake Stations.

Notes: See text for details

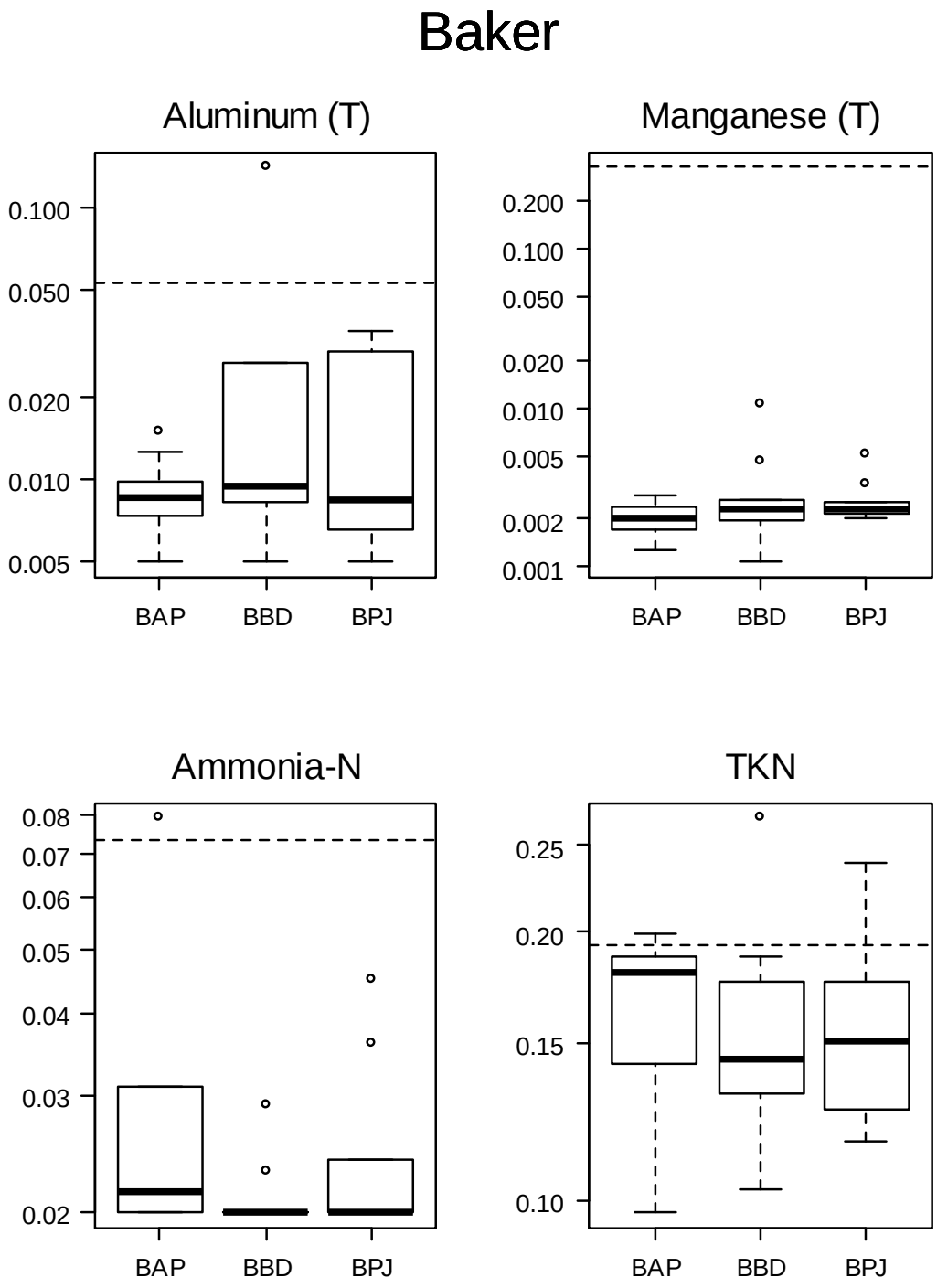


Figure 6 (page 2 of 4)

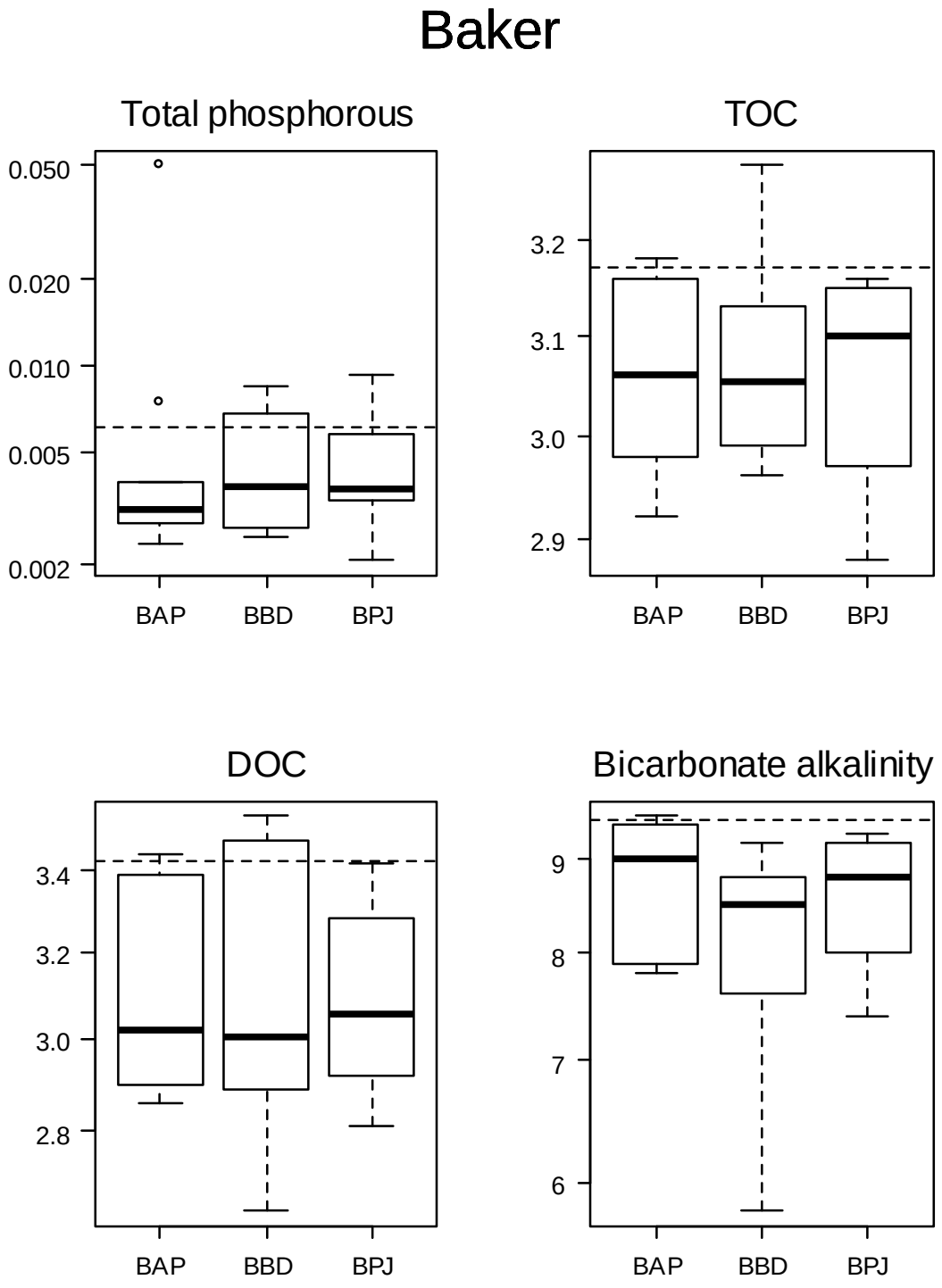


Figure 6 (page 3 of 4)

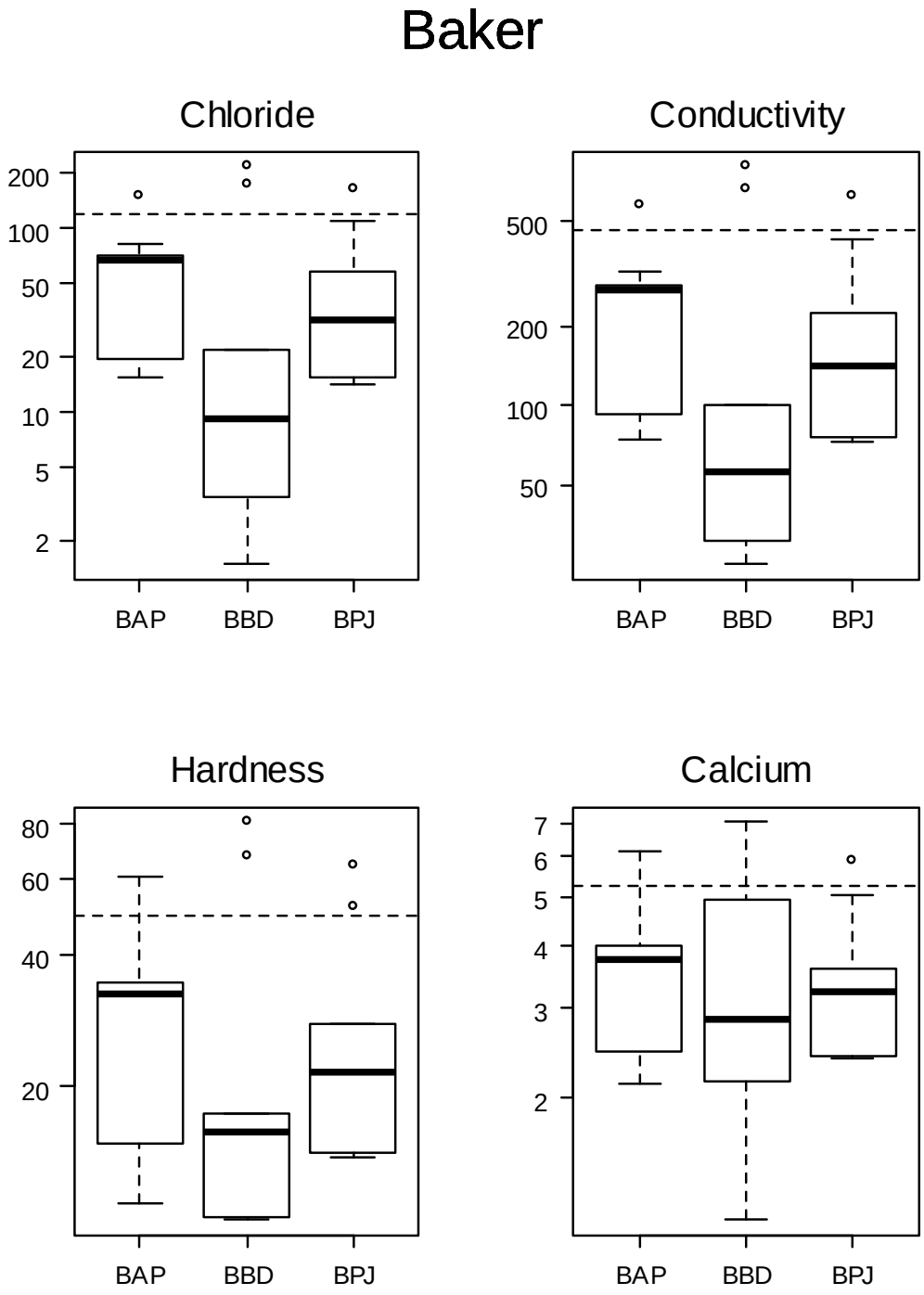


Figure 6 (page 4 of 4)

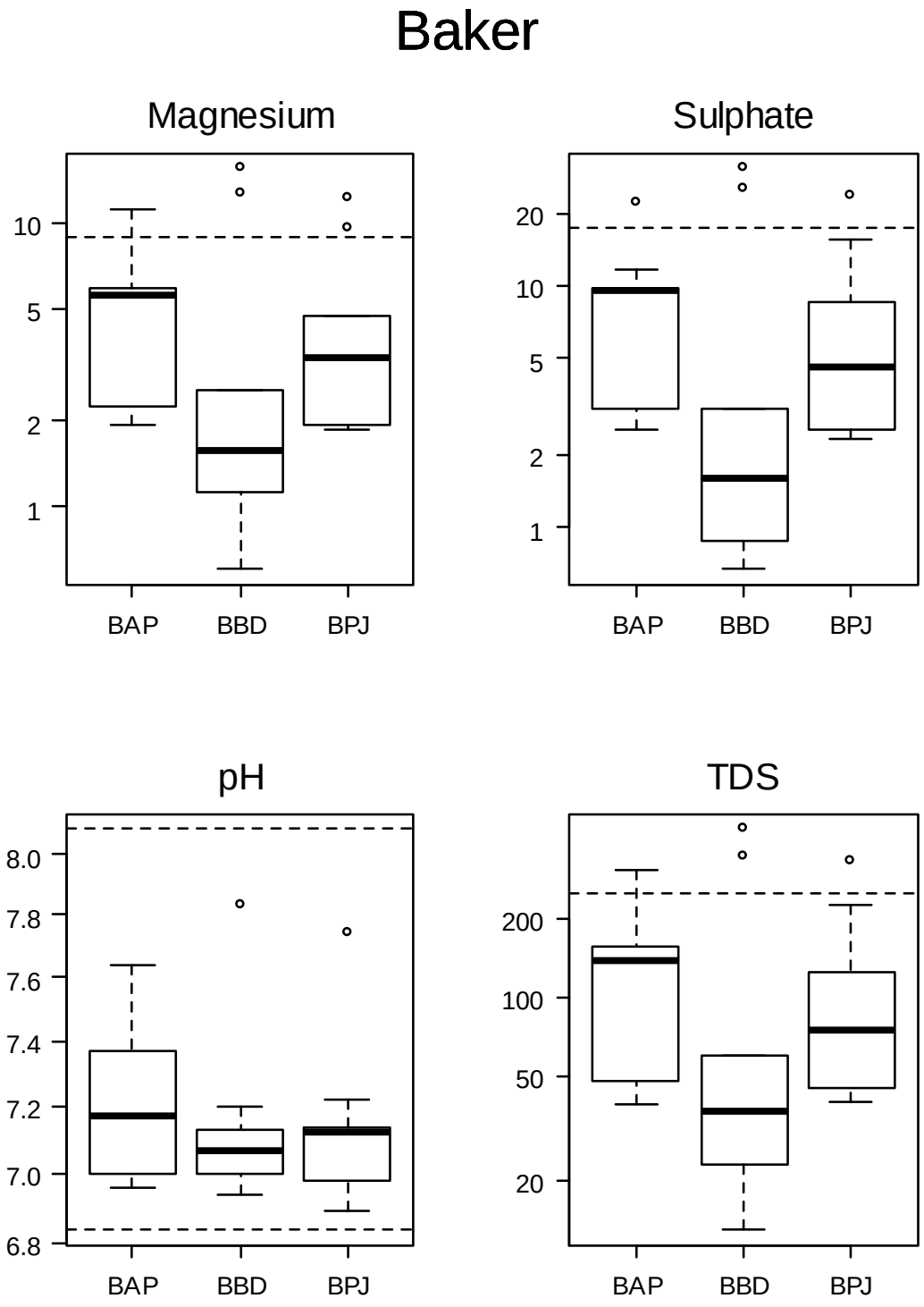


Figure 7. Box-plots of sample measurements for select variables for Baker Lake by Month

Notes: See text for details

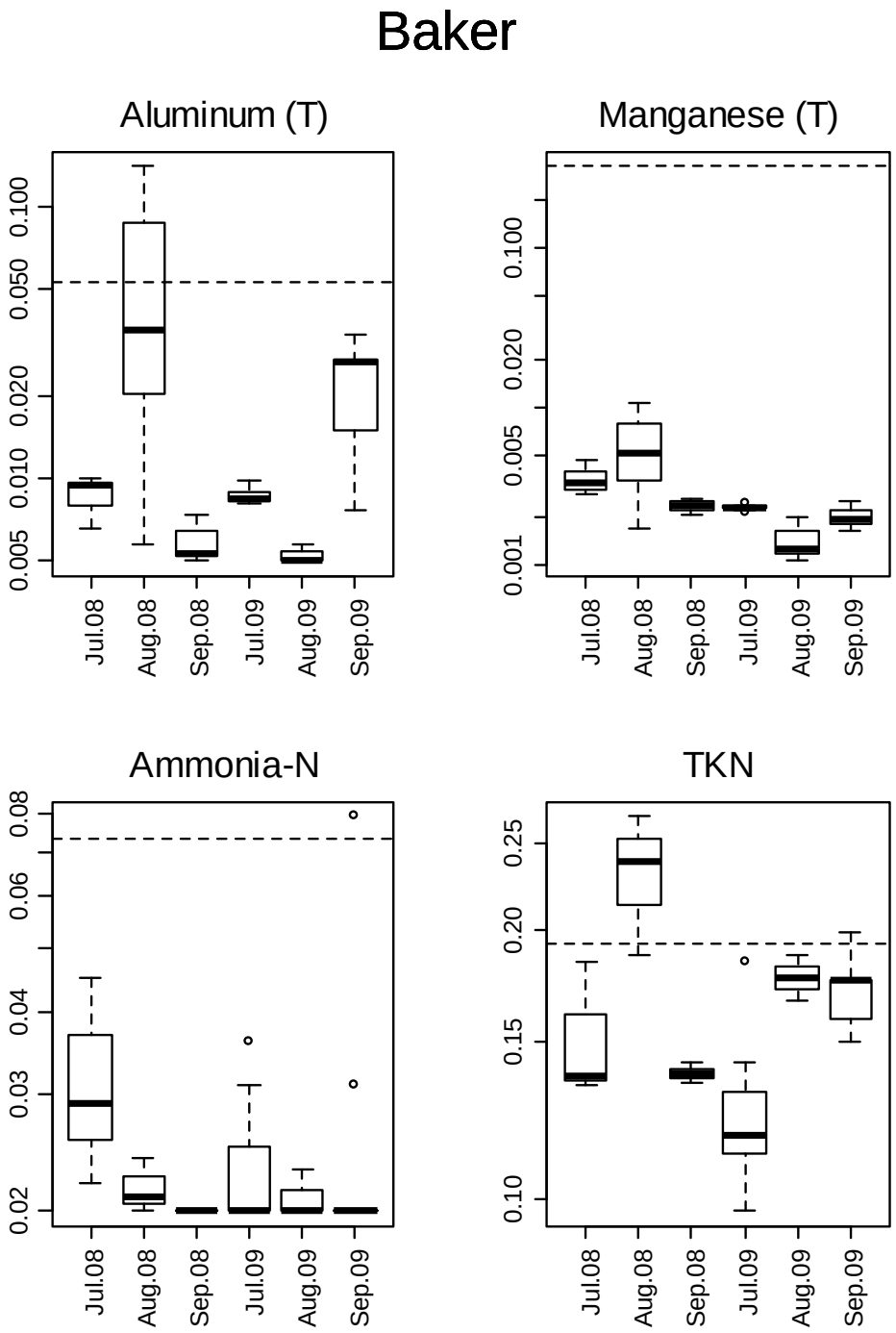


Figure 7 (page 2 of 4)

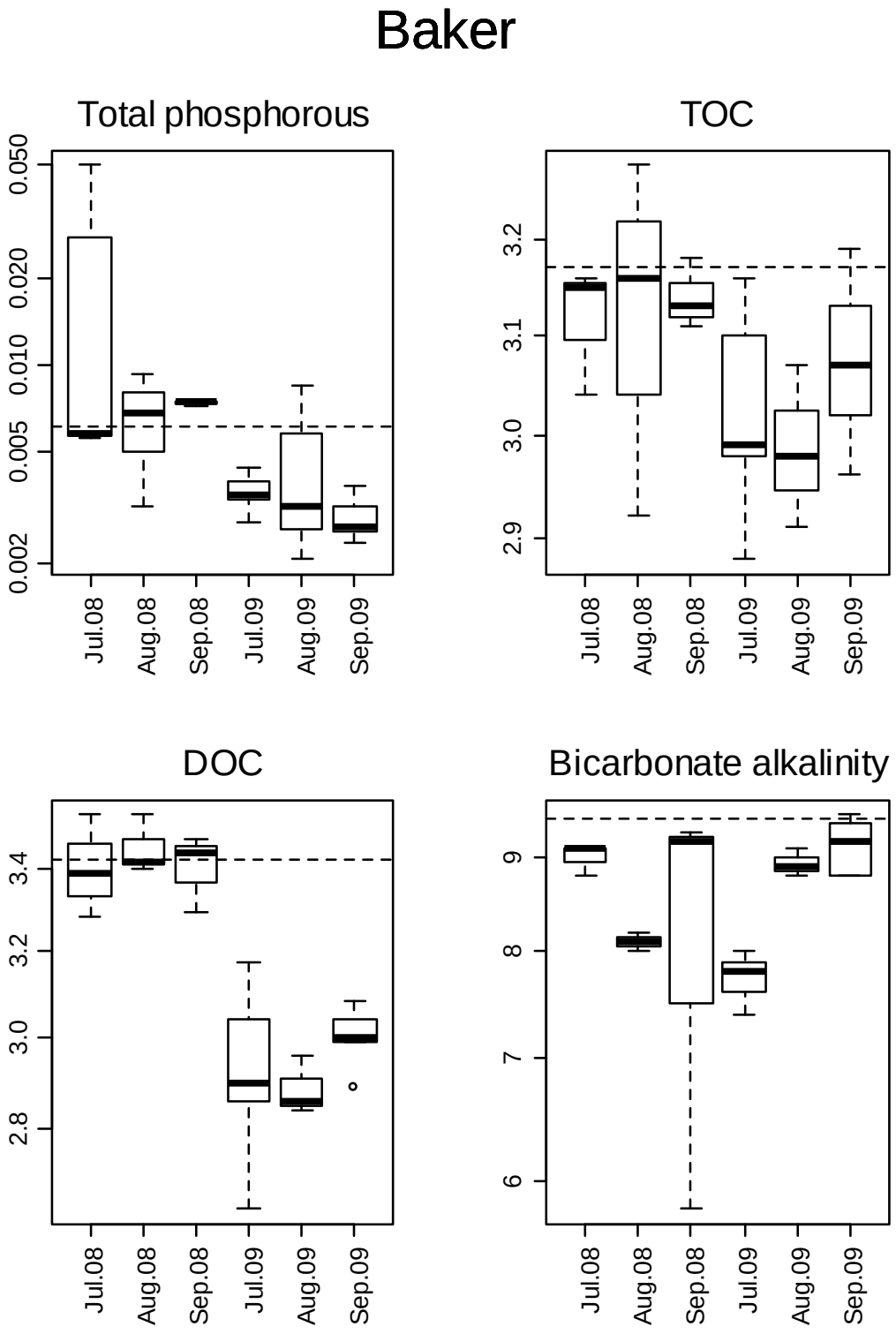


Figure 7 (page 3 of 4)

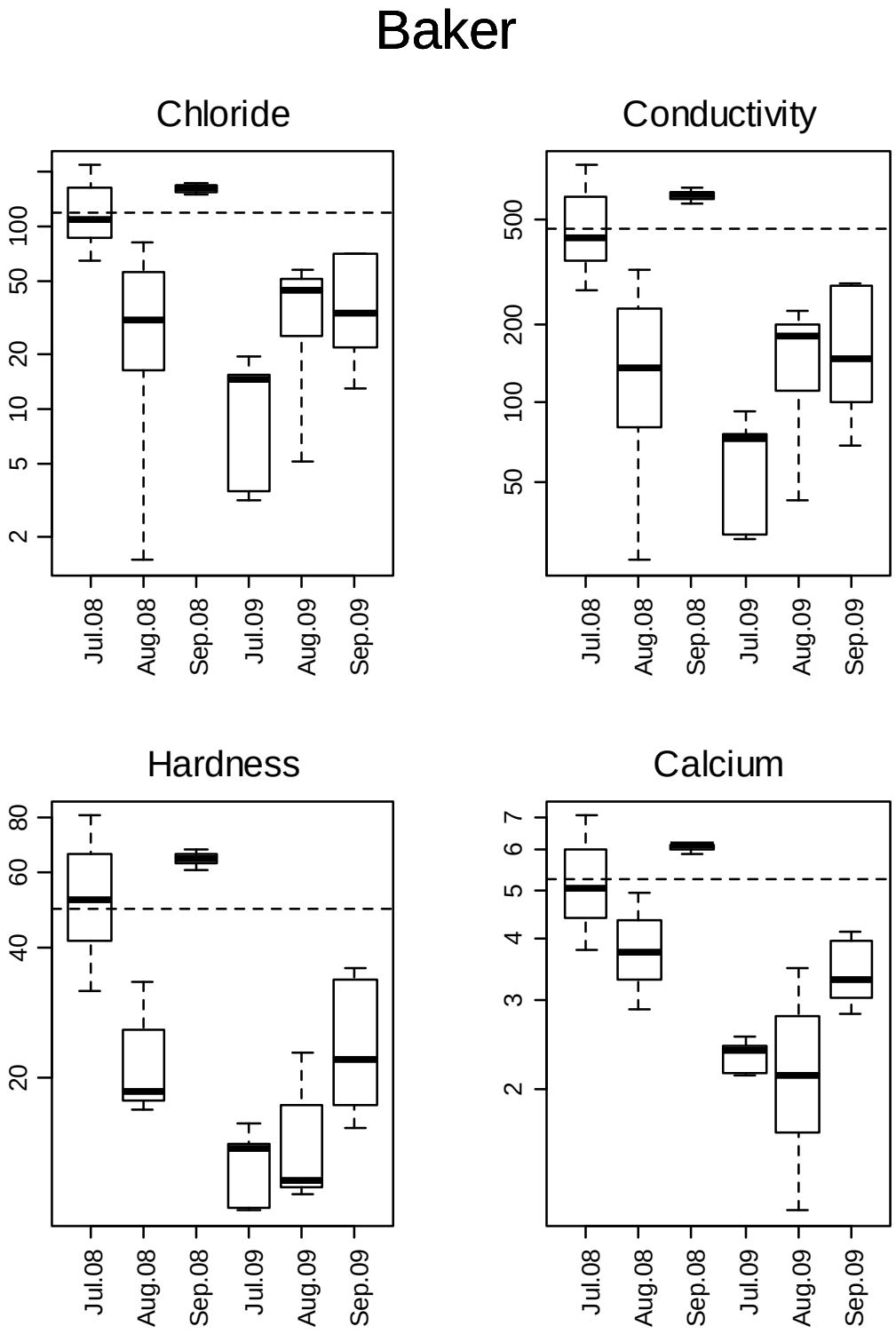


Figure 7 (page 4 of 4)

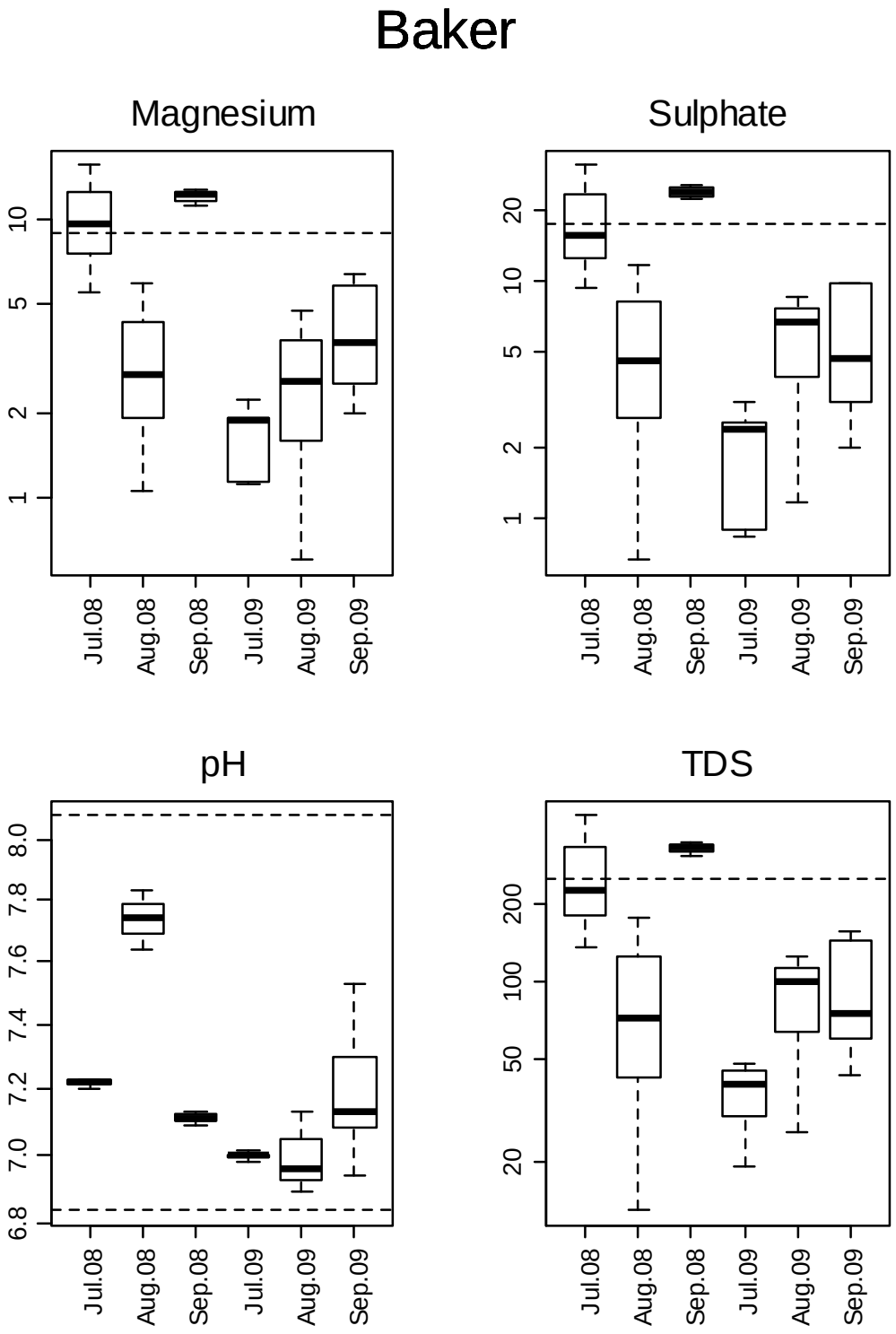


Figure 8. Box-plots of sample measurements for select variables for Meadowbank Stations.

Notes: See text for details

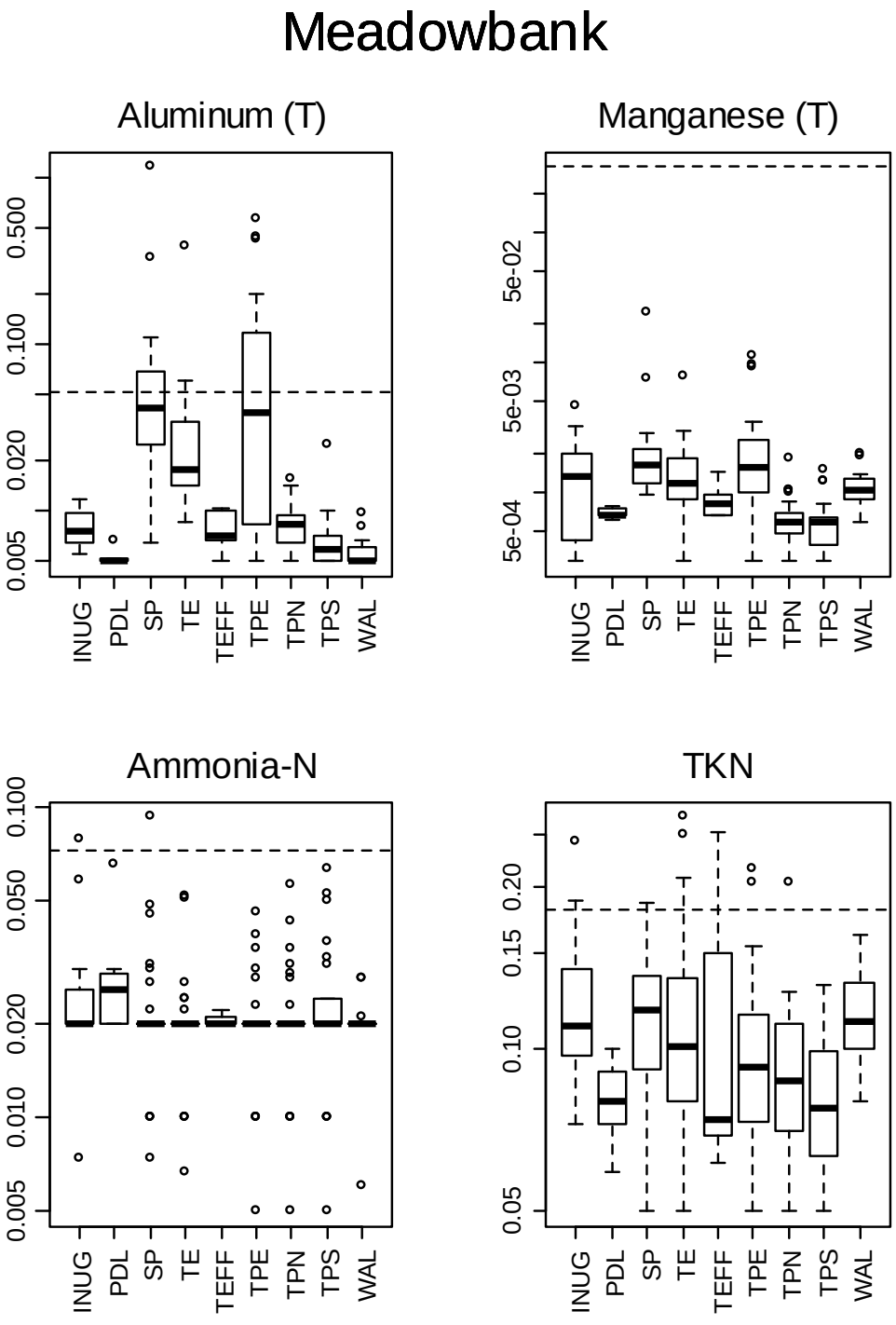


Figure 8 (page 2 of 4)

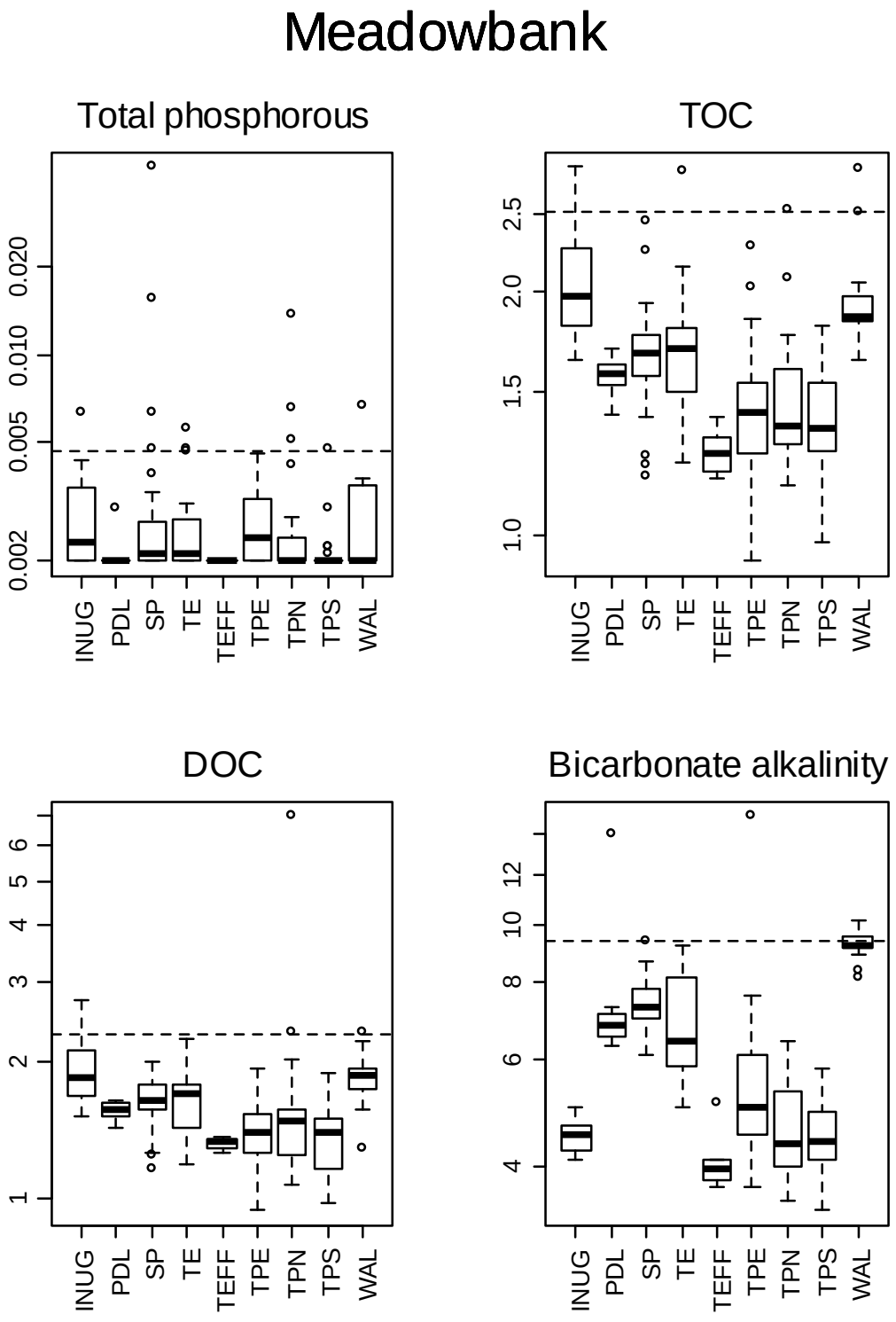


Figure 8 (page 3 of 4)

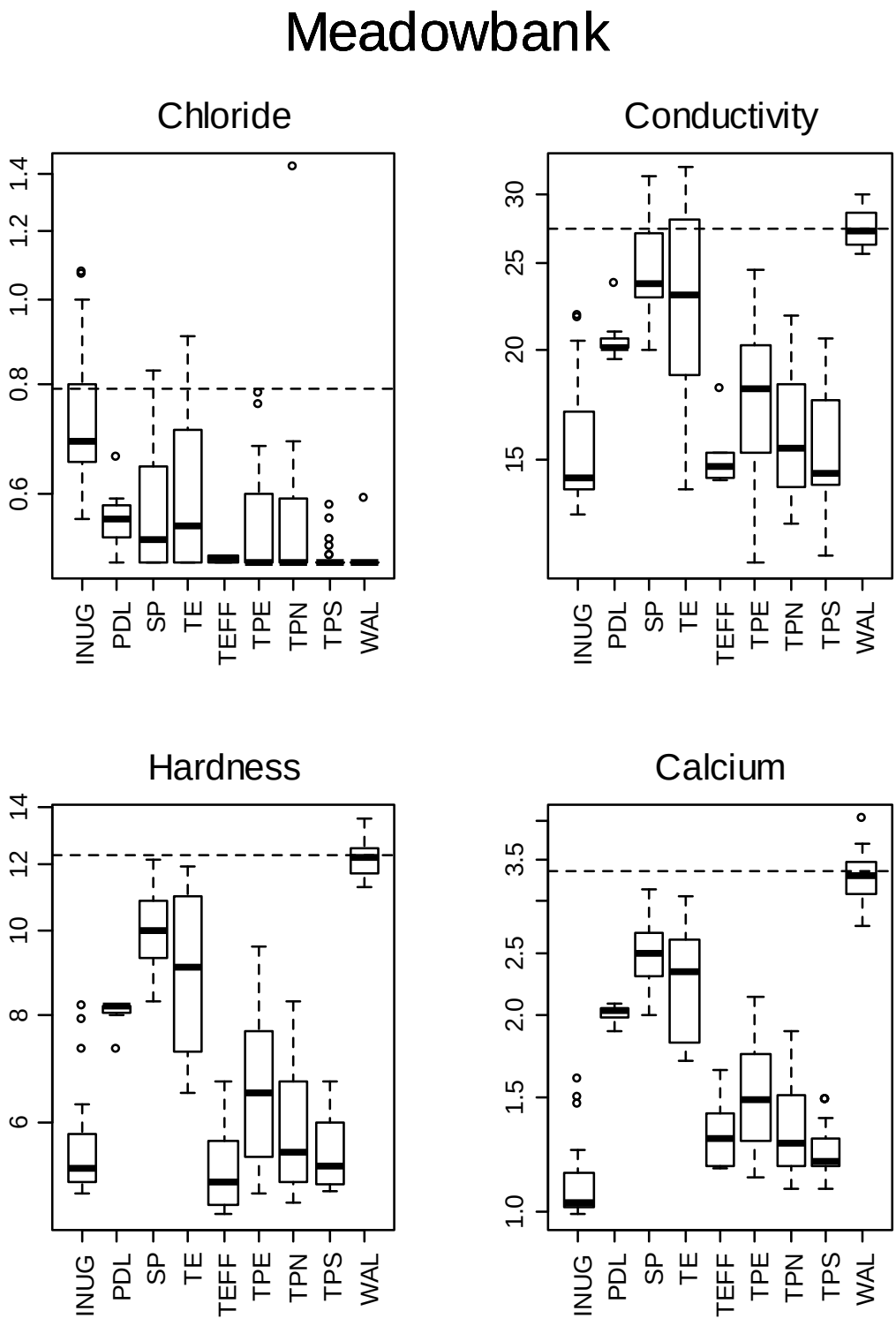


Figure 8 (page 4 of 4)

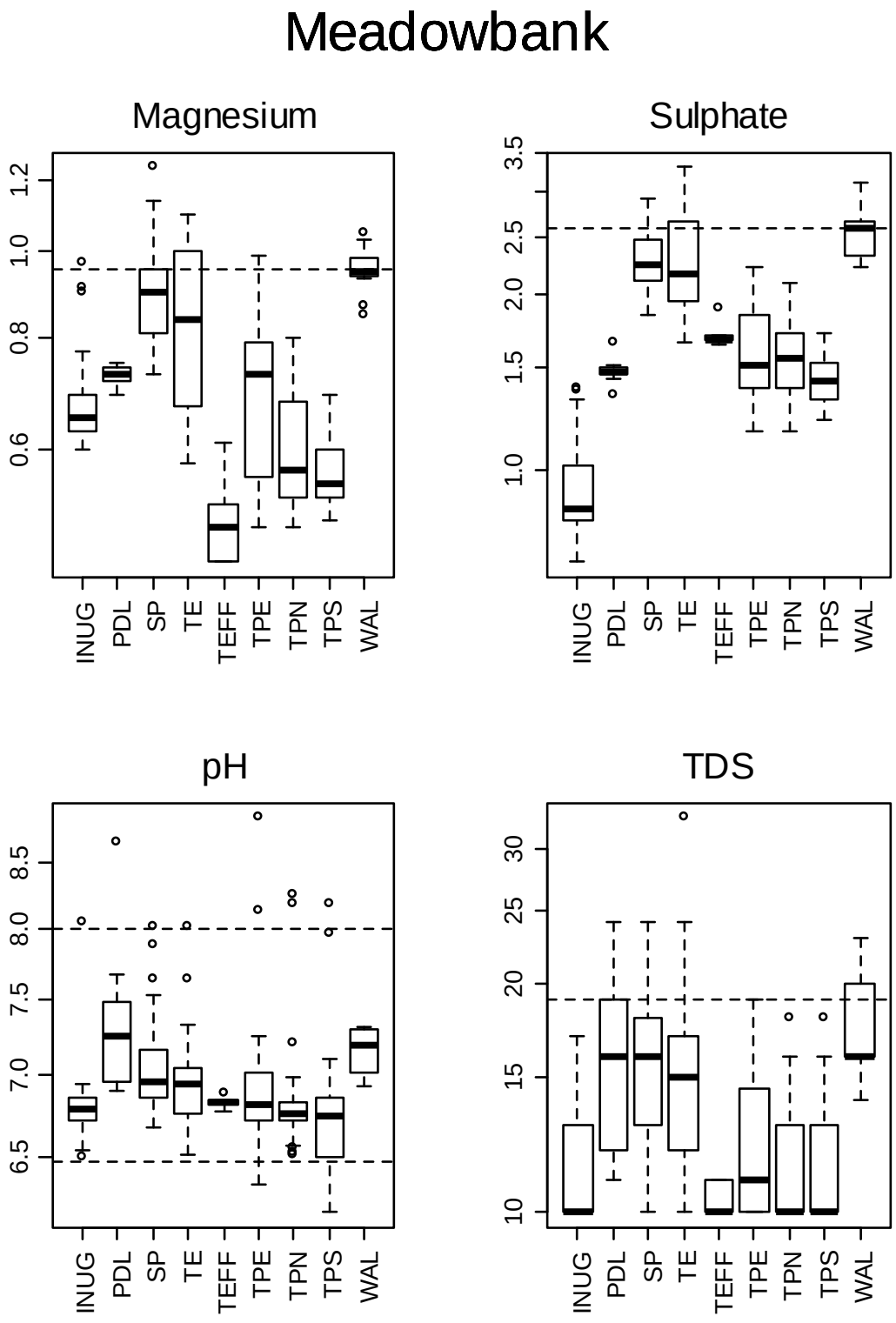


Figure 9. Box-plots of sample measurements for select variables for Meadowbank data by Month.

Notes: See text for details

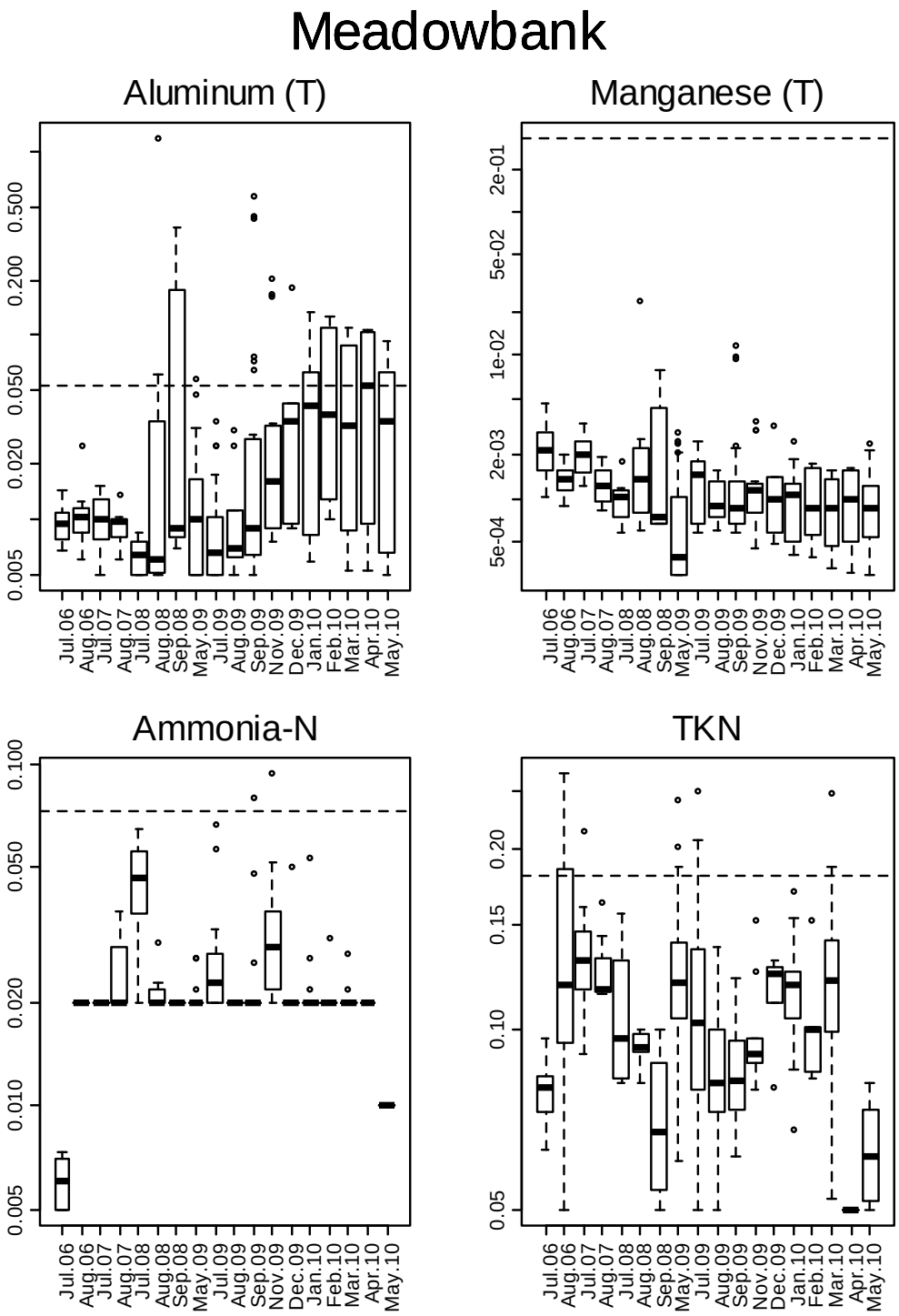


Figure 9 (page 2 of 4)

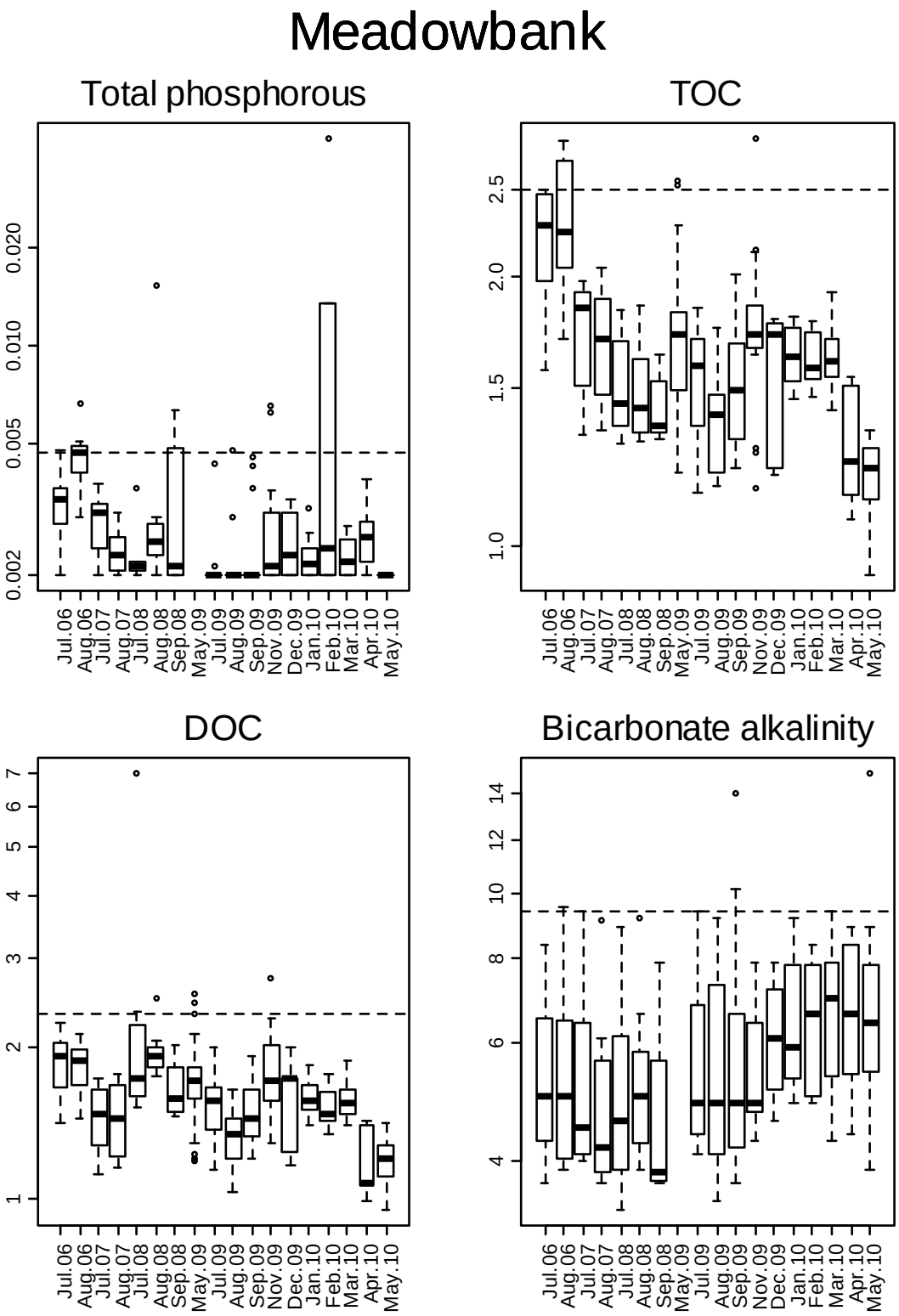


Figure 9 (page 3 of 4)

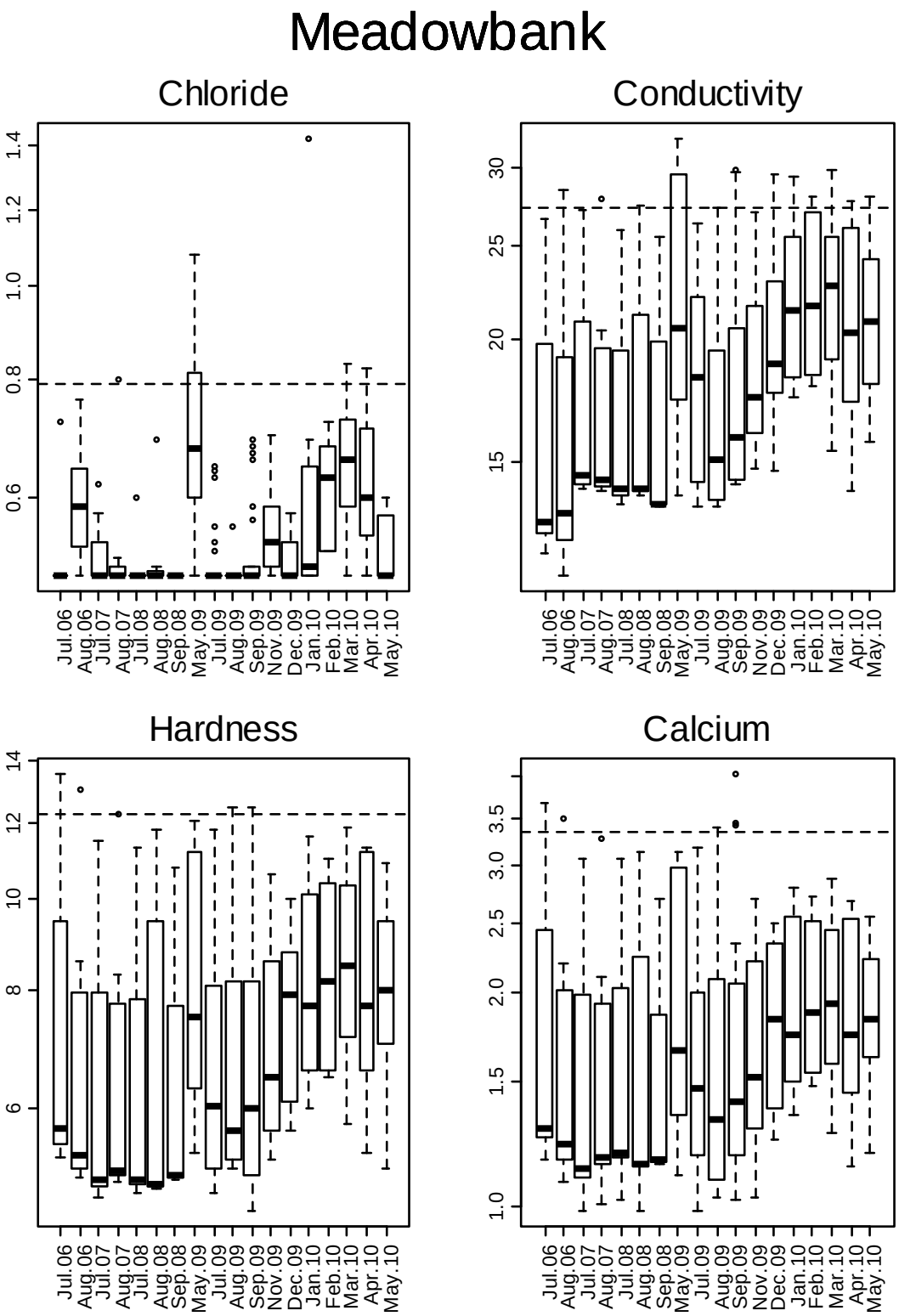


Figure 9 (page 4 of 4)

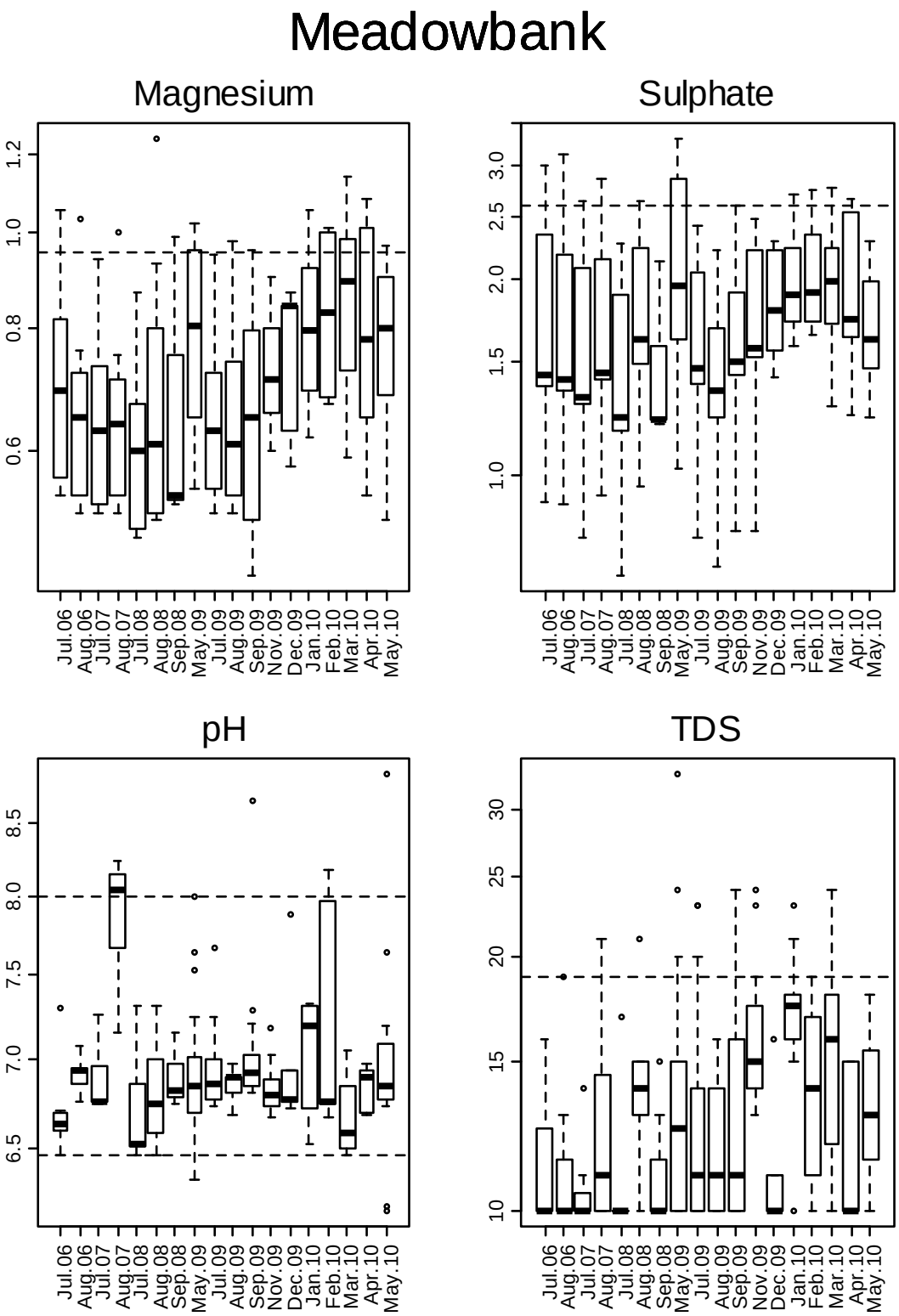


Figure 10. Estimates of BACI statistical power for detecting a significant increase (one-tailed test, alpha = 0.05) in a given variable (rows) by station (columns) as a function of sampling months (after period) and the number of sub-samples per month (see legend).

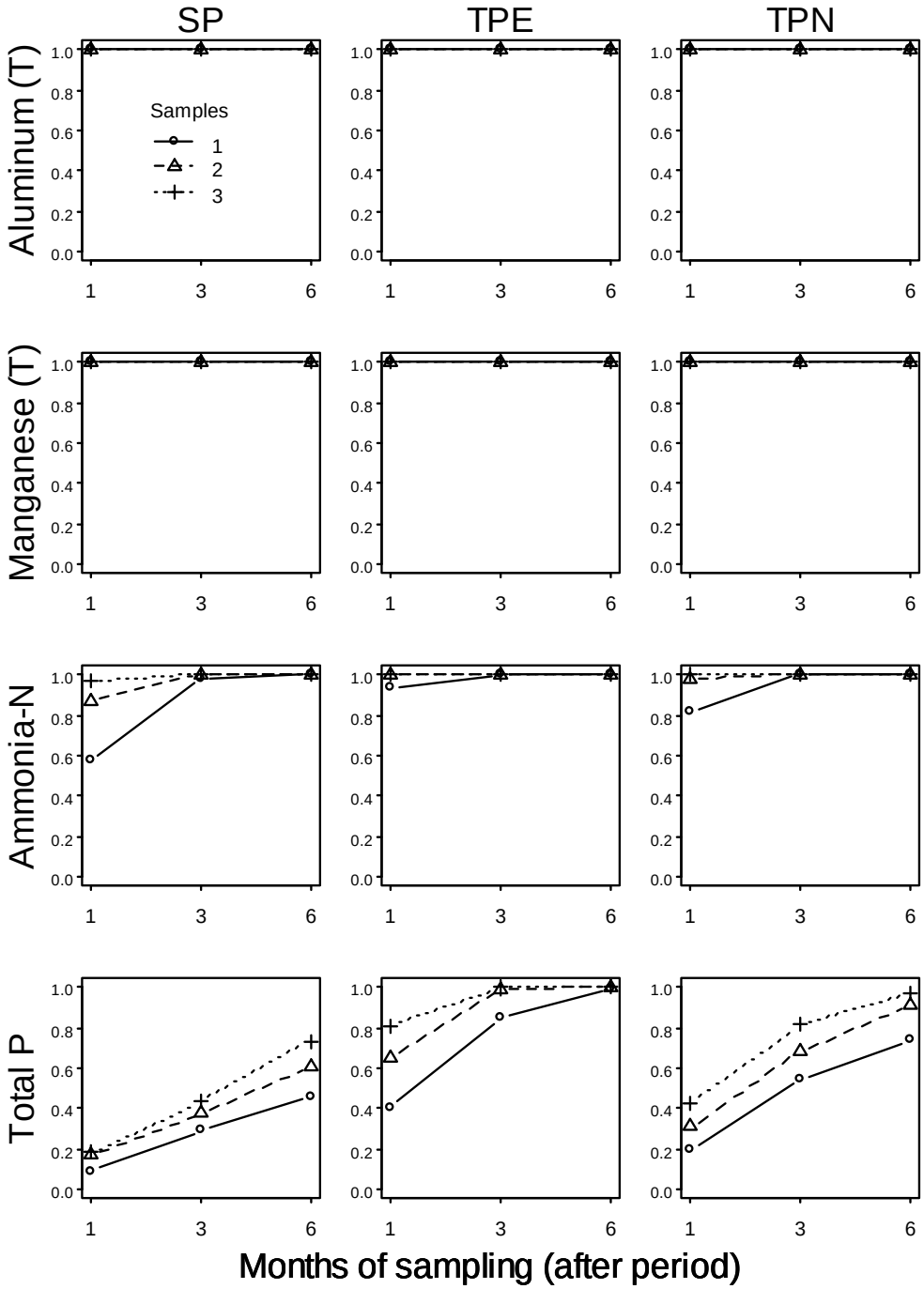


Figure 11. Estimates of BACI statistical power for detecting a significant increase in pH (top row) or decrease in pH (bottom row) by station (columns) for two-tailed tests (alpha = 0.05) as a function of sampling months (after period) and the number of sub-samples per month (see legend).

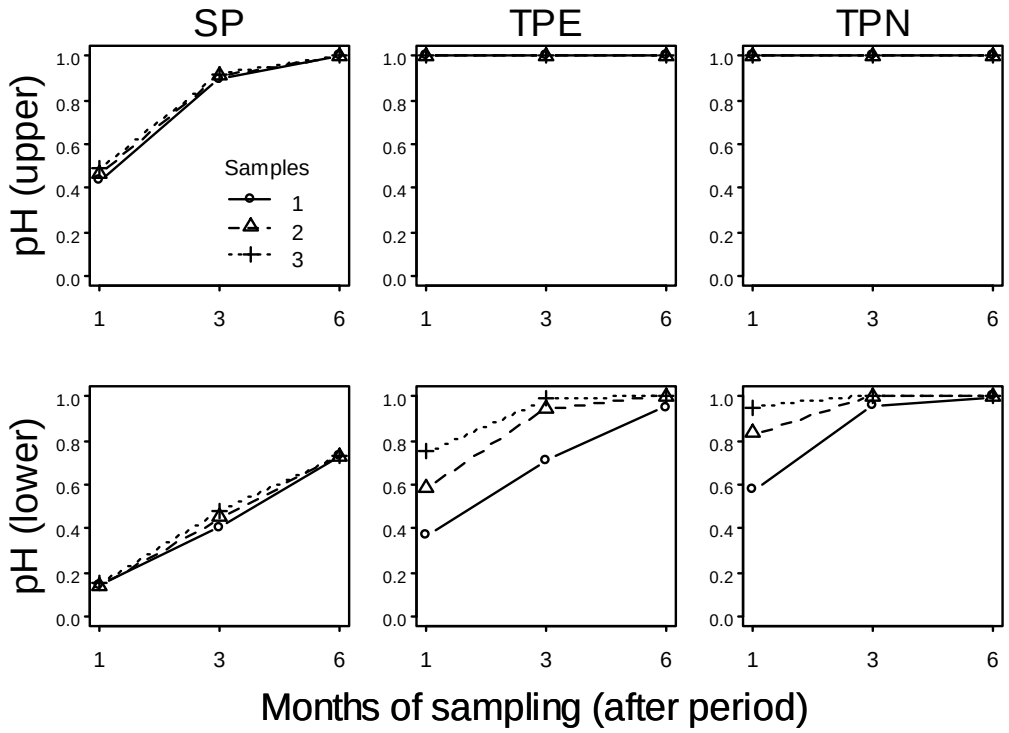


Figure 12. Coefficients of variation (%) for BACI estimates by variable (row) and station (column) as a function of sampling months (after period) and the number of sub-samples per month (see legend).

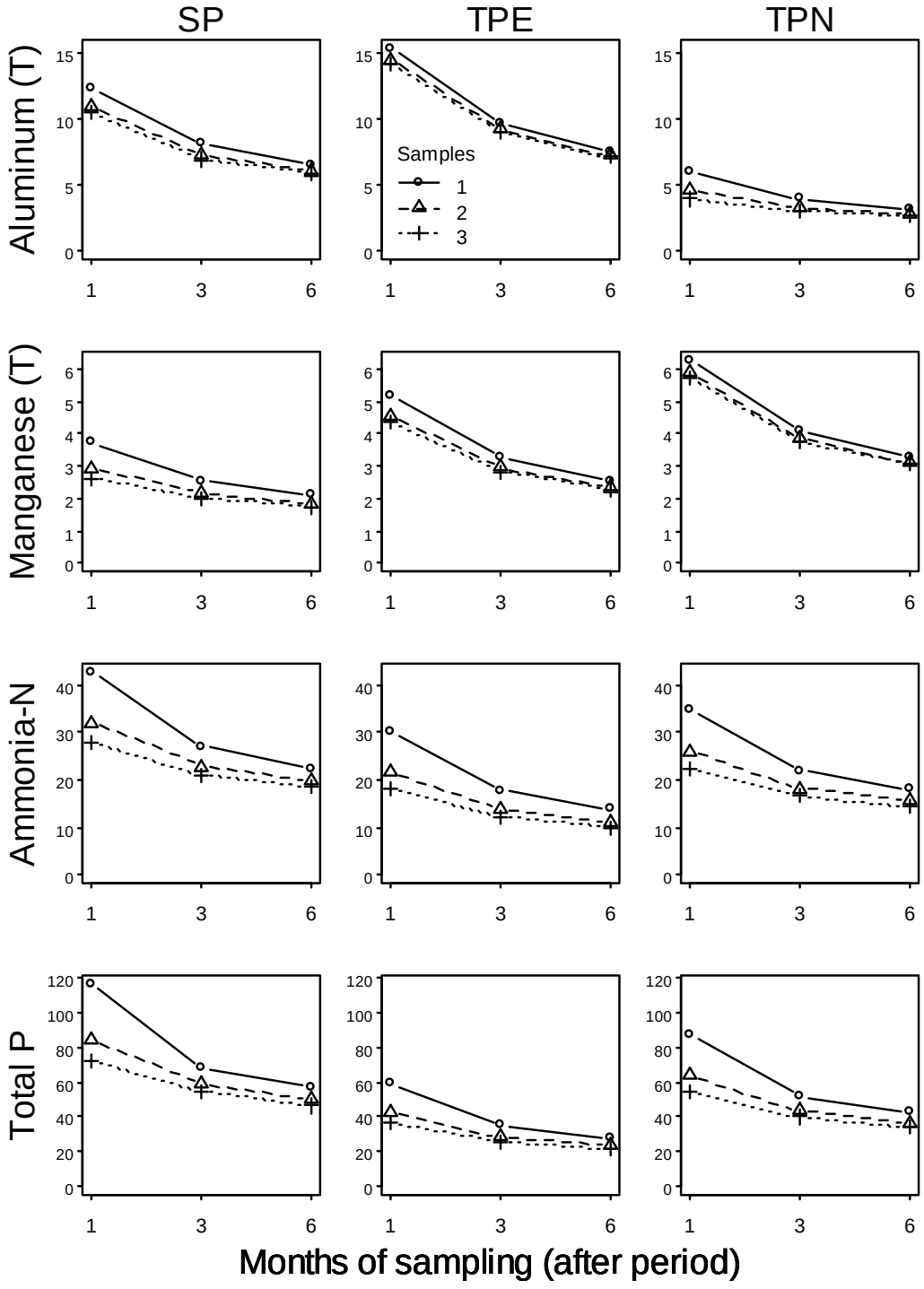
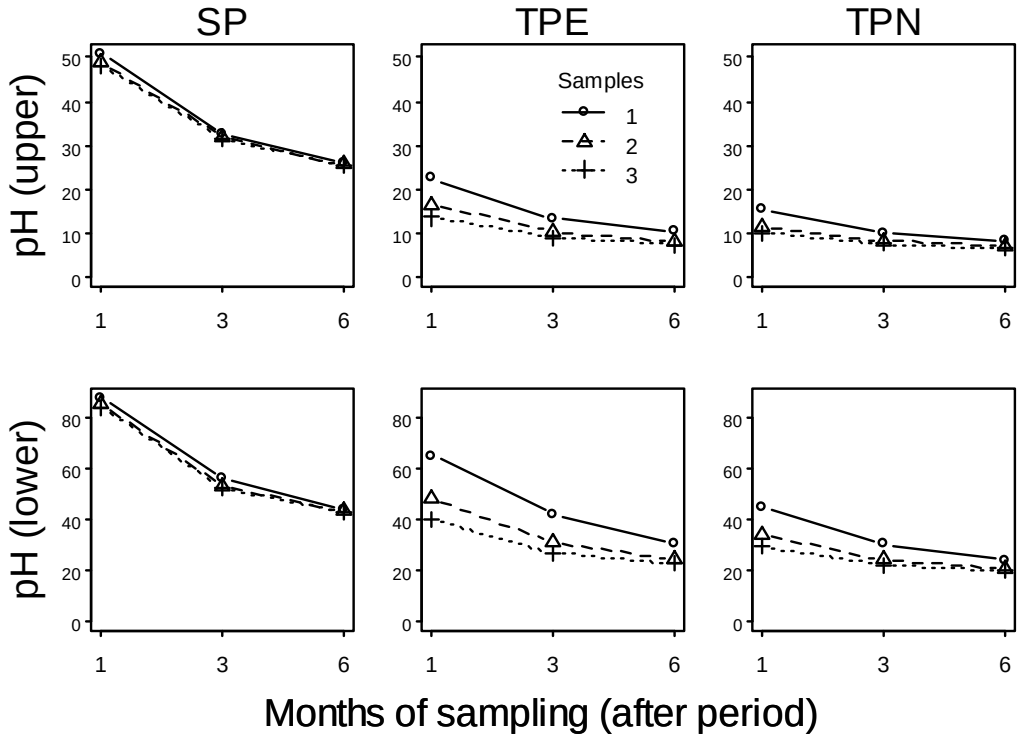


Figure 13. Coefficients of variation (%) for BACI estimates of increases in pH (top row) or decreases in pH (bottom row) by station (column) as a function of sampling months (after period) and the number of sub-samples per month (see legend).



APPENDIX B – STATISTICAL ANALYSES FOR SEDIMENT CHEMISTRY

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This appendix contains the following analyses.

- Summary of CREMP sediment samples (plus relevant samples from East Dike Effects Assessment Study)
- Analysis of field duplicates
- Comparison of core versus grab samples
- Analysis of Baker/Meadowbank differences
- Determination of triggers
- Analysis of sampling design

Some text, tables and figures are repeated from the main document so that the analyses contained in this appendix can be read and understood without reference to the main document. The material in this appendix assumes understanding of basic statistical methods (Venables and Ripley 2002), mixed-effects models (Pinheiro and Bates 2000), before-after-control-impact (BACI) experimental design (Stewart-Oaten et al. 1986; Underwood 1994; Smith 2002), and use of simulation in statistical analysis (Gelman and Hill 2006).

Analysis of sediment chemistry is limited to seven metals for which there are CCME sediment guidelines (As, Cd, Cr, Cu, Pb, Hg, Zn).

2. SUMMARY OF CREMP SEDIMENT SAMPLES

There are three basic types of samples:

1. Core samples – the top 1 cm specific to a discrete sample at a given location and month
2. Grab samples – composite of three spatial recplicates
3. Duplicate samples – occasionally collected; these consisted of a second sample collected in very close proximity to the base sample (in the case of core samples) or split samples from the same sample (in the case of grab samples).

For many station-month combinations, more than one core or grab sample was collected. These “sub-samples” represent spatial replication (a different location within the designated station). In addition, samples were designated as “control” or “impact” depending on operation activities near a given station (see main text for differentiation of control and impact).

The data set for core samples was expanded to include the July 2009 core samples collected as part of the East Dike Effects Assessment Study – those data are directly relevant as 'control' data. In total, there were 201 core samples from the CREMP program, an additional 60 from the East Dike program, plus 60 grab samples collected across the years 2006-2009 at Meadowbank and Baker Lake (see **Table 1**). An additional 17 duplicate samples were collected (13 core, 4 grab) – see **Table 2**.

3. ANALYSIS OF FIELD DUPLICATES

Field duplicates are different for core and grab samples. In the case of core samples, the duplicates are separate samples collected at the same location. In the case of grab samples, the duplicates are split samples – that is, a different aliquot of sample is taken from the composite sample. Thus, differences among core duplicates will reflect fine-scale spatial variation and sampling/lab precision (measurement error), whereas for grab duplicates, differences should reflect homogenization of sediment and sampling/lab precision.

Data – We would not expect control and impact samples to differ in terms of duplicate measures, so all duplicates were examined (for cores, there were 10 control and 3 impact; for grabs, all 4 duplicates were control). **Table 3** summarize the number of measurements and values above detection limits (DL) for all duplicates (core and grab). With 17 pairs of sample/duplicate values, N = 34. Lead was excluded from analysis because only one measure > DL.

Methods – For each variable, xy plots were produced using all data, with values below DL set equal to DL. For remaining analyses, data pairs (sample/duplicate) were limited to those pairs with both values above DL. For each variable, we computed the median across all samples, the median absolute difference (MAD) between samples and duplicates, and Spearman rank correlation between samples and duplicates. In addition, we fit a mixed-effects model (Pinheiro and Bates 2000) of the form:

$$\text{Log}(y) \sim \text{Location} + \text{Error}$$

where y is the sediment quality variable, Location denotes random effects for sample pairs (variability among pairs due to sample location/month) and Error denotes the residual error (variability between samples and duplicates).

Results – Results of the analysis of duplicates are shown in **Figure 1** (xy plots) and **Table 4** (medians, MADs, and Spearman r). For all variables, sample/duplicate values were reasonably similar, and there were no obvious differences in variability between core and grab duplicates. The most variable measures were for As and Cd (but sample size was low for Cd). The mixed-effects SDs reported in **Table 4** show the relative variability due

to sample location/month (SD Location) versus duplicate variability (SD Error). In relative terms, SD Error was highest for As and Cd. In all other cases, SD Error is much smaller than SD Location, which confirms that duplicate measures are fairly precise compared to general sample variation.

4. COMPARISONS OF CORE AND GRAB SAMPLES

Core and grab samples were compared graphically by station for years 2008 and 2009 (**Figure 2**; Pb omitted because most data < DL). Control and impact data were considered if consistent for grab/core samples; however, 2009 data for station TPE were omitted because the core data was collected in July and designated "control," while grab data was collected in August and designated "impact". Remaining data appear comparable, but comparisons are not perfect – for example, grab and core samples for SP and TE in 2009 were collected in different months (July versus August) and therefore potential for impacts would be different (see **Table 1**). In any case, in general, grab samples were reasonably consistent with core samples (e.g., grab measures were within the box/whiskers of core data) for As, Cr, Cu and Zn. For Cd, measures for grab samples did not exceed the DL (0.5 mg/kg) in 2008 or 2009 for any station, while values for numerous core samples exceeded 0.5 for several stations. For Hg, grab-sample measures tended to be at the low end of core-sample distributions in 2008 and for INUG in 2009. These general observations were not tested statistically (results would be weak given low sample sizes for grab samples). Differences for cadmium are obviously significant, and we have no explanation for those (possibly some kind of lab error).

5. ANALYSIS OF BAKER/MEADOWBANK DIFFERENCES

Baker Lake and the Meadowbank area lakes may exhibit differences in sediment chemistry due to their different characteristics. Consequently, development of separate thresholds and triggers may be appropriate for the two areas for any variables that exhibit large differences.

Data – Potential differences between Baker/Meadowbank were examined by considering the control core samples (with duplicates excluded, N=15 for Baker station BAP and N=135 across Meadowbank stations – see **Table 1**).

Methods – Analysis excluded Cd and Pb (no values > DL for Maker and less than half of Meadowbank samples > DL). Values below DLs were set equal to DLs. Box-plots were produced for each variable to compare the Baker and Meadowbank samples. In addition, the Baker and Meadowbank data were compared using Mann-Whitney tests (on the raw

data) and t-tests (using log data). The Mann-Whitney tests is more robust in this case as it is appropriate for non-normal data (which occurs when there are numerous values = DL).

Results – There are large differences between Baker and Meadowbank samples for all variables (**Figure 3** and **Table 5**), with concentrations much lower in Baker Lake compared to Meadowbank lakes. Since all data were log-transformed, the coefficient (X) has to be translated into a proportional effect size (ES) relative to (untransformed) samples[1]. For example, for As, Est = -1.67 and ES (%) = -81%. In other words, the difference in sample means (Baker log data – Meadow log data) was -1.67, which translates as a proportional effect size (raw units) with Baker data being -81% lower on average than Meadowbank data. In summary, there are much higher concentrations among Meadowbank core samples for the five sediment metals examined. For Cd, we would reach a similar conclusion simply based on proportions > DL (0 of 15 for Baker; 64 of 135 for Meadowbank; see **Table 4**). For Lead, there was no apparent difference, with few samples > DL for either lake (**Table 4**).

6. THRESHOLDS AND TRIGGERS

Data – The data set upon which thresholds and triggers were developed was all core control samples (which included the July 2009 data for the East Dike Effects Assessment Study), with duplicates excluded. Remaining data points with values below the usual DL were set equal to the DL. For Cd and Pb, there were a few samples with recorded detection limits greater than both the current DL as well as other valid sample values. These samples were removed, and then the remaining "<DL" values were set equal to the DL before computing medians and 95th percentiles.

Methods – The main text has described the rationale and approach for development of thresholds and triggers. Since all sediment variables had thresholds, the trigger was set as the maximum of either (a) the value halfway between the baseline median and the threshold ("Method A"), or (b) the 95th percentile of the baseline data ("Method B"). Medians and 95th percentiles were chosen as metrics rather than means, standard deviations, or maximums, because the former are generally robust to skewed distributions and potential outliers. For variables with few values above DL (Pb, and As for Baker), the reported medians equal the DL. Different triggers were set for Meadowbank and Baker for all variables except Lead (see above analysis and 95th percentiles in **Table 6**).

[1] The following steps are needed for calculating ES: (1) X = log(Baker) – log(Mbk); (2) X = log(Baker / Mbk); (3) Exp(X) = Baker / Mbk; (4) ES = (Baker – Mbk)/ Mbk = Baker / Mbk – 1; (5) ES = Exp(X) – 1; (6) ES (%) = (Exp(X) – 1)*100.

Results – Triggers and thresholds are summarized in **Tables 6 and 7** respectively. It should be noted that in cases where Method B was used for trigger development (i.e., cases where baseline data already exceed the guideline for > 5% of cases), it is possible for the trigger to be more extreme than the guideline (e.g., this occurs for all variables except Pb and Hg at Meadowbank, and for As at Baker) – in such cases the guideline is reported as the threshold but is not used as a criterion for action; rather, the trigger is the only criterion for action. For Pb, 9 of the 11 values above the DL were for Wally Lake, three of which were slightly above the CCME guideline of 35 mg/kg. Thus, development of separate triggers for Wally for Pb may be warranted (similarly, for Cu and possibly other variables, Wally-specific triggers could be developed).

Evaluation of Triggers – We next examined the data more closely in relation to the proposed triggers. Specific box plots are shown for Baker Lake data by station (**Figure 4**) Meadowbank control data by station (**Figure 5**) and data by year for four particular stations (INUG, SP, TE and TPE; Pb excluded (**Figure 6**)).

For **Baker** stations, the general pattern is for higher concentrations at impact station BPJ. These do not exceed Baker triggers except seriously for As (BPJ median ~ 20 mg/kg, trigger = 8.3). Given the high As levels observed for Meadowbank controls, it seems reasonable to question the validity of the low Baker trigger. However, given limited sample sizes at Baker station BAP to date, it may be simpler to defer more in-depth analysis of the Baker trigger for As until more data are available.

For **Meadowbank**, triggers look reasonable for As, Cd, Hg, and Zn. Wally has high levels of Cu and Pb (near the trigger), which suggests that the Cu and Pb triggers based on 95th percentiles are only strictly valid if Wally data are considered representative control data for Meadowbank stations, otherwise, triggers are somewhat high. In the case of Cu, the high values for Wally are balanced by the low values for INUG, and there is no reason to suspect that the trigger is generally too high. For Pb, detection limits, the trigger is set just barely above the DL so we are not concerned that it is set too high. For Cr, there is a lot of station-specific variation, which may suggest that station-specific triggers would be more appropriate. This possibility is discussed in the main text.

The two years of data for INUG, SP, TE, and TPE (**Figure 6**) illustrate two things. First, in only one case did impact data (July 2009 for SP and TE) appear to exceed a trigger: the median of Zn values for SP slightly exceeded the proposed trigger. Thus, there are no alarming findings among the impact data. Second, the two years of control data for INUG and TPE provide the best insight into possible inter-annual variation in sediment measures. For INUG, there was a significant difference between years for Cd ($p = 0.0005$ for Mann-Whitney test). For TPE, there were differences for Cr ($p = 0.0005$) and Zn ($p = 0.04$).

7. ANALYSIS OF SAMPLING DESIGN

Impact hypotheses and statistical design – Two general classes of impacts are hypothesized for the Meadowbank mine for the case of sediments:

1. Pulse events for which potential impacts would be high for a short time but may then dissipate relatively quickly (e.g., in the case of water) or not (e.g., in the case of sediment). Pulse events could be associated with any phase of the mine, but are more likely to be associated with particular activities such as dike construction.
2. Long-term cumulative impacts that may be associated with ongoing activities. Long-term cumulative impacts are more likely to be associated with ongoing activities of mine operations.

As operations have just begun in 2010, the focus of monitoring to date has been on detecting pulse events associated with construction.

In the case of sediment, a key assumption is that natural temporal variation is very limited; that is, you do not expect important fluctuations in sediment variables from to occur naturally from one year to the next at a given station. Under this assumption, the comparison of interest is simply between baseline (before) and impact (after) data for a given station. There is no need or use for a "control station" under this assumption – the value of a control is to specifically "control" for potential temporal variation. Thus, the primary model for these analyses was a simple **"BA"** comparison (a t-test) between baseline data (**B**efore data) and a simulated new year of impact data (**A**fter data). This test will generally have much greater power to detect a difference (real or not) than a full BACI model (which inherently assumes that temporal variation may be present, both across stations and within stations). For comparison, estimates/power for BACI models using impact station TPE and control station INUG were examined, both of which were sampled for two baseline years (**Table 1**) and thus provide the minimum of data required for the BACI model.

The BA model has a simple design:

$$X_{is} = \mu + \alpha_i + \varepsilon_{is} \quad (3)$$

Where X is the sediment variable measured for period i and subsample s, and the estimate of interest is the difference in means for the before and after periods, represented by α_i, which is tested via a t-test.

The BACI model is the same as that developed for water (see separate appendix) and is not repeated here.

Methods – The key analysis assessed the expected precision/power of BA estimates for different numbers of sub-samples (**s** = 5, 10, 15, 20, or 25) for a particular impact year

(random spatial samples collected in the after year). In other words, the objective is to detect a difference from baseline in a given (single) year. The analysis here included separate analyses for the three primary impact stations SP, TPE, and TPN. For TPE in which a BACI model was evaluated as a comparator, INUG was used as the control station. Separate analyses were conducted for **five** variables – As, Cr, Cu, Hg and Zn (Cd and Pb were excluded because many or most values < DL). For a given variable, the effect size (**ES)** was equal to the change in mean from baseline (i.e., during the before period) to the trigger value. For example, if the before-period mean = 10 and trigger = 15, then ES = 5 (=15-10). Thus, for all reported results, we are assessing the power/precision of BA estimates under the assumption that the true effect size resulted in a new annual mean equal to the trigger value (of course, the observed after-period mean will differ somewhat from the trigger value due to simulated random variation). In all cases, only the baseline (pre-impact) data were used to represent the "before period" dataset (July 2008 for SP and TPN, pooled July 2008 and 2009 data for TPE). For BA models, after-period data were simulated for a single year based on the ES and variance of the observed (before-period) data. For the TPE BACI simulations, after-period data were simulated based on estimates derived from the TPE-INUG data. Specifically, the following mixed-effects model was fit to the "before data" for the TPE-INUG station pair:

$$X_{jks} = \beta_j + \tau_k + (\tau\beta)_{kj} + \varepsilon_{jks}$$

The fit provided estimates of before-period station means β_j, random-effects variances for year ($\sigma^2[\tau]$) and year-by-station ($\sigma^2[\tau\beta]$), and the residual variance for sub-samples ($\sigma^2[\varepsilon]$).

After-period data were simulated using after-period means (β_control, β _impact + **ES**) and the above variance estimates. In all cases, log-transformed data were used. For each scenario, 500 simulations were used.

Data Summary – The focus was on sediment core samples for Meadowbank near-field impact stations (SP, TPE, TPN). For SP and TPN, the baseline ("before period") data consist of 15 samples for July 2008; for TPE, there are an additional 15 samples for July 2009 (pre-impact for that station; see **Table 1**). For each variable and station, the before-period estimates are summarized in **Table 8**. The metrics of interest are the effect sizes for log data, ES(log), and estimates of standard deviations (SD, log units). In relative terms, power will be high when ES(log) > SD. There are two cases where ES(log) < SD (Cr for TPN and Zn for TPE), which turn out below to be the only low power cases for the BA design.

Results – Estimates of BA statistical power are shown in **Table 9 and Figure 7.** For all variables, the *a priori* hypothesis is that impacts will result in increases, so power is based on one-tailed tests (α = 0.05 and 0.10 are both shown in Table 9; α = 0.05 for

Figure 7). Power was basically 100% except for two cases: Cr for TPN and Zn for TPE. In both cases, the mean of the baseline data was close to the trigger relative to the sample SD – see **Table 8** and box plots in **Figure 5**. The number of sub-samples makes the biggest difference for Zn for TPE, where for example s = 5 has power of 0.44 while s = 15 has power = 0.83.

A measure of precision for BACI estimates is shown in **Figure 8**. For log data, a useful measure of precision is the coefficient of variation (CV = SE[BA estimate]/ [BA estimate]). Results are averages across 500 trials, and do not depend on tails or alpha. As a rough guide, power > 80% when CV < 40%.

These limited comparisons make it difficult to recommend an optimal sub-sample size, in part because there is potential evidence that differences in "random selection" of sampling location resulted in significant differences between years (see results for INUG and TPE in **Figure 5**). This is potentially explained by reviewing the field sampling methods between the two years for INUG (in one year, the field crew clearly followed a linear path and did not randomly sample). Nevertheless, in our view, it seems reasonable to choose **s = 10**, which provides a decent sample number in case of severe spatial heterogeneity, high power in most cases, and marginal power in only two cases.

The BACI power estimates are compared to the BA estimates for TPE in **Table 10**. The results seem to be rather strange, with low power at alpha = 0.05 for all variables (except Cu) and dramatically higher power for some variables at alpha = 0.10. With only two years of before data and one year of after data, the test of the interaction effect (time period as before/after, and station type as control/impact) has only one degree of freedom, therefore the estimated effect must be very large in comparison to its standard error in order to achieve a low p-value.

Summary – Results of BA analysis show that subsampling to a moderate degree (e.g., 10) is worthwhile, given apparent spatial variability in sediment chemistry within lakes or basins. For sediment samples, the assumption that temporal variation does not matter is critical to the assessment of changes in variables by station. Without considerable replication across years, BACI models cannot be reliably applied. However, while the control data is not used in the BA analysis, the continued sampling of INUG is of high value to maintain the option of applying BACI models over time.

8. REFERENCES

CCME (Canadian Council of Ministers of the Environment) 2002. Canadian Sediment Quality Guidelines for the Protection of Freshwater Aquatic Life, 1999, updated 2002.

Gelman, A, and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models – Analytical Methods for Social Research. Cambridge University Press, New York, 625 pp.

Pinheiro, J., and D. Bates. 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York, 528 pp.

Smith, E. 2002. BACI design. Pages 141-148, in: A.H. El-Shaarawi and W.W. Piegorsch [Eds], Encyclopedia of Environmetrics, John Wiley & Sons, Chicherster.

Stewart-Oaten, A., W. Murdoch, and K. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67: 929-940.

Venables, W., and B. Ripley. 2002. Modern Applied Statistics. Springer, New York, 495 pp.

Underwood, A. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications 4(1): 3-15.

Table 1. Core and grab sediment samples (excluding duplicates) collected for the CREMP program (core samples for July 2009 are for the East Dike program).

Note: Shading denotes "Impact" periods.

CORE		Baker			Meadowbank									
Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul													0
2007	Aug													0
2008	Jul				15		15	15		15	15	15	15	105
	Aug	15	6	15										36
2009	Jul				15		15	15		15				60
	Aug													0
Total		**15**	**6**	**15**	**30**	**0**	**30**	**30**	**0**	**30**	**15**	**15**	**15**	**201**

GRAB		Baker			Meadowbank									
Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul				1		1	1		1	1	1	1	7
2007	Aug				1		1	1		1	1	1	1	7
2008	Jul													0
	Aug	1	1	1	1		1	1		1	1	1	1	10
2009	Jul													0
	Aug	3	3	3	3	3	3	3	3	3	3	3	3	36
Total		**4**	**4**	**4**	**6**	**3**	**6**	**6**	**3**	**6**	**6**	**6**	**6**	**60**

Table 2. Duplicate sediment samples collected for the AEMP program (core samples for July 2009 are for the East Dike program).

Note: Shading denotes "Impact" periods.

CORE		Baker			Meadowbank									
Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul													0
2007	Aug													0
2008	Jul				1		1	1		1	1	1	1	7
	Aug	1		1										2
2009	Jul				1		1	1		1				4
	Aug													0
Total		**1**	**0**	**1**	**2**	**0**	**2**	**2**	**0**	**2**	**1**	**1**	**1**	**13**

GRAB		Baker			Meadowbank									
Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul													0
2007	Aug										1			1
2008	Jul													0
	Aug										1			1
2009	Jul													0
	Aug					1						1		2
Total		**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**2**	**1**	**0**	**4**

Table 3. Summary of sample measurements (N) and values above detection limits (>DL) for duplicate samples (core/grab), Meadowbank core control samples, and Baker core control samples.

	Duplicates (core/grab)		Meadowbank (core control)		Baker (core control)	
Variable	N	>DL	N	>DL	N	>DL
Arsenic (T)	34	32	135	133	15	3
Cadmium (T)	34	13	135	64	15	0
Chromium (T)	34	34	135	135	15	15
Copper (T)	34	34	135	135	15	15
Lead (T)	34	1	135	11	15	0
Mercury (T)	34	33	135	135	15	13
Zinc (T)	34	34	135	135	15	15

Table 4. Summary of statistical comparisons between sediment core sample/duplicate pairs

Notes: Limited to pairs with both values above detection limits). N = total samples (pairs = N/2); Median = median of all samples; MAD = median absolute difference among sample/duplicate pairs; r = Spearman rank correlation among sample/duplicate pairs; SD = standard deviation of random effects for Location (variability among sample pairs) and residual errors (variability between samples and duplicates) fit to log-transformed data.

					Model SD (log data)	
Variable	N	Median	MAD	r	Location	Res. Error
Arsenic (T)	24	29.1	5.0	0.60	0.412	0.290
Cadmium (T)	8	0.85	0.18	0.80	0.245	0.155
Chromium (T)	26	72.9	3.5	0.97	0.563	0.075
Copper (T)	26	64.20	3.40	0.94	0.896	0.079
Mercury (T)	26	0.040	0.003	0.85	0.593	0.114
Zinc (T)	26	87.6	3.0	0.95	0.477	0.074

Table 5. Comparisons between core samples (controls) for Baker and Meadowbank programs.

Notes: Ratio = median(Baker)/median(Meadow); MW test P = P-value for the Mann-Whitney test; Est = difference in means (Baker – Meadow) for log-transformed data; P = P-value based on t-test; ES (%) = proportional effect size in untransformed units relative to Meadowbank samples, where ES (%) = (exp[Est] – 1)*100.

	Sample size (N)		Medians			MW	t-test (log data)		
Variable	Baker	Meadow	Baker	Meadow	Ratio	test P	Est	P	ES (%)
Arsenic (T)	15	135	5.0	26.5	0.19	<0.0001	-1.67	<0.0001	-81%
Chromium (T)	15	135	16.7	76.5	0.22	<0.0001	-1.48	<0.0001	-77%
Copper (T)	15	135	5.3	61.8	0.09	<0.0001	-2.46	<0.0001	-91%
Mercury (T)	15	135	0.006	0.038	0.17	<0.0001	-1.66	<0.0001	-81%
Zinc (T)	15	135	24.1	91.0	0.26	<0.0001	-1.29	<0.0001	-73%

Table 6. Summary of trigger values for sediment variables for Meadowbank and Baker stations.

Notes: For each variable, thresholds (guidelines) are shown (details for thresholds are in Table 7); DL = current detection limit; N = sample measurements; >DL = number of measurements above DL; medians and 95th percentiles for variables with at least one value >DL (with all values <DL set equal to DL); Method = method used to determine the trigger, where A = halfway from median (or DL if all values <DL) to threshold, and B = 95th percentile.

			Meadowbank (or both for Lead)						Baker					
Variable	Threshold	DL	N	>DL	Med	95th	Trigger	Method	N	>DL	Med	95th	Trigger	Method
Arsenic (As)	5.9	5.0	135	133	26.5	120.0	120.0	B	15	3	5.0	8.3	8.3	B
Cadmium (Cd)	0.6	0.50	135	64	0.50	1.10	1.10	B	15	0			0.55	A
Chromium (Cr)	37.3		135	135	76.5	114.3	114.3	B	15	15	16.7	20.3	27.0	A
Copper (Cu)	35.7		135	135	61.8	126.0	126.0	B	15	15	5.3	7.8	20.5	A
Lead (Pb)	35	30	150	11	30	32	32.5	A						
Mercury (Hg)	0.17	0.005	135	135	0.038	0.064	0.104	A	15	13	0.006	0.011	0.088	A
Zinc (Zn)	123		135	135	91.0	121.3	121.3	B	15	15	24.1	32.1	73.6	A

Table 7. Summary of CCME guidelines for sediment variables; (CCME, 2002).

Variable	Source	Description of guidelines
t-Arsenic (As)	CCME, 2002	The CCME ISQG for As is 5.9 mg/kg.
t-Cadmium (Cd)	CCME, 2002	The CCME ISQG for Cd is 0.6 mg/kg.
t- Chromium (Cr)	CCME, 2002	The CCME ISQG for Cr is 37.3 mg/kg.
t-Copper (Cu)	CCME, 2002	The CCME ISQG for Cu is 35.7 mg/kg
t-Lead (Pb)	CCME, 2002	The CCME ISQG for Pb is 35 mg/kg.
t-Mercury (Hg)	CCME, 2002	The CCME ISQG for Hg is 0.17 mg/kg.
t-Zinc (Zn)	CCME, 2002	The CCME ISQG for Zn is 123 mg/kg.

Table 8. Summary of "before-period" estimates by variable and impact station.

Notes: The station mean, trigger, and effect size (ES = Trigger – Mean) are in raw units (e.g., mg/kg). ES(log) denotes the effect size for log data (ES(log) = log[Trigger/Mean]). Estimates of the standard deviation (SD) for observations are for log-transformed data.

Variable	Station	Mean	Trigger	ES	ES(log)	SD
Arsenic	SP	31.3	120.0	88.7	1.345	1.012
	TPE	17.9	120.0	102.1	1.905	0.912
	TPN	27.3	120.0	92.7	1.480	1.046
Chromium	SP	64.9	114.3	49.4	0.566	0.162
	TPE	81.7	114.3	32.6	0.336	0.154
	TPN	98.4	114.3	15.9	0.150	0.222
Copper	SP	82.6	126.0	43.4	0.422	0.151
	TPE	57.6	126.0	68.4	0.782	0.190
	TPN	57.0	126.0	69.0	0.794	0.255
Mercury	SP	0.045	0.104	0.059	0.843	0.253
	TPE	0.029	0.104	0.075	1.271	0.302
	TPN	0.026	0.104	0.078	1.388	0.335
Zinc	SP	97.0	121.3	24.3	0.223	0.151
	TPE	107.3	121.3	14.0	0.123	0.161
	TPN	75.2	121.3	46.1	0.478	0.205

Table 9. Estimates of BA statistical power for detecting a significant increase (one-tailed test) in a given variable and impact year by station (SP, TPE, TPN) as a function of the number of sub-samples for the impact year.

Note: Power is shown for two levels of alpha (0.05 and 0.10).

		alpha = 0.05			alpha = 0.10		
Variable	Samples	SP	TPE	TPN	SP	TPE	TPN
Arsenic	5	1.00	1.00	1.00	1.00	1.00	1.00
	10	1.00	1.00	1.00	1.00	1.00	1.00
	15	1.00	1.00	1.00	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00
	25	1.00	1.00	1.00	1.00	1.00	1.00
Chromium	5	1.00	1.00	0.18	1.00	1.00	0.33
	10	1.00	1.00	0.29	1.00	1.00	0.45
	15	1.00	1.00	0.34	1.00	1.00	0.56
	20	1.00	1.00	0.40	1.00	1.00	0.61
	25	1.00	1.00	0.41	1.00	1.00	0.65
Copper	5	1.00	1.00	1.00	1.00	1.00	1.00
	10	1.00	1.00	1.00	1.00	1.00	1.00
	15	1.00	1.00	1.00	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00
	25	1.00	1.00	1.00	1.00	1.00	1.00
Mercury	5	1.00	1.00	1.00	1.00	1.00	1.00
	10	1.00	1.00	1.00	1.00	1.00	1.00
	15	1.00	1.00	1.00	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00	1.00	1.00
	25	1.00	1.00	1.00	1.00	1.00	1.00
Zinc	5	0.99	0.44	0.98	1.00	0.61	1.00
	10	1.00	0.67	1.00	1.00	0.84	1.00
	15	1.00	0.83	1.00	1.00	0.92	1.00
	20	1.00	0.90	1.00	1.00	0.95	1.00
	25	1.00	0.93	1.00	1.00	0.98	1.00

Table 10. Estimates of power for the BA versus BACI design for station TPE (with INUG as control) for detecting a significant increase (one-tailed test) in a given variable and impact year as a function of the number of sub-samples for the impact year.

Note: Power is shown for two levels of alpha (0.05 and 0.10).

		alpha = 0.05		alpha = 0.10	
Variable	Samples	BA	BACI	BA	BACI
Arsenic	5	1.00	0.00	1.00	0.26
	10	1.00	0.00	1.00	0.41
	15	1.00	0.00	1.00	0.53
	20	1.00	0.00	1.00	0.64
	25	1.00	0.00	1.00	0.66
Chromium	5	1.00	0.00	1.00	0.29
	10	1.00	0.00	1.00	0.45
	15	1.00	0.00	1.00	0.55
	20	1.00	0.00	1.00	0.67
	25	1.00	0.00	1.00	0.68
Copper	5	1.00	0.40	1.00	1.00
	10	1.00	0.97	1.00	1.00
	15	1.00	1.00	1.00	1.00
	20	1.00	1.00	1.00	1.00
	25	1.00	1.00	1.00	1.00
Mercury	5	1.00	0.01	1.00	0.96
	10	1.00	0.04	1.00	0.99
	15	1.00	0.08	1.00	0.99
	20	1.00	0.11	1.00	1.00
	25	1.00	0.11	1.00	1.00
Zinc	5	0.44	0.00	0.61	0.03
	10	0.67	0.00	0.84	0.03
	15	0.83	0.00	0.92	0.04
	20	0.90	0.00	0.95	0.08
	25	0.93	0.00	0.98	0.06

Figure 1. Sample versus duplicate values of selected sediment variables.

Notes: Values below detection limits were set equal to the given detection limit. Solid line is the 1:1 line.

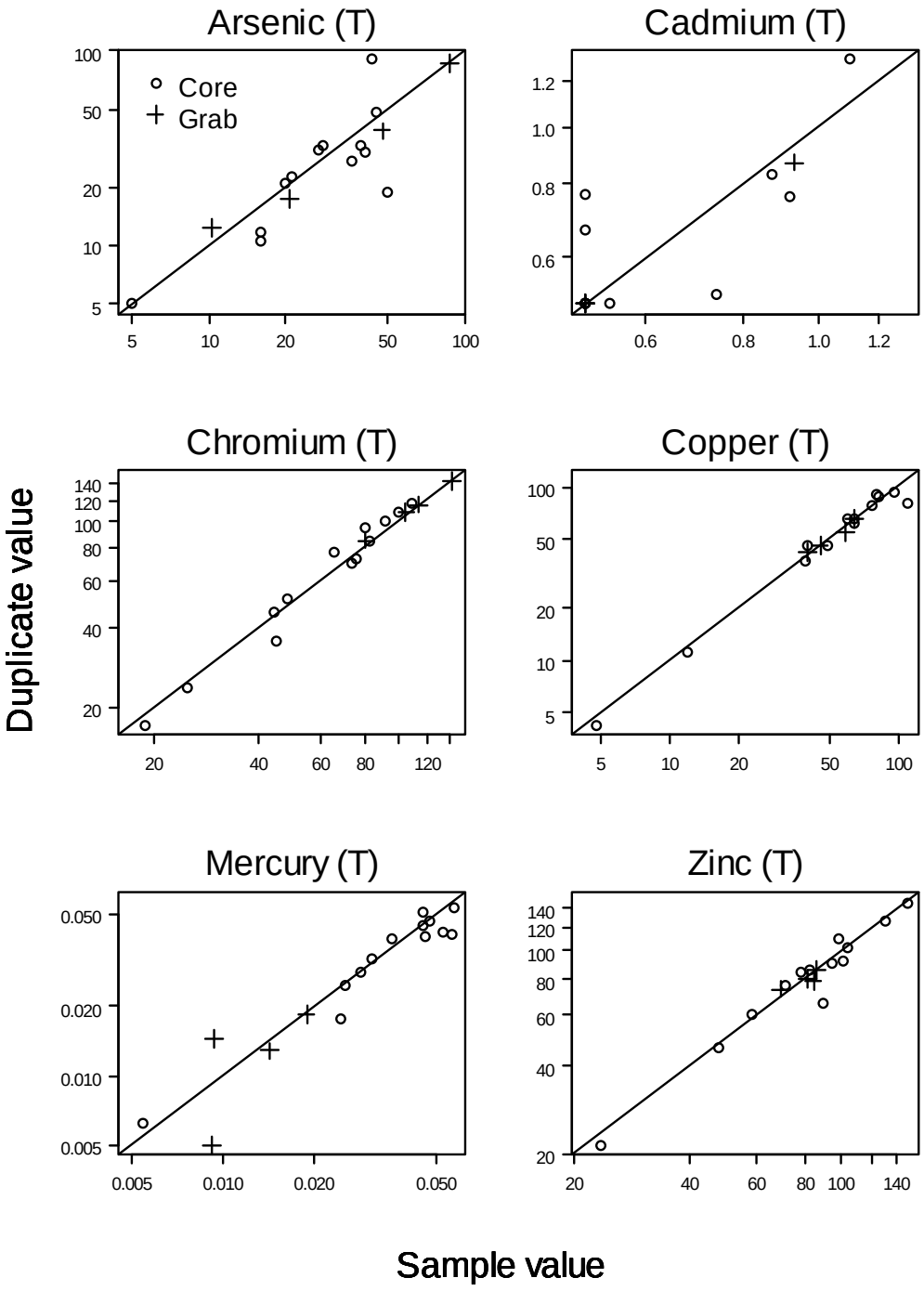


Figure 2. Box-plots of core samples by station for 2008 (top) and 2009 (bottom), with values for grab samples are superimposed (red diamonds).

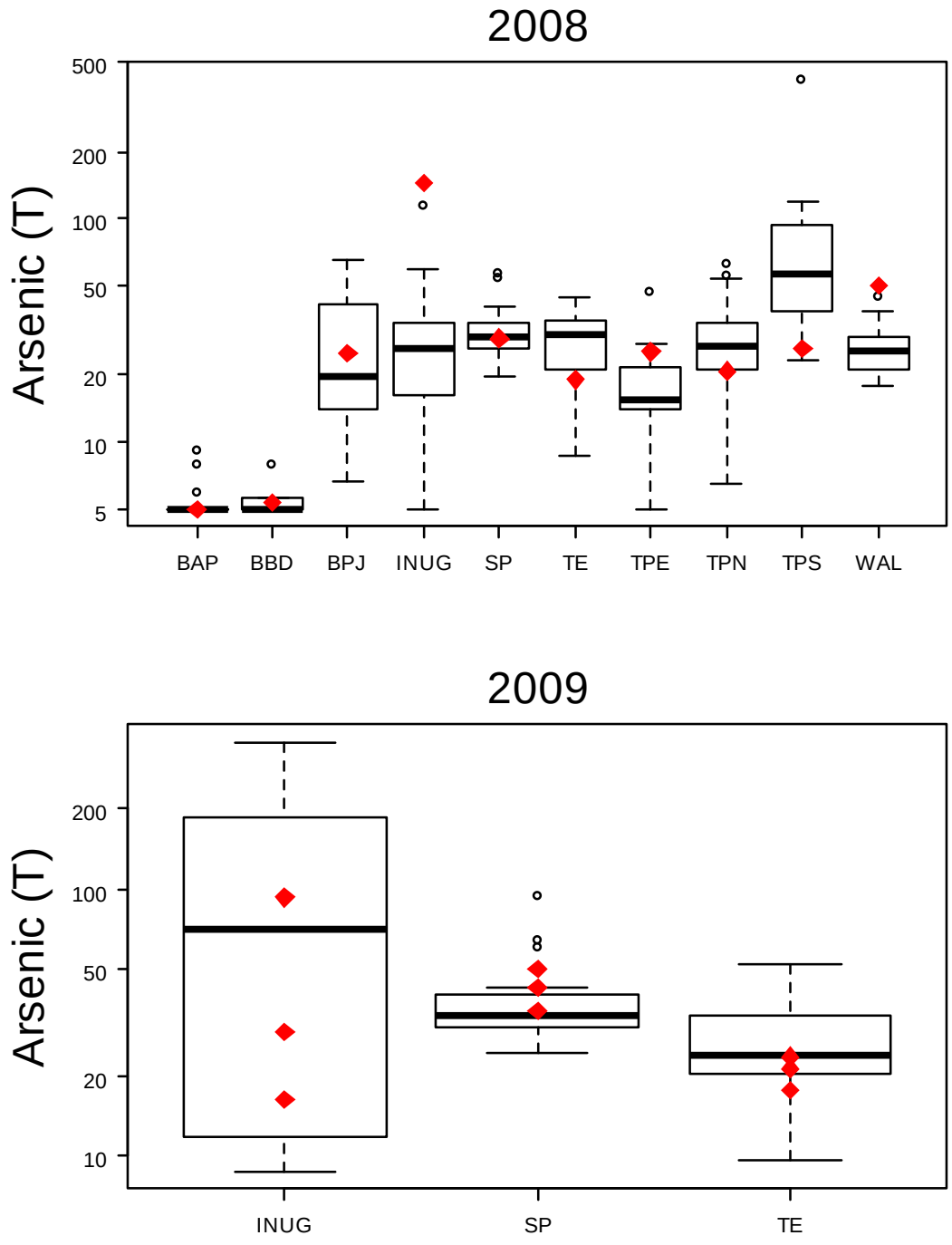
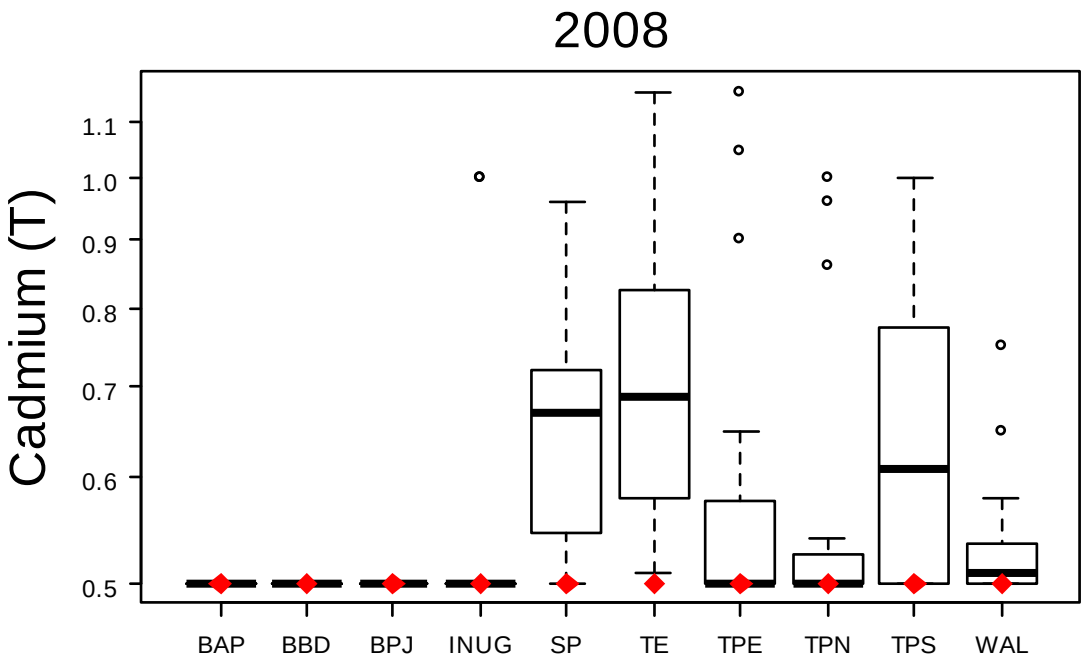


Figure 2 (page 2 of 6)



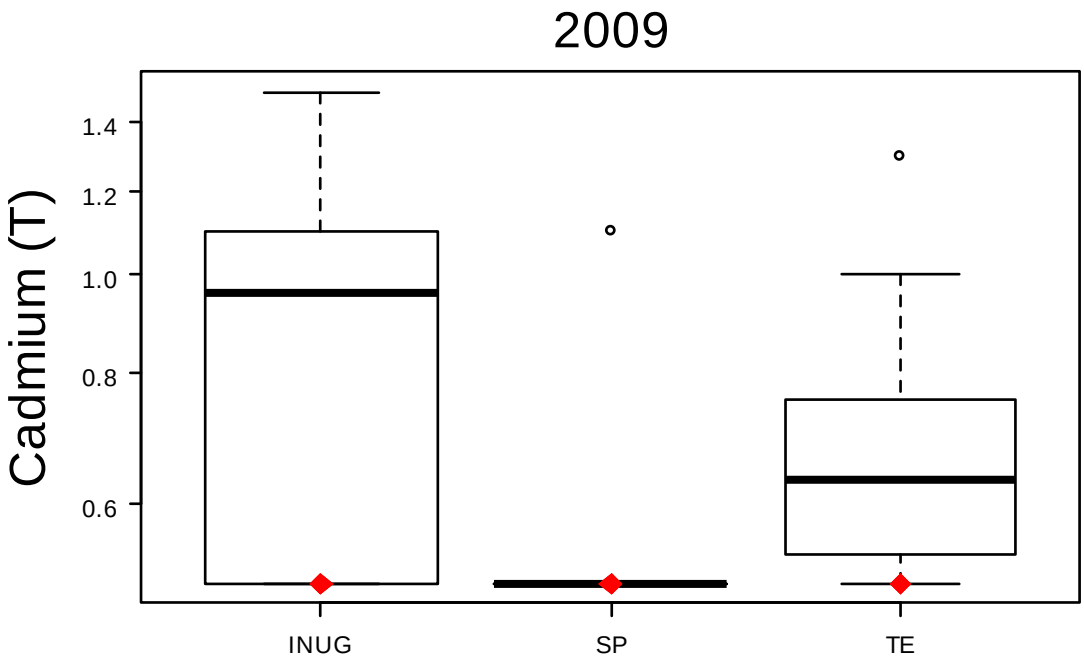
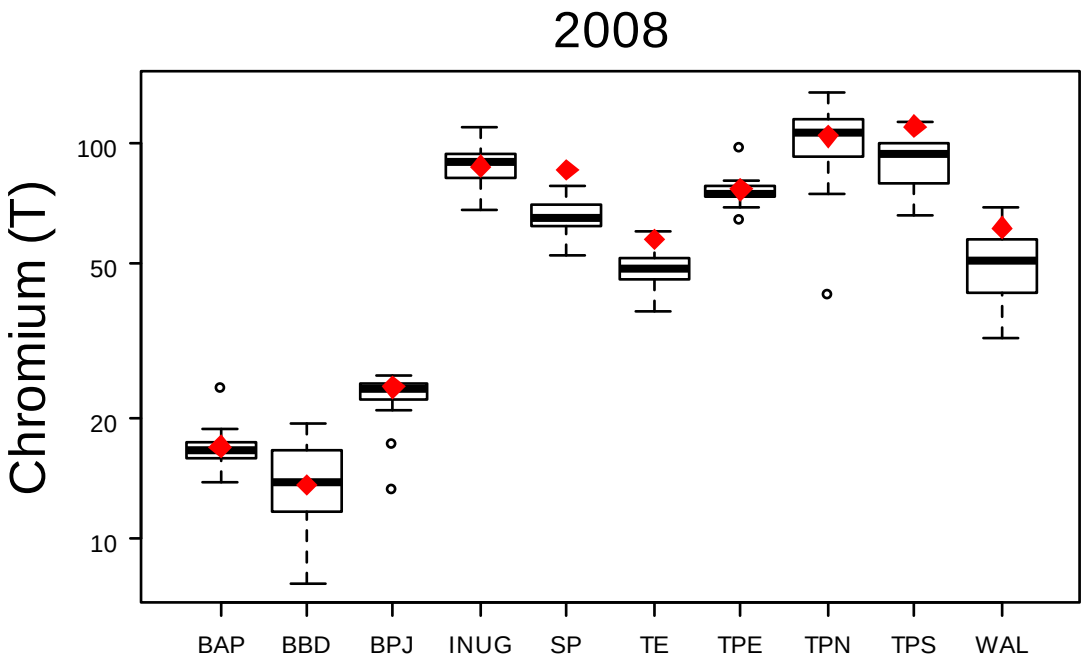


Figure 2 (page 3 of 6)



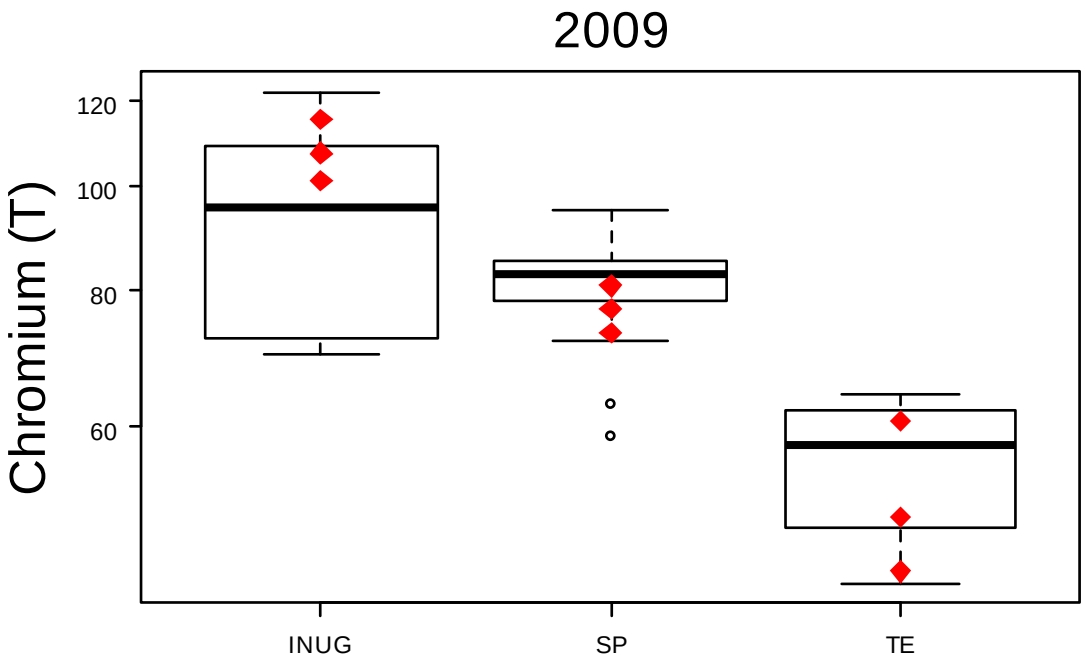
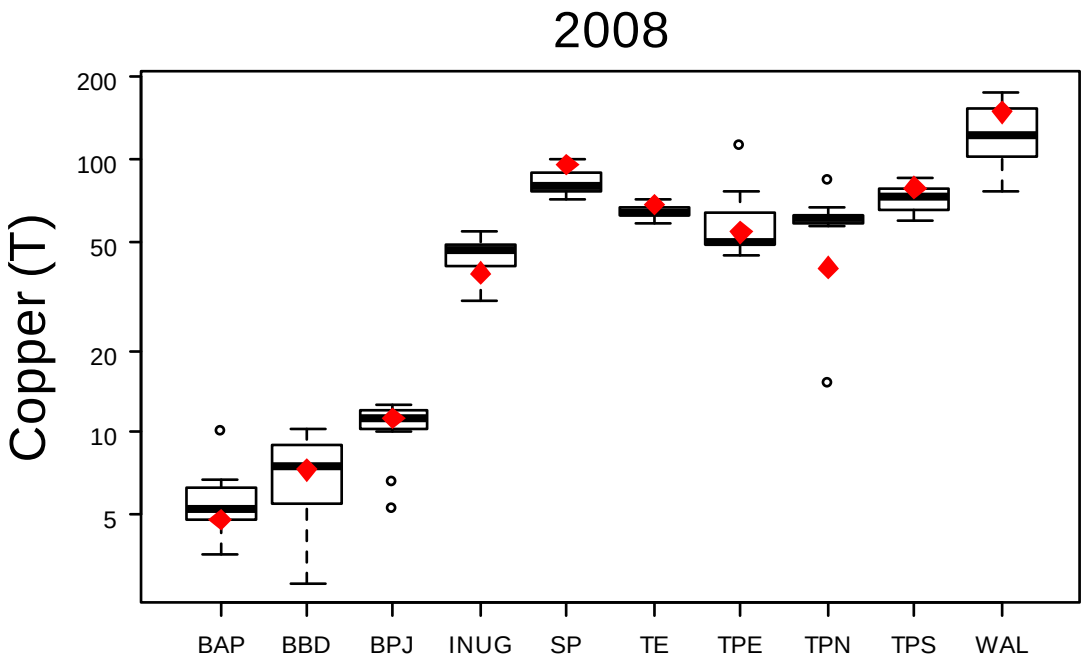


Figure 2 (page 4 of 6)



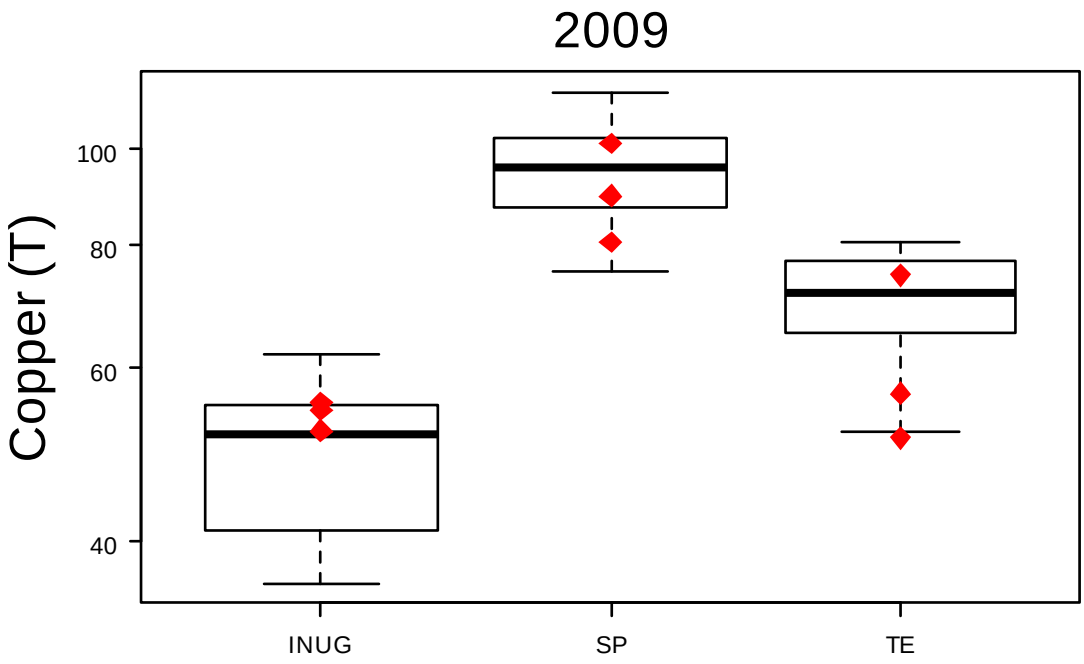
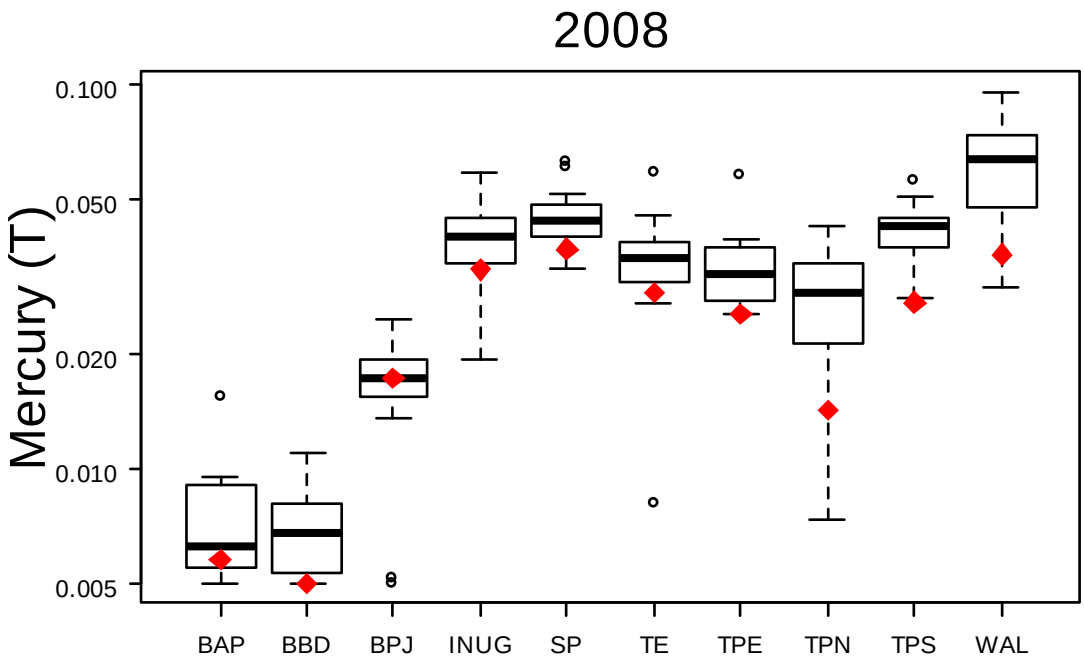


Figure 2 (page 5 of 6)



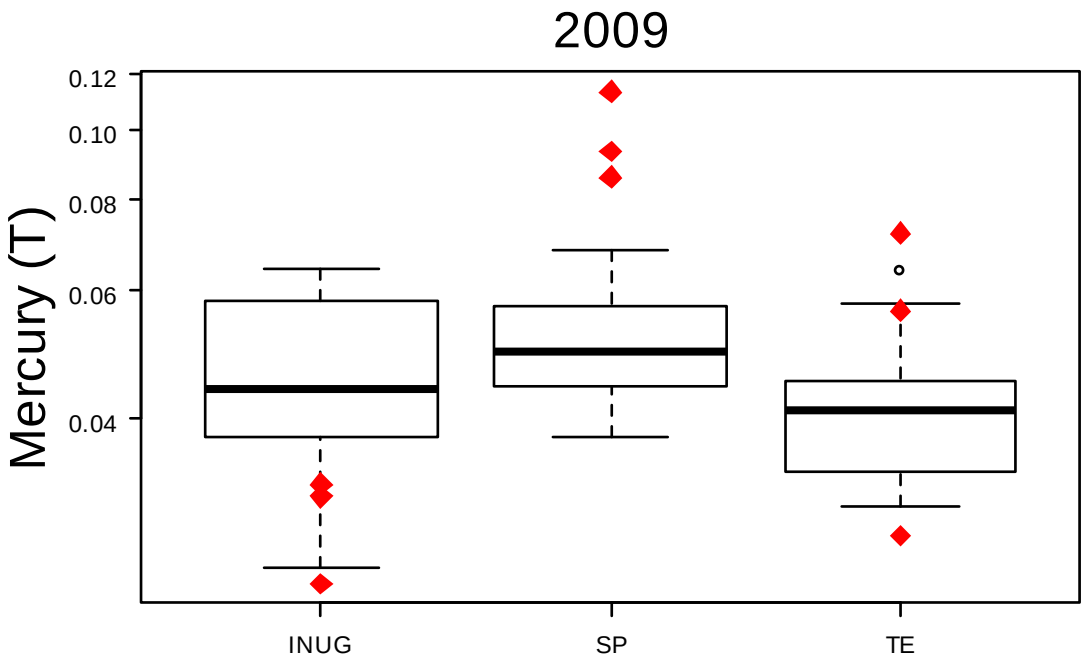
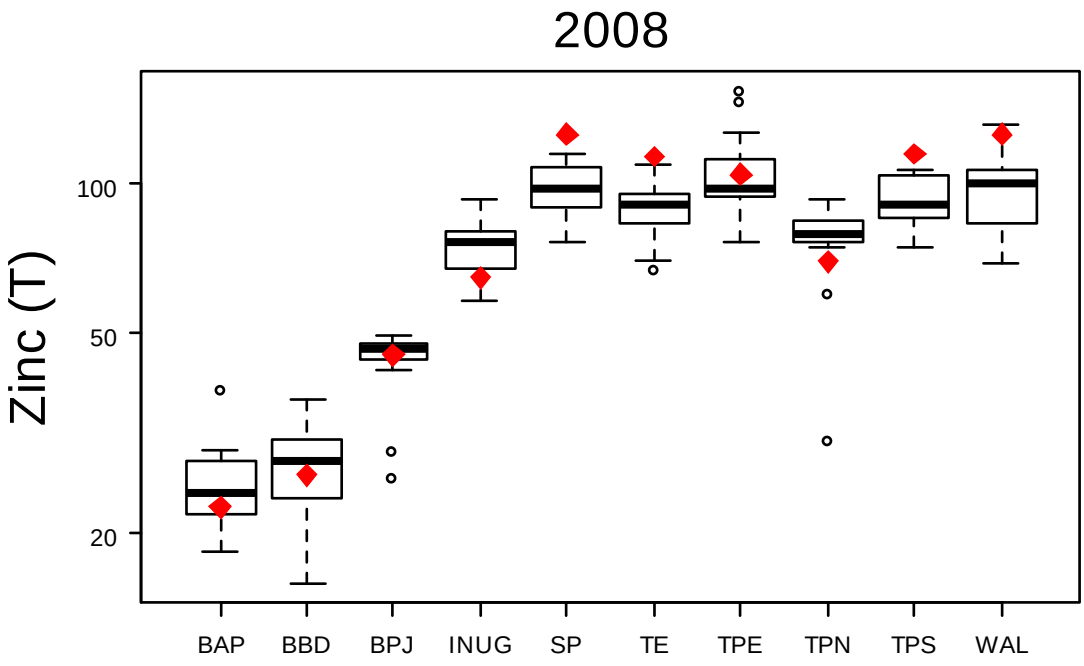


Figure 2 (page 6 of 6)



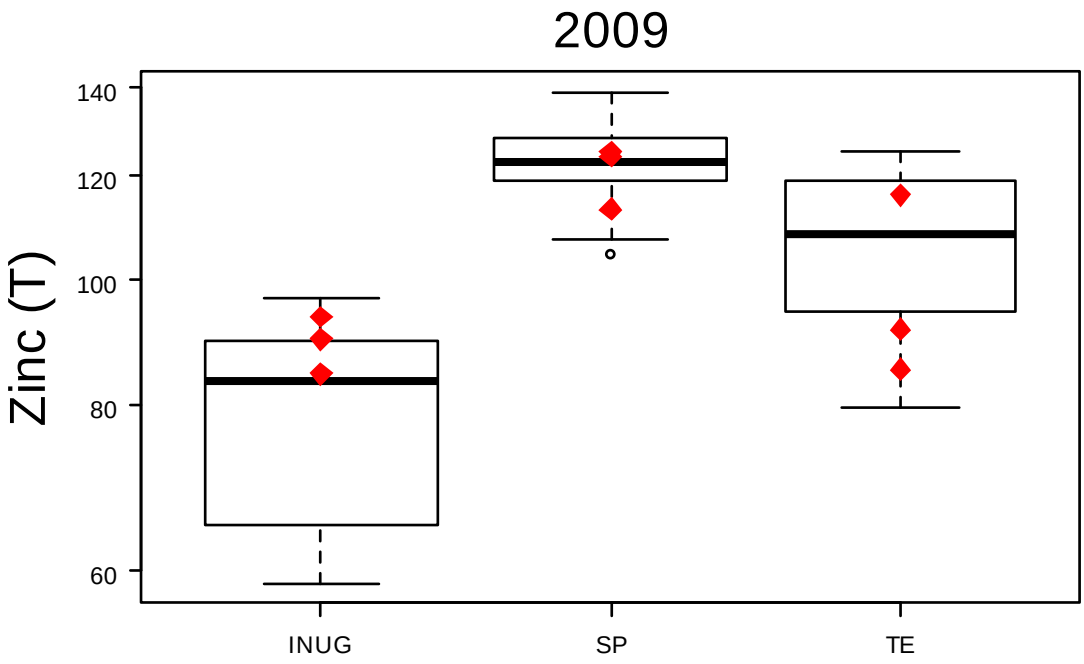


Figure 3. Box -plots of sample measurements (core control samples) by project lake for selected sediment variables (log-log scale).

Note: values < DL were set = DL.

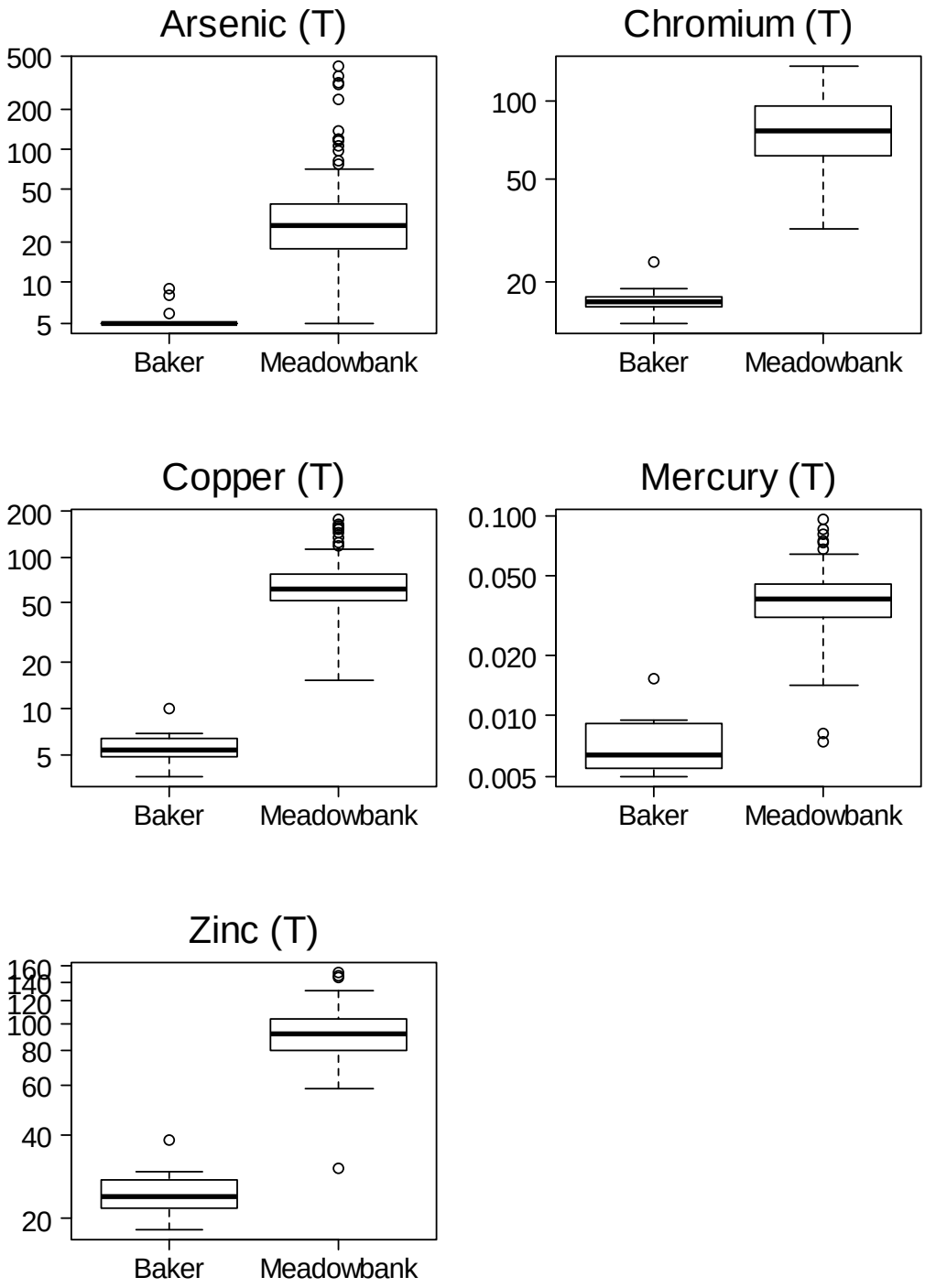


Figure 4. Box-plots of sample measurements by station for Baker Lake.

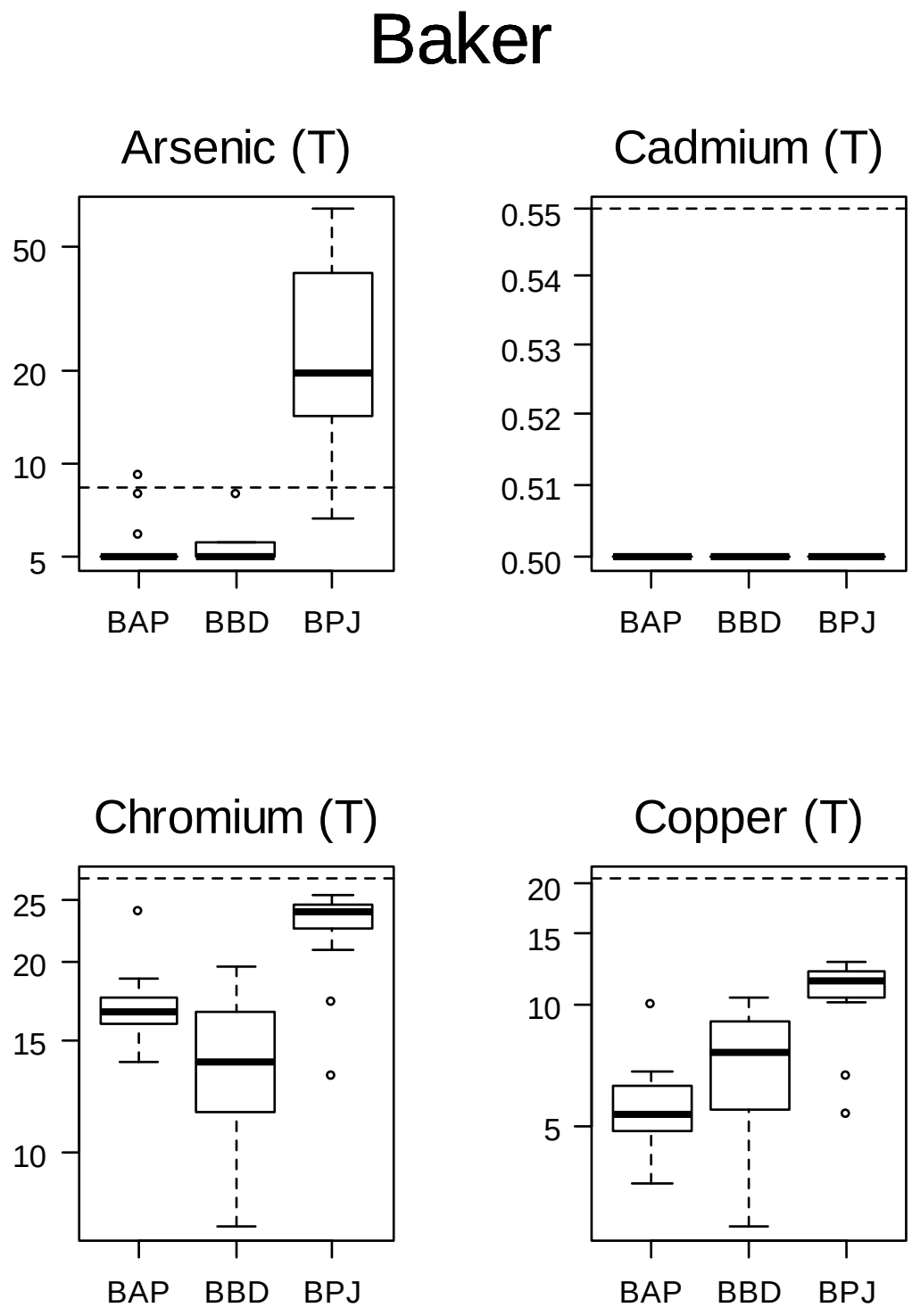


Figure 4 (page 2 of 2)

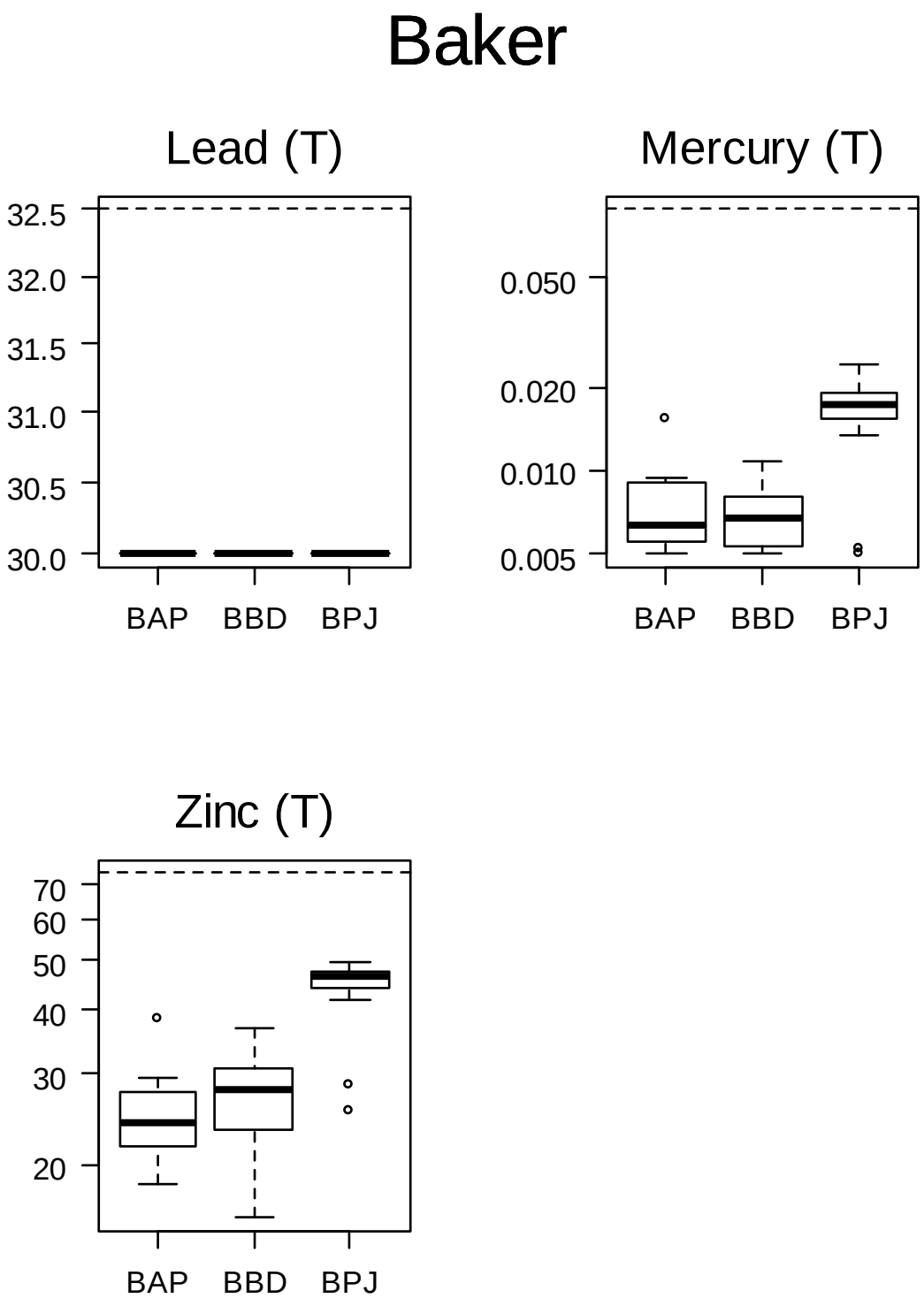


Figure 5. Box-plots of sample measurements by station for Meadowbank stations (control data only).

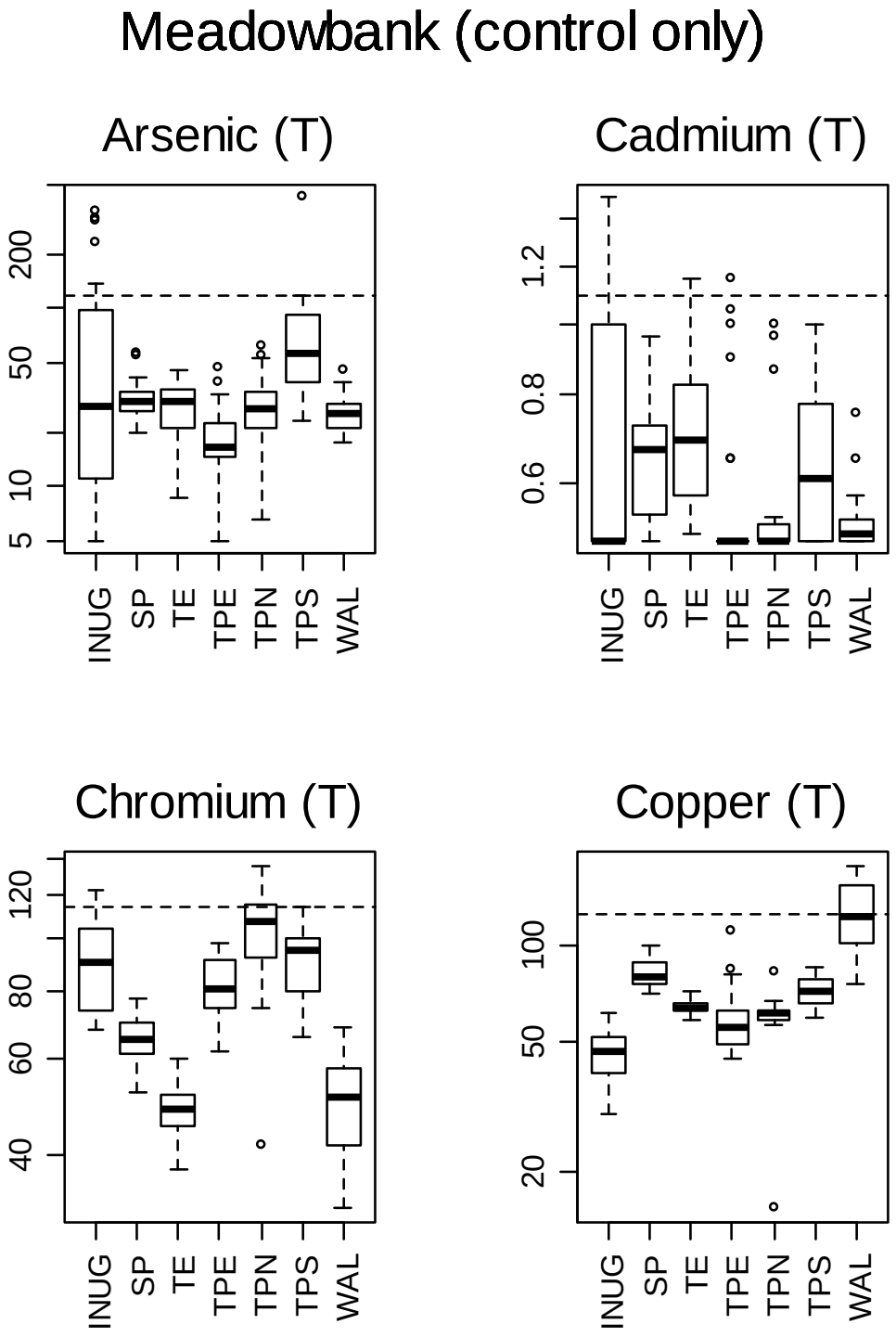


Figure 5 (page 2 of 2)

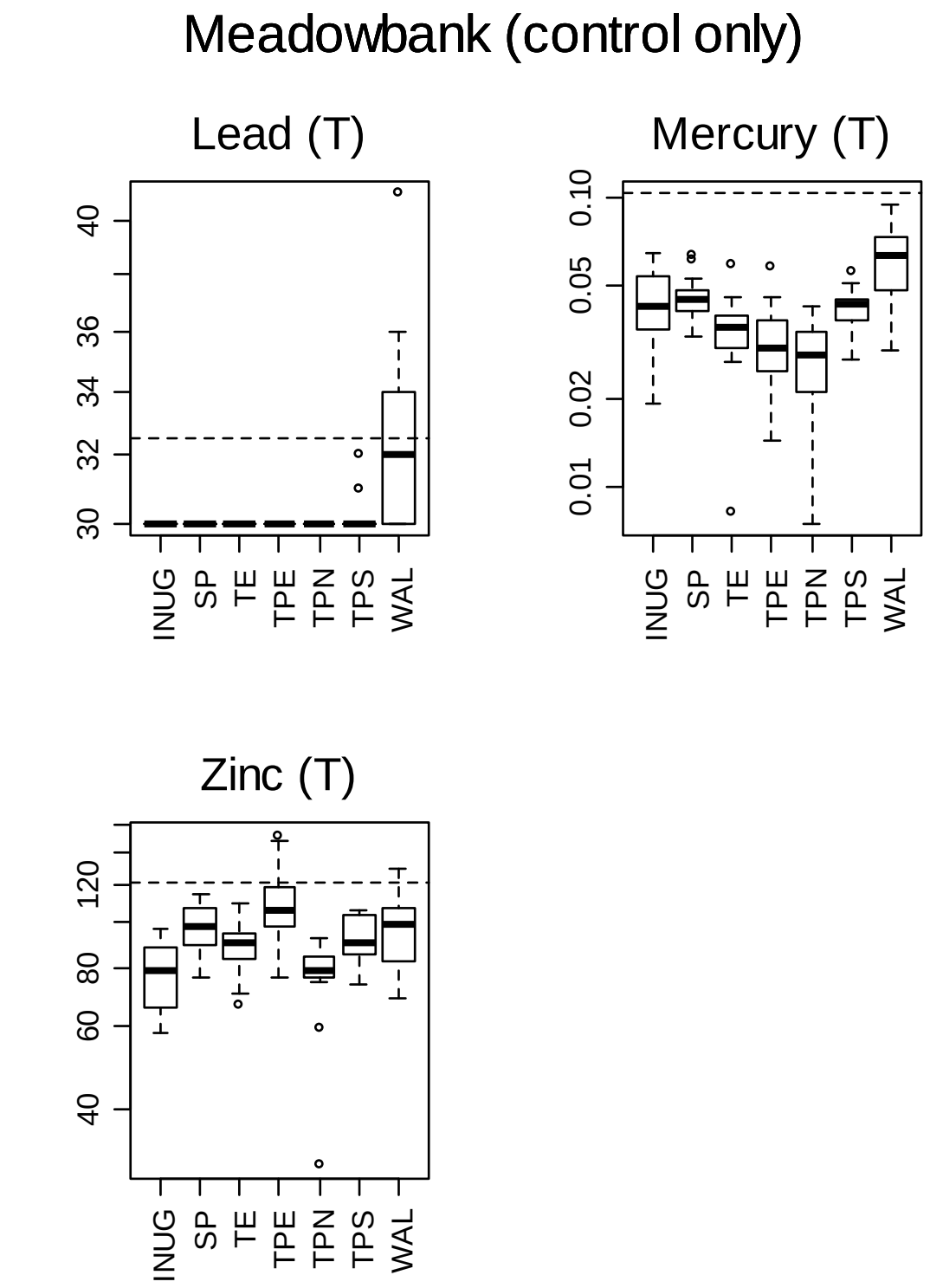


Figure 6. Box-plots of sample measurements by year for select Meadowbank stations.

Note: Pb excluded.

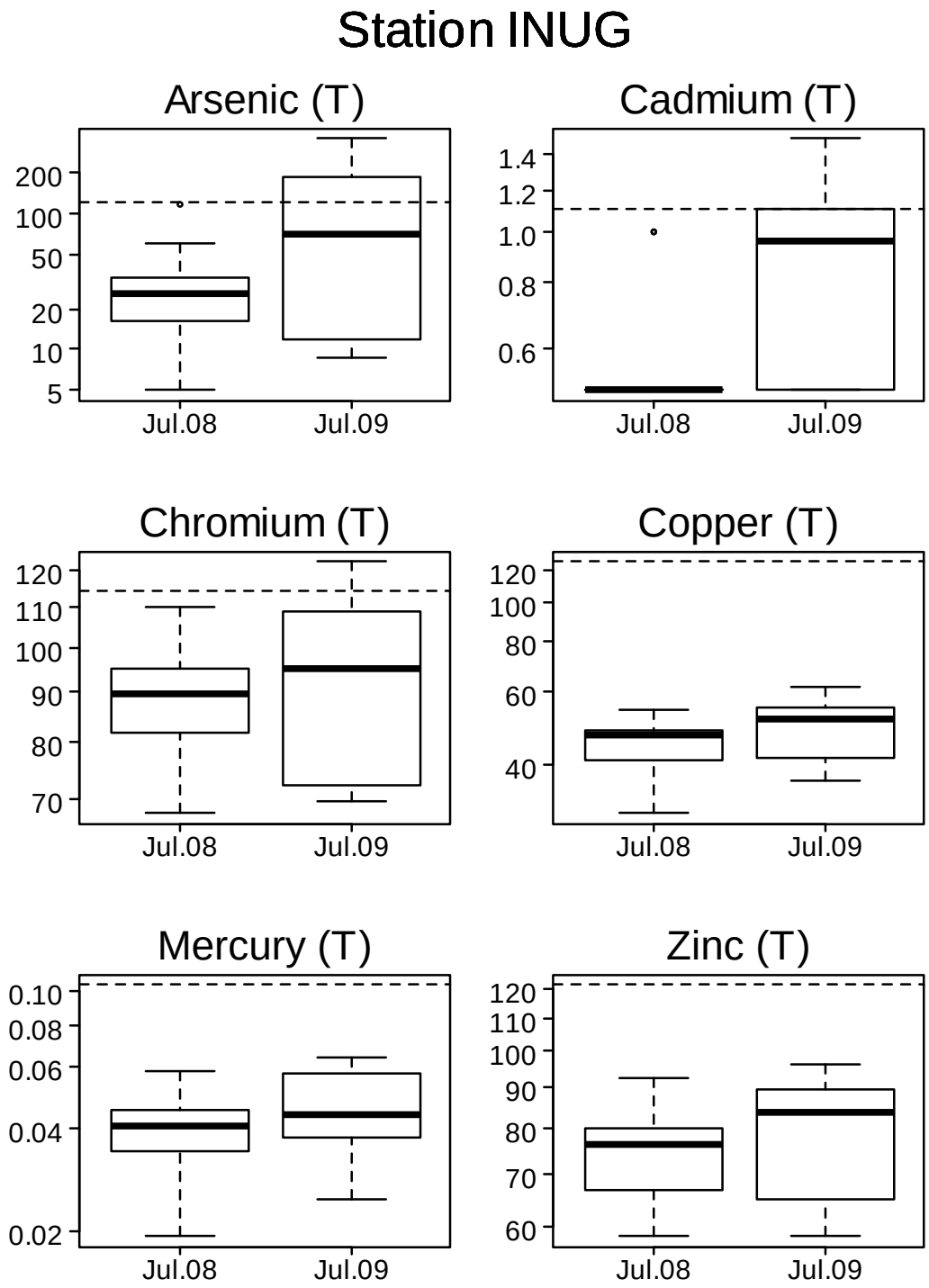


Figure 6 (page 2 of 4)

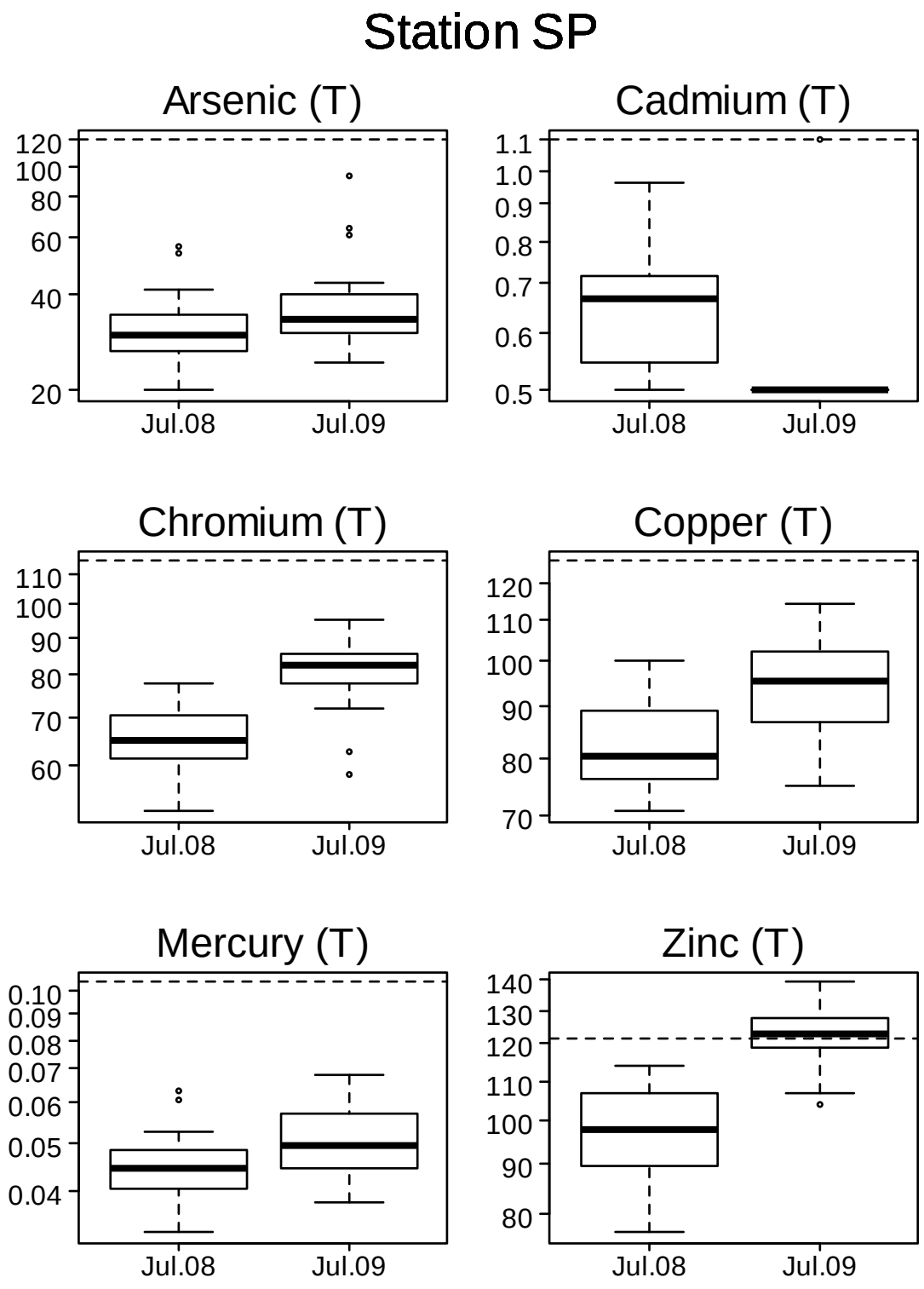


Figure 6 (page 3 of 4)

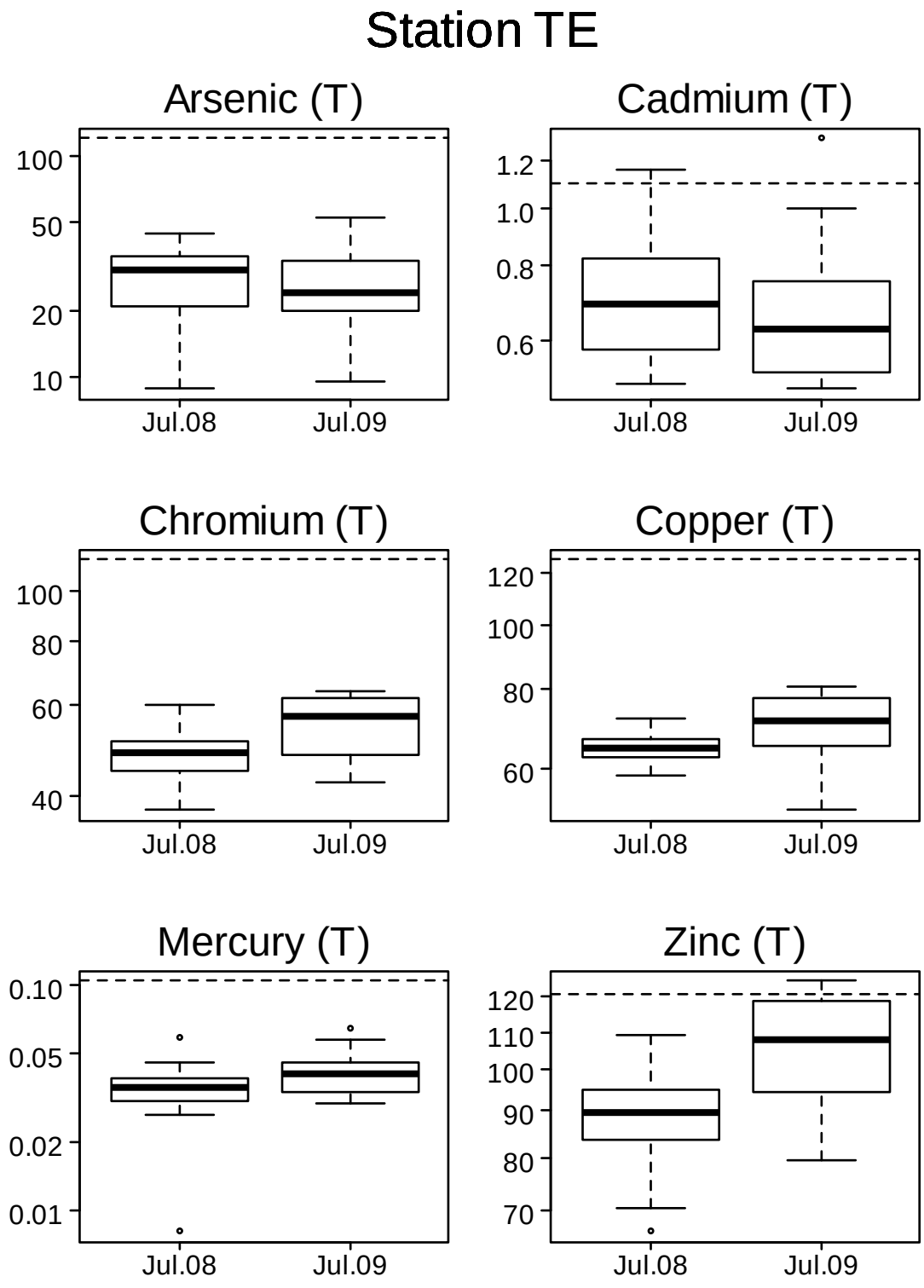


Figure 6 (page 4 of 4)

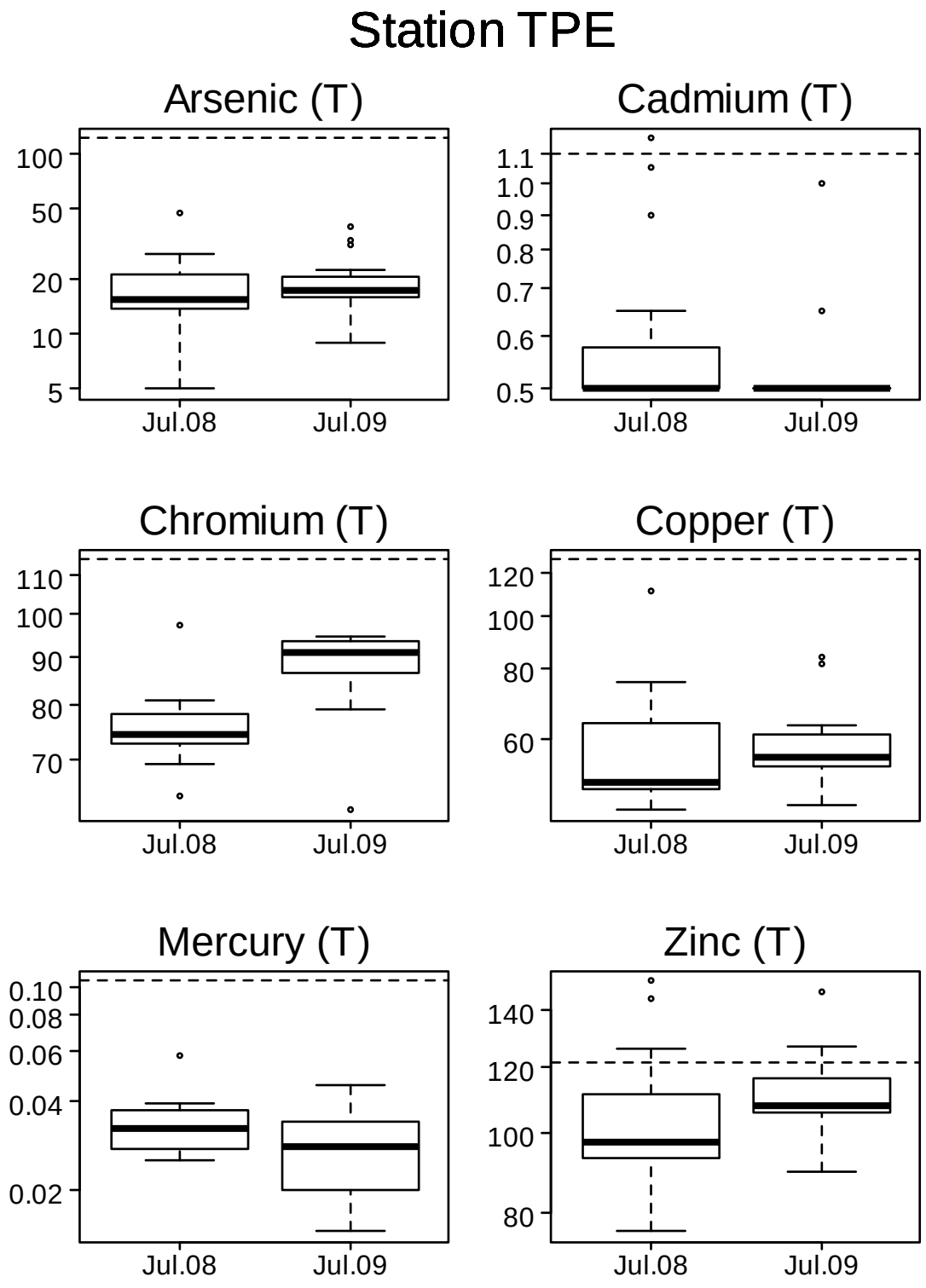


Figure 7. Estimates of BA statistical power for detecting a significant increase (one-tailed test, alpha = 0.05) in a given variable (rows) by station (columns) as a function of the number of sub-samples in the impact year (after period).

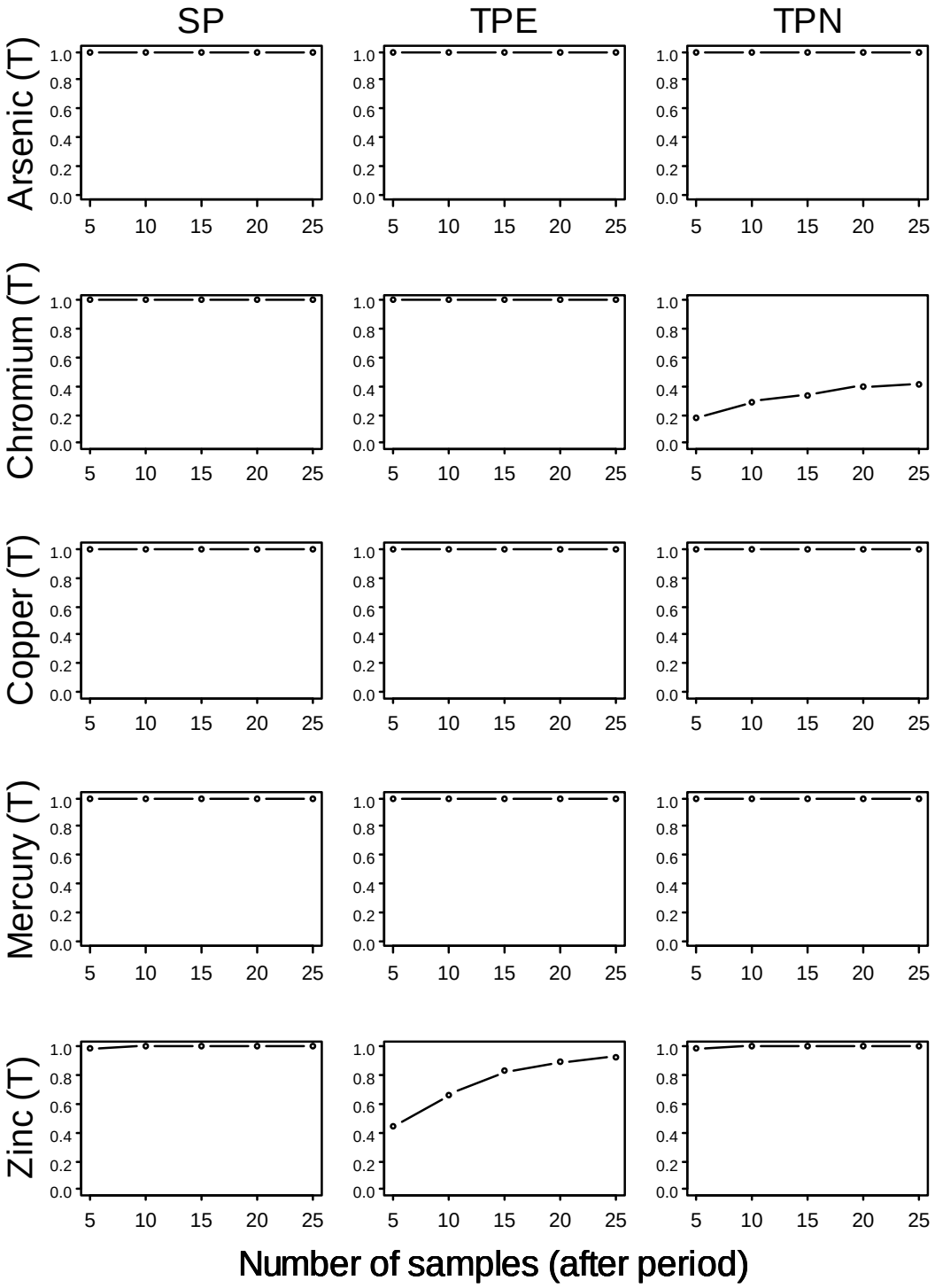
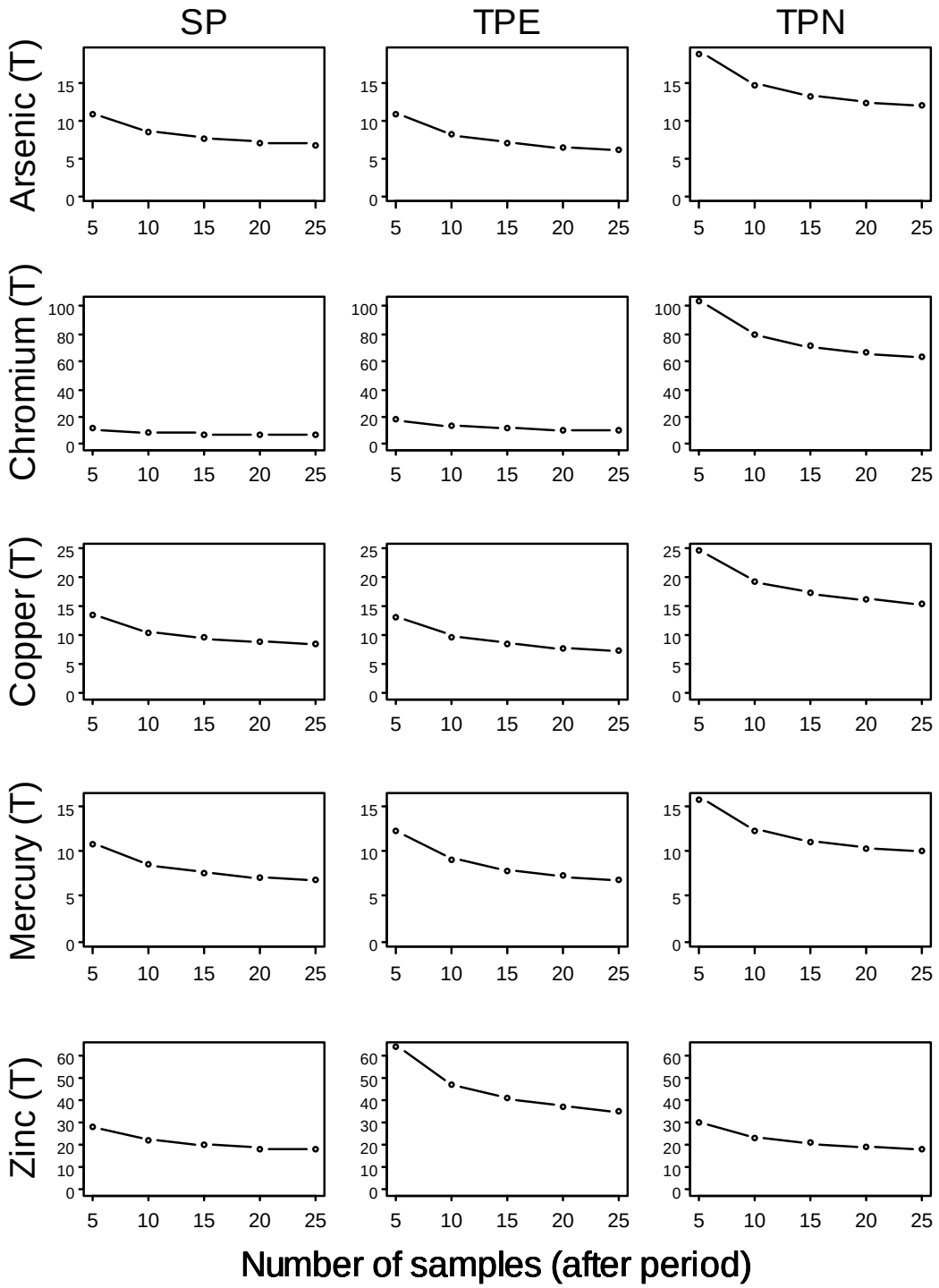


Figure 8. Coefficients of variation (%) for BA estimates by variable (row) and station (column) as a function of the number of sub-samples in the impact year (after period).



APPENDIX C – STATISTICAL ANALYSES FOR PHYTOPLANKTON

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This appendix contains the following analyses.

- Summary of CREMP phytoplankton samples
- Analysis of field duplicates
- Analysis of depth
- Data evaluation
- Analysis of sampling design

Some text, tables and figures are repeated from the main document so that the analyses contained in this appendix can be read and understood without reference to the main document. The material in this appendix assumes understanding of basic statistical methods (Venables and Ripley 2002), mixed-effects models (Pinheiro and Bates 2000), before-after-control-impact (BACI) experimental design (Stewart-Oaten et al. 1986; Underwood 1994; Smith 2002), and use of simulation in statistical analysis (Gelman and Hill 2006).

2. SUMMARY OF CREMP PHYTOPLANKTON SAMPLES

Phytoplankton data include estimates of biomass by taxonomic group (in mg/m3), total biomass, species richness (i.e., a count of the number of unique species in a sample) and an estimate of diversity (Simpson's diversity). For the analysis here, matching chlorophyll-a data were also compiled from available water chemistry data (i.e., chlorophyll-a data collected at the same location and time as phytoplankton), and were used for some analyses.

There are three basic types of phytoplankton samples:

1. Standard samples – specific to a given location, month (a single sampling day), and depth.
2. Field duplicate samples – occasionally collected; these consisted of a second sample specific to a given location, time, and depth.
3. Depth replicates – collected at a given location and time (typically one surface, one bottom, and one integrated sample).

For many station-month combinations, more than one "standard" sample was collected. These "sub-samples" represent spatial replication (a different location within the designated station). In addition, samples were designated as "control" or "impact" depending on operation activities near a given station (see main text for differentiation of control and impact).

The field duplicate samples and depth replicates are evaluated in subsequent sections of this appendix. Not including the duplicates and depth replicates, there have been 308 samples collected for the CREMP including samples collected in the summer of 2010 (**Table 1**). Of these, 161 samples were designated as 'control' and 147 as 'impact' (shaded cells).

3. ANALYSIS OF FIELD DUPLICATES

Field duplicates are samples collected at the same location, time and depth as a standard sample. Presumably, differences among duplicates largely reflect field sampling error and lab precision (measurement error). This variation will be a component of the variation observed among station/month sub-samples. Thus, comparing the two provides insight into the extent of real sub-sample variation (due to differing locations within a station).

Data – Duplicates have been collected for 37 samples (14 control and 23 impact). Analysis excluded Euglenophytes because almost all counts of that group are zero.

Methods – For each variable, xy plots were produced and we computed the median across all samples, the median absolute difference (MAD) between samples and duplicates, and Spearman rank correlation between samples and duplicates. In addition, we fit a mixed-effects model (Pinheiro and Bates 2000) of the form:

$$\text{Log}(y) \sim \text{Location} + \text{Error}$$

where y is the water quality variable, Location denotes random effects for sample pairs (variability among pairs due to sample location/month) and Error denotes the residual error (variability between samples and duplicates). For all plots and analyses, if a variable had zero counts, then 0.1 was added to all values (this applied to a few biomass variables – see **Table 5**).

Results – Results of the analysis of duplicates are shown in **Figure 1** (xy plots) and **Table 2** (medians, MADs, and Spearman r). In general, there was more variability between samples and duplicates than was observed for water chemistry variables. The standard deviations (SDs) resulting from the mixed-effect model show the relative variability due to sample location/month (SD.Location) versus duplicate variability (SD.Error). In most cases SD.Error is smaller than SD.Location, which just reaffirms that duplicate measures are generally precise compared to general sample variation (i.e., minimal measurement error or fine-scale spatial variation). These results suggests that sample/duplicate measures are reasonably consistent for key measures such as Total biomass and species counts, and should thus provide useful data for tracking and comparing phytoplankton conditions across stations and months.

4. ANALYSIS OF DEPTH

Evaluation of the importance of depth is important to determine if phytoplankton samples collected at different depths are comparable, and if not, to determine whether sampling at different depths is warranted in future. The primary null hypothesis is that a given variable is the same across sample types for depth (i.e., surface, bottom, integrated); however, depth differences may be related to station and especially season. Potential stratification of the water column is expected to be much greater during "winter" months (November through May) than for "summer" months (July, August, September), though July may be more similar to winter conditions.

Data – There were 41 cases or "depth tests" where phytoplankton samples were simultaneously collected at differing depths. For one cases, no "bottom" sample was collected, therefore for consistency of comparisons this case was removed, providing 40 depth-test locations with total samples size of 120. These 40 cases can be summarized as follows:

	INUG	TE	TPS	Total
Control	12		9	21
Impact		19		19
Total	12	19	9	40

	INUG	TE	TPS	Total
Jul.09	2	2		4
Aug.09	1	1		2
Sep.09	2	2		4
Nov.09	2	2		4
Dec.09		1	1	2
Jan.10			2	2
Feb.10		1	1	2
Mar.10		2	2	4
Apr.10		1	1	2
May.10		2	2	4
Jul.10	2	2		4
Aug.10	1	1		2
Sep.10	2	2		4
Total	12	19	9	40

Methods – Analysis included all variables except Euglenophytes (that group had mostly zero counts). For each variable, xy plots were produced. In addition, we fit three mixed-effects models (Pinheiro and Bates 2000) of the following forms:

Model 1: Log(y) ~ Type + Location + Error

Model 2: Log(y) ~ Type + Station + Type*Station + Location + Error

Model 3: Log(y) ~ Type + Season + Type*Season + Location + Error

where "Type" (with fixed effects) is the key factor variable of interest denoting depth types (surface, bottom, integrated), "Station" is the CREMP sampling station (e.g., TPE) that is actually a broad area, and "Location" denotes unique sample locations (with random effects) that were sampled within a given station at a given time. In Models 2 and 3, the terms of interest are the interactions (fixed effects) between Type and Station (i.e., potential station-specific effects of depth) and Type and Season (potential season-specific effects of depth). Two options were examined for season models. First," Season" was defined as a factor variable with two levels separating Summer months (July, August, September) and Winter months (all others). In the second analysis, July was considered a Winter month. For Model 3 (Season) analysis, data for station TPS were excluded because no "summer" months were sampled.

Results – The statistical significance for the key depth terms of models 1 to 3 is presented in **Table 3** (p-values, based on F tests). There were significant differences ($P < 0.05$) among depth samples (Type) for only 2 of the 10 variables (Chrysophyte and Diatom biomass), only one significant interaction between depth type and station (Diatom biomass), and one significant interaction for season (Simpson's diversity). **Table 4** shows depth type coefficients for Model 1. The coefficients were defined as "Bottom – Surface" and "Integrated – Surface" to make explicit comparisons with surface samples (the standard sample type). Since all data were log-transformed, the coefficient (X) has to be translated into a proportional effect size (ES) relative to (untransformed) surface samples[1]. The percentage ES for each variable is also shown in **Table 4**. For example, for Diatom biomass, the estimated ES = 49.5% for bottom samples (i.e., biomass in bottom samples was 49.5% higher than that observed in surface samples, on average). In general, there was no consistent pattern across variables. **Figure 2** shows bottom v. surface and integrated v. surface measures for each variable. Also shown are the implied differences (ES in **Table 4**) for which $P < 0.1$ (dashed lines, which in log-log scale are parallel lines to the 1:1 line).

The key question of interest is how large are the significant depth differences compared to variation among locations and among depth replicates (sample types). As shown in **Figure 2**, differences between solid and dashed lines are minimal compared to the full data ranges. Even for Diatoms, where there is a significant 50% increase for bottom samples, it implies a change of 2 to 3, for example, while across locations, data range from 2 to 20 or more, and these differences are tracked in both bottom and surface samples.

[1] The following steps are needed for calculating ES: (1) X = log(Bottom) – log(Surface); (2) X = log(Bottom/Surface); (3) Exp(X) = Bottom/Surface; (4) ES = (Bottom – Surface)/Surface = Bottom/Surface – 1; (5) ES = Exp(X) – 1; (6) ES (%) = (Exp(X) – 1)*100.

For model 2 (Station), there was a significant depth interaction for Diatoms (p=0.002), while the next strongest station effect was for Dinoflagellatees (p-0.15; **Table 3**). Box-plots capture the marginal station-depth differences found in the Diatom data (**Figure 3**). There were clear differences at station TPS (bottom > surface or integrated), but no patterns for the other stations. Dinoflagellates were included in **Figure 3** for an additional comparison – minimal station-specific differences are observed.

For model 3 (Season), there was a significant depth interaction with season for Simpson's diversity (Table 3) when July was treated as a summer month or as a winter month. The next strongest season effects were for Diatoms and Dinoflagellates (P close to 0.05). Again, box-plots capture the season-depth differences found for Simpson's diversity (**Figure 4**), with Dinoflagellates also shown for an additional comparison. For Simpson's diversity, the data suggest lower diversity in bottom samples during winter months but little difference during summer months. For Dinoflagellates, the marginally significant seasonal differences among surface, bottom and integrated samples are subtle (minimal).

Conclusion – Analysis shows that for the few variables where there are statistically significant differences in values by depth, the magnitude of differences associated with depth is minimal compared to the magnitude of differences that occur naturally between samples and stations. We conclude that future sampling should focus on surface samples, to take advantage of the most baseline data. However, in cases where there is reason to suspect potential for large depth-specific effects, a bottom sample could be added occasionally to test for a hypothesized depth effect. Given that the CREMP is intended to monitor large-scale basin-wide changes, depth-specific sampling is less likely to be warranted for the CREMP than for other programs that have a finer scale of resolution. Depth samples were not included in power analyses later in this appendix – as with duplicate samples, the depth samples are pseudo-replicates of the surface samples with which they are associated.

5. DATA EVALUATION

Data are summarized here prior to analysis of the sampling design in the next section. Summary statistics for the data are provided in **Table 5** (excluding field duplicates and depth replicates). Biomass data are quite variable, while counts of total species and Simpson's diversity are much less variable. Total biomass is dominated by Chrysophytes. **Table 6** shows Spearman correlations among phytoplankton variables and Chlorophyll-a. Variation in total biomass reflects variation in Chrysophytes (r= 0.96), followed by Diatoms (r=0.73) and Dinoflagellates (r=0.68). Total biomass and species counts well correlated (0.77). Chlorophyll-a was only moderately correlated with Total biomass (r = 0.56) and Species counts (r = 0.55). **Tables 7-10** shows medians by station/month for Chlorophyll-a, total biomass, species counts, and Simpsons diversity, while **Figures 5-8**

show corresponding box-plots for stations INUG, SP, TPE, and TPN. Key findings include:

- Chlorophyll-a and total biomass are highest across Baker stations.
- In addition, total biomass and species counts are lower across winter months.
- There is generally high monthly variation relative to within-month sub-sample variation.
- There is some tendency for lower diversity in winter months
- Phytoplankton were clearly affected at SP in August 2008, as indicated by low values for total biomass, species counts and Simpson's diversity.

6. ANALYSIS OF SAMPLING DESIGN

Impact hypotheses and statistical design – Two general classes of impacts are hypothesized for the Meadowbank mine:

1. Pulse events for which potential impacts would be high for a short time but would then (in the case of phytoplankton) dissipate relatively quickly. Pulse events could be associated with any phase of the mine, but are more likely to be associated with particular activities such as dike construction.
2. Long-term cumulative impacts that may be associated with ongoing activities. Long-term cumulative impacts are more likely to be associated with ongoing activities of mine operations.

As operations have just begun in 2010, the focus of monitoring to date has been on detecting pulse events associated with construction. The appropriate framework for analysis is a before-after-control-impact (BACI) that is aimed at detecting a potential impact in a particular lake or basin in a particular time period. The BACI framework can also be used to evaluate long-term impacts, but other tools such as time series regression analysis may also be appropriate for evaluating long-term trends. For this design document, we focus on the use of the BACI framework, recognizing that other tools such as time series regressions may be useful at a future date once sufficient time series data are available[2].

The classic BACI (paired) design has before/after periods α_i (i = B, A; I = 2), control/impact sites β_j (j = C, I; J = 2), and a total of K paired sampling times τ_k that are

[2] In theory, a BACI analysis that is appropriately framed should be capable of detecting changes associated with long-term trends.

nested within period. A statistical model for this design is given by (Smith 2002, equation 2):

$$(1) \qquad X_{ijk} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

The key term is the interaction ($\alpha\beta$), which can be tested using an F test with F = MS[($\alpha\beta$)]/MS[Resid] and degrees of freedom = 1, $K - 2$. As discussed by Smith (2002), this is equivalent to simply taking the differences between the control and impact values across times and using a two-sample (before-after) t test (Stewart-Oaten et al. 1986).

Model (1) can be extended to include additional control sites (e.g., "asymmetric" designs; Underwood 1994) and/or additional impact sites. To be valid, the additional sites must be replicates rather than subsamples (i.e., as controls, they should be spatially independent of each other but representative of the impact sites, while replicates for impacts need to be spatially independent and (ideally) affected by independent disturbances). So whereas j = (C, I) in the classic BACIP, j may compose any combination of J total sites, for example $J = 4$ where j = (C_1, C_2, C_3, I). The general test of ($\alpha\beta$) still applies, but with degrees of freedom = $(J - 1)$, $(K - 2)(J - 1)$ (e.g., see Table 1 of Underwood (1994) and Table 9 of Smith (2002)).

In addition, there may be n replicate subsamples s at each site/time combination (jk), as assumed in Table 1 of Underwood (1994). In this case, we modify equation (1) as:

$$(2) \qquad X_{ijks} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + (\tau\beta)_{k(i)j} + \varepsilon_{ijks} \ ,$$

where subsamples now permit estimation of times-by-site interactions ($\tau\beta$). The appropriate F ratio for ($\alpha\beta$) is now F = MS[($\alpha\beta$)]/MS[($\tau\beta$)] with df = $(J - 1)$, $(K - 2)(J - 1)$. As Underwood demonstrates, specific comparisons (interaction terms, such as the impact site versus either "period" or a specific "time" unit) can be examined by partitioning variation accordingly (e.g., Underwood Table 2[3]).

Methods – The analysis here assessed the expected precision and power of BACI estimates for different after-period (impact) durations and different numbers of sub-samples (random spatial samples collected each month). Formal analysis of the sampling design uses Meadowbank data. Baker Lake data have only been collected since 2008, thus inferences would be limited based on the small data set. Results from Meadowbank should be generalizable to Baker Lake given that the BACI analyses below for Meadowbank compare a single control to each potential impact station individually (the same scenario as for Baker, which has one control and two impact stations). Separate analyses were conducted for the three primary impact stations SP, TPE, and TPN. In each case, INUG was used as the control station. Separate analyses were conducted for

[3] We note that there are mistakes in the tables presented by Underwood (1994).

two variables – total biomass and species counts. For each variable, the effect size or ES (fixed across months) was set at either a 20% reduction from baseline or a 50% reduction in baseline.

For purposes of the BACI analyses, the before-period data included all available control-impact paired data to represent the "before period" dataset. This assumes that these data, regardless of station or month, are reasonably representative of "natural" conditions. This provides a "worst-case" scenario because additional month-specific variation in the data caused by potential impacts at SP and TPE in particular are incorporated and assumed to be representative of baseline. If those impact cases were removed, month-specific variation would be expected be lower, and estimates of power would be expected to be higher. Therefore results for TPN should provide the best representation of power/precision.

After-period data were simulated using variances consistent with observed data. Specifically, the following mixed-effects model was fit to the "before data" for a given control-impact station pair:

$$X_{jks} = \beta_j + \tau_k + (\tau\beta)_{kj} + \varepsilon_{jks}$$

The fit provided estimates of before-period station means β_j, random-effects variances for month ($\sigma^2[\tau]$) and month-by-station ($\sigma^2[\tau\beta]$), and the residual variance for sub-samples ($\sigma^2[\varepsilon]$).

After-period data were simulated using after-period means (β_control, β _impact + **ES**) and the above variance estimates for three durations (1, 3, and 6 months) and three sub-sample scenarios (1, 2, and 3 subsamples per month). **Thus, a total of 9 scenarios of after-period duration and sub-sampling were examined.** In all cases, log-transformed data were used. For each scenario, 500 simulations were used.

<u>*Data summary*</u> – Given the methods above, when data were limited to paired data (months) with INUG, there were 25 samples for each station (**Table 11**). As noted above, values used in BACI simulations were obtained from fits of mixed-effects models to current data (the "before period"). These estimates are summarized in **Table 19**. In each case, data were paired with INUG samples (means for INUG, which were used as the after-period means, are of minor importance and are not shown). The metrics of interest are the effect sizes for log data, ES(log), and estimates of standard deviations (SD, log units). In relative terms, power will be high when ES(log) >> SD(M x S) and SD(Error), so we expect higher power for biomass and species counts at TPE.

<u>*Results*</u> – Estimates of statistical power are shown in **Table 13** and repeated in **Figure 9**. In all cases, the *a priori* hypothesis is that impacts will result in decreases, so power is

based on one-tailed tests (alpha = 0.05 and 0.10 are both shown in **Table 13**; alpha = 0.05 for **Figure 9**). For total biomass and ES = 20%, power was < 0.6 after 6 months of data (all stations). For Total biomass and ES = **50%**, power was > 0.6 after 3 months, and basically > 0.9 after 6 months. Power was higher for species counts than for total biomass. For ES = 20% and N = 2 sub-samples, power was > 0.8 after 3 months for TPE and TPN (but only around 0.5 for SP after 6 months). For ES = 50%, power was basically 100% at all stations after 3 months. In general, the number of sub-samples was not important in comparison to the number of months of sampling.

A measure of precision for BACI estimates is shown in **Figure 10**. For log data, a useful measure of precision is the coefficient of variation (CV = SE[BACI estimate]/ [BACI estimate]). Results are averages across 500 trials, and do not depend on tails or alpha. These results mirror those for power. As a rough guide, power > 80% when CV < 40%.

Conclusion – Results show that the power to detect 20% reductions in phytoplankton variables is much lower than power to detect 50% reductions. This is particularly true for total biomass, and we would expect power to be even lower for all other biomass variables (except Chrysophytes) given their higher relatively variability. Results also show that the benefits of subsampling are limited, with very little gain in power associated with more than two subsamples. In terms of duration of sampling, the results show notable increases in power associated with 3 months of sampling, and smaller increases associated with moving from 3 to 6 months.

7. REFERENCES

Gelman, A, and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models – Analytical Methods for Social Research. Cambridge University Press, New York, 625 pp.

Pinheiro, J., and D. Bates. 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York, 528 pp.

Smith, E. 2002. BACI design. Pages 141-148, in: A.H. El-Shaarawi and W.W. Piegorsch [Eds], Encyclopedia of Environmetrics, John Wiley & Sons, Chicherster.

Stewart-Oaten, A., W. Murdoch, and K. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67: 929-940.

Venables, W., and B. Ripley. 2002. Modern Applied Statistics. Springer, New York, 495 pp.

Underwood, A. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications 4(1): 3-15.

Table 1. Phytoplankton samples collected for the CREMP.

Notes: Duplicates and depth replicates removed. Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006	Jul				1		1	1		1	1	1	1	7
	Aug				2		2	2		2	2	2	2	14
	Sep				1		1	1		1	1	1	1	7
2007	Jul				1		1	1		1	1	1	1	7
	Aug				1		1	1		1	1	1	1	7
2008	Jul	1	1	1	1		1	1		1	1	1	1	10
	Aug	1	1	1	1		1	1		1	1	1	1	10
	Sep	1	1	1			1	1		1	1	1		8
2009	Jul	3	3	3	3	3	3	3	2	3	3	3	3	35
	Aug	1	1	1	1	1	1	1	1	1	1	1	1	12
	Sep	3	3	3	3	3	3	3	3	3	3	3	3	36
	Nov				3		3	3		3	3			15
	Dec						1	1		1	1	1		5
2010	Jan						3	2		3		3		11
	Feb						1	1		1	1	1		5
	Mar						3	3		3	3	3		15
	Apr						1	1		1	1	1		5
	May						3	3		3	3	3		15
	Jul	3	3	3	3	3	3	3	3	3	3	3	3	36
	Aug	1	1	1	1	1	1	1	1	1	1	1	1	12
	Sep	3	3	3	3	3	3	3	3	3	3	3	3	36
Total		17	17	17	25	14	38	37	13	38	35	35	22	308

Table 2. Summary of statistical comparisons between sample/duplicate pairs for phytoplankton biomass variables.

Notes: N = total samples (pairs = N/2); Median = median of all samples; MAD = median absolute difference among sample/duplicate pairs; r = Spearman rank correlation among sample/duplicate pairs; SD = standard deviation of random effects for Location (variability among sample pairs) and residual errors (variability between samples and duplicates) fit to log-transformed data.

					Model SD (log data)	
Variable	N	Median	MAD	r	Location	Res. Error
Cyanophyte	74	0.19	0.17	0.46	0.728	1.058
Chlorophyte	74	4.2	1.4	0.71	0.892	0.716
Chrysophyte	74	99.2	15.4	0.87	0.820	0.265
Diatoms	74	8.6	1.5	0.93	1.672	0.343
Cryptophytes	74	8.6	3.3	0.71	0.672	0.415
Dinoflagellates	74	9.7	4.2	0.78	1.347	0.729
Total.biomass	74	148.1	14.7	0.89	0.659	0.221
Species	74	35	2	0.84	0.336	0.095
Simpsons.diversity	74	0.84	0.02	0.83	0.076	0.030

Table 3. P-values for F tests of the key fixed-effect terms for depth: Type (Model 1), Type*Station (Model 2), and Type*Season (Model 3; Seas 1).

Notes: Shaded cells denote $0.05 \leq P < 0.01$ (light shade) and $P \leq 0.01$ (dark shade). N = number of samples (locations = N/3).

		Type	Type*Station	Type*Season (Model 3)		
Variable	N	(Model 1)	(Model 2)	N	July summer	July winter
Chlorophyte	120	0.879	0.838	93	0.925	0.980
Chrysophyte	120	0.014	0.726	93	0.248	0.889
Diatoms	120	0.013	0.002	93	0.061	0.052
Cryptophytes	120	0.809	0.197	93	0.188	0.393
Dinoflagellates	120	0.098	0.154	93	0.057	0.727
Total.biomass	120	0.089	0.362	93	0.152	0.827
Species	120	0.098	0.963	93	0.456	0.097
Simpsons.diversity	120	0.471	0.836	93	0.021	0.019

Table 4. Summary of coefficients for Model 1 (log[y] ~ Type + Location).

Notes: Est = estimate; SE = standard error; P = P-value based on t-test; ES (%) = proportional effect size in untransformed units relative to surface samples, where ES (%) = (exp[Est] – 1)*100. Shaded cells denote $0.05 \leq P < 0.01$ (light shade) and $P \leq 0.01$ (dark shade).

	Bottom – Surface				Integrated - Surface			
Variable	Est	SE	P	ES (%)	Est	SE	P	ES (%)
Chlorophyte	0.037	0.163	0.823	3.7%	0.083	0.163	0.614	8.6%
Chrysophyte	-0.013	0.062	0.833	-1.3%	-0.167	0.062	0.009	-15.4%
Diatoms	0.402	0.133	0.003	49.5%	0.188	0.133	0.162	20.7%
Cryptophytes	-0.064	0.099	0.522	-6.2%	-0.023	0.099	0.820	-2.2%
Dinoflagellates	-0.309	0.213	0.151	-26.6%	-0.458	0.213	0.035	-36.7%
Total.biomass	-0.007	0.052	0.891	-0.7%	-0.104	0.052	0.049	-9.9%
Species	0.026	0.025	0.311	2.6%	-0.029	0.025	0.247	-2.9%
Simpsons div.	0.004	0.013	0.742	0.4%	0.015	0.013	0.236	1.5%

Table 5. Summary statistics for phytoplankton biomass variables across CREMP samples (control and impact).

Variable	N	N=0	Mean	Median	SD	CV
Chlorophyll a	243	0	0.46	0.38	0.31	0.67
Cyanophyte	308	145	0.62	0.05	1.19	1.93
Chlorophyte	308	1	6.62	4.77	9.04	1.37
Euglenophyte	308	302	0.02	0.00	0.20	8.23
Chrysophyte	308	0	110.7	100.0	72.8	0.66
Diatoms	308	8	15.5	11.3	17.3	1.12
Cryptophytes	308	0	10.7	7.5	10.5	0.99
Dinoflagellates	308	5	13.5	10.7	12.3	0.91
Total biomass	308	0	157.6	146.7	88.9	0.56
Total species	308	0	33.6	36.0	8.5	0.25
Simpsons diversity	308	0	0.82	0.85	0.09	0.11

Table 6. Spearman correlations among phytoplankton biomass and species variables (N = 308).

	Chloro.a	Cyano.	Chloro.	Eugleno.	Chryso.	Diatoms	Crypto.	Dino.	Tot.bio.	Species
Chlorophyll.a										
Cyanophyte	0.21									
Chlorophyte	0.41	0.42								
Euglenophyte	0.04	0.16	0.08							
Chrysophyte	0.50	0.08	0.29	0.03						
Diatoms	0.58	0.32	0.49	0.05	0.54					
Cryptophytes	0.09	-0.08	-0.09	-0.07	-0.14	-0.01				
Dinoflagellates	0.32	0.05	0.17	0.02	0.59	0.39	-0.12			
Total.biomass	0.56	0.15	0.39	0.01	0.95	0.68	0.00	0.61		
Total species	0.55	0.39	0.64	0.08	0.68	0.66	-0.09	0.43	0.74	
Simpsons div.	0.37	0.41	0.62	0.10	0.33	0.57	-0.19	0.16	0.38	0.64

Table 7. Medians of Chlorophyll-a by station and month (N = 243).

Note: Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL
2006	Jul				0.40		0.19	0.13		0.24			0.40
	Aug				0.21		0.30	0.21		0.22	0.28	0.17	0.41
2007	Jul				0.52		0.60	0.46		0.36	0.37	0.40	0.53
	Aug				0.57		0.54	0.58		0.33	0.39	0.33	0.69
2008	Jul	0.93	0.44	0.80	0.51		0.34	0.31		0.40	0.36	0.39	0.56
	Aug	1.04	1.34	1.76	0.62		0.56	0.60		0.40	0.34	0.33	0.66
	Sep	0.97	0.94	0.85			0.65	1.11		0.71	0.52	0.60	
2009	May				0.34		0.56	0.34		0.13	0.37		
	Jul	0.82	1.11	1.21	0.36	0.47	0.69	0.51	0.33	0.38	0.21	0.32	0.62
	Aug	1.19	1.03	1.01	0.34	0.23	0.49	0.45	0.44	0.34	0.37	0.34	0.32
	Sep	1.19	1.05	1.41	0.54	0.47	0.52	0.28	0.16	0.20	0.29	0.27	0.33
	Nov				0.44		0.74	0.57		0.51	0.56		
	Dec						0.39	0.37		0.28	0.32	0.23	
2010	Jan						0.30	0.31		0.21	0.12	0.14	
	Feb						0.20	0.30		0.25	0.14	0.12	
	Mar						0.12	0.49		0.22	0.23	0.11	
	Apr						0.35	0.30		0.11	0.31	0.19	
	May						0.51	0.42		0.15	0.30	0.35	

Table 8. Medians of Total biomass by station and month (N = 308).

Note: Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL
2006	Jul				230		142	103		142	109	107	141
	Aug				151		108	133		128	114	106	134
	Sep				216		283	202		170	127	146	326
2007	Jul				241		154	121		154	110	94	230
	Aug				121		99	111		145	109	154	203
2008	Jul	199	145	361	281		254	210		178	251	119	216
	Aug	241	213	337	173		37	144		185	174	122	168
	Sep	141	139	160			77	105		196	144	135	
2009	Jul	209	304	237	329	287	262	289	269	240	168	145	310
	Aug	316	468	184	190	122	248	144	196	262	168	224	141
	Sep	235	262	207	148	191	167	123	188	77	137	150	196
	Nov				161		144	122		97	49		
	Dec						60	105		36	48	23	
2010	Jan						46	52		39		34	
	Feb						25	85		19	28	15	
	Mar						37	91		56	33	23	
	Apr						29	36		28	30	26	
	May						89	52		35	85	22	
	Jul	337	253	283	190	296	175	179	172	98	73	73	201
	Aug	174	298	249	99	130	135	120	114	126	143	157	194
	Sep	149	146	131	162	215	246	151	110	168	162	123	167

Table 9. Medians of Species counts by station and month (N = 308).

Note: Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL
2006	Jul				38		33	26		33	30	30	31
	Aug				38		35	37		34	28	35	39
	Sep				48		44	45		41	35	38	41
2007	Jul				41		37	33		36	30	30	41
	Aug				35		31	36		39	35	39	39
2008	Jul	34	32	39	37		34	36		36	37	33	38
	Aug	41	42	50	45		17	37		41	34	39	41
	Sep	37	39	37			41	36		38	38	39	
2009	Jul	39	40	39	40	39	38	36	37	38	37	34	40
	Aug	45	44	38	44	32	46	41	44	34	36	36	40
	Sep	41	42	44	40	39	44	40	40	31	37	33	45
	Nov				33		36	32		31	24		
	Dec						26	30		22	23	26	
2010	Jan						20	21		22		19	
	Feb						12	28		14	14	15	
	Mar						20	25		19	17	19	
	Apr						19	21		14	15	13	
	May						15	18		15	19	15	
	Jul	39	35	39	34	35	33	36	34	25	30	28	38
	Aug	35	37	37	38	31	34	41	31	28	34	34	44
	Sep	36	38	38	35	39	44	41	35	34	40	40	47

Table 10. Medians of Simpsons diversity by station and month (N = 308).

Note: Shading denotes "Impact" periods.

Year	Month	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL
2006	Jul				0.87		0.85	0.84		0.87	0.82	0.87	0.85
	Aug				0.91		0.83	0.85		0.85	0.84	0.87	0.85
	Sep				0.88		0.87	0.89		0.84	0.89	0.84	0.85
2007	Jul				0.80		0.79	0.84		0.78	0.80	0.82	0.79
	Aug				0.87		0.81	0.80		0.80	0.80	0.80	0.84
2008	Jul	0.83	0.87	0.89	0.84		0.77	0.87		0.86	0.64	0.85	0.89
	Aug	0.89	0.91	0.88	0.84		0.23	0.90		0.86	0.86	0.84	0.89
	Sep	0.87	0.88	0.90			0.92	0.83		0.87	0.85	0.86	
2009	Jul	0.85	0.84	0.85	0.88	0.90	0.80	0.80	0.83	0.85	0.85	0.85	0.82
	Aug	0.92	0.93	0.89	0.85	0.88	0.92	0.90	0.87	0.83	0.89	0.90	0.88
	Sep	0.92	0.89	0.91	0.90	0.88	0.90	0.89	0.89	0.89	0.90	0.88	0.89
	Nov				0.89		0.84	0.87		0.90	0.75		
	Dec						0.86	0.85		0.81	0.85	0.87	
2010	Jan						0.69	0.72		0.77		0.76	
	Feb						0.74	0.76		0.65	0.64	0.61	
	Mar						0.77	0.80		0.75	0.69	0.61	
	Apr						0.74	0.68		0.72	0.60	0.67	
	May						0.71	0.73		0.77	0.61	0.76	
	Jul	0.78	0.76	0.78	0.82	0.76	0.69	0.76	0.81	0.77	0.81	0.76	0.87
	Aug	0.88	0.85	0.82	0.89	0.84	0.86	0.89	0.83	0.84	0.86	0.84	0.90
	Sep	0.89	0.89	0.89	0.87	0.88	0.88	0.87	0.89	0.86	0.87	0.90	0.91

Table 11. Samples by station used as the "before-period" in BACI simulations.

Note: Shaded months denote periods currently designated as "impact" months (but treated as "control" data in the BACI analysis).

	INUG	SP	TPE	TPN
Jul.06	1	1	1	1
Aug.06	2	2	2	2
Sep.06	1	1	1	1
Jul.07	1	1	1	1
Aug.07	1	1	1	1
Jul.08	1	1	1	1
Aug.08	1	1	1	1
Jul.09	3	3	3	3
Aug.09	1	1	1	1
Sep.09	3	3	3	3
Nov.09	3	3	3	3
Jul.10	3	3	3	3
Aug.10	1	1	1	1
Sep.10	3	3	3	3
	25	**25**	**25**	**25**

Table 12. Summary of "before-period" mixed-effects model estimates by variable and impact station (paired with control INUG).

Notes: Station means are in raw units. Effect sizes (ES) are shown in raw and log units for ES = 20% and 50% reductions in mean. Estimates of the standard deviations (SD) for random-effects terms (Month and Month-by-Station, M x S) and residuals errors are for log-transformed data.

			ES = 20%		ES = 50%		SD (log data)		
Variable	Station	Mean	ES(raw)	ES(log)	ES(raw)	ES(log)	Month	MxS	Error
Total	SP	159.9	-32.0	-0.223	-80.0	-0.693	0.241	0.278	0.263
Biomass	TPE	144.5	-28.9	-0.223	-72.2	-0.693	0.252	0.162	0.197
	TPN	124.4	-24.9	-0.223	-62.2	-0.693	0.185	0.279	0.190
Species	SP	35.3	-7.1	-0.223	-17.7	-0.693	0.000	0.173	0.079
Counts	TPE	33.5	-6.7	-0.223	-16.8	-0.693	0.102	0.047	0.092
	TPN	33.0	-6.6	-0.223	-16.5	-0.693	0.084	0.068	0.082

Table 13. Estimates of BACI statistical power for detecting a significant decrease (one-tailed test) in a given variable by station (SP, TPE, TPN) as a function of sampling months (after period) and the number of sub-samples per month.

Note: Power is shown for two levels of alpha (0.05 and 0.10).

			alpha = 0.05			alpha = 0.10		
Variable	Months	Samples	SP	TPE	TPN	SP	TPE	TPN
Total biomass (ES = **20%**)	1	1	0.09	0.11	0.09	0.19	0.26	0.19
		2	0.11	0.15	0.11	0.20	0.25	0.21
		3	0.09	0.14	0.09	0.20	0.28	0.21
	3	1	0.11	0.18	0.13	0.21	0.31	0.23
		2	0.15	0.26	0.16	0.28	0.43	0.29
		3	0.16	0.28	0.16	0.27	0.46	0.28
	6	1	0.17	0.31	0.19	0.28	0.49	0.32
		2	0.18	0.40	0.19	0.35	0.59	0.39
		3	0.20	0.43	0.21	0.36	0.59	0.39
Total biomass (ES = **50%**)	1	1	0.32	0.55	0.37	0.47	0.71	0.51
		2	0.34	0.67	0.38	0.54	0.82	0.57
		3	0.38	0.74	0.41	0.56	0.86	0.58
	3	1	0.61	0.93	0.73	0.77	0.97	0.86
		2	0.70	0.97	0.81	0.87	1.00	0.91
		3	0.80	0.99	0.84	0.90	1.00	0.93
	6	1	0.89	0.99	0.94	0.95	1.00	0.98
		2	0.94	1.00	0.97	0.98	1.00	0.99
		3	0.95	1.00	0.98	0.98	1.00	0.99
Species counts (ES = **20%**)	1	1	0.16	0.43	0.38	0.29	0.57	0.55
		2	0.16	0.57	0.47	0.29	0.72	0.64
		3	0.16	0.66	0.52	0.30	0.80	0.70
	3	1	0.28	0.78	0.73	0.48	0.90	0.86
		2	0.36	0.93	0.87	0.50	0.97	0.95
		3	0.33	0.97	0.92	0.54	0.99	0.96
	6	1	0.51	0.97	0.96	0.71	0.99	0.98
		2	0.54	1.00	0.99	0.72	1.00	1.00
		3	0.54	1.00	1.00	0.71	1.00	1.00
Species counts (ES = **50%**)	1	1	0.77	1.00	1.00	0.88	1.00	1.00
		2	0.82	1.00	1.00	0.91	1.00	1.00
		3	0.83	1.00	1.00	0.92	1.00	1.00
	3	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	0.99	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00
	6	1	1.00	1.00	1.00	1.00	1.00	1.00
		2	1.00	1.00	1.00	1.00	1.00	1.00
		3	1.00	1.00	1.00	1.00	1.00	1.00

Figure 1. Sample versus duplicate values of selected phytoplankton biomass variables (log-log scale).

Notes: If zeros were present, 0.1 was added to all values. Solid line is the 1:1 line.

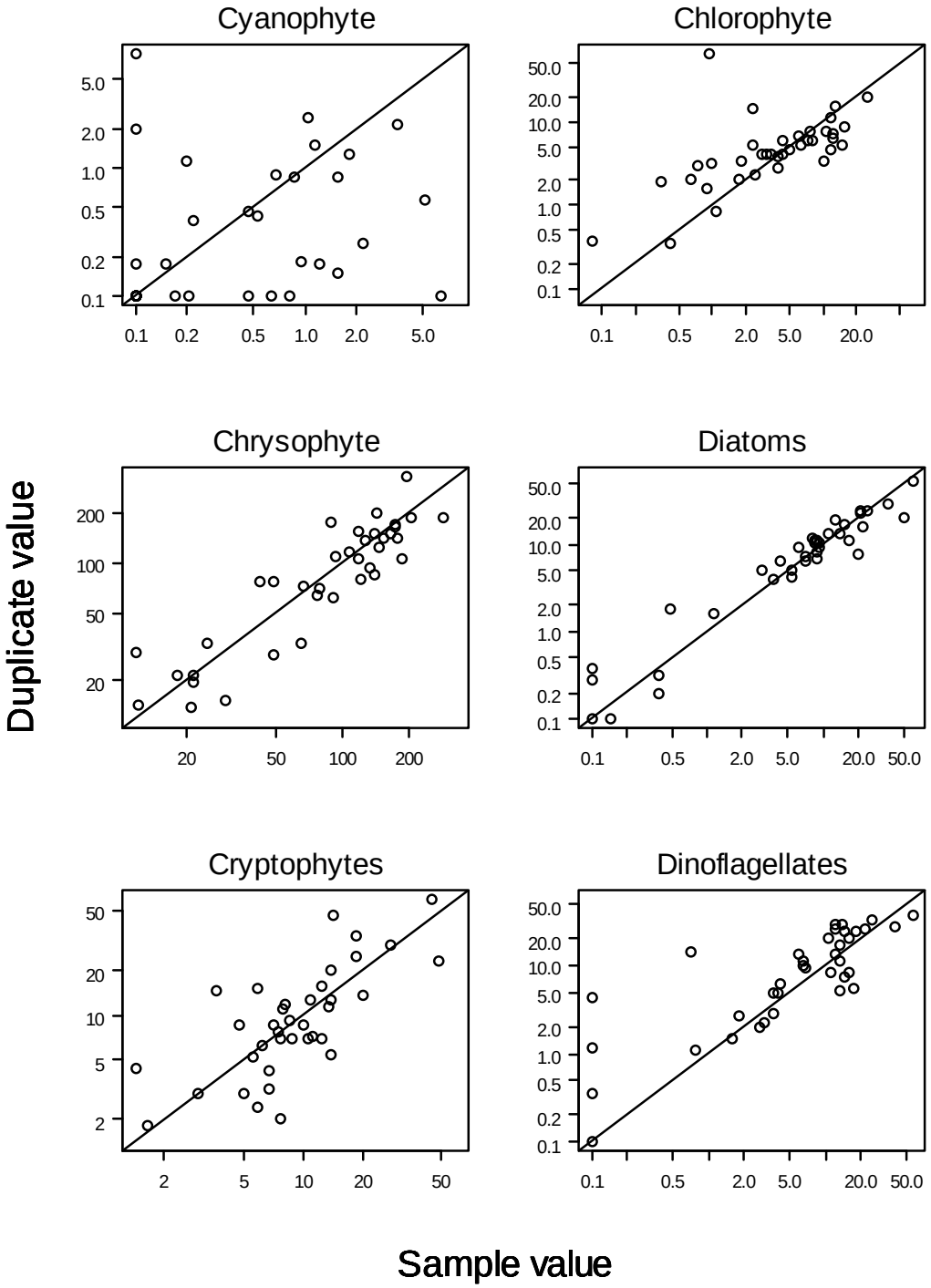


Figure 1 (page 2 of 2)

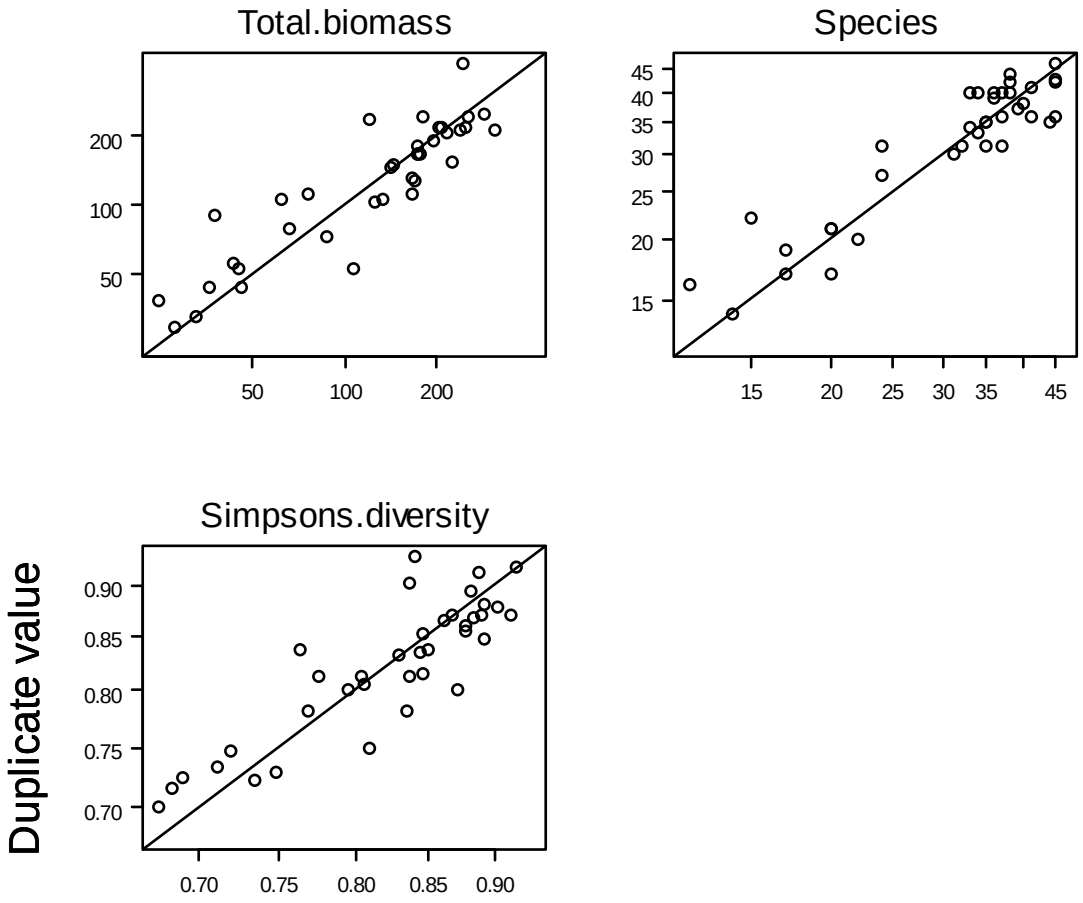


Figure 2. Surface versus Bottom samples (left side; circles) and Integrated samples (right side; triangles) for selected phytoplankton variables (log-log scale).

Notes: Dashed lines denote estimated differences (if $P < 0.1$) relative to Surface samples (see text). Solid line is the 1:1 line.

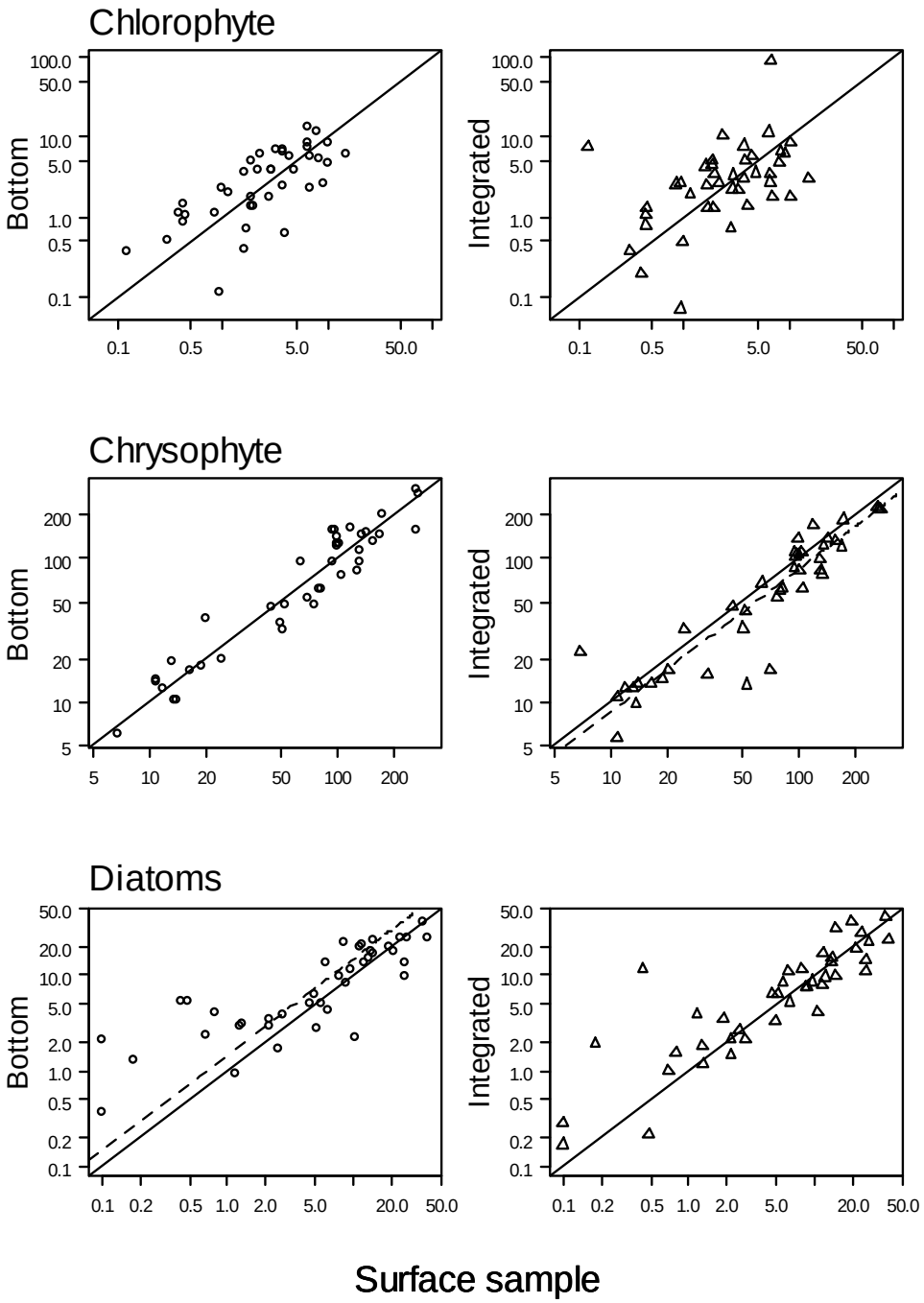


Figure 2 (page 2 of 3)

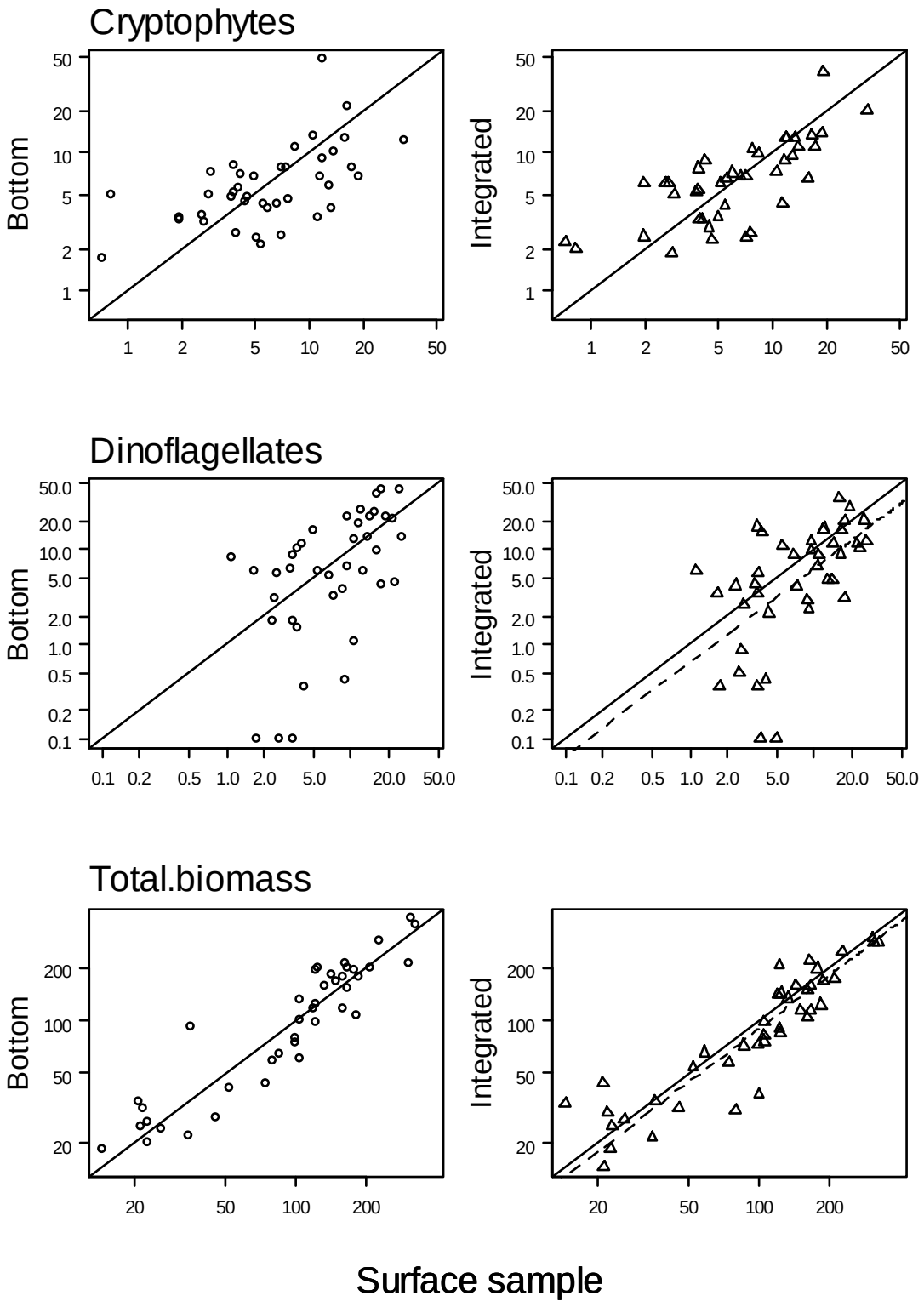
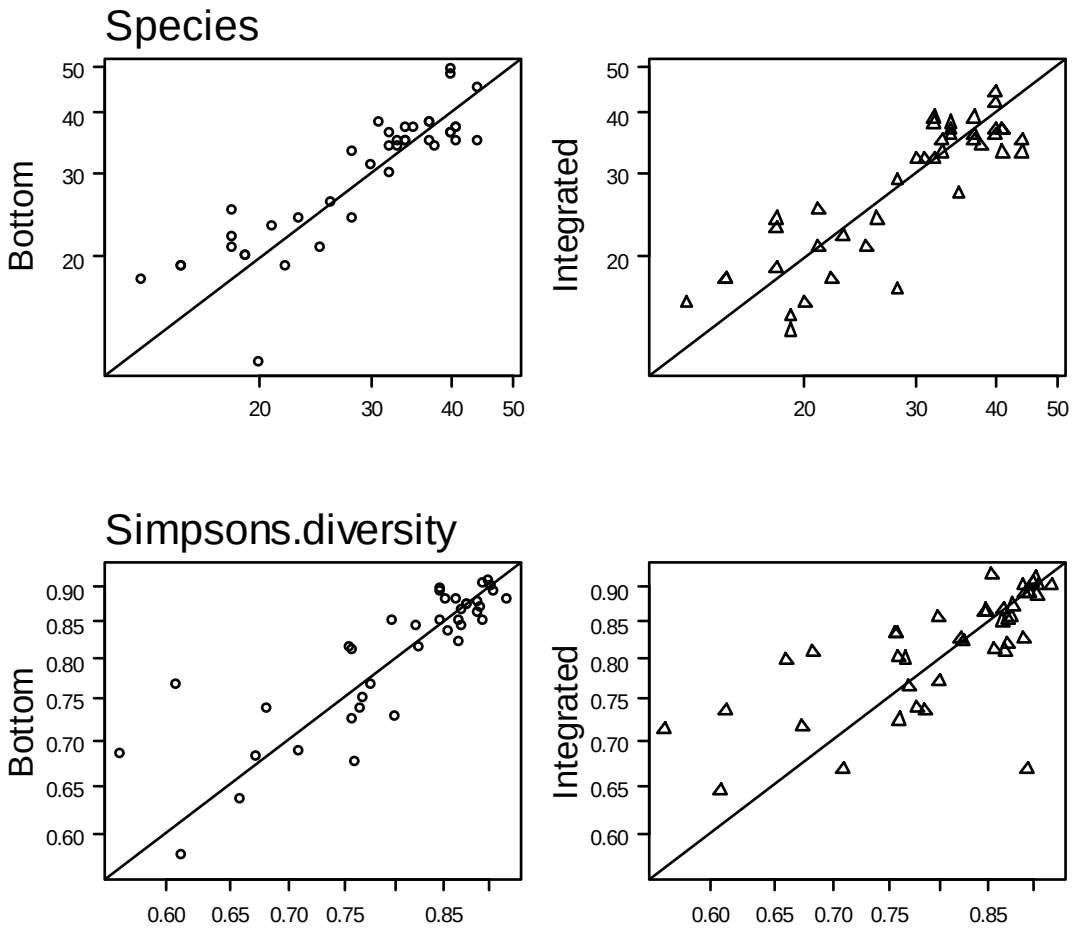


Figure 2 (page 3 of 3)



Surface sample

Figure 3. Box-plots of sample measurements by station (INUG, TE, and TPS) and depth type (.S = surface, .B = bottom, and .Int = integrated) for Diatom biomass (top panel) and Dinoflagellate biomass (bottom panel).

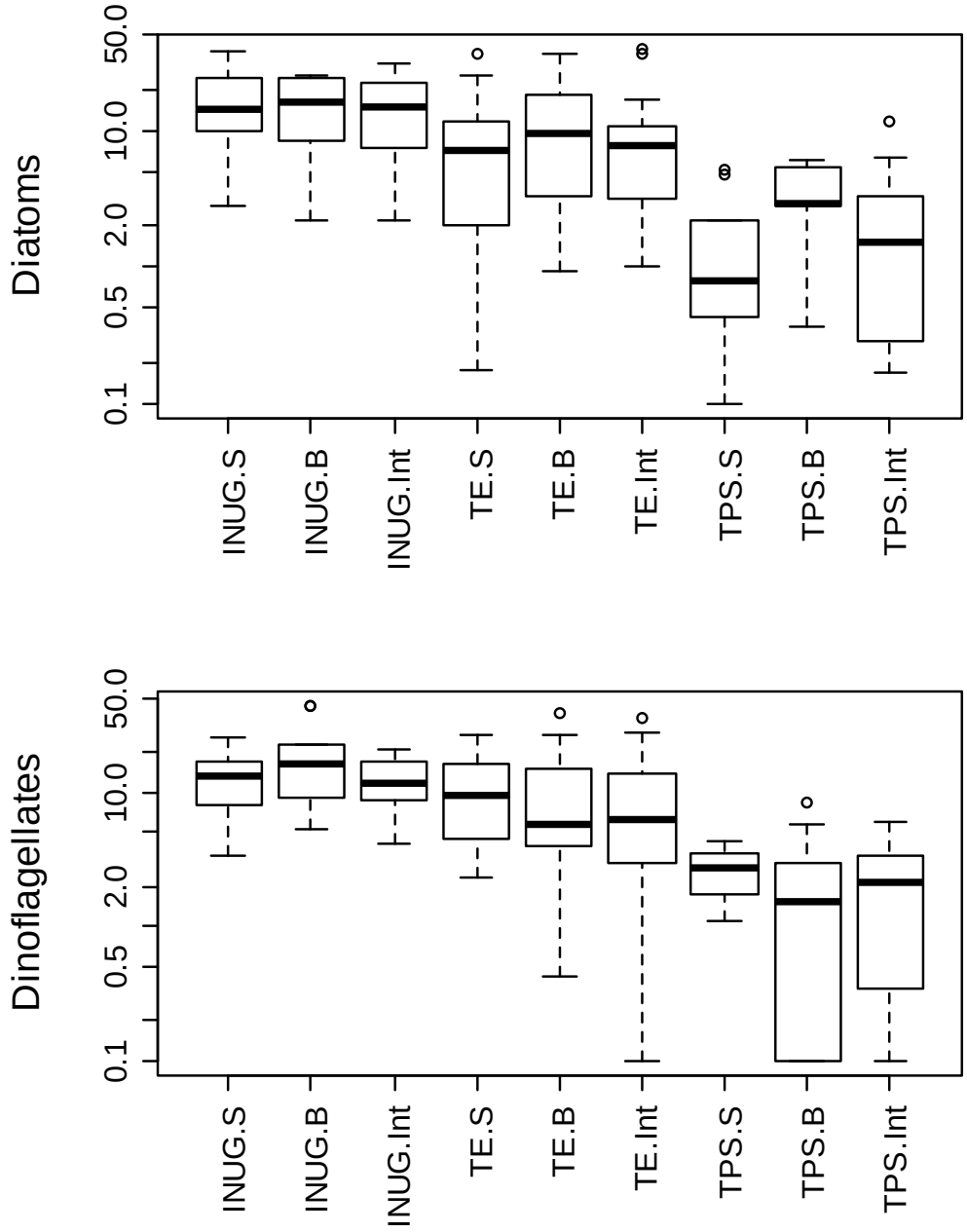


Figure 4. Box-plots of sample measurements by season (Summer or Winter) and depth type (.S = surface, .B = bottom, and .Int = integrated) for Simpson's diversity (top panels) and Dinoflagellate biomass (bottom panels).

Notes: Separate plots are shown for stations INUG (left side) and TE (right side).

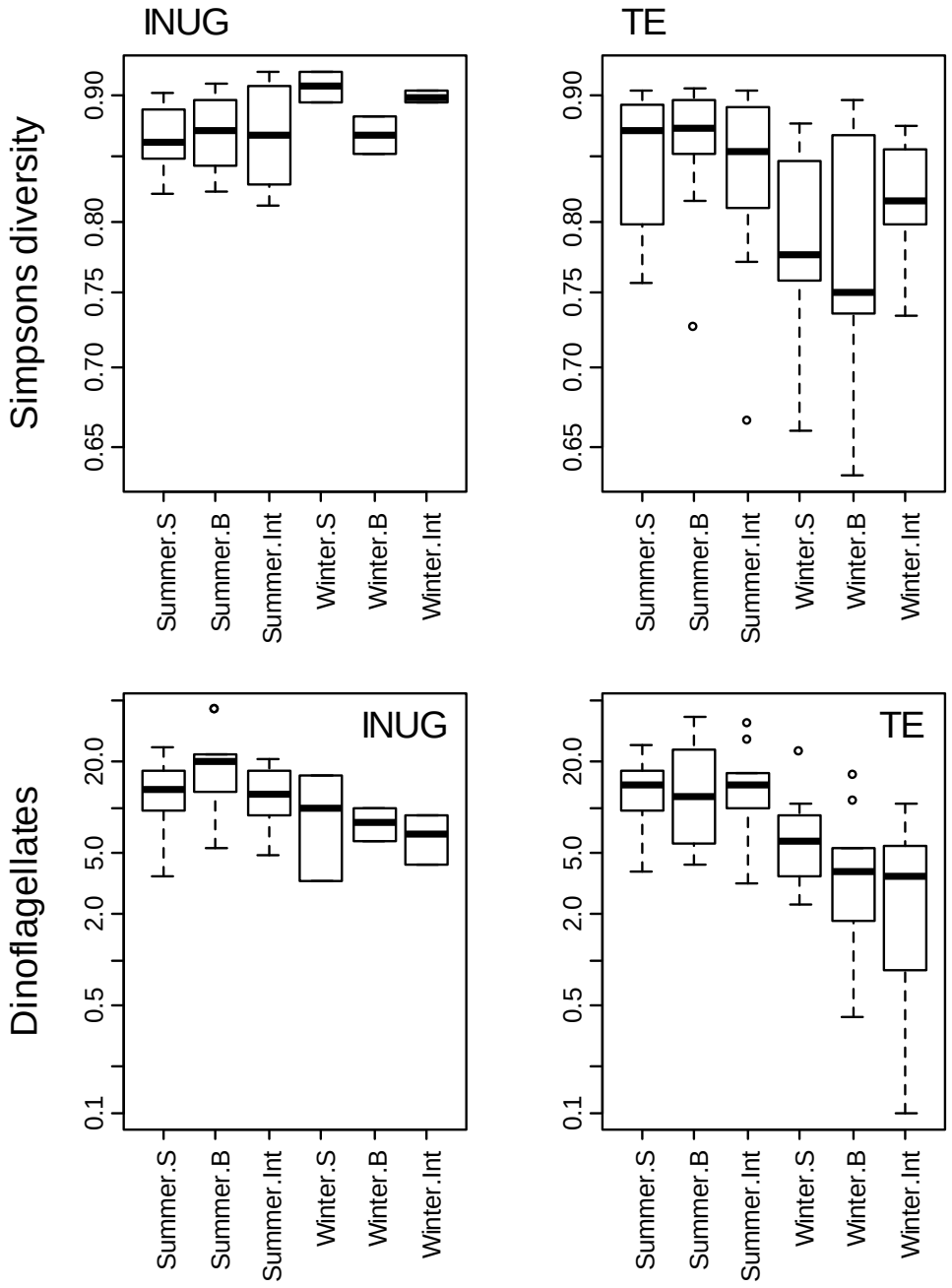


Figure 5. Box-plots of Chlorophyll-a by month for four stations (from top: INUG, SP, TPE, TPN).

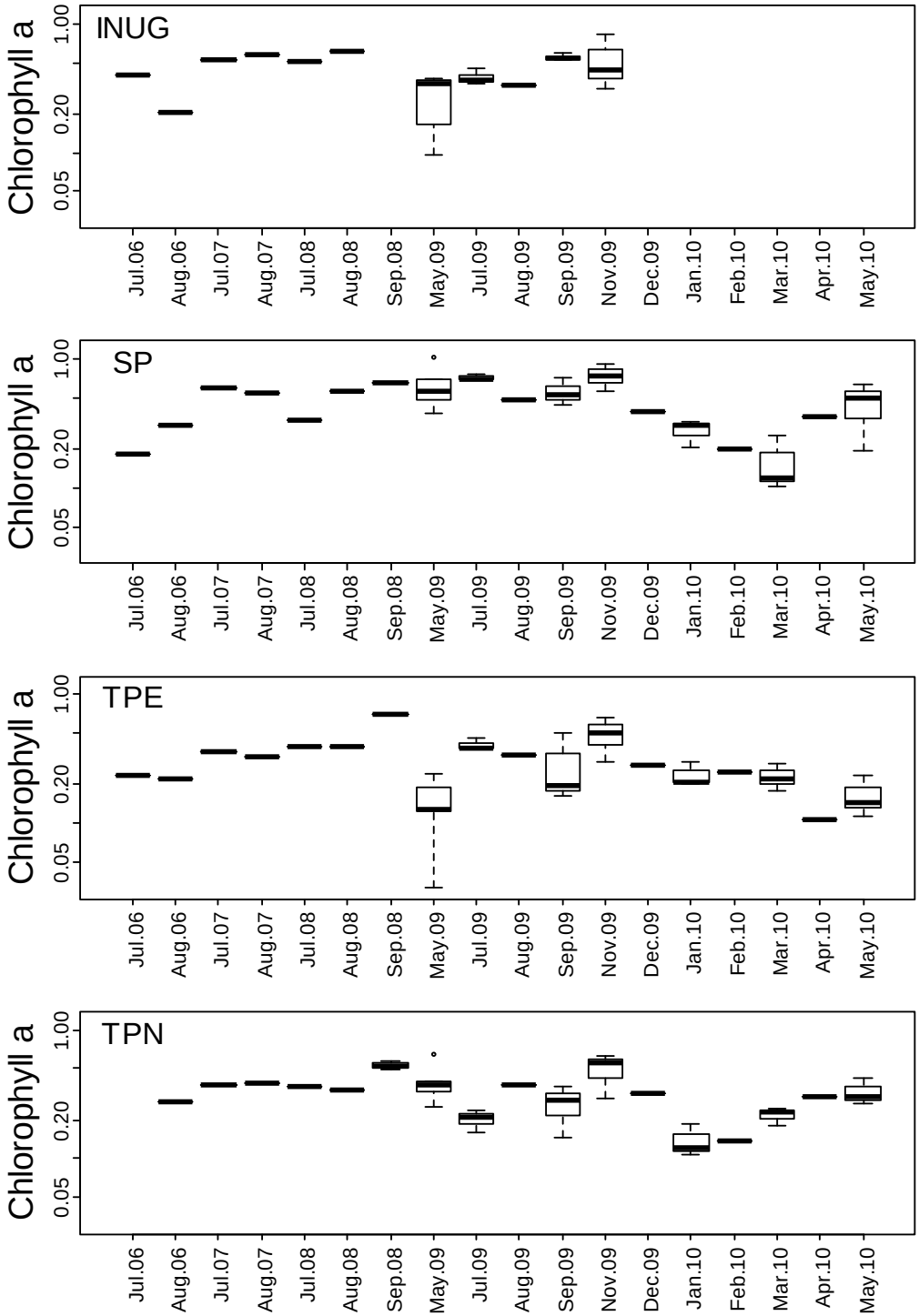


Figure 6. Box-plots of total biomass by month for four stations (from top: INUG, SP, TPE, TPN).

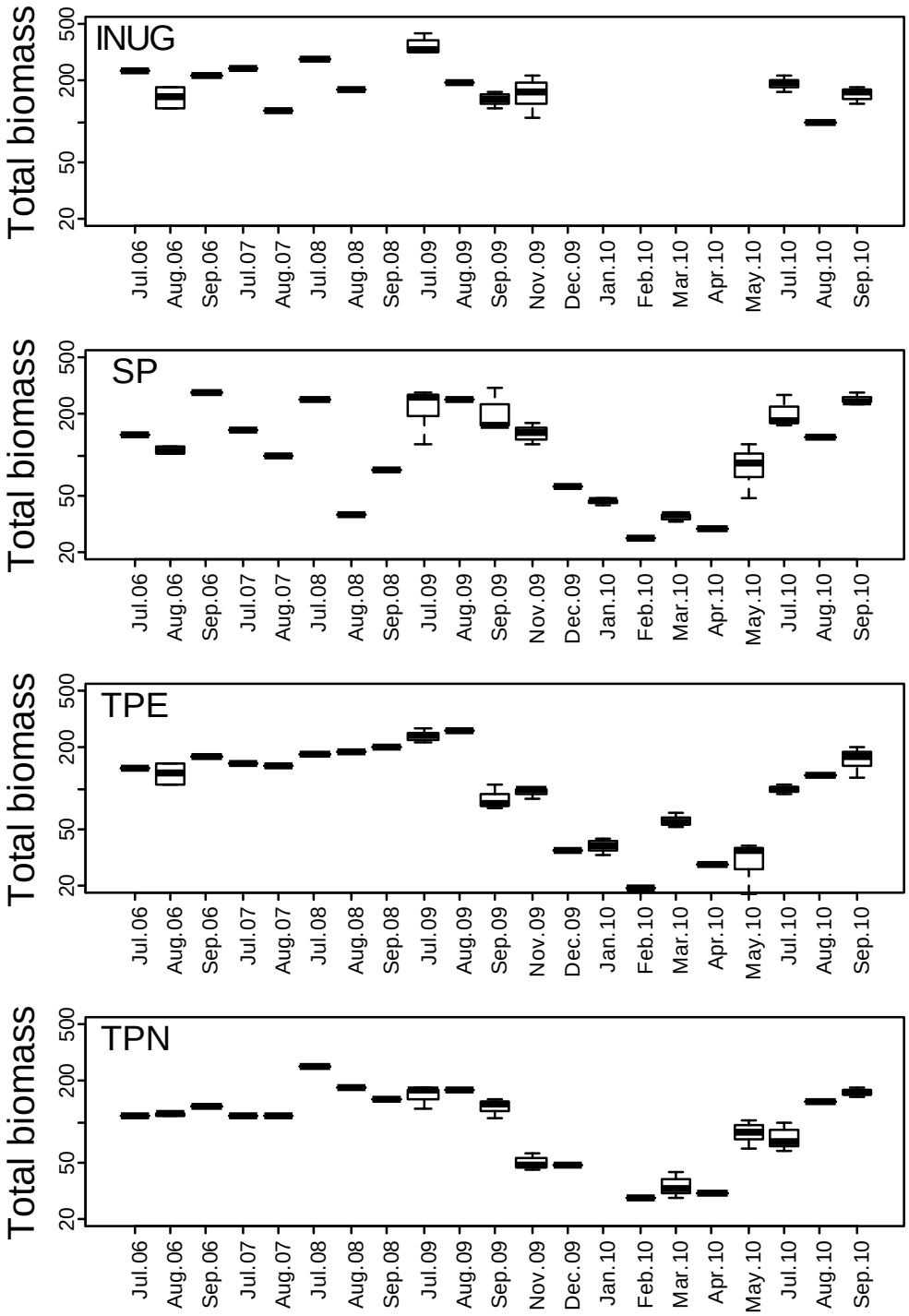


Figure 7. Box-plots of species counts by month for four stations (from top: INUG, SP, TPE, TPN).

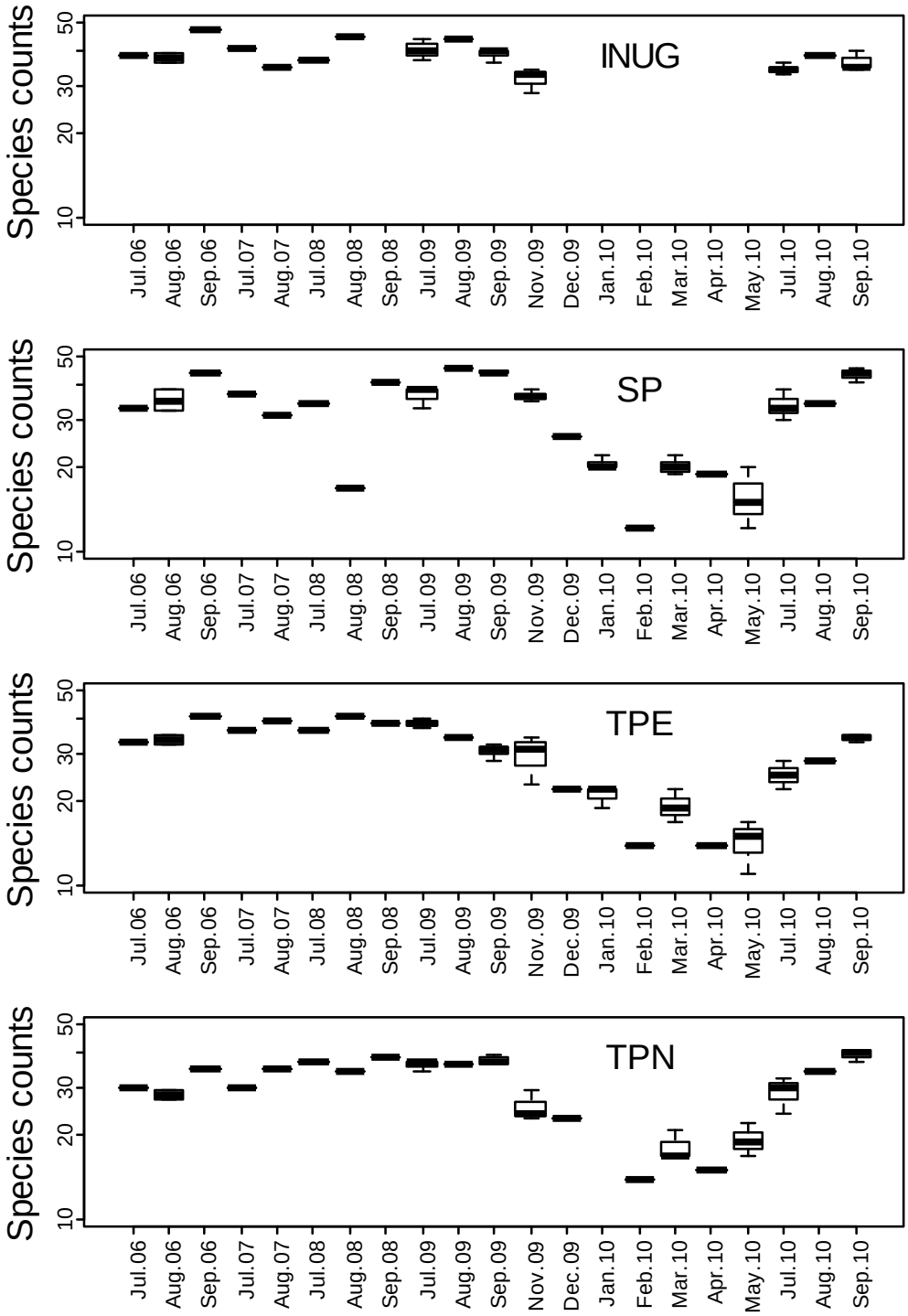


Figure 8. Box-plots of Simpsons diversity index by month for four stations (from top: INUG, SP, TPE, TPN).

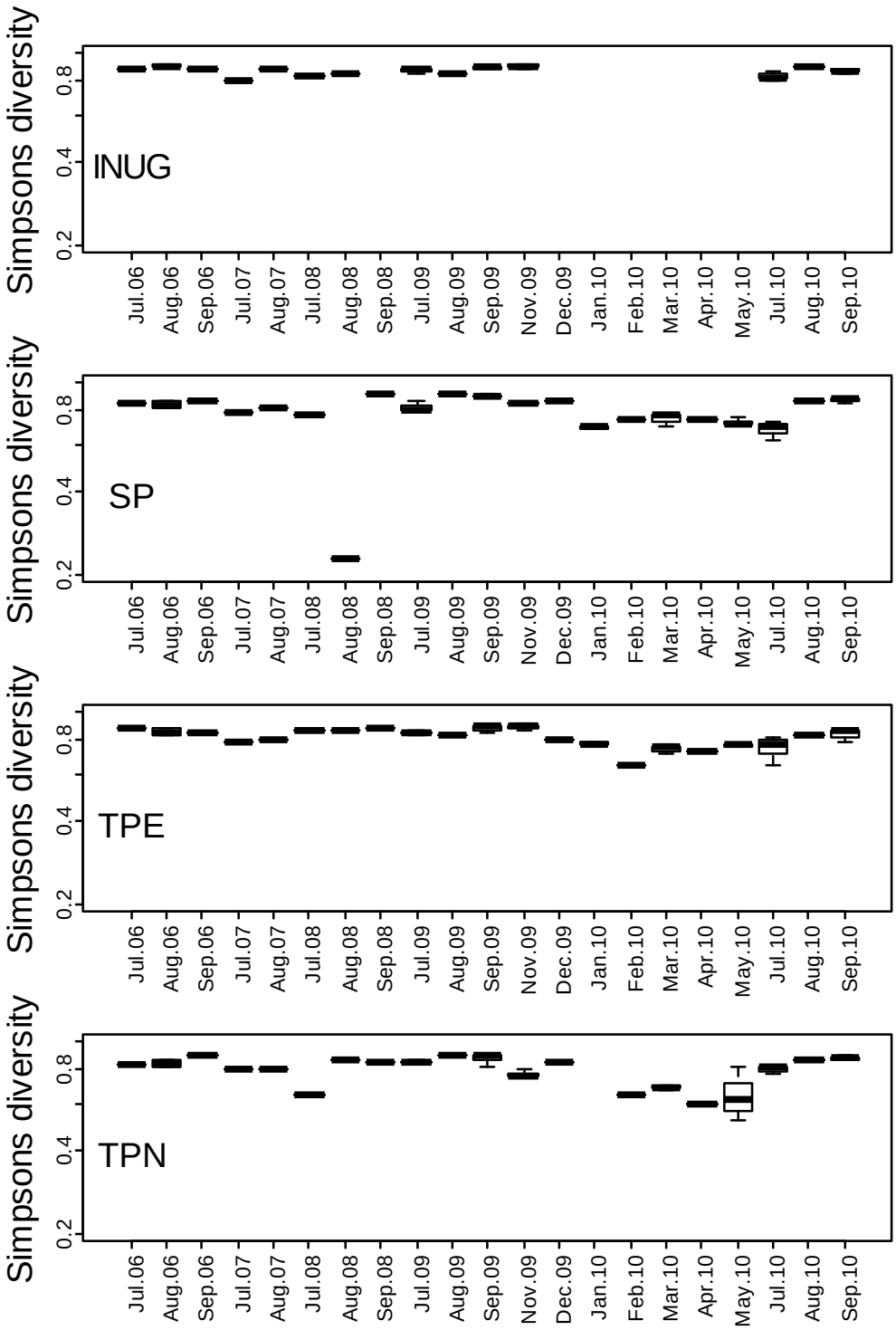


Figure 9. Estimates of BACI power for detecting a significant decrease (one-tailed, alpha = 0.05) in a given variable (Total biomass or Simpsons diversity) and effect size (rows) by station (columns) as a function of sampling months (after period) and the number of sub-samples per month (see legend).

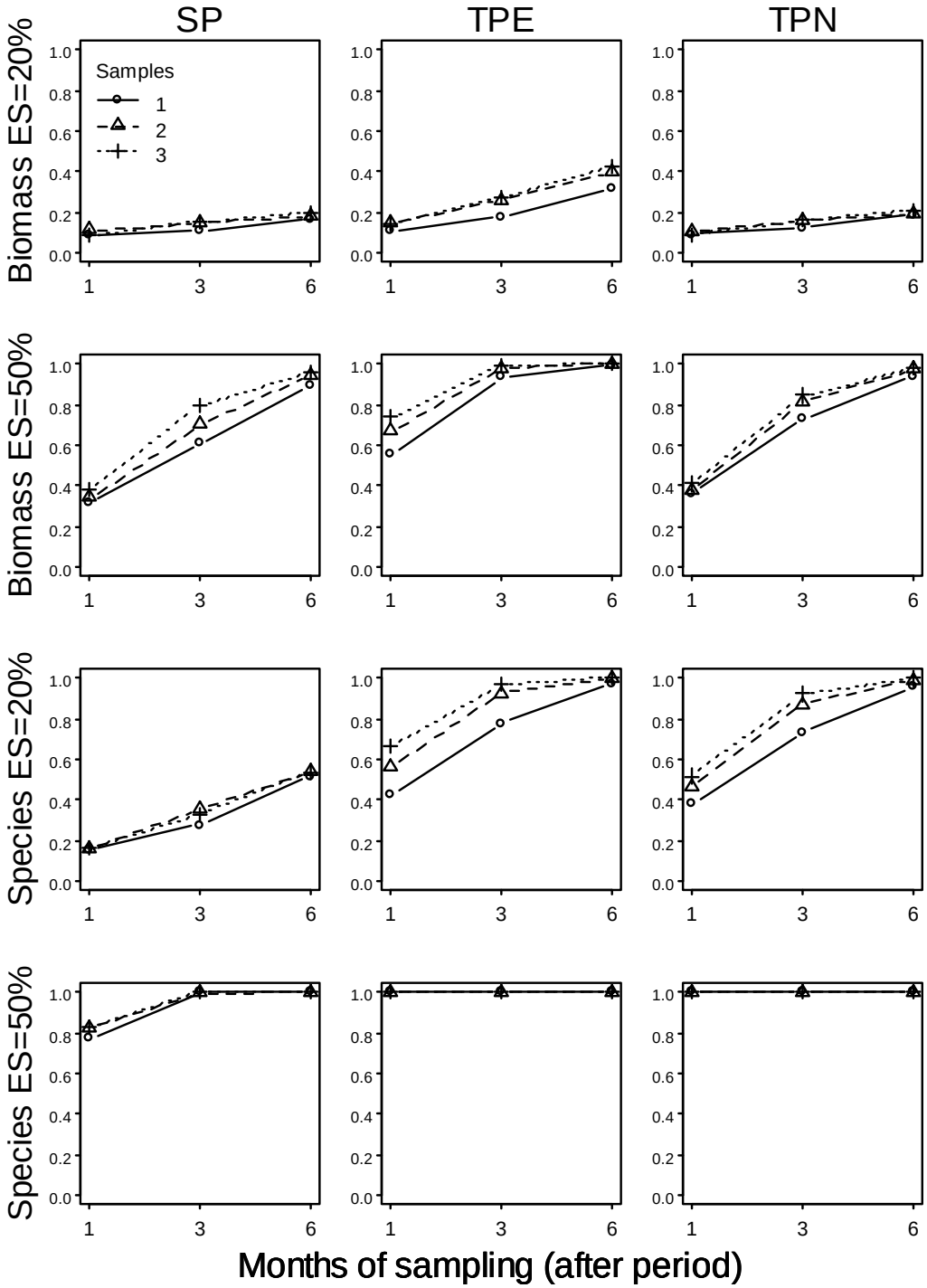
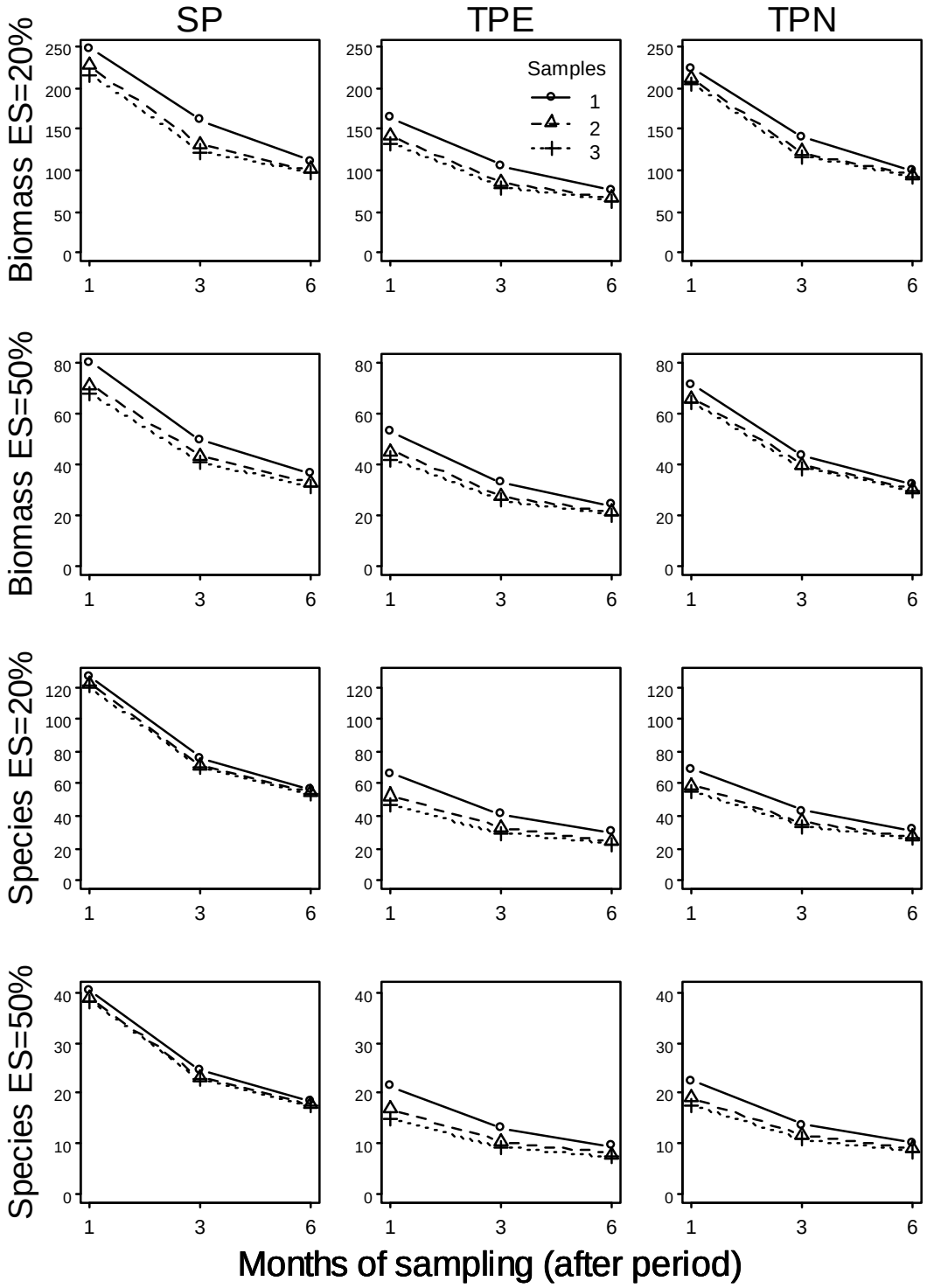


Figure 10. Coefficients of variation (%) for BACI estimates by variable (row) and station (column) as a function of sampling months (after period) and the number of sub-samples per month (see legend).



APPENDIX D – STATISTICAL ANALYSES FOR THE BENTHIC INVERTEBRATE COMMUNITY

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This appendix contains the following analyses.

- Summary of CREMP benthic invertebrate samples
- Data evaluation
- Analysis of sampling design

Some text, tables and figures are repeated from the main document so that the analyses contained in this appendix can be read and understood without reference to the main document. The material in this appendix assumes understanding of basic statistical methods (Venables and Ripley 2002), mixed-effects models (Pinheiro and Bates 2000), before-after-control-impact (BACI) experimental design (Stewart-Oaten et al. 1986; Underwood 1994; Smith 2002), and use of simulation in statistical analysis (Gelman and Hill 2006).

2. SUMMARY OF CREMP BENTHIC INVERTEBRATE SAMPLES

Benthic invertebrates have been collected as part of the CREMP once each year in August from 2006 to 2010. Each sample is a composite of two Petite Ponar grabs, sieved using a 500 um mesh screen. In 2006, additional samples were collected to examine mesh size (250 um, 25 samples) and seasonal patterns (28 samples in July; 25 in September). An additional 20 replicates were collected in 2009 consisting of 1 grab per sample.

For the analyses here, we used only standard samples (August samples that were composites of two grabs and sieved with a 500 um mesh screen), except that five samples for 2006 were removed because they had been poorly preserved and were flagged in the benthos database as "do not use data.".

Thus, the data set includes 242 samples as summarized in **Table 1**. It should be noted that 2006 data include both early and late August samples. Of the total samples, 162 were designated as control and 80 samples were designated as impact (shaded cells in the table). There were no duplicate samples collected for benthos.

3. DATA EVALUATION

A summary of the variables used to represent the benthic invertebrate community is provided in **Table 2**. Abundance data (N, number/m^2) were quite variable (and dominated by insects), whereas counts of total taxa (i.e., richness) and Simpsons diversity index were much less variable. **Table 3** shows Spearman correlations among variables – across samples, variation in total abundance reflects mainly the abundance of insects (r= 0.96),

and variation in total taxa reflects mainly variation in insect taxa (r = 0.91). Total abundance and total taxa are well correlated (r = 0.74). **Tables 4-6** shows medians by station/year for total abundance, total taxa, and Simpson's diversity index, while **Figures 1-3** show corresponding box-plots for stations INUG, SP, TPE, and TPN. Baker control station (BAP) appears to have higher abundances than BBD and BPJ, but the key feature is high inter-annual variability across Baker stations (**Table 4**). BAP had a considerably higher median taxa count in all three years (**Table 5**), whereas Simpson's index was low across Baker stations in 2008 (**Table 6**). Similarly, for Meadowbank stations there is a lot of variability for total abundance and total taxa, particularly for Wally. Total abundance appears lower in 2010 than earlier years for many Meadowbank stations including INUG, SP, TE and TPE. In general, however, it is difficult to discern common patterns as there is clear inter-annual variation within and between stations.

4. ANALYSIS OF SAMPLING DESIGN

Impact hypotheses and statistical design – Two general classes of impacts are hypothesized for the Meadowbank mine:

1. Pulse events for which potential impacts would be high for a short time and then may or may not dissipate relatively quickly. Pulse events could be associated with any phase of the mine, but are more likely to be associated with particular activities such as dike construction.
2. Long-term cumulative impacts that may be associated with ongoing activities. Long-term cumulative impacts are more likely to be associated with ongoing activities of mine operations.

As operations have just begun in 2010, the focus of monitoring to date has been on detecting pulse events associated with construction. The appropriate framework for analysis is a before-after-control-impact (BACI) that is aimed at detecting a potential impact in a particular lake or basin in a particular time period. The BACI framework can also be used to evaluate long-term impacts, but other tools such as time series regression analysis may also be appropriate for evaluating long-term trends. For this design document, we focus on the use of the BACI framework, recognizing that other tools such as time series regressions may be useful at a future date once sufficient time series data are available[1].

The classic BACI (paired) design has before/after periods α_i (i = B, A; I = 2), control/impact sites β_j (j = C, I; J = 2), and a total of K paired sampling times τ_k that are

[1] In theory, a BACI analysis that is appropriately framed should be capable of detecting changes associated with long-term trends.

nested within period. A statistical model for this design is given by (Smith 2002, equation 2):

$$(1) \qquad X_{ijk} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

The key term is the interaction ($\alpha\beta$), which can be tested using an F test with F = MS[($\alpha\beta$)]/MS[Resid] and degrees of freedom = 1, $K - 2$. As discussed by Smith (2002), this is equivalent to simply taking the differences between the control and impact values across times and using a two-sample (before-after) t test (Stewart-Oaten et al. 1986).

Model (1) can be extended to include additional control sites (e.g., "asymmetric" designs; Underwood 1994) and/or additional impact sites. To be valid, the additional sites must be replicates rather than subsamples (i.e., as controls, they should be spatially independent of each other but representative of the impact sites, while replicates for impacts need to be spatially independent and (ideally) affected by independent disturbances). So whereas j = (C, I) in the classic BACIP, j may compose any combination of J total sites, for example $J = 4$ where j = (C_1, C_2, C_3, I). The general test of ($\alpha\beta$) still applies, but with degrees of freedom = $(J - 1)$, $(K - 2)(J - 1)$ (e.g., see Table 1 of Underwood (1994) and Table 9 of Smith (2002)).

In addition, there may be n replicate subsamples s at each site/time combination (jk), as assumed in Table 1 of Underwood (1994). In this case, we modify equation (1) as:

$$(2) \qquad X_{ijks} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + (\tau\beta)_{k(i)j} + \varepsilon_{ijks} \ ,$$

where subsamples now permit estimation of times-by-site interactions ($\tau\beta$). The appropriate F ratio for ($\alpha\beta$) is now F = MS[($\alpha\beta$)]/MS[($\tau\beta$)] with df = $(J - 1)$, $(K - 2)(J - 1)$. As Underwood demonstrates, specific comparisons (interaction terms, such as the impact site versus either "period" or a specific "time" unit) can be examined by partitioning variation accordingly (e.g., Underwood Table 2[2]).

Methods – The analysis here assessed the expected precision and power of BACI estimates for different after-period (impact) durations and different numbers of sub-samples (random spatial replicates collected at the same station during the sampling event each August). Formal analysis of the sampling design uses Meadowbank data. Baker Lake data have only been collected since 2008, thus inferences would be limited based on the small data set. Results from Meadowbank should be generalizable to Baker Lake given that the BACI analyses below for Meadowbank compare a single control to each potential impact station individually (the same scenario as for Baker, which has one control and two impact stations). Separate analyses were conducted for the three primary impact stations SP, TPE, and TPN. In each case, INUG was used as the control station. Separate analyses were conducted for two variables – total abundance and total taxa. For

[2] We note that there are mistakes in the tables presented by Underwood (1994).

each variable, the effect size or ES (fixed across months) was set at either a 20% reduction from baseline or a 50% reduction in baseline.

For purposes of the BACI analyses, the before-period data included all available control-impact paired data to represent the "before period" dataset. This assumes that these data, regardless of station or year, are reasonably representative of "natural" conditions. This provides a "worst-case" scenario because additional year-specific variation in the data caused by potential impacts at SP and TPE in particular (e.g., reduced abundance at SP in 2008 – see **Figure 1**) are incorporated and assumed to be representative of baseline. If those impact cases were removed, month-specific variation would be expected be lower, and estimates of power would be expected to be higher.

After-period data were simulated using variances consistent with observed data. Specifically, the following mixed-effects model was fit to the "before data" for a given control-impact station pair:

$$X_{jks} = \beta_j + \tau_k + (\tau\beta)_{kj} + \varepsilon_{jks}$$

The fit provided estimates of before-period station means β_j, random-effects variances for year ($\sigma^2[\tau]$) and year-by-station ($\sigma^2[\tau\beta]$), and the residual variance for sub-samples ($\sigma^2[\varepsilon]$).

After-period data were simulated using after-period means (β_control, β _impact + **ES**) and the above variance estimates for three durations (1, 2 and 3 years) and four sub-sample scenarios (3, 5, 10 and 20 subsamples per year). **Thus, a total of 12 scenarios of after-period duration and sub-sampling were examined.** In all cases, log-transformed data were used. For each scenario, 500 simulations were used.

Data summary – Given the methods above, the available before data included 26 samples for INUG, SP and TPE, and 23 samples for TPN (**Table 1**). As noted above, values used in BACI simulations were obtained from fits of mixed-effects models to current data (the "before period"). These estimates are summarized in **Table 7**. The metrics of interest are the effect sizes for log data, ES(log), and estimates of standard deviations (SD, log units). In relative terms, power will be high when ES(log) >> SD(Y x S) and SD(Error), so we expect relatively higher power for total abundance for station TPN, and for total taxa at stations SP and TPE.

Results – Estimates of statistical power are shown in **Table 8** for total abundance and **Table 9** for total taxa, and are repeated in **Figure 4**. In all cases, the *a priori* hypothesis is that impacts will result in decreases, so power is based on one-tailed tests (alpha = 0.05 and 0.10 are both shown in **Tables 8 and 9**; alpha = 0.05 for **Figure 4**).

For total abundance and ES = 20%, power was < 0.35 after 3 years of data regardless of station, number of samples, or α level (Table 8). For ES = 50% and $\alpha = 0.05$, power only exceeded 0.8 for TPN. For example, power = 0.82 after 2 years and samples = 10, or 0.92 for samples = 20. For ES = 50% but with a less stringent $\alpha = 0.10$, power exceeded > 0.8 after one year for TPN (but only for samples = 20) and after three years for SP and TPE (for most sample cases).

Power was higher for total taxa counts (**Table 9**, **Figure 4**). Nevertheless, it is still very difficult to detect an ES of 20% even after 3 years. For TPE, power > 0.8 after two years when $\alpha = 0.10$ and samples = 10. For ES = 50%, power > 0.8 or 0.9 after one year in most cases when samples =5 or more.

A measure of precision for BACI estimates is shown in **Figure 5**. For log data, a useful measure of precision is the coefficient of variation (CV = SE[BACI estimate]/ [BACI estimate]). Results are averages across 500 trials, and do not depend on tails or alpha. These results mirror those for power. As a rough guide, power > 80% ($\alpha = 0.05$) when CV < 40%.

Conclusion – Under the BACI model, the presence of year-by-station effects (i.e., natural station-specific inter-annual variation) will make it very difficult to distinguish between impacts and natural variations given only one year of "impact" data (or after several years for relatively minor effects such as ES = 20%). Additional sub-sampling (e.g., n = 10) may improve precision and power considerably in specific situations: that is, when year-by-station effects are relatively low compared to sub-sample variation (e.g., total abundance for TPN, where SD(YxS) = 0.098 and SD(Error) = 0.562; **Table 7**). Many plots for power and CV show useful improvements for n = 10 compared to n = 5. In general, the benthic program has limited power compared to some other programs, but it does have reasonable power to detect large changes in a given year (e.g., ES = 50% for total taxa).

5. REFERENCES

Gelman, A, and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models – Analytical Methods for Social Research. Cambridge University Press, New York, 625 pp.

Pinheiro, J., and D. Bates. 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York, 528 pp.

Smith, E. 2002. BACI design. Pages 141-148, in: A.H. El-Shaarawi and W.W. Piegorsch [Eds], Encyclopedia of Environmetrics, John Wiley & Sons, Chicherster.

Stewart-Oaten, A., W. Murdoch, and K. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67: 929-940.

Venables, W., and B. Ripley. 2002. Modern Applied Statistics. Springer, New York, 495 pp.

Underwood, A. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications 4(1): 3-15.

Table 1. Summary of August benthic invertebrate samples for the CREMP (mesh size = 500 um)

Note: Shading denotes "Impact" periods.

Year	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL	Total
2006				6		6	4		6	3	6	6	37
2007				5		5	5		5	5	5	5	35
2008	5	5	5	5		5	5		5	5	5	5	50
2009	5	5	5	5	5	5	5	5	5	5	5	5	60
2010	5	5	5	5	5	5	5	5	5	5	5	5	60
Total	15	15	15	26	10	26	24	10	26	23	26	26	242

Table 2. Summary statistics for benthic invertebrate variables across August CREMP samples (control and impact).

Note: CV = Coefficient of variation (SD/Mean).

Variable	N	N=0	Mean	Median	SD	CV
Total abundance (N)	242	0	2065	1174	2512	1.22
Oligochaete N	242	99	163	22	770	4.74
Insect N	242	0	1488	728	2181	1.47
Mollusc N	242	15	355	326	281	0.79
Total taxa	242	1	10.4	10.0	4.0	0.38
Oligochaete taxa	242	112	0.8	1.0	1.0	1.23
Insect taxa	242	1	7.5	7.0	3.2	0.42
Mollusc taxa	242	15	1.2	1.0	0.6	0.52
Simpson's diversity	240	0	0.8	0.8	0.1	0.16

Table 3. Spearman correlations among benthic invertebrate variables (N = 242).

	Total.N	Oligo.N	Insect.N	Mollusc.N	Tot.taxa	Oligo.taxa	Insect.taxa	Mollusc.taxa
Total.N								
Oligochaete.N	0.43							
Insect.N	0.96	0.37						
Mollusc.N	0.52	0.17	0.35					
Total.taxa	0.74	0.47	0.71	0.34				
Oligochaete.taxa	0.34	0.84	0.29	0.10	0.49			
Insect.taxa	0.71	0.29	0.73	0.24	0.91	0.26		
Mollusc.taxa	0.20	0.06	0.12	0.41	0.29	0.01	0.07	
Simpson's div.	-0.12	0.14	-0.08	-0.12	0.29	0.20	0.29	-0.05

Table 4. Medians of Total N (abundance/m^2) by station and year (N = 242).

Note: Shading denotes "Impact" periods.

Station	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL
2006				1109		674	1000		2620	2043	1163	11848
2007				848		978	870		1609	1326	1130	4478
2008	2935	6717	1891	1239		457	761		5261	1326	1674	587
2009	4761	174	326	1130	1348	804	848	1196	1826	1174	2022	1478
2010	8217	1000	457	630	1043	217	630	935	1152	1435	1587	1065

Table 5. Medians of Total taxa counts by station and year (N = 242).

Note: Shading denotes "Impact" periods.

Station	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL
2006				12		8	9		11	13	12	11
2007				12		10	10		10	11	9	13
2008	16	10	9	14		7	10		15	9	10	7
2009	20	6	7	15	11	7	7	11	12	9	11	10
2010	22	10	8	7	8	5	7	11	10	11	9	11

Table 6. Medians of Simpson's diversity by station and year (N = 242).

Note: Shading denotes "Impact" periods.

Station	BAP	BBD	BPJ	INUG	PDL	SP	TE	TEFF	TPE	TPN	TPS	WAL
2006				0.83		0.76	0.68		0.76	0.86	0.87	0.54
2007				0.86		0.82	0.77		0.81	0.81	0.83	0.72
2008	0.54	0.46	0.57	0.86		0.81	0.83		0.72	0.66	0.78	0.79
2009	0.89	0.88	0.90	0.88	0.81	0.77	0.79	0.84	0.83	0.78	0.84	0.79
2010	0.89	0.82	0.85	0.77	0.77	0.76	0.74	0.81	0.81	0.75	0.77	0.78

Table 7. Summary of "before-period" mixed-effects model estimates by variable and impact station (paired with control INUG).

Notes: Station means are in raw units. Effect sizes (ES) are shown in raw and log units for ES = 20% and 50% reductions in mean. Estimates of the standard deviations (SD) for random-effects terms (Year and Year-by-Station, Y x S) and residuals errors are for log-transformed data.

			ES = 20%		ES = 50%		SD (log data)		
Variable	Station	Mean	ES(raw)	ES(log)	ES(raw)	ES(log)	Year	YxS	Error
Total N	SP	548	-110	-0.223	-274	-0.693	0.278	0.268	0.447
	TPE	2086	-417	-0.223	-1043	-0.693	0.359	0.246	0.409
	TPN	1201	-240	-0.223	-600	-0.693	0.000	0.098	0.562
Total taxa	SP	6.6	-1.3	-0.223	-3.3	-0.693	0.214	0.000	0.310
	TPE	11.0	-2.2	-0.223	-5.5	-0.693	0.141	0.000	0.232
	TPN	9.5	-1.9	-0.223	-4.8	-0.693	0.000	0.135	0.298

Table 8. Total abundance: Estimates of BACI statistical power for detecting a significant decrease (one-tailed test) in total abundance/m^2 by station (SP, TPE, TPN) as a function of sampling years (after period) and the number of sub-samples per year.

Notes: Power is shown for two simulated effect sizes (ES = 20% and 50% reductions) and two levels of alpha (0.05 and 0.10). Light shading denotes power ≥ 0.8; dark shading denotes power ≥ 0.9.

Effect			**α = 0.05**			**α = 0.10**		
Size	Years	Samples	SP	TPE	TPN	SP	TPE	TPN
20%	1	3	0.03	0.04	0.05	0.12	0.13	0.16
		5	0.03	0.03	0.03	0.11	0.11	0.12
		10	0.03	0.03	0.05	0.13	0.14	0.21
		20	0.04	0.04	0.05	0.12	0.12	0.18
	2	3	0.05	0.05	0.05	0.15	0.16	0.19
		5	0.06	0.07	0.08	0.18	0.20	0.20
		10	0.09	0.10	0.09	0.18	0.19	0.25
		20	0.06	0.07	0.08	0.20	0.22	0.26
	3	3	0.06	0.07	0.08	0.18	0.19	0.21
		5	0.07	0.08	0.09	0.18	0.19	0.25
		10	0.06	0.07	0.09	0.20	0.23	0.28
		20	0.10	0.11	0.10	0.20	0.24	0.33
50%	1	3	0.15	0.19	0.22	0.36	0.41	0.42
		5	0.16	0.19	0.29	0.41	0.46	0.55
		10	0.22	0.27	0.55	0.47	0.53	0.76
		20	0.22	0.28	0.68	0.46	0.52	0.88
	2	3	0.33	0.40	0.43	0.58	0.66	0.67
		5	0.42	0.48	0.63	0.65	0.73	0.84
		10	0.45	0.55	0.82	0.71	0.76	0.95
		20	0.51	0.59	0.92	0.74	0.79	0.99
	3	3	0.47	0.54	0.62	0.74	0.80	0.80
		5	0.56	0.64	0.76	0.80	0.84	0.92
		10	0.64	0.71	0.91	0.82	0.88	0.98
		20	0.68	0.76	0.99	0.87	0.93	1.00

Table 9. Total taxa: Estimates of BACI statistical power for detecting a significant decrease (one-tailed test) in total taxa counts by station (SP, TPE, TPN) as a function of sampling years (after period) and the number of sub-samples per year.

Notes: Power is shown for two simulated effect sizes (ES = 20% and 50% reductions) and two levels of alpha (0.05 and 0.10). Light shading denotes power ≥ 0.8; dark shading denotes power ≥ 0.9.

Effect			**α = 0.05**			**α = 0.10**		
Size	Years	Samples	SP	TPE	TPN	SP	TPE	TPN
20%	1	3	0.10	0.15	0.07	0.23	0.32	0.19
		5	0.10	0.19	0.05	0.27	0.43	0.17
		10	0.22	0.41	0.08	0.47	0.65	0.22
		20	0.27	0.58	0.07	0.62	0.83	0.21
	2	3	0.16	0.28	0.11	0.35	0.51	0.24
		5	0.23	0.45	0.14	0.49	0.67	0.31
		10	0.33	0.61	0.16	0.61	0.84	0.36
		20	0.43	0.79	0.19	0.76	0.97	0.40
	3	3	0.24	0.43	0.15	0.47	0.64	0.31
		5	0.31	0.56	0.17	0.56	0.78	0.38
		10	0.42	0.72	0.21	0.69	0.92	0.46
		20	0.60	0.94	0.25	0.88	0.99	0.47
50%	1	3	0.71	0.94	0.46	0.88	0.98	0.73
		5	0.88	0.99	0.57	0.97	1.00	0.81
		10	0.99	1.00	0.74	1.00	1.00	0.90
		20	1.00	1.00	0.79	1.00	1.00	0.93
	2	3	0.96	1.00	0.82	0.99	1.00	0.92
		5	0.99	1.00	0.92	1.00	1.00	0.97
		10	1.00	1.00	0.96	1.00	1.00	1.00
		20	1.00	1.00	0.98	1.00	1.00	1.00
	3	3	0.99	1.00	0.94	1.00	1.00	0.99
		5	1.00	1.00	0.99	1.00	1.00	1.00
		10	1.00	1.00	0.99	1.00	1.00	1.00
		20	1.00	1.00	1.00	1.00	1.00	1.00

Figure 1. Box-plots of Total abundance for August benthic samples (stations INUG, SP, TPE, TPN).

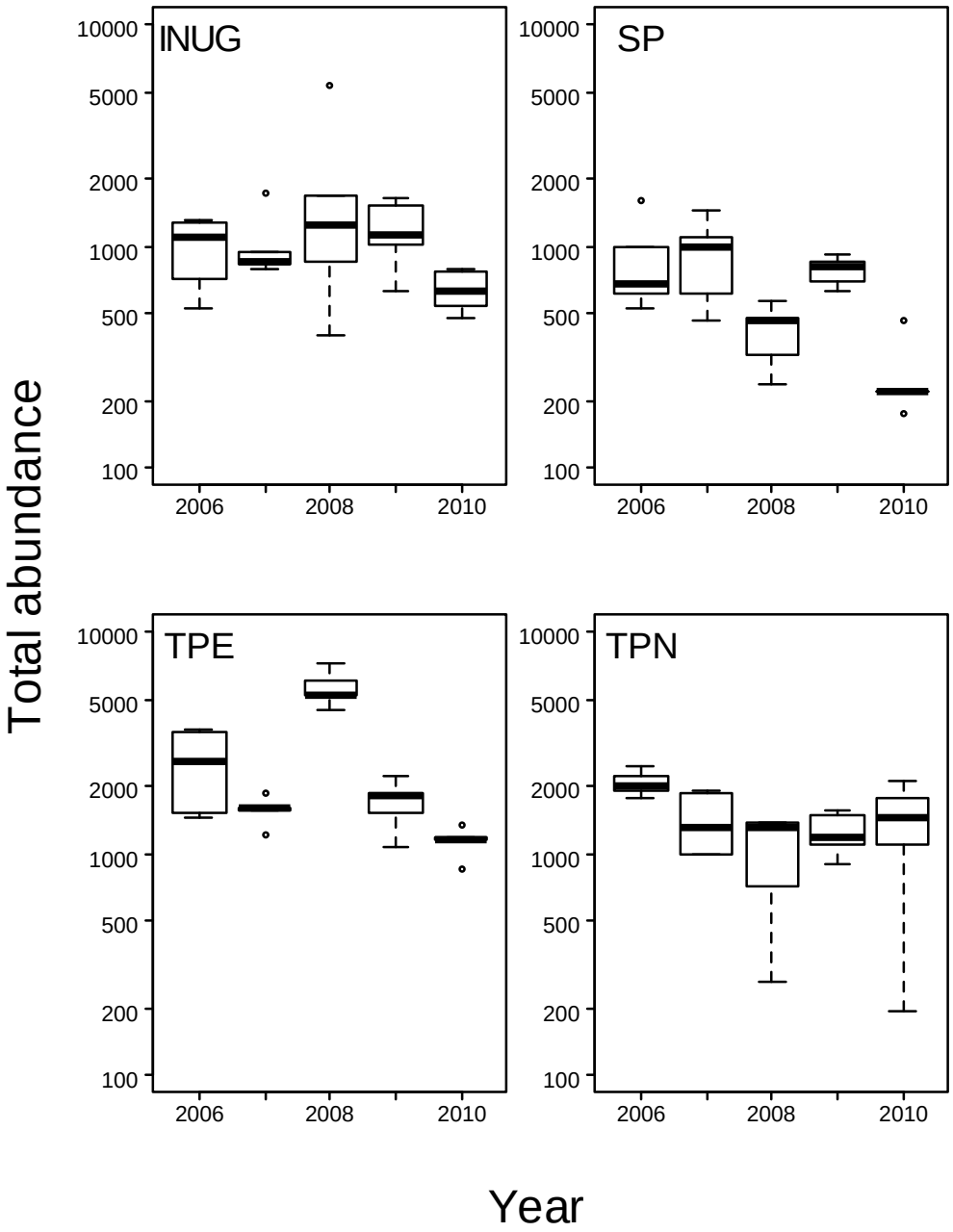


Figure 2. Box -plots of total taxa for August benthic samples (stations INUG, SP, TPE, TPN).

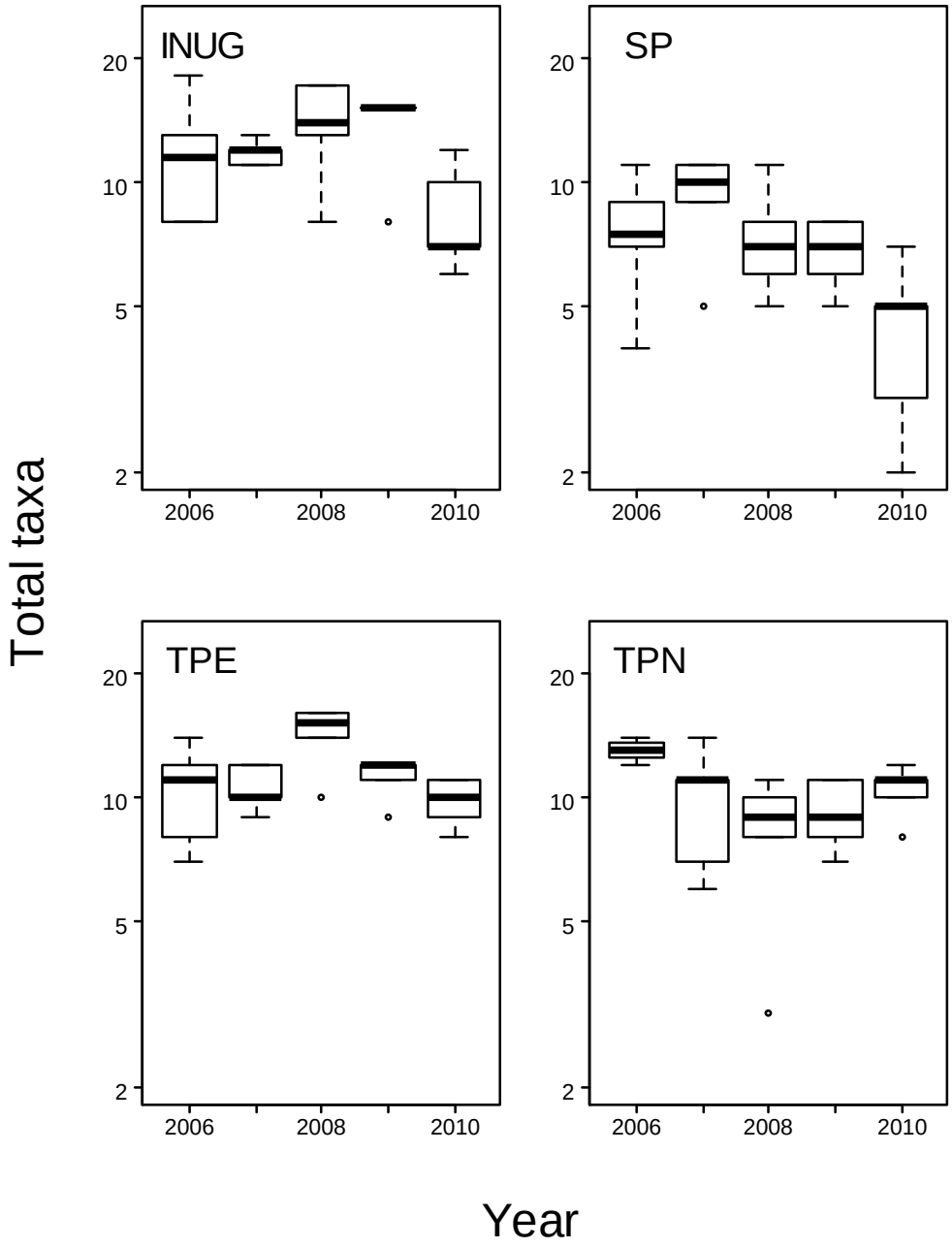


Figure 3. Box-plots of Simpson's diversity for August benthic samples (stations INUG, SP, TPE, TPN).

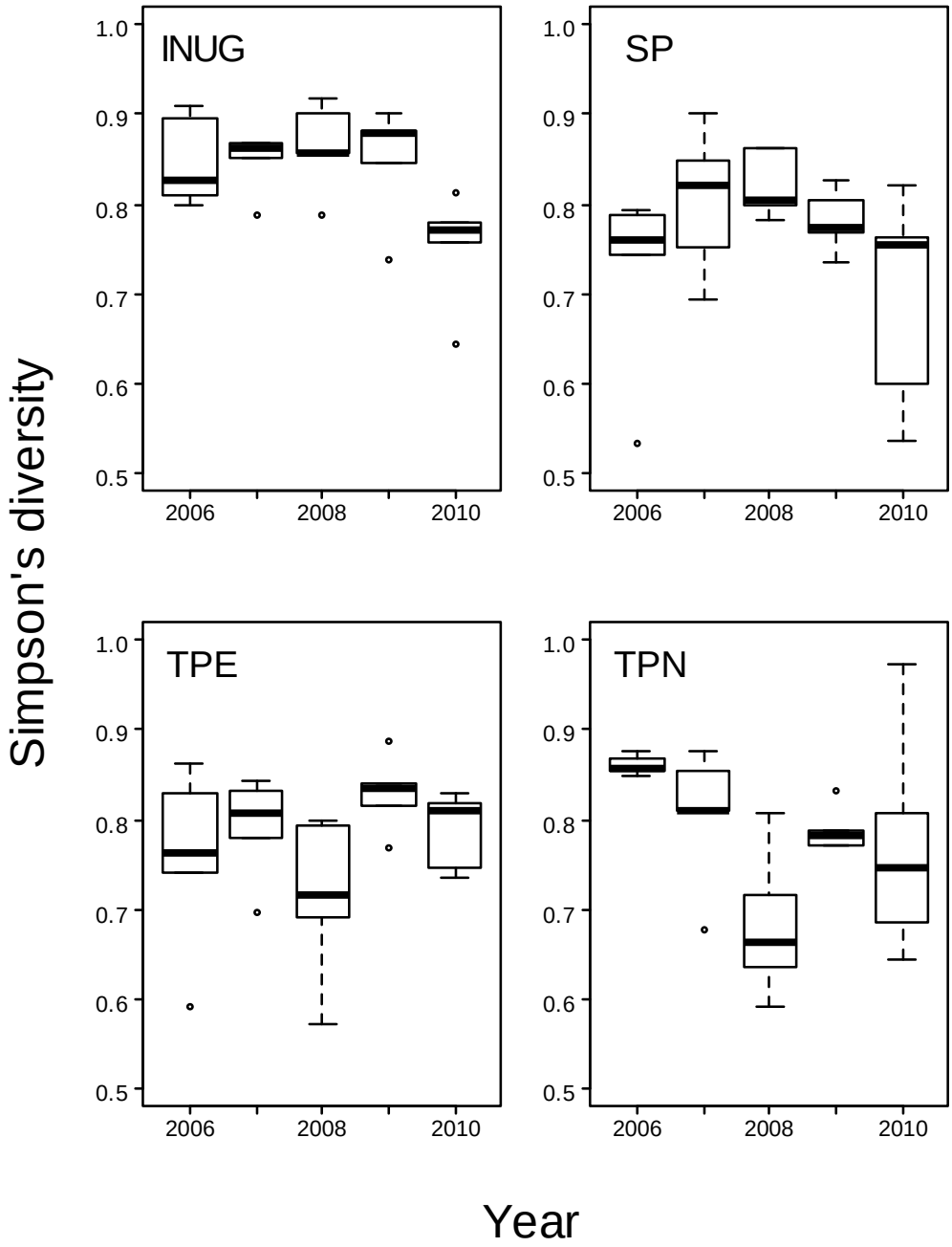


Figure 4. Estimates of BACI power for detecting a significant decrease (one-tailed, alpha = 0.05) in a given variable (Total N or Total taxa) and effect size (rows) by station (columns) as a function of sampling years (after period) and the number of sub-samples per year (see legend).

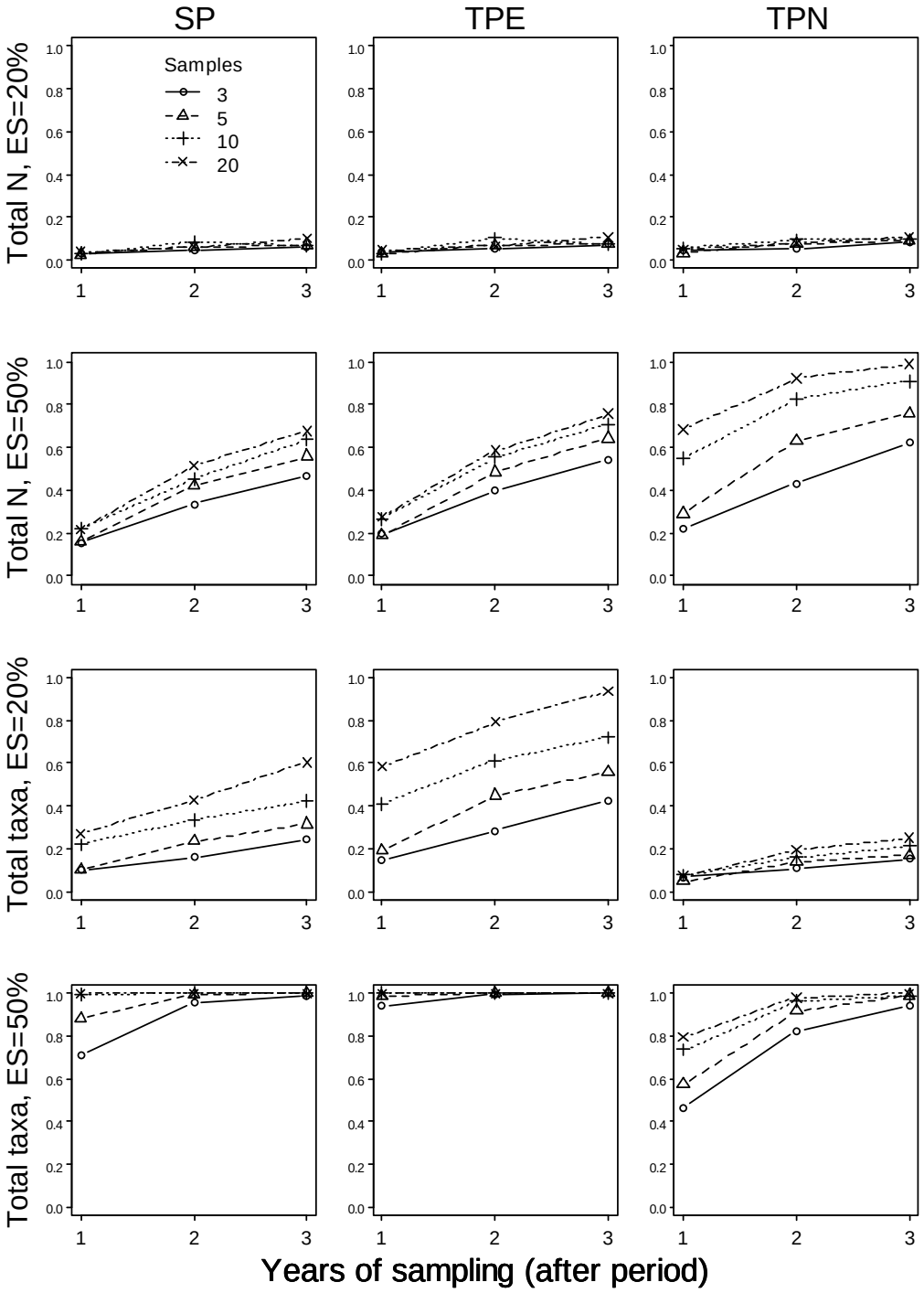
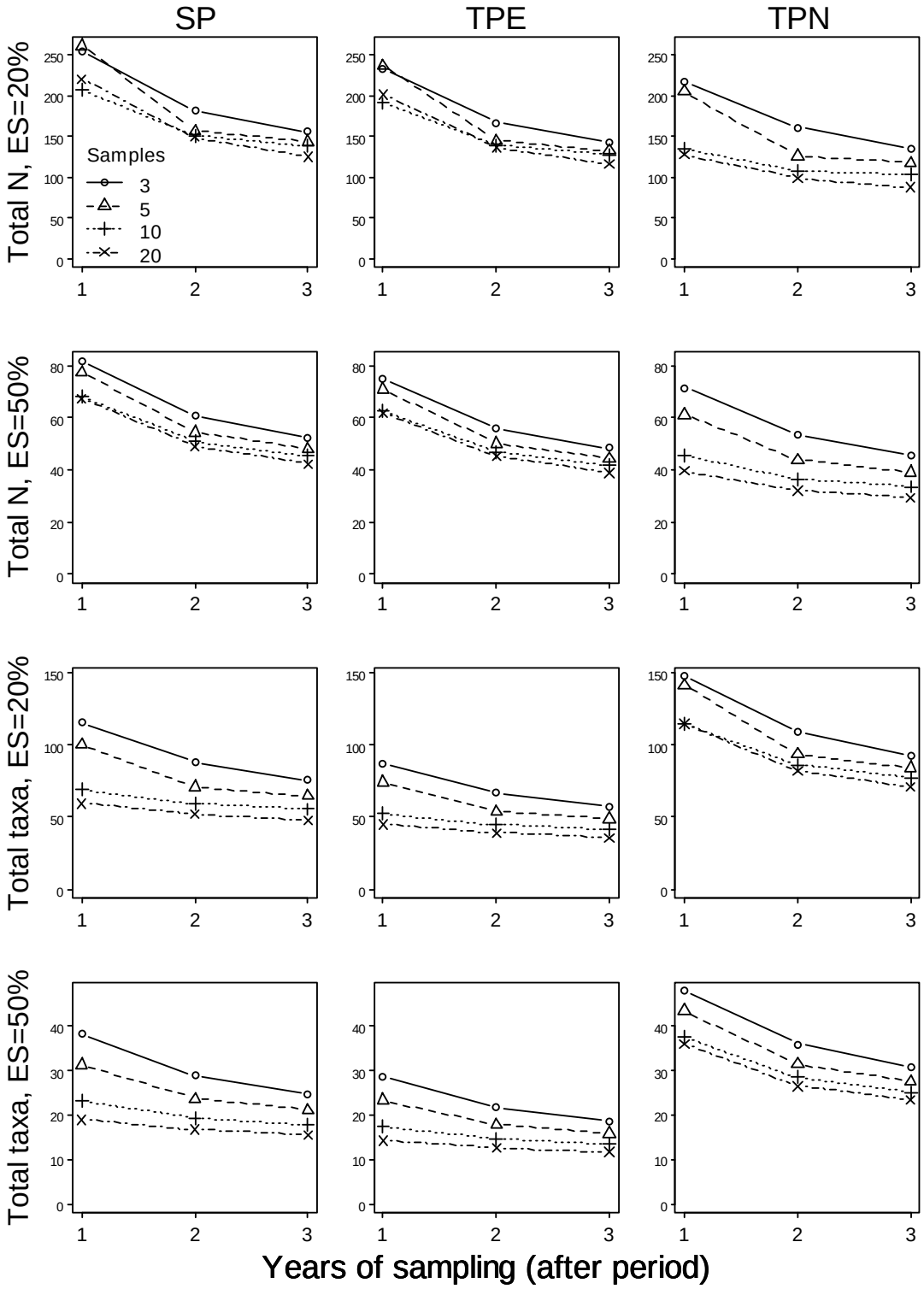


Figure 5. Coefficients of variation (%) for BACI estimates by variable (row) and station (column) as a function of sampling years (after period) and the number of sub-samples per year (see legend).



APPENDIX E – STATISTICAL ANALYSES FOR ZOOPLANKTON

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This appendix contains the following analyses:

- Summary of CREMP zooplankton data
- Analysis of sampling design

Some text, tables and figures are repeated from the main document so that the analyses contained in this appendix can be read and understood without reference to the main document. The material in this appendix assumes understanding of basic statistical methods (Venables and Ripley 2002), mixed-effects models (Pinheiro and Bates 2000), before-after-control-impact (BACI) experimental design (Stewart-Oaten et al. 1986; Underwood 1994; Smith 2002), and use of simulation in statistical analysis (Gelman and Hill 2006).

2. SUMMARY AND EVALUATION OF CREMP ZOOPLANKTON DATA

The zooplankton data for the CREMP include 45 samples collected in August 2010 (5 spatial replicates at each of the 9 Meadowbank stations) and 90 samples collected in August 2011 (10 spatial replicates at each of the 9 stations). Three variables were examined (across all taxa): wet biomass (mg/m^3); dry biomass (mg/m^3); and abundance (total counts). The data can be summarized as follows:

Variable	N	N=0	Mean	Med	SD
wet.biomass	135	0	687	519	550
dry.biomass	135	0	27	25	13
abundance	135	0	7391	5701	6637

Across all samples (N = 135), these variables were moderately correlated (Spearman r):

Samples	wet.biomass	dry.biomass	abundance
wet.biomass			
dry.biomass	0.76		
abundance	0.53	0.64	

Figure 1 shows box-plots by station for each variable, while **Table 1** provides medians (raw data; same as thick line in each box-plot). As demonstrated in **Figure 1** and **Table 1**, there is considerable variation among stations (indicating the usefulness of the data to identify differences) but also among replicates (indicating the level of noise or sampling error in the data). There are general patterns as well as anomalies. All three measures

were generally higher across stations in 2011, with relatively high measures for INUG and WAL in both years. Values for PDL appear the least consistent: relatively low biomass in 2010, high biomass in 2011, but similar abundances for both years (**Figure 1**; **Table 1**).

An important consideration in the experimental design (discussed in next section) is the extent to which variation in a given zooplankton measure is due to differences among stations, years, and replicates. To estimate these components of variation, we fit a mixed-effects model to each variable of the form:

$$X_{jks} = \beta_j + \tau_k + (\tau\beta)_{kj} + \varepsilon_{jks}$$

where all Station (β) and Year (τ) terms (including the interaction term $\tau\beta$) were treated as random effects, X is the measured value of subsample s at Station j in Year k, and ε is the residual error for subsamples.

There were notable differences among variables in the results (**Table 2**). For wet biomass, little of the variation (7%) was accounted for by Station (i.e., differences among stations that were consistent in both years), while roughly equal proportions of variation (~30%) were accounted for by Year (differences between the two years), Station x Year (differences specific to each station-year combination), and Error (the residual variation among replicate samples). In contrast, there were much larger Station effects (>30%) and lower Station x Year effects (< 5%) estimated for dry biomass and wet biomass. For the experimental design explored in the next section, we would expect relatively low power to detect change for wet biomass due to high Station x Year variation (SD = 0.42), and the highest power to detect change for dry biomass abundance given low Station x Year variation (SD = 0.11). These results are uncertain given that there are only two years of data.

3. ANALYSIS OF SAMPLING DESIGN

Impact hypotheses and statistical design – Two general classes of impacts are hypothesized for the Meadowbank mine for the case of zooplankton:

1. Pulse events for which potential impacts would be high for a short time but may then (for zooplankton) dissipate. Pulse events could be associated with any phase of the mine, but are more likely to be associated with particular activities such as dike construction.

2. Long-term cumulative impacts that may be associated with ongoing activities. Long-term cumulative impacts are more likely to be associated with ongoing activities of mine operations.

As operations have just begun in 2010, the focus of CREMP monitoring to date has been on detecting pulse events associated with construction. The appropriate framework for analysis is a before-after-control-impact (BACI) that is aimed at detecting a potential impact in a particular lake or basin in a particular time period. The BACI framework can also be used to evaluate long-term impacts, but other tools such as time series regression analysis may also be appropriate for evaluating long-term trends. For this design document, we focus on the use of the BACI framework, recognizing that other tools such as time series regressions may be useful at a future date once sufficient time series data are available[1].

The classic BACI (paired) design has before/after periods α_i (i = B, A; I = 2), control/impact sites β_j (j = C, I; J = 2), and a total of K paired sampling times τ_k that are nested within period. A statistical model for this design is given by (Smith 2002, equation 2):

$$(1) \qquad X_{ijk} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

The key term is the interaction ($\alpha\beta$), which can be tested using an F test with F = MS[($\alpha\beta$)]/MS[Resid] and degrees of freedom = 1, K – 2. As discussed by Smith (2002), this is equivalent to simply taking the differences between the control and impact values across times and using a two-sample (before-after) t test (Stewart-Oaten et al. 1986).

Model (1) can be extended to include additional control sites (e.g., "asymmetric" designs; Underwood 1994) and/or additional impact sites. To be valid, the additional sites must be replicates rather than subsamples (i.e., as controls, they should be spatially independent of each other but representative of the impact sites, while replicates for impacts need to be spatially independent and (ideally) affected by independent disturbances). So whereas j = (C, I) in the classic BACIP, j may compose any combination of J total sites, for example J = 4 where j = (C_1, C_2, C_3, I). The general test of ($\alpha\beta$) still applies, but with degrees of freedom = (J – 1), (K – 2)(J – 1) (e.g., see Table 1 of Underwood (1994) and Table 9 of Smith (2002)).

In addition, there may be n replicate subsamples s at each site/time combination (jk), as assumed in Table 1 of Underwood (1994). In this case, we modify equation (1) as:

$$(2) \qquad X_{ijks} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + (\tau\beta)_{k(i)j} + \varepsilon_{ijks} \ ,$$

[1] In theory, a BACI analysis that is appropriately framed should be capable of detecting changes associated with long-term trends.

where subsamples now permit estimation of times-by-site interactions ($\tau\beta$). The appropriate F ratio for ($\alpha\beta$) is now F = MS[($\alpha\beta$)]/MS[($\tau\beta$)] with df = (J – 1), (K – 2)(J – 1). As Underwood demonstrates, specific comparisons (interaction terms, such as the impact site versus either "period" or a specific "time" unit) can be examined by partitioning variation accordingly (e.g., Underwood Table 2[2]).

Methods – To test for potential impacts on zooplankton in future years, a BACI design is warranted given natural inter-annual variation that is station-specific to some extent. The analysis assessed the expected precision and power of BACI estimates for different after-period (impact) durations and different numbers of sub-samples (random spatial replicates collected at the same station during the sampling event each August). Separate analyses were conducted for the three primary impact stations SP, TPE, and TPN, compared against either one control (INUG) or two controls (INUG and PDL). Separate analyses were conducted for wet biomass, dry biomass and abundance (all data log-transformed). For each variable, the effect size or ES (fixed across months) was set at either a 20% reduction from baseline or a 50% reduction in baseline.

For purposes of the BACI analyses, the two years (2010-2011) of available control-impact paired data were used to represent the "before period" dataset. We also examined scenarios in which one or two additional "before" years of data were simulated prior to impacts, providing three scenarios (before years = 2, 3, and 4). The additional before-period data were simulated with variances equal to those estimated for observed data across all stations (**Table 2**), using the model specified in **Section 2** above[3]. After-period data were simulated using after-period means (β_control, β_impact + **ES**) for three durations (1, 2, and 3 years) and three sub-sample scenarios (5, 10, and 20 replicates per station per year), again using variances derived from the observed data (**Table 2**). In the case of the BACI model with two control stations (INUG and PDL), an additional random-effects term (station) is added to the model to allow for differences in means between control stations. For each scenario, 500 simulations were used. Power was computed for one-tailed tests (based on the a priori assumption that impacts will reduce biomass or abundance) using two alpha levels (0.05 and 0.10).

Results for Single Control Station – Estimates of statistical power for the case of a single control (INUG) are reported in **Tables 3-5** and displayed in **Figures 2-7**. The tables report power for both alpha levels (0.05 and 0.10) but omit results for "Before" years = 2 (very low power). The figures only show power for alpha = 0.05 but include "Before" years = 2. Power was low for wet biomass across all scenarios (**Table 3**; **Figures 2-3**).

[2] We note that there are mistakes in the tables presented by Underwood (1994).

[3] Data were insufficient to justify estimation of variances specific to each impact station paired to control (as was done for water chemistry), therefore variances were derived based on pooled data for all stations.

High power (> 0.8) was estimated for dry biomass for ES = 50% under various scenarios (**Table 4**; **Figures 4-5**). These included conditions not shown: "Before" years = 2, "After" years = 2, and alpha = 0.10 (power > 0.8 for all stations and sample replicates). Two key points are illustrated by the results for dry biomass (ES = 50%). First, when "Before" years = 2 (i.e., the current data), it is virtually impossible to detect a difference with only one "After" year of impact data. In this scenario, the degrees of freedom (DF) associated with the estimated effect is DF = 1, and the t-statistic must exceed 6.3 for a one-tailed test with alpha = 0.05. However, with an extra before year or after year, DF = 2 and the critical t-value drops to 2.9. So at these low sample sizes (years), power is driven as much or more by degrees of freedom than by improved precision. Second, increasing sample replication (n) from 5 to 10 appears important to improve power in certain circumstances. For example, high power is achieved for n = 10 (but not n = 5) when before years = 3, after years = 1, and alpha = 0.10, or when before years = 3, after years = 2, and alpha = 0.05. Power was intermediate for abundance (**Table 5**, **Figures 6-7**). With few exceptions, high power was only achieved for ES = 50% when the combination of before/after years totaled 5 or more years and alpha = 0.10 (**Table 5**).

Results for Two Control Stations – In theory, adding a second control station (PDL) can improve power by increasing sample sizes. In particular, the degrees of freedom for testing the BACI interaction increases because additional station-by-year units exist for estimating this random-effects component (the denominator in the F-test of the BACI interaction). However, data for PDL exhibit strong differences between 2010 and 2011 (especially for wet and dry biomass), and such Station x Year effects (high variance term) may offset or reduce power despite increases in sample size. Estimates of BACI statistical power for the case of two controls case are reported in **Table 6-8** and displayed in **Figures 8-13**. Again, the tables report power for both alpha levels (0.05 and 0.10) but omit results for "Before" years = 2, and the figures only show power for alpha = 0.05 but include "Before" years = 2. As before, power was low for wet biomass across all scenarios (**Table 6**; **Figures 8-9**). For dry biomass and abundance, power was generally higher under the two-control test than the single-control case (more specifically, there was a slightly broader range of scenarios under which high power was achieved). For example, power was greatly improved for dry biomass (ES = 50%) when "Before" years =2 and "After" years = 1 (compare top panels **of Figures 5 and 11**). For these scenarios, power was roughly 0.8 or greater when $\alpha = 0.10$ and sample replication = 10 or 20.

Conclusions – Estimates of power for short-term experimental designs are most promising for dry biomass (ES = 50%), in particular with n = 10 or more replicates, but less so for abundance and particularly for wet biomass. The key factor driving these results was the observed/simulated variances for Station x Year effects, which was lowest for dry biomass (**Table 2**). If this variance term is underestimated, then estimates of power are too optimistic. For example, the results for PDL are concerning (**Table 1** and

Figure 1) because between-year fluctuations in dry biomass were much greater than observed across other stations. However, incorporating PDL as a second control (with 2010 and 2011 data present in the "before" period) did not adversely affect estimates for power for dry biomass or abundance.

4. REFERENCES

Gelman, A, and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models – Analytical Methods for Social Research. Cambridge University Press, New York, 625 pp.

Pinheiro, J., and D. Bates. 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York, 528 pp.

Smith, E. 2002. BACI design. Pages 141-148, in: A.H. El-Shaarawi and W.W. Piegorsch [Eds], Encyclopedia of Environmetrics, John Wiley & Sons, Chicherster.

Stewart-Oaten, A., W. Murdoch, and K. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67: 929-940.

Venables, W., and B. Ripley. 2002. Modern Applied Statistics. Springer, New York, 495 pp.

Underwood, A. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications 4(1): 3-15.

Table 1. Summary of medians (raw data) by station and year for zooplankton variables collected in August 2010 (5 samples per station) and August 2011 (10 samples per station).

	Wet biomass		Dry biomass		Abundance	
Station	2010	2011	2010	2011	2010	2011
INUG	834	932	45.1	41.6	3752	14120
PDL	186	1485	8.0	25.7	1298	1591
SP	358	528	21.2	27.9	2500	9404
TE	298	420	17.7	17.3	1944	6423
TEFF	303	698	12.4	21.2	2488	8496
TPE	194	554	11.5	29.6	4672	10942
TPN	325	433	15.9	23.0	1672	5496
TPS	211	497	21.2	27.4	1458	5639
WAL	528	1251	36.3	38.9	6976	16533
Mean	360	756	21.0	28.1	2973	8738
SD	207	385	12.2	7.9	1865	4632
CV	0.57	0.51	0.58	0.28	0.63	0.53

Table 2. Estimates of standard deviation (SD) and percent of variation explained for each term in the mixed-effects models fit to (log) zooplankton variables.

	SD Estimate (log data)			Percent of variation		
Term	Wet biomass	Dry biomass	Abundance	Wet biomass	Dry biomass	Abundance
Station	0.21	0.33	0.64	7%	31%	35%
Year	0.48	0.28	0.70	36%	23%	42%
Station x Year	0.42	0.11	0.21	28%	3%	4%
Error (reps)	0.44	0.39	0.46	30%	43%	19%
Sum				100%	100%	100%

Table 3. Wet biomass: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year. Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10).

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.19	0.06	0.09	0.38	0.18	0.22
			10	0.20	0.06	0.11	0.38	0.23	0.29
			20	0.25	0.12	0.16	0.42	0.26	0.31
	3	2	5	0.16	0.08	0.10	0.27	0.20	0.22
			10	0.16	0.11	0.13	0.31	0.22	0.24
			20	0.16	0.13	0.14	0.30	0.25	0.27
	3	3	5	0.15	0.11	0.11	0.29	0.24	0.24
			10	0.17	0.13	0.14	0.28	0.24	0.26
			20	0.14	0.11	0.12	0.27	0.22	0.23
	4	1	5	0.19	0.12	0.13	0.35	0.25	0.26
			10	0.22	0.15	0.17	0.32	0.25	0.28
			20	0.22	0.15	0.17	0.36	0.29	0.32
	4	2	5	0.22	0.15	0.16	0.38	0.28	0.30
			10	0.20	0.16	0.18	0.35	0.29	0.30
			20	0.23	0.18	0.19	0.34	0.31	0.32
	4	3	5	0.15	0.12	0.13	0.31	0.26	0.28
			10	0.17	0.14	0.14	0.29	0.27	0.28
			20	0.17	0.14	0.14	0.31	0.29	0.30
50%	3	1	5	0.35	0.14	0.18	0.51	0.38	0.40
			10	0.38	0.19	0.23	0.59	0.42	0.47
			20	0.41	0.23	0.29	0.58	0.45	0.49
	3	2	5	0.38	0.26	0.29	0.56	0.46	0.48
			10	0.40	0.31	0.33	0.57	0.52	0.52
			20	0.41	0.35	0.37	0.58	0.53	0.55
	3	3	5	0.47	0.41	0.43	0.64	0.59	0.60
			10	0.45	0.39	0.40	0.61	0.57	0.58
			20	0.46	0.42	0.43	0.61	0.59	0.59
	4	1	5	0.34	0.23	0.27	0.52	0.41	0.45
			10	0.37	0.28	0.30	0.54	0.45	0.48
			20	0.39	0.31	0.33	0.56	0.49	0.51
	4	2	5	0.46	0.40	0.42	0.62	0.56	0.57
			10	0.44	0.39	0.40	0.59	0.55	0.56
			20	0.43	0.40	0.40	0.60	0.56	0.57
	4	3	5	0.46	0.43	0.44	0.63	0.59	0.60
			10	0.46	0.43	0.44	0.65	0.61	0.62
			20	0.48	0.44	0.45	0.64	0.61	0.62

Table 4. Dry biomass: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year. Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10). Light shading denotes power ≥ 0.8; dark shading denotes power ≥ 0.9.

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.07	0.03	0.00	0.23	0.17	0.11
			10	0.11	0.07	0.05	0.29	0.24	0.20
			20	0.18	0.15	0.11	0.41	0.37	0.33
	3	2	5	0.12	0.09	0.05	0.33	0.26	0.20
			10	0.17	0.15	0.11	0.36	0.33	0.29
			20	0.27	0.25	0.23	0.49	0.46	0.45
	3	3	5	0.23	0.18	0.12	0.41	0.36	0.32
			10	0.25	0.25	0.22	0.46	0.43	0.42
			20	0.34	0.33	0.31	0.52	0.51	0.50
	4	1	5	0.12	0.07	0.05	0.29	0.23	0.18
			10	0.18	0.14	0.12	0.38	0.35	0.31
			20	0.27	0.24	0.22	0.44	0.42	0.41
	4	2	5	0.24	0.19	0.13	0.40	0.37	0.31
			10	0.25	0.23	0.19	0.43	0.39	0.39
			20	0.35	0.33	0.31	0.53	0.50	0.50
	4	3	5	0.25	0.22	0.17	0.43	0.40	0.34
			10	0.30	0.28	0.26	0.51	0.50	0.46
			20	0.41	0.40	0.39	0.58	0.56	0.56
50%	3	1	5	0.51	0.34	0.23	0.76	0.67	0.63
			10	0.72	0.59	0.56	0.90	0.85	0.85
			20	0.85	0.77	0.79	0.96	0.93	0.95
	3	2	5	0.81	0.73	0.67	0.92	0.89	0.90
			10	0.90	0.84	0.85	0.97	0.95	0.95
			20	0.93	0.91	0.92	0.98	0.98	0.98
	3	3	5	0.91	0.87	0.87	0.98	0.96	0.97
			10	0.98	0.96	0.96	1.00	0.99	0.99
			20	0.98	0.98	0.98	1.00	1.00	1.00
	4	1	5	0.64	0.53	0.47	0.82	0.76	0.73
			10	0.81	0.73	0.74	0.92	0.88	0.90
			20	0.89	0.84	0.87	0.96	0.94	0.95
	4	2	5	0.88	0.83	0.83	0.96	0.93	0.94
			10	0.95	0.92	0.92	0.99	0.99	0.98
			20	0.98	0.97	0.97	1.00	1.00	1.00
	4	3	5	0.95	0.92	0.94	0.97	0.97	0.97
			10	0.98	0.97	0.98	0.99	0.99	0.99
			20	0.99	0.99	0.99	1.00	1.00	1.00

Table 5. Abundance: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year. Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10). Light shading denotes power ≥ 0.8; dark shading denotes power ≥ 0.9.

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.08	0.01	0.08	0.23	0.10	0.24
			10	0.10	0.02	0.10	0.26	0.13	0.28
			20	0.16	0.06	0.16	0.36	0.21	0.37
	3	2	5	0.10	0.04	0.10	0.26	0.13	0.27
			10	0.14	0.07	0.15	0.27	0.18	0.29
			20	0.19	0.12	0.19	0.32	0.25	0.34
	3	3	5	0.15	0.07	0.16	0.32	0.21	0.33
			10	0.20	0.09	0.21	0.31	0.27	0.34
			20	0.19	0.11	0.20	0.36	0.29	0.38
	4	1	5	0.12	0.06	0.13	0.27	0.17	0.28
			10	0.15	0.10	0.16	0.29	0.22	0.31
			20	0.21	0.13	0.21	0.35	0.30	0.36
	4	2	5	0.18	0.11	0.20	0.34	0.25	0.35
			10	0.19	0.12	0.19	0.33	0.27	0.34
			20	0.24	0.17	0.25	0.37	0.31	0.38
	4	3	5	0.17	0.11	0.18	0.32	0.24	0.33
			10	0.21	0.14	0.22	0.36	0.28	0.37
			20	0.25	0.20	0.26	0.40	0.37	0.42
50%	3	1	5	0.32	0.10	0.33	0.58	0.37	0.60
			10	0.46	0.20	0.48	0.69	0.54	0.70
			20	0.55	0.37	0.57	0.75	0.63	0.76
	3	2	5	0.53	0.36	0.57	0.74	0.65	0.76
			10	0.61	0.50	0.63	0.78	0.71	0.80
			20	0.67	0.59	0.70	0.82	0.78	0.83
	3	3	5	0.67	0.57	0.69	0.82	0.78	0.84
			10	0.74	0.68	0.76	0.88	0.85	0.90
			20	0.76	0.71	0.79	0.88	0.87	0.90
	4	1	5	0.41	0.27	0.44	0.61	0.51	0.62
			10	0.54	0.41	0.55	0.72	0.65	0.73
			20	0.58	0.50	0.60	0.74	0.70	0.75
	4	2	5	0.64	0.53	0.66	0.81	0.74	0.82
			10	0.71	0.63	0.73	0.84	0.80	0.85
			20	0.76	0.71	0.78	0.88	0.86	0.88
	4	3	5	0.75	0.69	0.78	0.88	0.85	0.89
			10	0.77	0.74	0.79	0.90	0.89	0.91
			20	0.81	0.79	0.83	0.91	0.90	0.92

Table 6. Two controls; Wet biomass: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year. Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10). Light shading denotes power ≥ 0.8; dark shading denotes power ≥ 0.9.

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.00	0.01	0.00	0.05	0.06	0.06
			10	0.01	0.02	0.01	0.05	0.08	0.07
			20	0.00	0.01	0.01	0.06	0.08	0.08
	3	2	5	0.01	0.01	0.01	0.07	0.08	0.08
			10	0.02	0.02	0.02	0.07	0.09	0.09
			20	0.02	0.02	0.02	0.06	0.09	0.08
	3	3	5	0.02	0.03	0.03	0.11	0.13	0.12
			10	0.03	0.04	0.04	0.11	0.13	0.13
			20	0.03	0.05	0.05	0.11	0.13	0.13
	4	1	5	0.02	0.03	0.03	0.07	0.08	0.08
			10	0.02	0.03	0.02	0.06	0.08	0.08
			20	0.01	0.02	0.02	0.08	0.09	0.09
	4	2	5	0.04	0.06	0.06	0.14	0.17	0.16
			10	0.05	0.06	0.06	0.13	0.16	0.15
			20	0.05	0.06	0.06	0.14	0.18	0.17
	4	3	5	0.04	0.05	0.05	0.14	0.16	0.16
			10	0.05	0.06	0.06	0.14	0.16	0.15
			20	0.06	0.06	0.06	0.13	0.16	0.16
50%	3	1	5	0.05	0.08	0.08	0.16	0.21	0.20
			10	0.05	0.08	0.08	0.18	0.24	0.22
			20	0.06	0.09	0.09	0.18	0.24	0.24
	3	2	5	0.14	0.17	0.17	0.34	0.39	0.38
			10	0.14	0.21	0.20	0.36	0.42	0.42
			20	0.15	0.19	0.19	0.37	0.42	0.41
	3	3	5	0.27	0.29	0.29	0.49	0.52	0.52
			10	0.28	0.31	0.31	0.49	0.53	0.52
			20	0.29	0.33	0.33	0.52	0.56	0.56
	4	1	5	0.09	0.12	0.12	0.27	0.31	0.31
			10	0.08	0.13	0.12	0.26	0.30	0.29
			20	0.10	0.15	0.14	0.28	0.32	0.32
	4	2	5	0.23	0.26	0.26	0.41	0.45	0.45
			10	0.25	0.32	0.32	0.45	0.48	0.47
			20	0.28	0.32	0.31	0.45	0.49	0.48
	4	3	5	0.32	0.34	0.34	0.53	0.54	0.54
			10	0.36	0.39	0.39	0.54	0.57	0.56
			20	0.37	0.41	0.40	0.58	0.61	0.60

Table 7. Two controls; Dry biomass: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year. Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10). Light shading denotes power ≥ 0.8; dark shading denotes power ≥ 0.9.

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.05	0.07	0.06	0.16	0.20	0.18
			10	0.11	0.15	0.14	0.23	0.27	0.27
			20	0.18	0.25	0.25	0.34	0.43	0.42
	3	2	5	0.11	0.15	0.13	0.27	0.32	0.29
			10	0.17	0.22	0.19	0.37	0.44	0.41
			20	0.24	0.30	0.30	0.45	0.50	0.50
	3	3	5	0.17	0.20	0.19	0.35	0.41	0.38
			10	0.25	0.29	0.28	0.43	0.47	0.46
			20	0.36	0.41	0.40	0.58	0.61	0.62
	4	1	5	0.12	0.14	0.11	0.22	0.27	0.23
			10	0.13	0.17	0.16	0.31	0.37	0.35
			20	0.22	0.29	0.28	0.42	0.48	0.48
	4	2	5	0.17	0.21	0.18	0.32	0.36	0.35
			10	0.23	0.30	0.29	0.43	0.48	0.47
			20	0.36	0.40	0.39	0.52	0.56	0.57
	4	3	5	0.19	0.23	0.20	0.37	0.42	0.39
			10	0.32	0.36	0.36	0.51	0.55	0.53
			20	0.46	0.52	0.50	0.62	0.65	0.65
50%	3	1	5	0.58	0.65	0.60	0.81	0.84	0.83
			10	0.75	0.80	0.80	0.92	0.94	0.94
			20	0.91	0.94	0.95	0.98	0.99	0.99
	3	2	5	0.88	0.91	0.91	0.98	0.98	0.98
			10	0.96	0.98	0.98	1.00	1.00	1.00
			20	1.00	1.00	1.00	1.00	1.00	1.00
	3	3	5	0.96	0.97	0.97	0.98	0.99	0.99
			10	0.99	0.99	0.99	1.00	1.00	1.00
			20	1.00	1.00	1.00	1.00	1.00	1.00
	4	1	5	0.70	0.75	0.73	0.87	0.90	0.87
			10	0.88	0.90	0.89	0.95	0.97	0.97
			20	0.97	0.98	0.98	0.99	0.99	0.99
	4	2	5	0.92	0.93	0.93	0.97	0.97	0.97
			10	0.97	0.98	0.97	1.00	1.00	1.00
			20	1.00	1.00	1.00	1.00	1.00	1.00
	4	3	5	0.98	0.99	0.99	1.00	1.00	1.00
			10	0.99	0.99	0.99	1.00	1.00	1.00
			20	1.00	1.00	1.00	1.00	1.00	1.00

Table 8. Two controls; Abundance: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year. Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10). Light shading denotes power ≥ 0.8; dark shading denotes power ≥ 0.9.

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.03	0.06	0.06	0.09	0.16	0.14
			10	0.04	0.09	0.09	0.13	0.23	0.19
			20	0.06	0.15	0.14	0.17	0.26	0.22
	3	2	5	0.05	0.10	0.09	0.17	0.22	0.21
			10	0.08	0.13	0.12	0.21	0.26	0.26
			20	0.10	0.17	0.16	0.23	0.29	0.28
	3	3	5	0.10	0.14	0.14	0.23	0.27	0.27
			10	0.14	0.17	0.18	0.28	0.30	0.31
			20	0.15	0.18	0.19	0.32	0.35	0.36
	4	1	5	0.07	0.10	0.10	0.16	0.20	0.21
			10	0.06	0.12	0.10	0.18	0.24	0.23
			20	0.10	0.17	0.16	0.22	0.29	0.28
	4	2	5	0.11	0.15	0.14	0.24	0.28	0.27
			10	0.12	0.18	0.17	0.29	0.34	0.33
			20	0.20	0.23	0.23	0.32	0.35	0.35
	4	3	5	0.13	0.15	0.15	0.25	0.30	0.29
			10	0.17	0.20	0.20	0.32	0.35	0.36
			20	0.20	0.22	0.23	0.39	0.42	0.42
50%	3	1	5	0.23	0.36	0.33	0.49	0.60	0.58
			10	0.32	0.47	0.44	0.57	0.68	0.67
			20	0.41	0.60	0.55	0.65	0.77	0.76
	3	2	5	0.53	0.63	0.62	0.74	0.79	0.78
			10	0.67	0.77	0.74	0.83	0.87	0.86
			20	0.75	0.81	0.80	0.88	0.91	0.90
	3	3	5	0.74	0.78	0.77	0.88	0.91	0.91
			10	0.80	0.84	0.83	0.92	0.94	0.94
			20	0.86	0.88	0.87	0.94	0.95	0.96
	4	1	5	0.38	0.47	0.45	0.60	0.67	0.66
			10	0.48	0.59	0.57	0.72	0.78	0.77
			20	0.60	0.70	0.68	0.78	0.84	0.82
	4	2	5	0.62	0.70	0.69	0.80	0.84	0.84
			10	0.74	0.78	0.78	0.86	0.88	0.88
			20	0.81	0.85	0.84	0.92	0.93	0.93
	4	3	5	0.72	0.76	0.75	0.86	0.89	0.88
			10	0.82	0.85	0.85	0.93	0.94	0.94
			20	0.89	0.91	0.90	0.95	0.95	0.95

Figure 1. Box-plots of zooplankton variables by station collected in August 2010 (5 samples per station) and August 2011 (10 samples per station).

Note: Dashed lines denote means across all samples

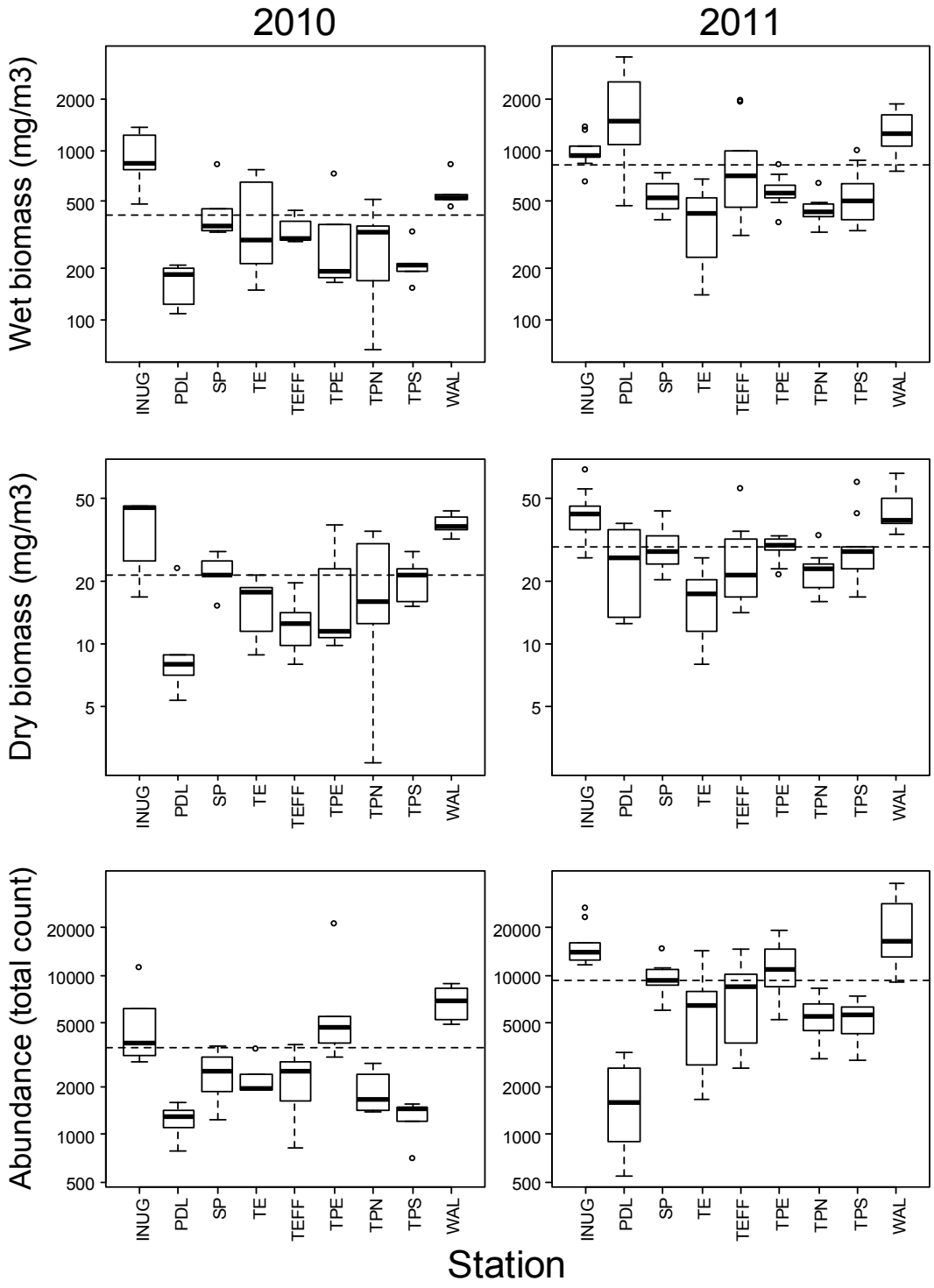


Figure 2. Wet biomass, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

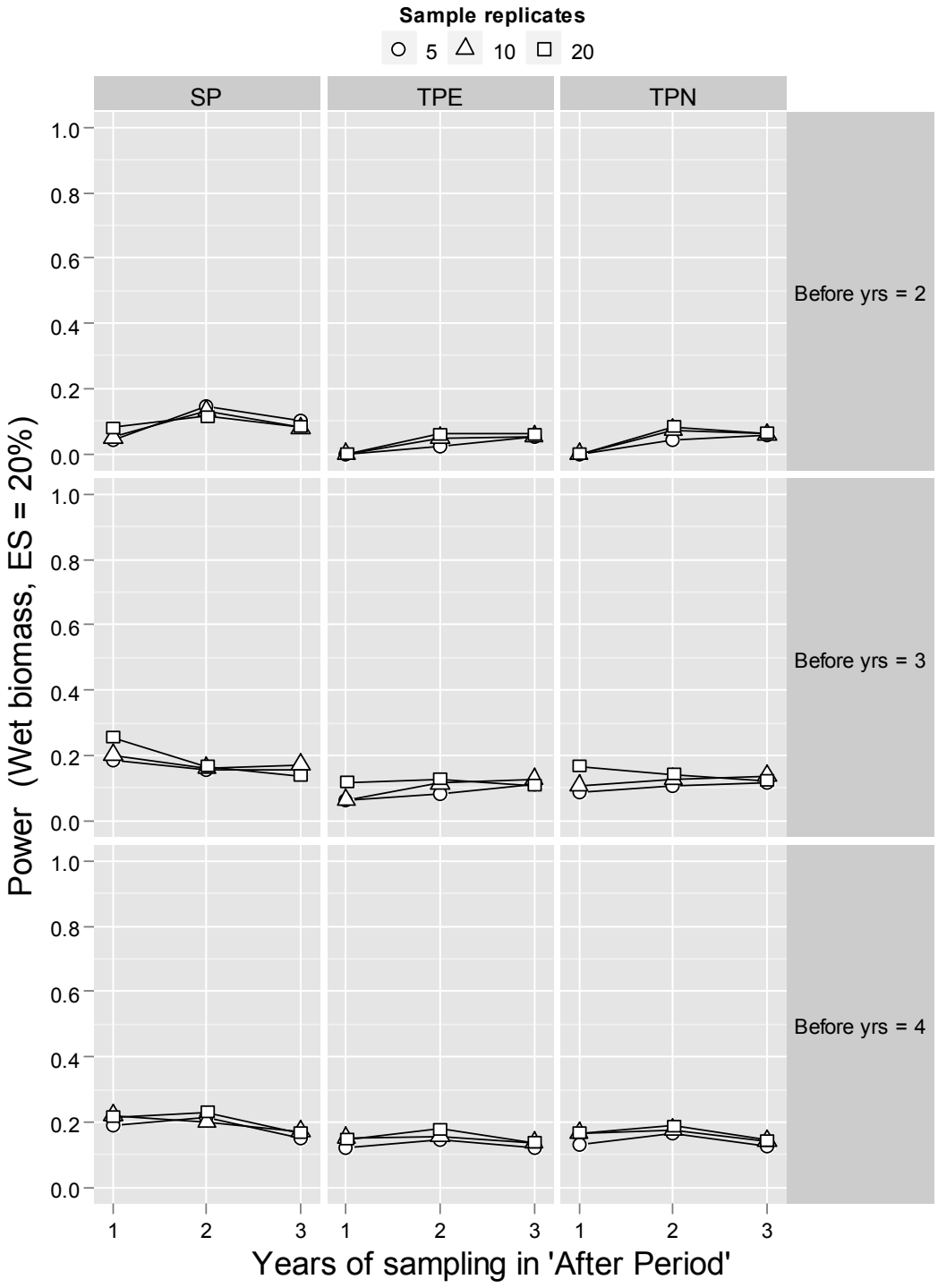


Figure 3. Wet biomass, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

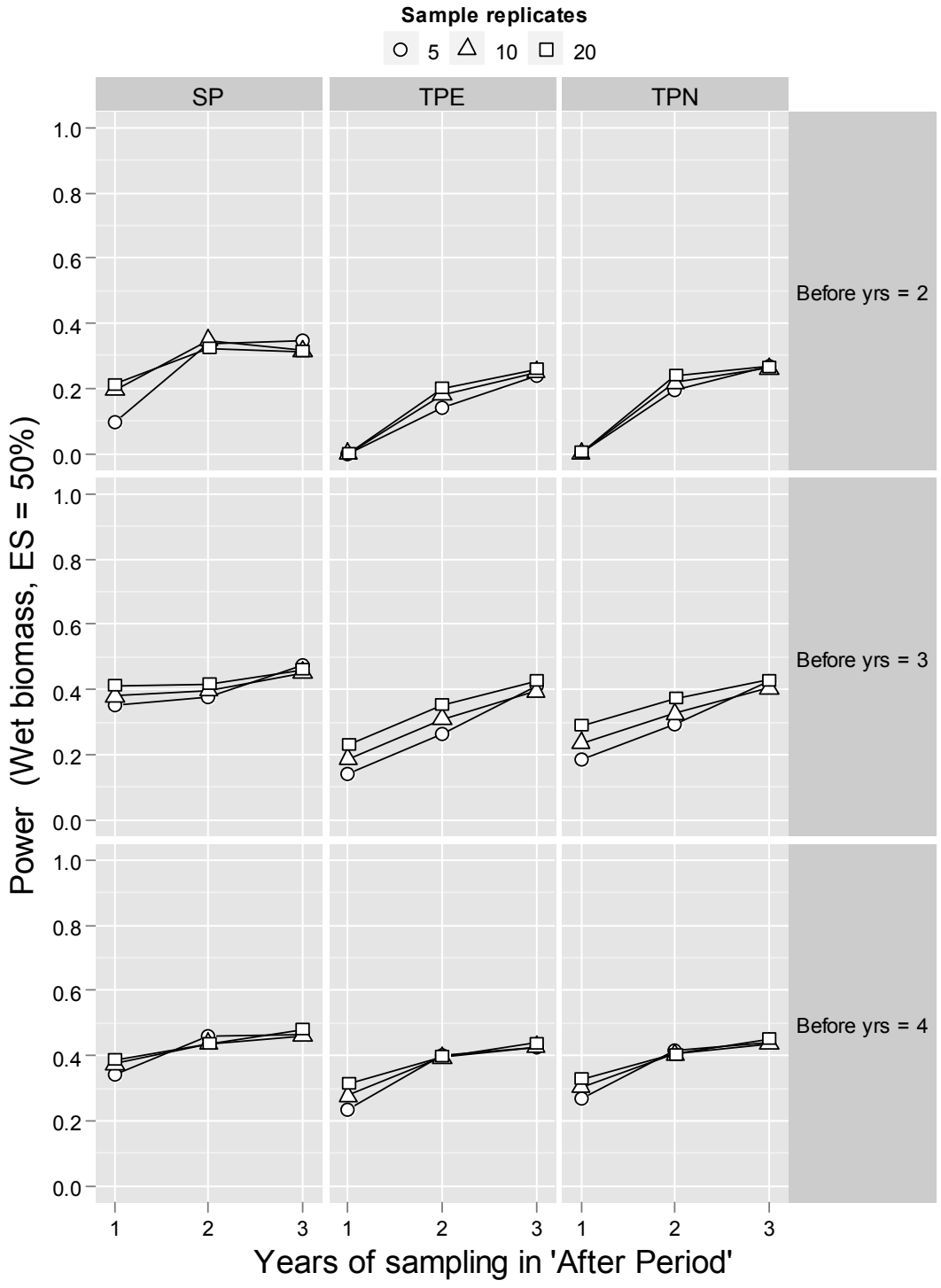


Figure 4. Dry biomass, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

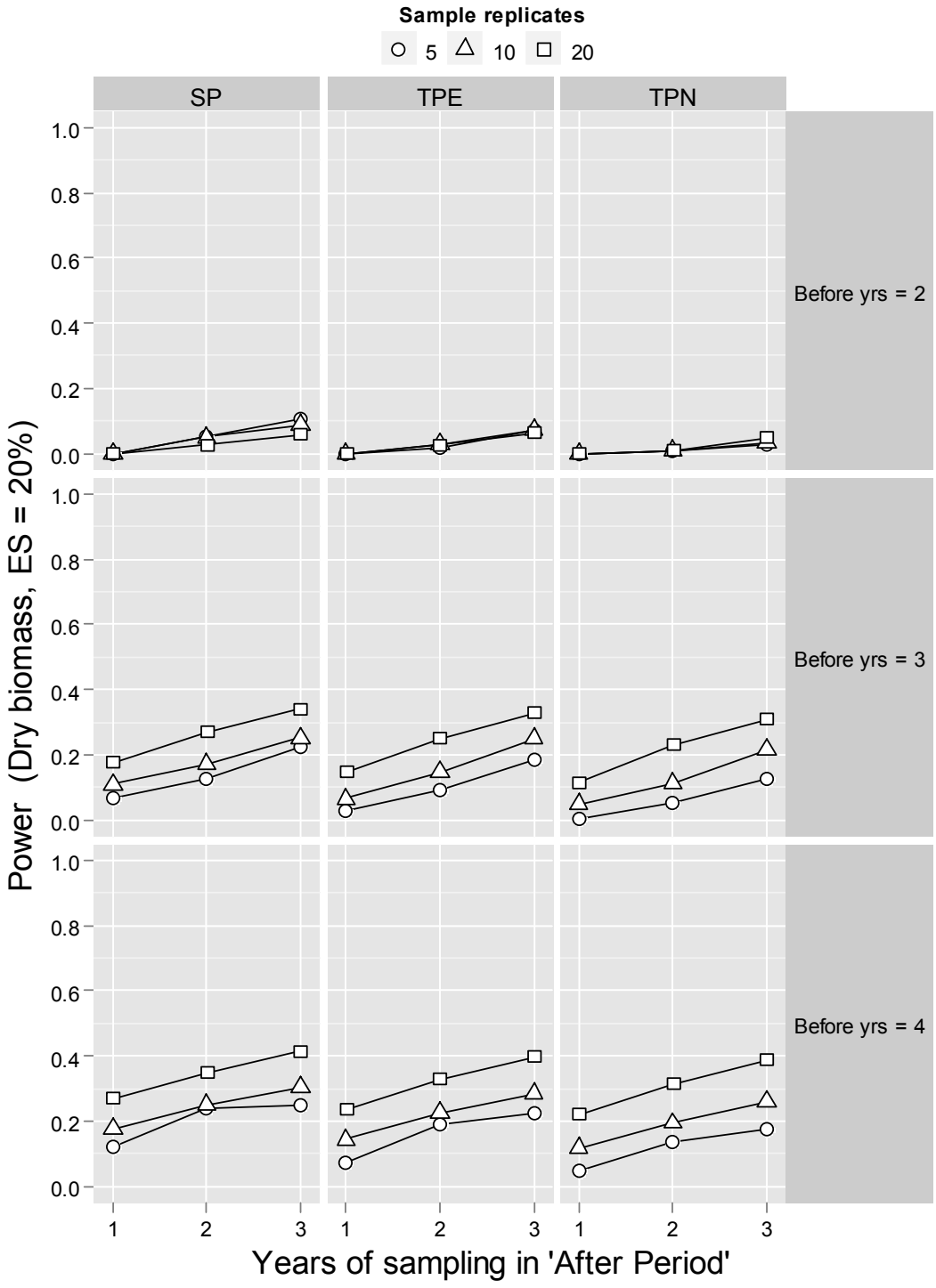


Figure 5. Dry biomass, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

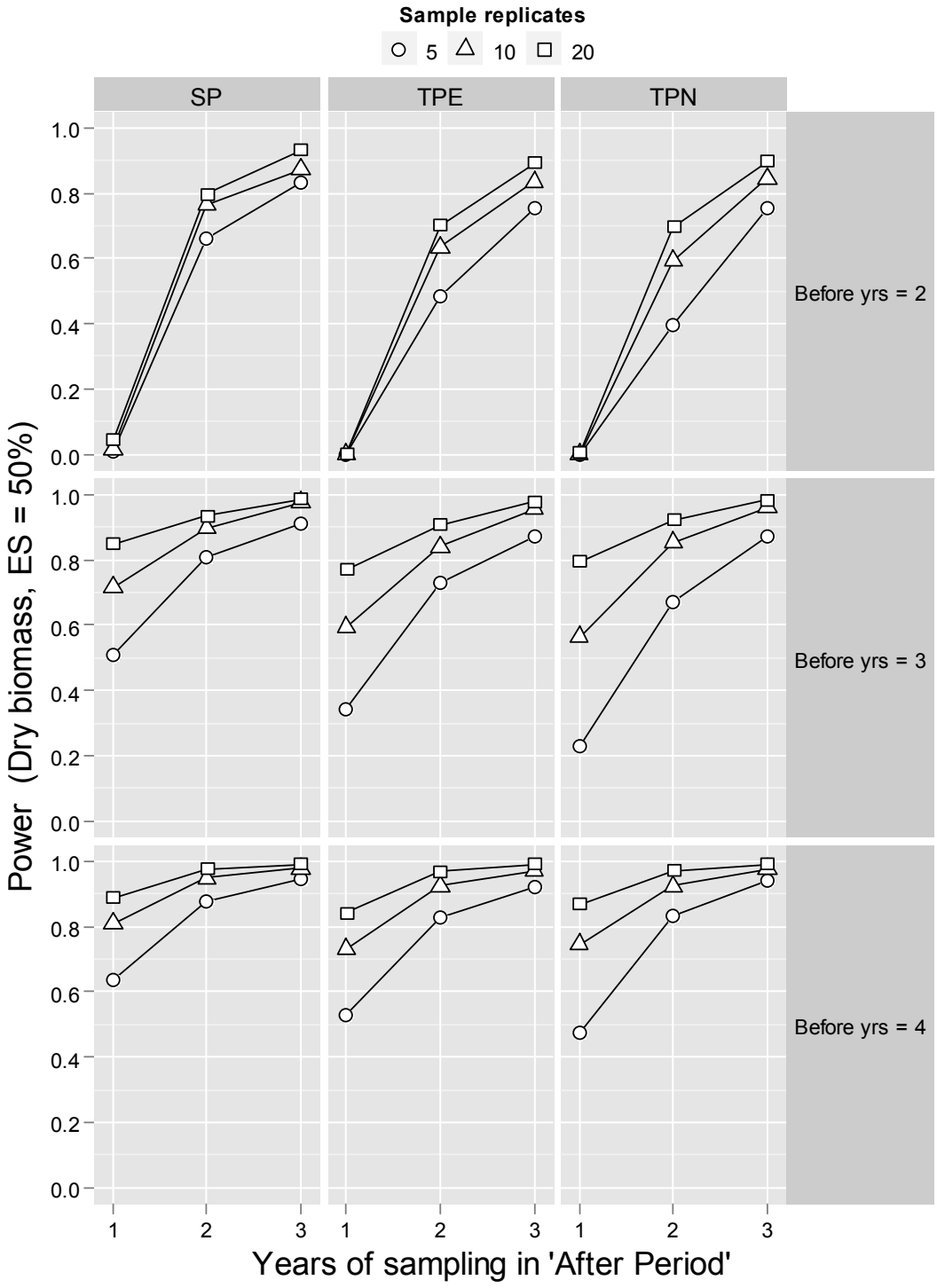


Figure 6. Abundance, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

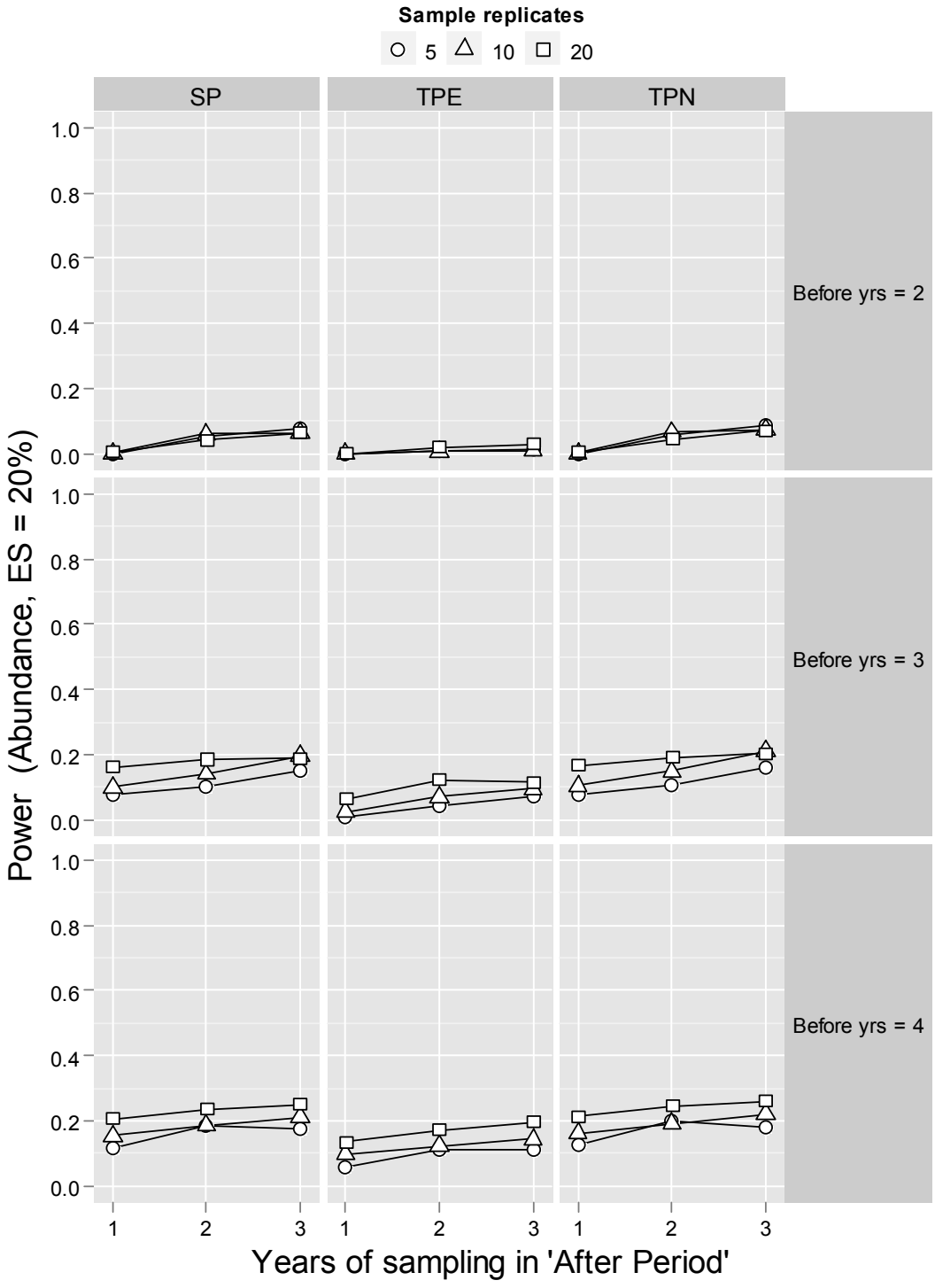


Figure 7. Abundance, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

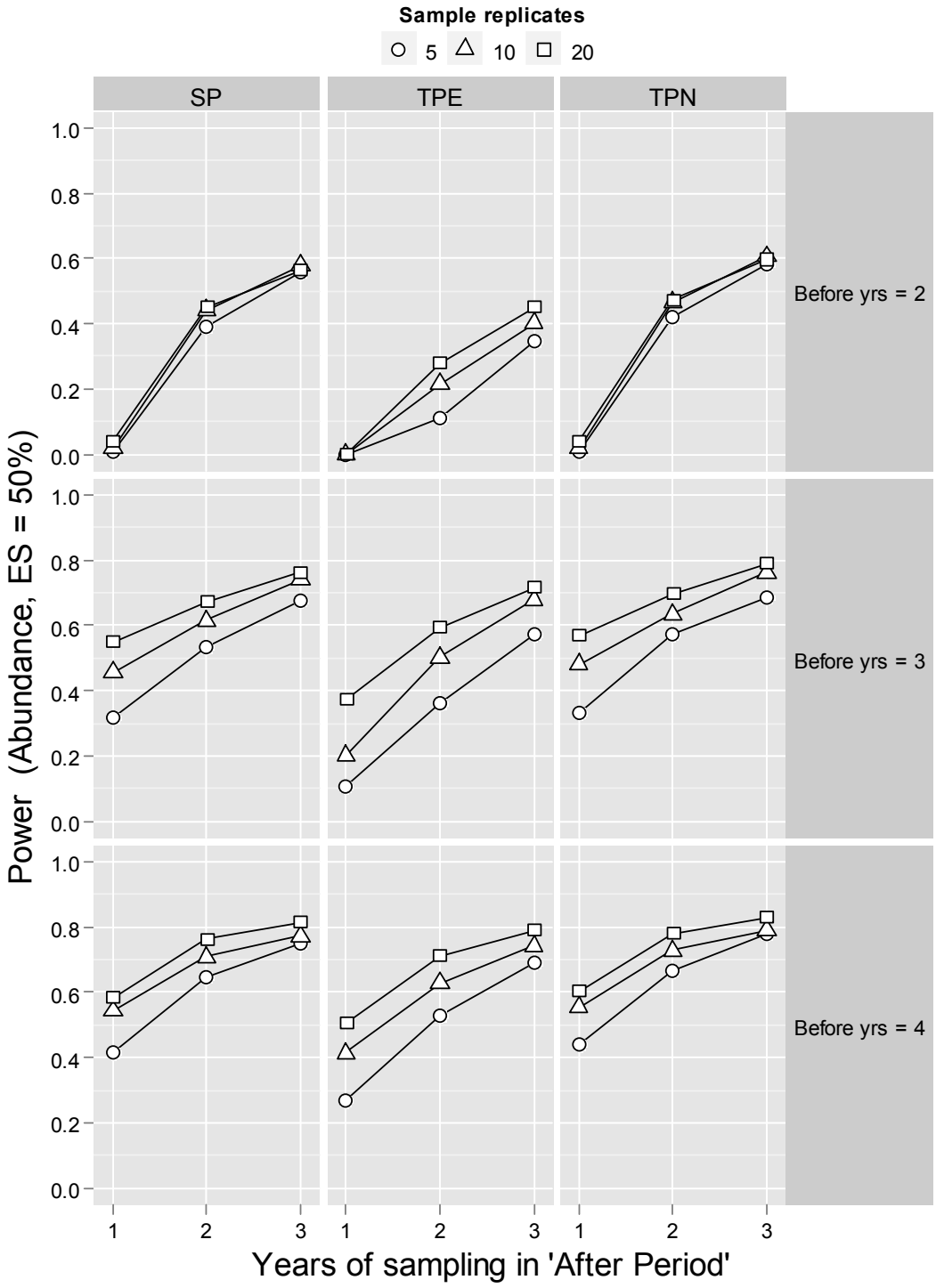


Figure 8. Two controls; Wet biomass, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of sampling years (before-period in rows, after-period along x-axis).

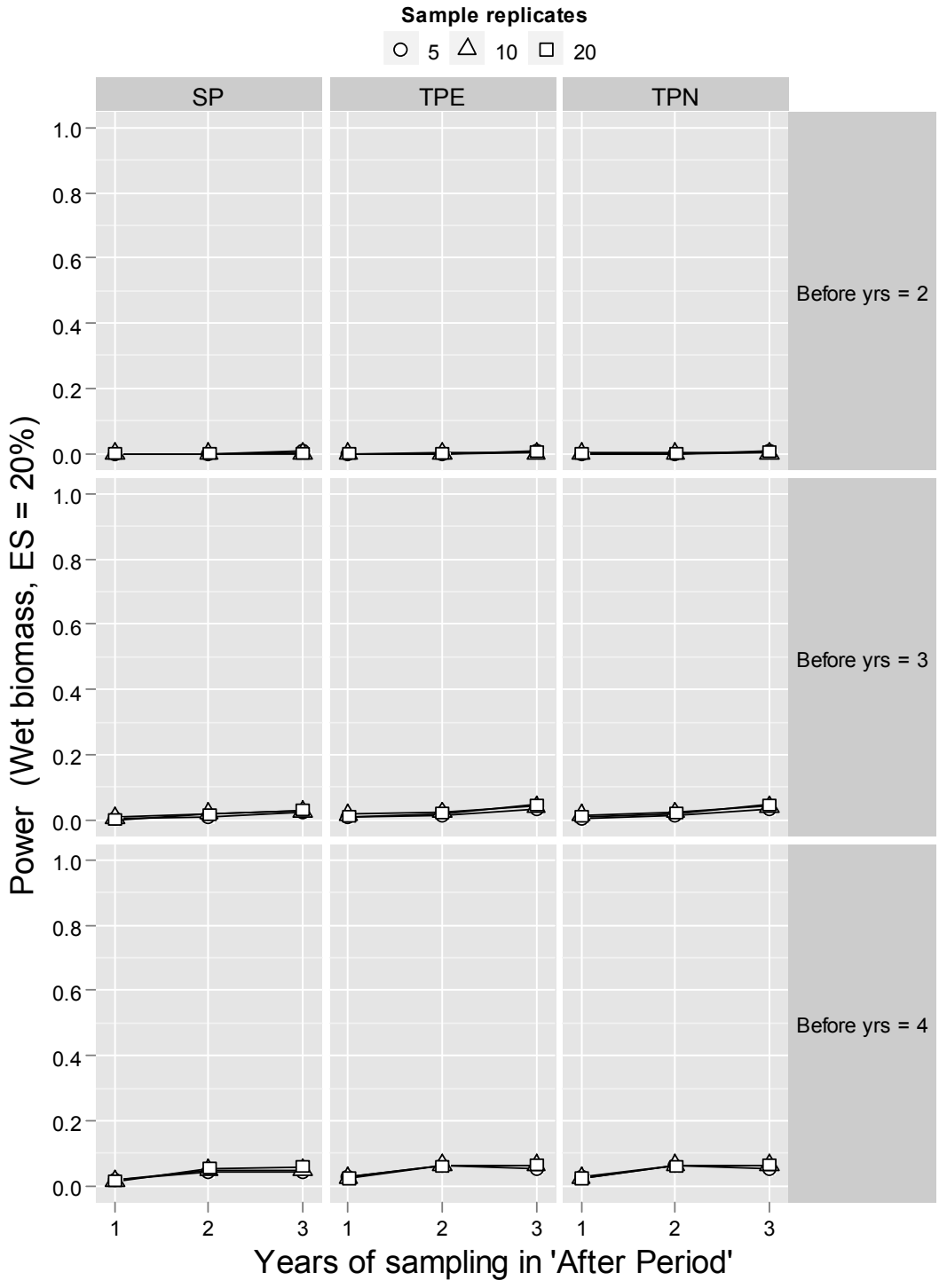


Figure 9. Two controls; Wet biomass, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of sampling years (before-period in rows, after-period along x-axis).

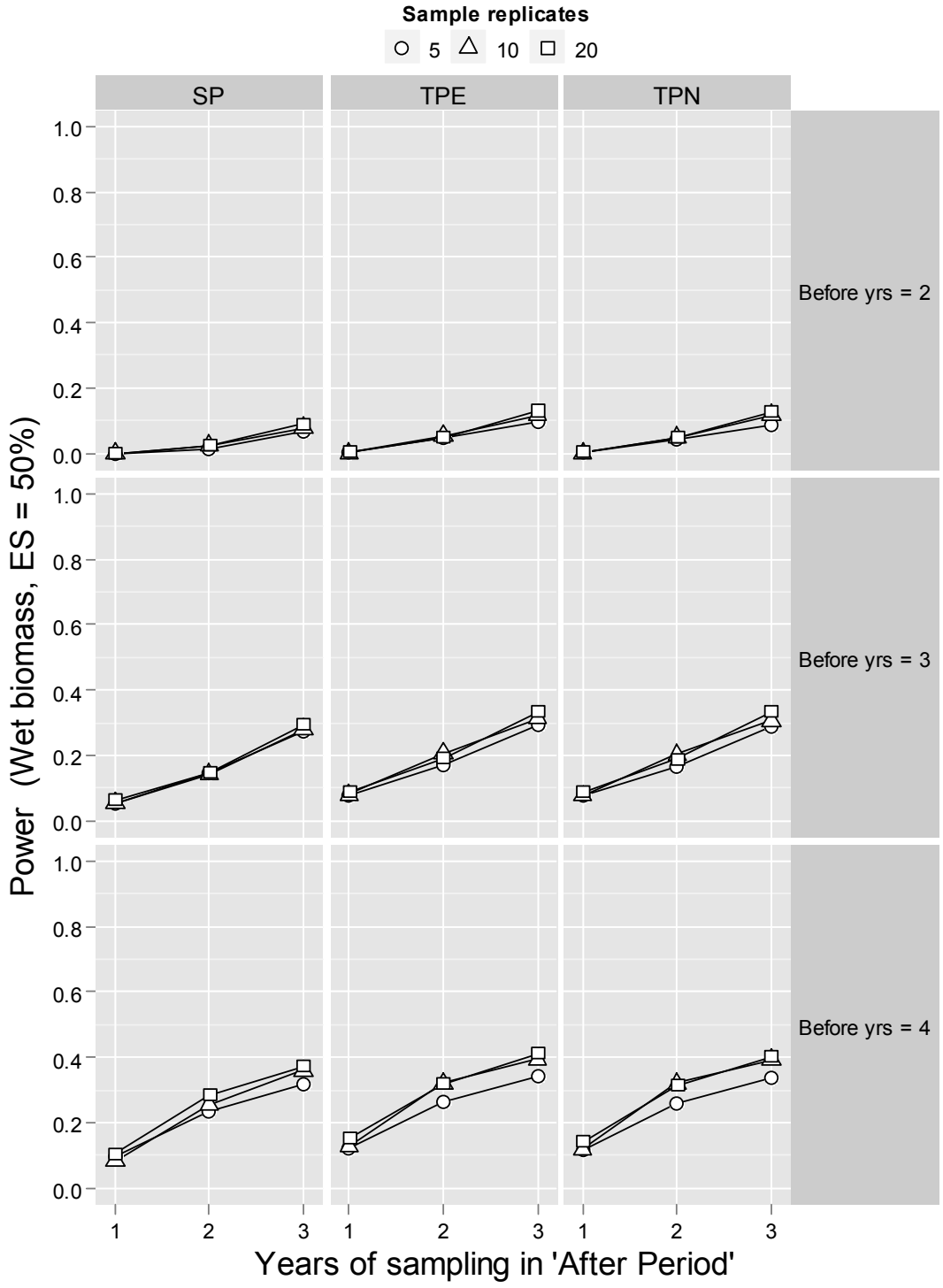


Figure 10. Two controls; Dry biomass, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of sampling years (before-period in rows, after-period along x-axis).

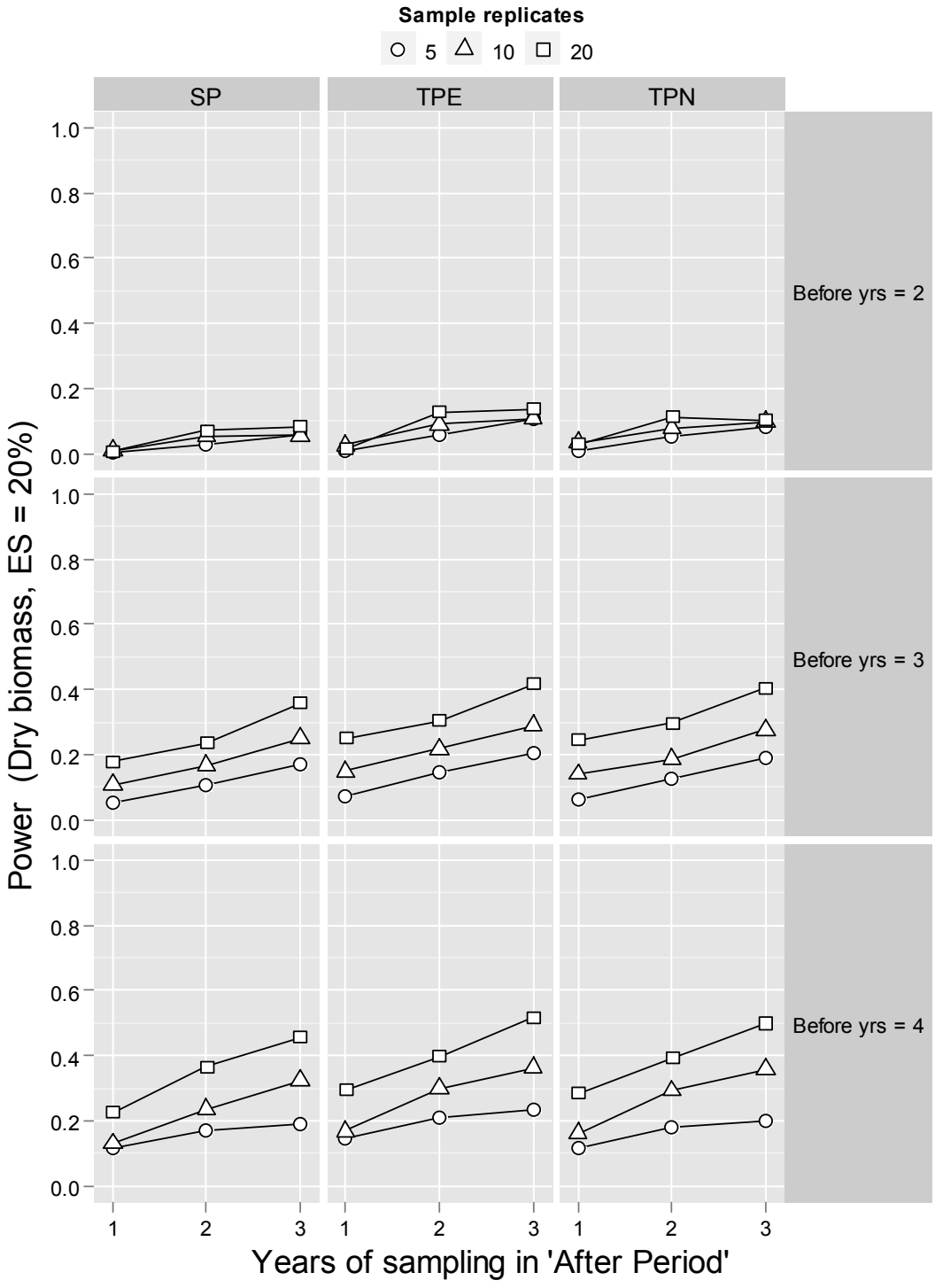


Figure 11. Two controls; Dry biomass, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of sampling years (before-period in rows, after-period along x-axis).

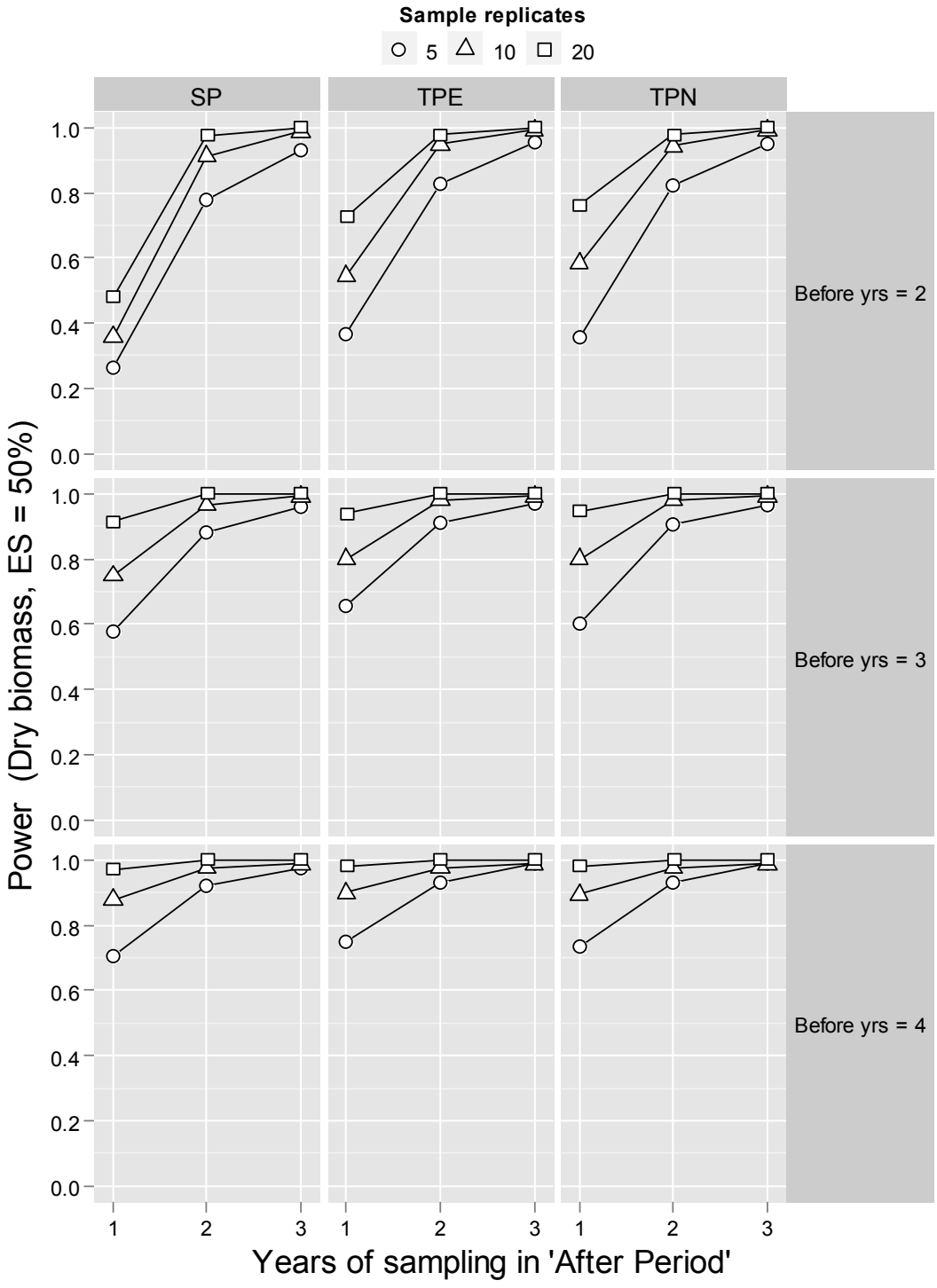


Figure 12. Two controls; Abundance, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of sampling years (before-period in rows, after-period along x-axis).

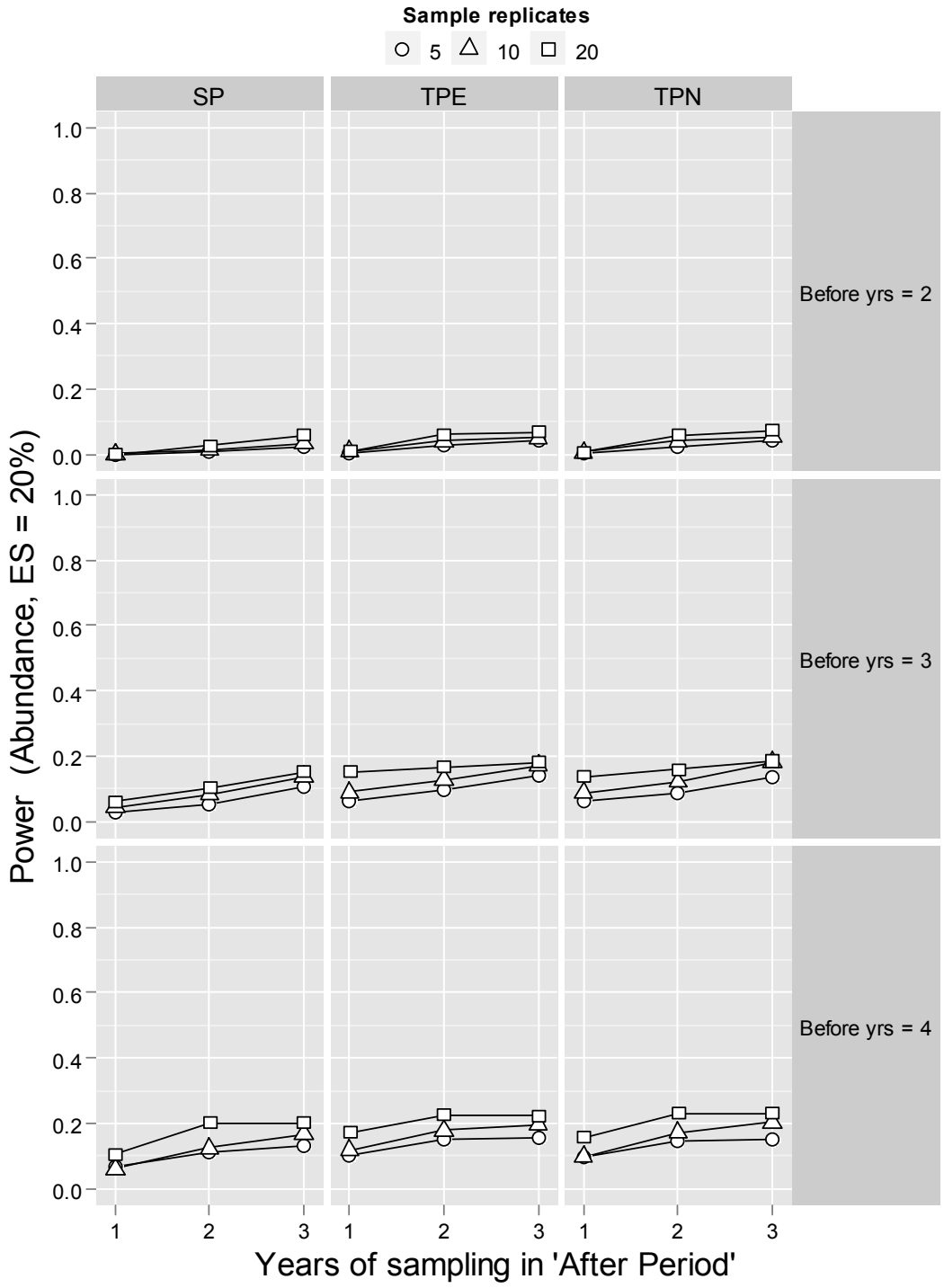
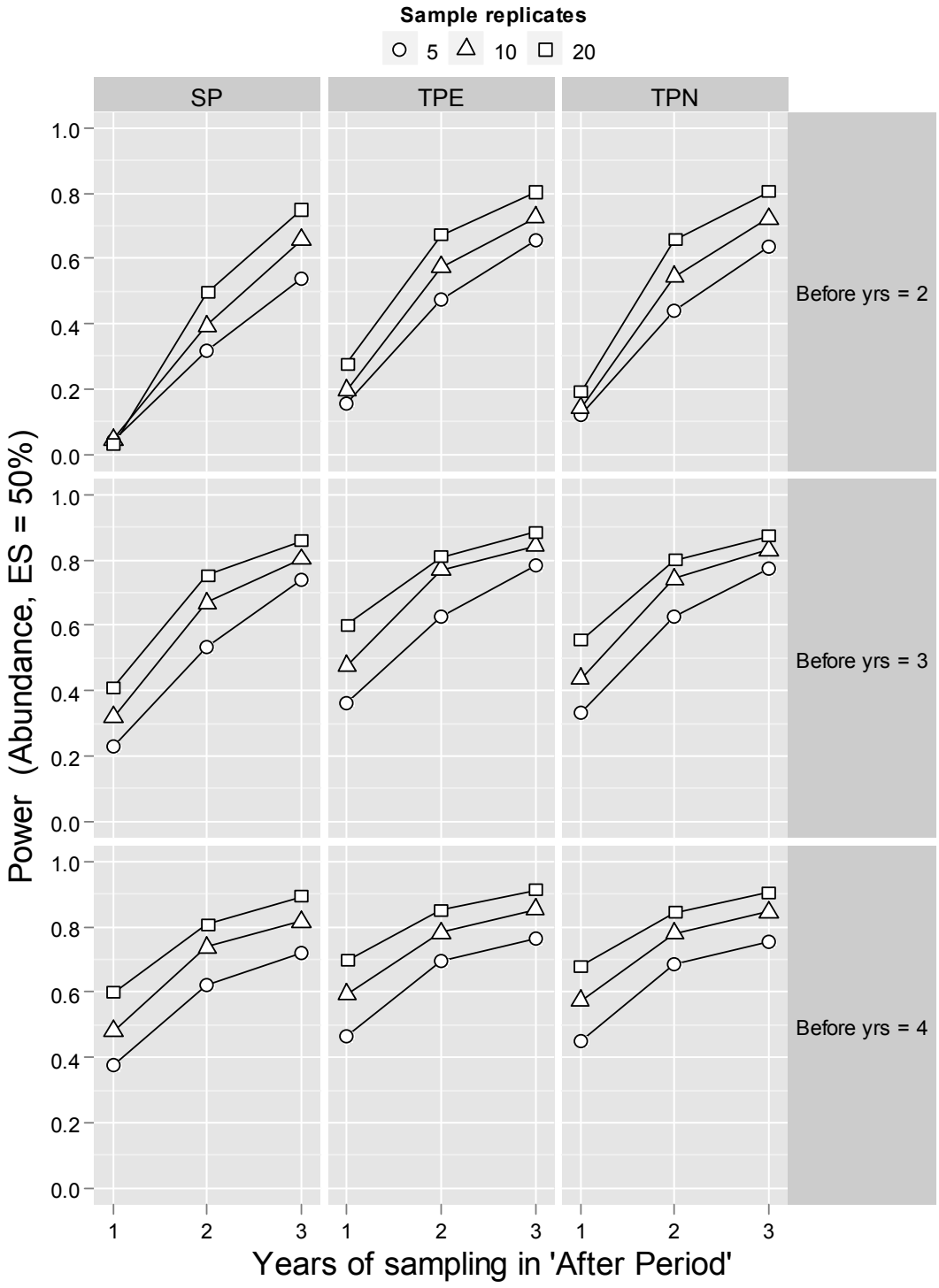


Figure 13. Two controls; Abundance, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of sampling years (before-period in rows, after-period along x-axis).



APPENDIX F – STATISTICAL ANALYSES FOR PERIPHYTON

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

This appendix contains the following analyses:

- Summary of CREMP periphyton data
- Analysis and implications for sampling design

The material in this appendix assumes understanding of basic statistical methods (Venables and Ripley 2002), mixed-effects models (Pinheiro and Bates 2000), before-after-control-impact (BACI) experimental design (Stewart-Oaten et al. 1986; Underwood 1994; Smith 2002), and use of simulation in statistical analysis (Gelman and Hill 2006).

2. SUMMARY AND EVALUATION OF CREMP PERIPHYTON DATA

Periphyton samples were collected as part of the CREMP from 2007 onwards in August of each year (**Table 1**). Most of the samples were collected in 2007 and 2008 (5 spatial replicates at each of 7 Meadowbank stations, for a total of 35 samples each year). For 2009 to 2011 there were 5 spatial replicates collected in SP and also at reference (SP-DT). We examined biomass variables (ug/cm^2), which are summarized in **Table 2** across all samples. Across samples, total biomass was dominated by Cyanobacteria, followed by Diatoms and Chlorophyte. Data appear highly variable across samples, with CV (coefficient of variation) = 0.57 for total biomass and > 0.7 for all other variables. **Table 3** shows Spearman correlations among variables. Across samples, variation in total biomass reflected variation in Cyanobacteria (r = 0.68) and Diatoms (r = 0.61). However, Cyanobacteria and Diatoms were not correlated (r = 0.0), suggesting these two components have distinct patterns across stations (or at some spatial/temporal scale).

The remaining analysis focused on three biomass variables: Cyanobacteria, Diatom, and total biomass. **Figure 1** shows box-plots for each variable by station (all years combined). Diatoms show considerably greater variation among stations than Cyanobacteria. From a sampling-design perspective, we are most interested in potential year-to-year differences. **Figures 2-4** show box-plots of sample measurements by station and year for each variable. In general, although there is some indication of variability among years (e.g., Cyanobacteria biomass and total biomass at SP), such patterns may be simply a result of limited sample sizes and high variability among spatial replicates.

An important consideration in the experimental design (discussed in next section) is the extent to which variation in a given periphyton measure is due to differences among stations, years, and replicates. To estimate these components of variation, we fit a mixed-effects model to each variable of the form:

$$X_{jks} = \beta_j + \tau_k + (\tau\beta)_{kj} + \varepsilon_{jks}$$

where all Station (β) and Year (τ) terms (including the interaction term $\tau\beta$) were treated as random effects, X is the measured value of subsample s at Station j in Year k, and ε is the residual error for subsamples.

There were notable differences among variables in the results (**Table 4**). For Cyanobacteria, most of the variation (76%) was accounted for by sampling variation (Error; the residual variation among replicate samples). In contrast, larger Station effects (i.e., differences among stations that were consistent across years) were estimated for Diatoms (52% of variation) and total biomass (31%). For all three variables, a similar proportion of variation (~10%) was accounted for by Station x Year (differences specific to each station-year combination). For the experimental design explored in the next section, we would expect relatively low power to detect change for Diatoms due to higher Station x Year variation (SD = 0.32), and the highest power to detect change for total biomass due to lower Station x Year variation (SD = 0.20) and sampling error (0.47).

3. ANALYSIS OF SAMPLING DESIGN

Impact hypotheses and statistical design – Two general classes of impacts are hypothesized for the Meadowbank mine for the case of periphyton:

1. Pulse events for which potential impacts would be high for a short time but may then dissipate. Pulse events could be associated with any phase of the mine, but are more likely to be associated with particular activities such as dike construction.
2. Long-term cumulative impacts that may be associated with ongoing activities. Long-term cumulative impacts are more likely to be associated with ongoing activities of mine operations.

As operations have just begun in 2010, the focus of CREMP monitoring to date has been on detecting pulse events associated with construction. The appropriate framework for analysis is a before-after-control-impact (BACI) that is aimed at detecting a potential impact in a particular lake or basin in a particular time period. The BACI framework can also be used to evaluate long-term impacts, but other tools such as time series regression analysis may also be appropriate for evaluating long-term trends. For this design document, we focus on the use of the BACI framework, recognizing that other tools such

as time series regressions may be useful at a future date once sufficient time series data are available[1].

The classic BACI (paired) design has before/after periods α_i (i = B, A; I = 2), control/impact sites β_j (j = C, I; J = 2), and a total of K paired sampling times τ_k that are nested within period. A statistical model for this design is given by (Smith 2002, equation 2):

$$(1) \qquad X_{ijk} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \ .$$

The key term is the interaction ($\alpha\beta$), which can be tested using an F test with F = MS[($\alpha\beta$)]/MS[Resid] and degrees of freedom = 1, K – 2. As discussed by Smith (2002), this is equivalent to simply taking the differences between the control and impact values across times and using a two-sample (before-after) t test (Stewart-Oaten et al. 1986).

Model (1) can be extended to include additional control sites (e.g., "asymmetric" designs; Underwood 1994) and/or additional impact sites. To be valid, the additional sites must be replicates rather than subsamples (i.e., as controls, they should be spatially independent of each other but representative of the impact sites, while replicates for impacts need to be spatially independent and (ideally) affected by independent disturbances). So whereas j = (C, I) in the classic BACIP, j may compose any combination of J total sites, for example J = 4 where j = (C_1, C_2, C_3, I). The general test of ($\alpha\beta$) still applies, but with degrees of freedom = (J – 1), (K – 2)(J – 1) (e.g., see Table 1 of Underwood (1994) and Table 9 of Smith (2002)).

In addition, there may be n replicate subsamples s at each site/time combination (jk), as assumed in Table 1 of Underwood (1994). In this case, we modify equation (1) as:

$$(2) \qquad X_{ijks} = \mu + \alpha_i + \tau_{k(i)} + \beta_j + (\alpha\beta)_{ij} + (\tau\beta)_{k(i)j} + \varepsilon_{ijks} \ ,$$

where subsamples now permit estimation of times-by-site interactions ($\tau\beta$). The appropriate F ratio for ($\alpha\beta$) is now F = MS[($\alpha\beta$)]/MS[($\tau\beta$)] with df = (J – 1), (K – 2)(J – 1). As Underwood demonstrates, specific comparisons (interaction terms, such as the impact site versus either "period" or a specific "time" unit) can be examined by partitioning variation accordingly (e.g., Underwood Table 2[2]).

Methods – The analysis assessed the expected precision and power of BACI estimates for different after-period (impact) durations and different numbers of sub-samples (random spatial replicates collected at the same station during the sampling event each August).

---

[1] In theory, a BACI analysis that is appropriately framed should be capable of detecting changes associated with long-term trends.

[2] We note that there are mistakes in the tables presented by Underwood (1994).

Separate analyses were conducted for the three primary impact stations SP, TPE, and TPN. In each case, INUG was used as the control station. Separate analyses were conducted for the three biomass variables. For each variable, the effect size or ES (fixed across months) was set at either a 20% reduction from baseline or a 50% reduction in baseline.

For purposes of the BACI analyses, the two years (2007-2008) of available control-impact paired data were used to represent the "before period" dataset. This assumes that these data, regardless of station or year, are reasonably representative of "natural" conditions. We also examined scenarios in which one or two additional "before" years of data were simulated prior to impacts, providing three scenarios (before years = 2, 3, and 4). The additional before-period data were simulated with variances equal to those estimated for observed data across all stations (**Table 4**), using the model specified in **Section 2** above[3]. After-period data were simulated using after-period means (β _control, β _impact + **ES**) for three durations (1, 2, and 3 years) and three sub-sample scenarios (5, 10, and 20 replicates per station per year), again using variances derived from the observed data (**Table 4**). For each scenario, 500 simulations were used. Power was computed for one-tailed tests (based on the a priori assumption that impacts will reduce periphyton biomass) using two alpha levels (0.05 and 0.10).

Results – Estimates of statistical power are shown in **Tables 5-7** and displayed in **Figures 5-10**. The tables report power for both alpha levels (0.05 and 0.10) but omit results for "Before" years = 2 (very low power). The figures only show power for alpha = 0.05 but include "Before" years = 2 Results were quite consistent across impact stations (SP, TPE, and TPN) for all variables and sampling designs. Power was low (< 0.8) for Cyanobacteria across all scenarios (**Table 5**; **Figures 5-6**) except for the most extensive sampling design (4 years of "Before" data, three years of "After" data, and 20 sample replicates for all additional sampling years) with ES = 50% and alpha = 0.10. Power was also low for Diatom biomass across all scenarios (**Table 6**; **Figures 7-8**). For total biomass with ES = 50%, high power was achieved for several scenarios (**Table 7**; **Figures 9-10**). Power was only high for alpha = 0.05 for the longest duration design (4 "Before" years and three "After" years) with 10 or 20 sample replicates, while for alpha = 0.10, high power was often achieved when the combination before/after years totaled 5 or more years.

Conclusion – Overall, periphyton variables are not realistically capable of detecting effects in a given year, and generally will not achieve high power even with high

[3] Data were insufficient to justify estimation of variances specific to each impact station paired to control (as was done for water chemistry), therefore variances were derived based on pooled data for all stations.

replication and several years of data. The only potentially useful variable, total biomass, does not achieve high (0.8) power until several years of data are available.

4. REFERENCES

Gelman, A, and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models – Analytical Methods for Social Research. Cambridge University Press, New York, 625 pp.

Pinheiro, J., and D. Bates. 2000. Mixed-Effects Models in S and S-PLUS. Springer-Verlag, New York, 528 pp.

Smith, E. 2002. BACI design. Pages 141-148, in: A.H. El-Shaarawi and W.W. Piegorsch [Eds], Encyclopedia of Environmetrics, John Wiley & Sons, Chicherster.

Stewart-Oaten, A., W. Murdoch, and K. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time? Ecology 67: 929-940.

Venables, W., and B. Ripley. 2002. Modern Applied Statistics. Springer, New York, 495 pp.

Underwood, A. 1994. On beyond BACI: sampling designs that might reliably detect environmental disturbances. Ecological Applications 4(1): 3-15.

Table 1. Summary of August periphyton samples by station.

Note: Shading denotes "Impact" periods.

Year	INUG	SP	SP-DT	TE	TPE	TPN	TPS	WAL	Total
2007	5	5		5	5	5	5	5	35
2008	5	5		5	5	5	5	5	35
2009		5	5						10
2010		5	5						10
2011		5	5		5				15
Total	10	25	15	10	15	10	10	10	105

Table 2. Summary statistics for periphyton biomass (ug/cm^2) variables across all samples.

Note: CV = Coefficient of variation (SD/Mean).

Variable	N	N=0	Mean	Median	SD	CV
Cyanobacteria	105	0	231	180	171	0.74
Chlorophyte	105	14	62	34	84	1.36
Chrysophyte	105	98	0.1	0.0	0.4	5.64
Diatom	105	0	129	103	111	0.86
Dinoflagellate	105	98	0.9	0.0	4.7	5.26
Total biomass	105	0	423	353	240	0.57

Table 3. Spearman correlations among periphyton biomass variables (N = 105 samples).

	Cyanobacteria	Chlorophyte	Chrysophyte	Diatom	Dinoflagellate
Cyanobacteria					
Chlorophyte	0.08				
Chrysophyte	-0.04	-0.06			
Diatom	0.00	0.30	-0.28		
Dinoflagellate	-0.12	0.06	0.09	-0.04	
Total biomass	0.68	0.43	-0.13	0.61	-0.02

Table 4. Estimates of standard deviation (SD) and percent of variation explained for each term in the mixed-effects models fit to (log) periphyton biomass variables.

	SD Estimate (log data)			Percent of variation		
Term	Cyanobact.	Diatom	Total biomass	Cyanobact.	Diatom	Total biomass
Station	0.19	0.73	0.34	6%	52%	31%
Year	0.20	0.17	0.00	7%	3%	0%
Station x Year	0.24	0.32	0.20	10%	10%	11%
Error (reps)	0.66	0.60	0.47	76%	35%	58%
Sum				100%	100%	100%

Table 5. Cyanobacteria biomass: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year.

Note: Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10). Light shading denotes power ≥ 0.8.

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.04	0.04	0.01	0.18	0.19	0.13
			10	0.08	0.08	0.05	0.24	0.25	0.19
			20	0.14	0.15	0.09	0.32	0.32	0.27
	3	2	5	0.08	0.08	0.06	0.19	0.20	0.15
			10	0.10	0.11	0.09	0.25	0.25	0.21
			20	0.17	0.17	0.15	0.33	0.33	0.29
	3	3	5	0.10	0.11	0.08	0.25	0.26	0.21
			10	0.14	0.15	0.11	0.29	0.29	0.27
			20	0.18	0.19	0.16	0.32	0.32	0.30
	4	1	5	0.09	0.10	0.07	0.24	0.25	0.19
			10	0.14	0.14	0.11	0.28	0.28	0.24
			20	0.18	0.19	0.17	0.36	0.36	0.32
	4	2	5	0.15	0.16	0.13	0.29	0.30	0.25
			10	0.14	0.15	0.12	0.28	0.29	0.26
			20	0.21	0.22	0.19	0.33	0.33	0.31
	4	3	5	0.13	0.13	0.11	0.27	0.28	0.25
			10	0.19	0.19	0.17	0.31	0.31	0.29
			20	0.22	0.22	0.20	0.38	0.38	0.36
50%	3	1	5	0.16	0.18	0.11	0.41	0.44	0.34
			10	0.28	0.29	0.21	0.58	0.59	0.51
			20	0.46	0.46	0.38	0.68	0.68	0.62
	3	2	5	0.35	0.37	0.28	0.57	0.58	0.53
			10	0.44	0.45	0.38	0.66	0.66	0.62
			20	0.55	0.55	0.52	0.74	0.74	0.70
	3	3	5	0.50	0.50	0.45	0.69	0.70	0.65
			10	0.55	0.55	0.51	0.75	0.75	0.73
			20	0.61	0.61	0.59	0.78	0.79	0.77
	4	1	5	0.28	0.30	0.21	0.49	0.50	0.45
			10	0.42	0.43	0.37	0.61	0.61	0.57
			20	0.50	0.50	0.48	0.66	0.66	0.63
	4	2	5	0.46	0.47	0.42	0.66	0.66	0.60
			10	0.55	0.56	0.51	0.74	0.74	0.71
			20	0.62	0.62	0.60	0.79	0.79	0.77
	4	3	5	0.57	0.57	0.53	0.74	0.74	0.71
			10	0.64	0.64	0.61	0.79	0.79	0.77
			20	0.71	0.71	0.68	0.81	0.81	0.81

Table 6. Diatom biomass: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year.

Note: Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10). Light shading denotes power ≥ 0.8.

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.02	0.04	0.04	0.14	0.17	0.18
			10	0.06	0.09	0.09	0.22	0.25	0.25
			20	0.13	0.16	0.16	0.33	0.34	0.35
	3	2	5	0.06	0.08	0.08	0.16	0.19	0.20
			10	0.11	0.11	0.12	0.22	0.24	0.24
			20	0.16	0.17	0.17	0.31	0.30	0.30
	3	3	5	0.08	0.10	0.10	0.21	0.25	0.26
			10	0.13	0.15	0.14	0.28	0.28	0.29
			20	0.16	0.17	0.16	0.30	0.29	0.29
	4	1	5	0.08	0.10	0.10	0.21	0.23	0.24
			10	0.14	0.15	0.15	0.27	0.29	0.29
			20	0.18	0.19	0.19	0.33	0.34	0.35
	4	2	5	0.13	0.15	0.15	0.26	0.29	0.29
			10	0.15	0.17	0.17	0.30	0.31	0.32
			20	0.21	0.22	0.22	0.33	0.33	0.34
	4	3	5	0.13	0.14	0.13	0.24	0.26	0.26
			10	0.18	0.19	0.19	0.29	0.30	0.30
			20	0.20	0.20	0.21	0.34	0.35	0.35
50%	3	1	5	0.10	0.15	0.15	0.32	0.38	0.39
			10	0.22	0.26	0.27	0.49	0.53	0.53
			20	0.38	0.39	0.39	0.58	0.59	0.60
	3	2	5	0.27	0.30	0.32	0.50	0.53	0.55
			10	0.38	0.40	0.39	0.59	0.60	0.60
			20	0.47	0.47	0.47	0.64	0.64	0.64
	3	3	5	0.43	0.47	0.48	0.63	0.66	0.66
			10	0.47	0.49	0.49	0.68	0.68	0.68
			20	0.53	0.54	0.54	0.69	0.69	0.69
	4	1	5	0.22	0.26	0.26	0.43	0.46	0.47
			10	0.36	0.37	0.38	0.54	0.55	0.55
			20	0.44	0.45	0.45	0.60	0.59	0.60
	4	2	5	0.42	0.45	0.45	0.60	0.61	0.62
			10	0.48	0.49	0.49	0.65	0.66	0.67
			20	0.52	0.52	0.52	0.71	0.71	0.71
	4	3	5	0.49	0.51	0.52	0.70	0.70	0.71
			10	0.56	0.57	0.58	0.71	0.72	0.72
			20	0.59	0.60	0.60	0.75	0.75	0.75

Table 7. Total biomass: Statistical power for BACI tests (one tailed) by station (SP, TPE, TPN) as a function of years sampled in the before/after periods and the number of replicate samples (Reps) per year.

Note: Power is shown for two effect sizes (ES = 20% and 50% reductions) and alpha levels (0.05 and 0.10). Light shading denotes power ≥ 0.8.

Effect	Years sampled			Power (α = 0.05)			Power (α = 0.10)		
Size	Before	After	Reps	SP	TPE	TPN	SP	TPE	TPN
20%	3	1	5	0.07	0.07	0.04	0.21	0.22	0.18
			10	0.10	0.10	0.09	0.27	0.27	0.24
			20	0.20	0.19	0.17	0.40	0.40	0.38
	3	2	5	0.11	0.11	0.09	0.25	0.25	0.23
			10	0.15	0.15	0.14	0.26	0.27	0.25
			20	0.23	0.22	0.21	0.36	0.35	0.35
	3	3	5	0.13	0.14	0.11	0.30	0.30	0.29
			10	0.16	0.17	0.15	0.35	0.34	0.33
			20	0.21	0.21	0.20	0.36	0.35	0.34
	4	1	5	0.12	0.13	0.10	0.27	0.28	0.24
			10	0.19	0.19	0.16	0.36	0.36	0.34
			20	0.23	0.23	0.21	0.37	0.37	0.36
	4	2	5	0.17	0.17	0.15	0.33	0.34	0.31
			10	0.20	0.19	0.18	0.35	0.35	0.33
			20	0.27	0.27	0.25	0.42	0.42	0.41
	4	3	5	0.20	0.21	0.18	0.34	0.35	0.34
			10	0.23	0.22	0.22	0.38	0.38	0.37
			20	0.31	0.31	0.30	0.44	0.44	0.43
50%	3	1	5	0.29	0.30	0.24	0.55	0.56	0.52
			10	0.48	0.47	0.43	0.70	0.71	0.68
			20	0.60	0.60	0.57	0.77	0.76	0.75
	3	2	5	0.57	0.57	0.54	0.76	0.77	0.74
			10	0.67	0.67	0.65	0.85	0.84	0.83
			20	0.72	0.71	0.70	0.85	0.85	0.85
	3	3	5	0.71	0.71	0.67	0.86	0.86	0.85
			10	0.77	0.77	0.76	0.92	0.92	0.91
			20	0.79	0.79	0.78	0.90	0.90	0.90
	4	1	5	0.43	0.44	0.39	0.62	0.63	0.60
			10	0.58	0.57	0.54	0.75	0.75	0.74
			20	0.67	0.67	0.65	0.79	0.79	0.78
	4	2	5	0.66	0.66	0.63	0.82	0.83	0.80
			10	0.76	0.76	0.73	0.89	0.89	0.88
			20	0.80	0.80	0.79	0.90	0.90	0.90
	4	3	5	0.78	0.78	0.77	0.90	0.89	0.89
			10	0.84	0.84	0.83	0.93	0.93	0.93
			20	0.89	0.89	0.88	0.94	0.94	0.94

Figure 1. Box-plots of periphyton variables by station (all available years combined).

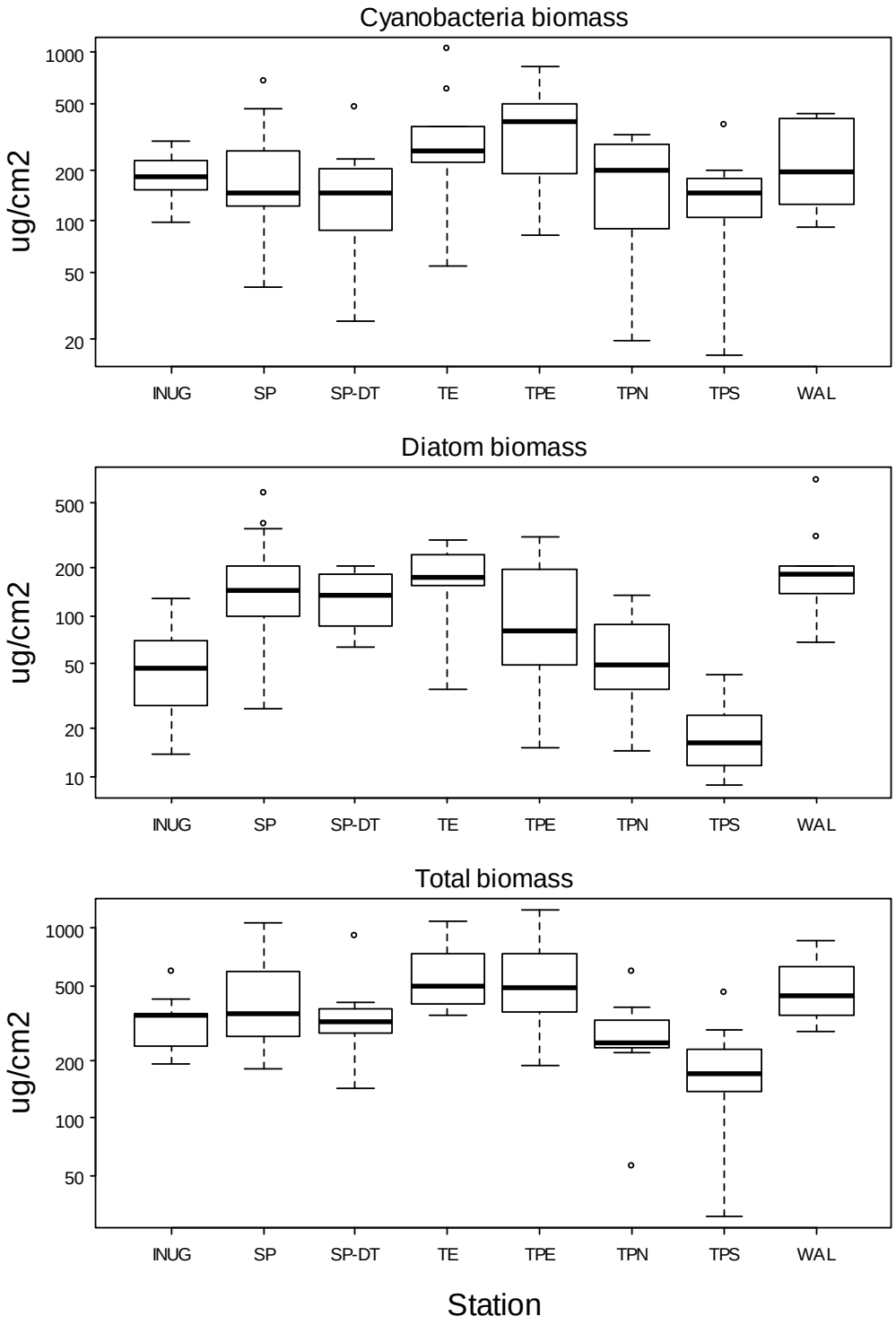


Figure 2. Box-plots of Cyanobacteria biomass by station and year (5 samples per station/year).

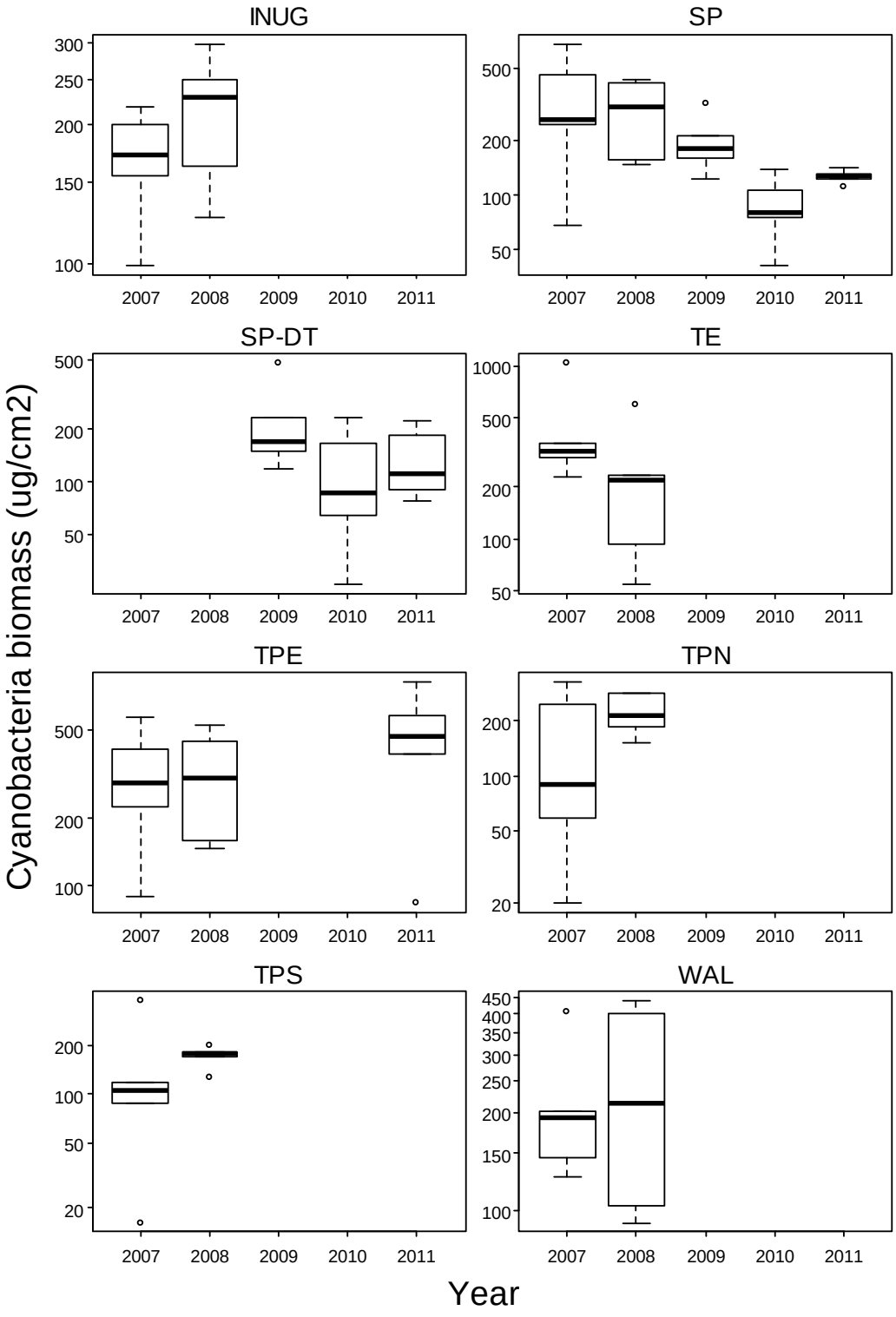


Figure 3. Box-plots of Diatom biomass by station and year (5 samples per station/year).

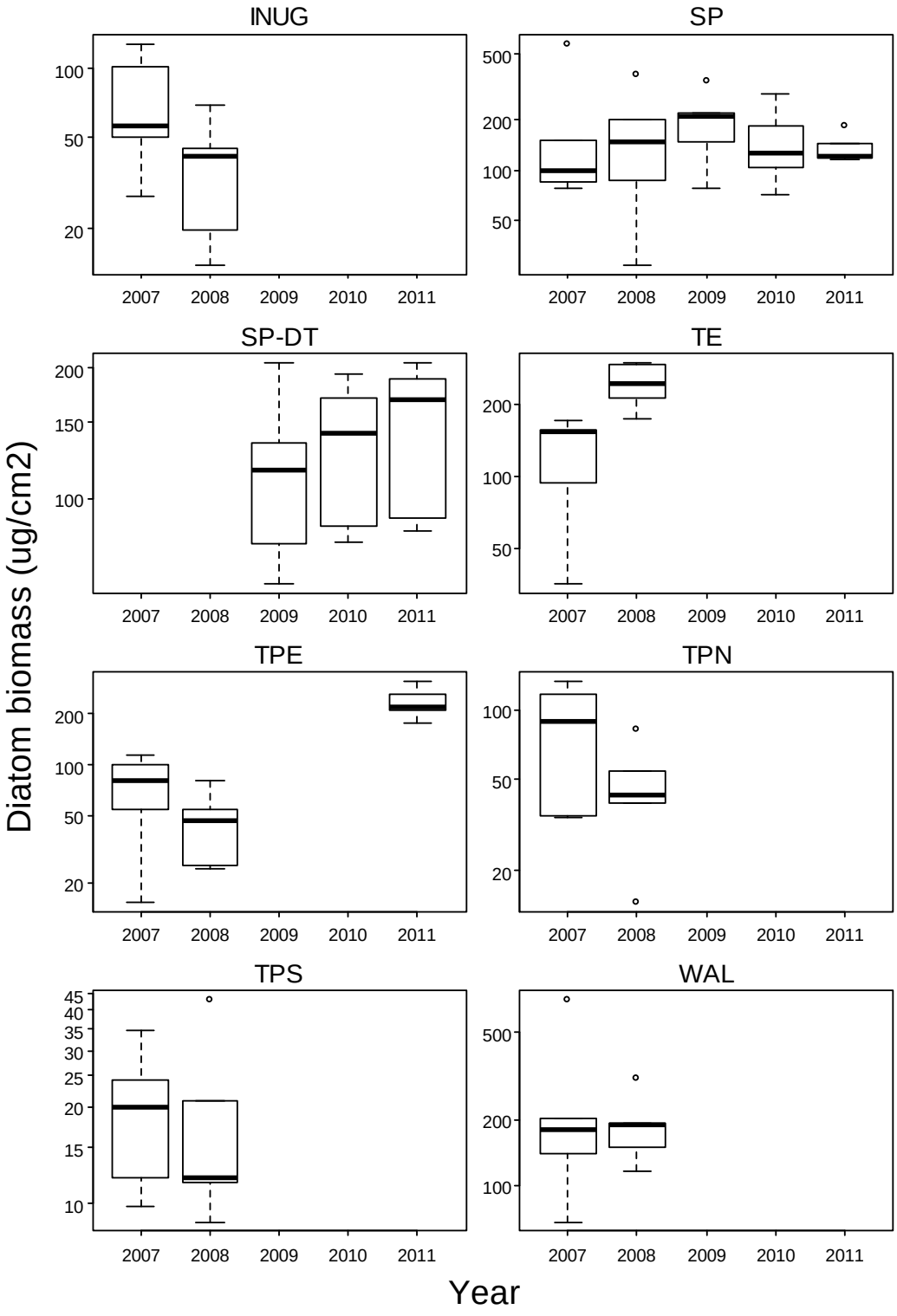


Figure 4. Box-plots of total biomass by station and year (5 samples per station/year).

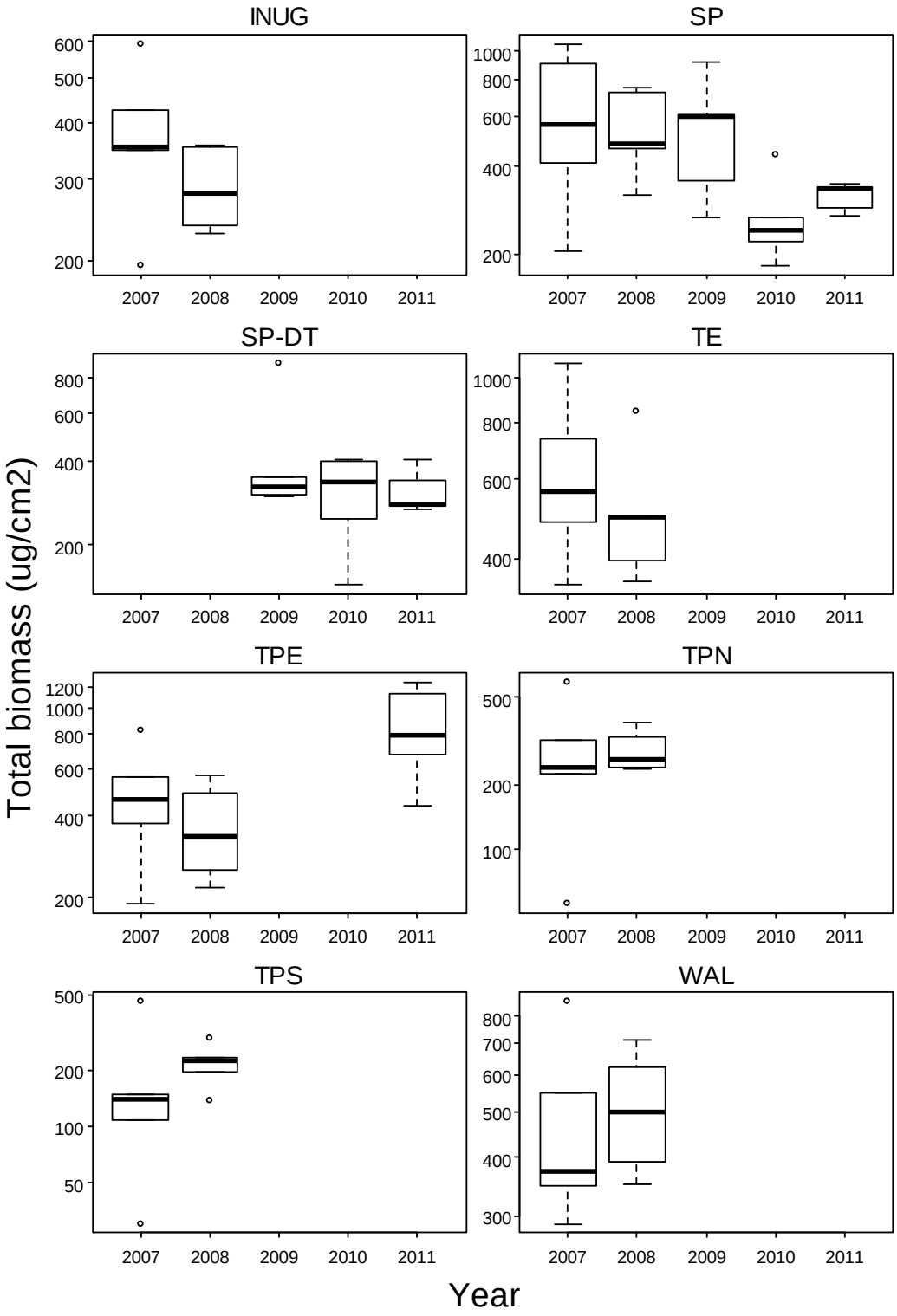


Figure 5. Cyanobacteria, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

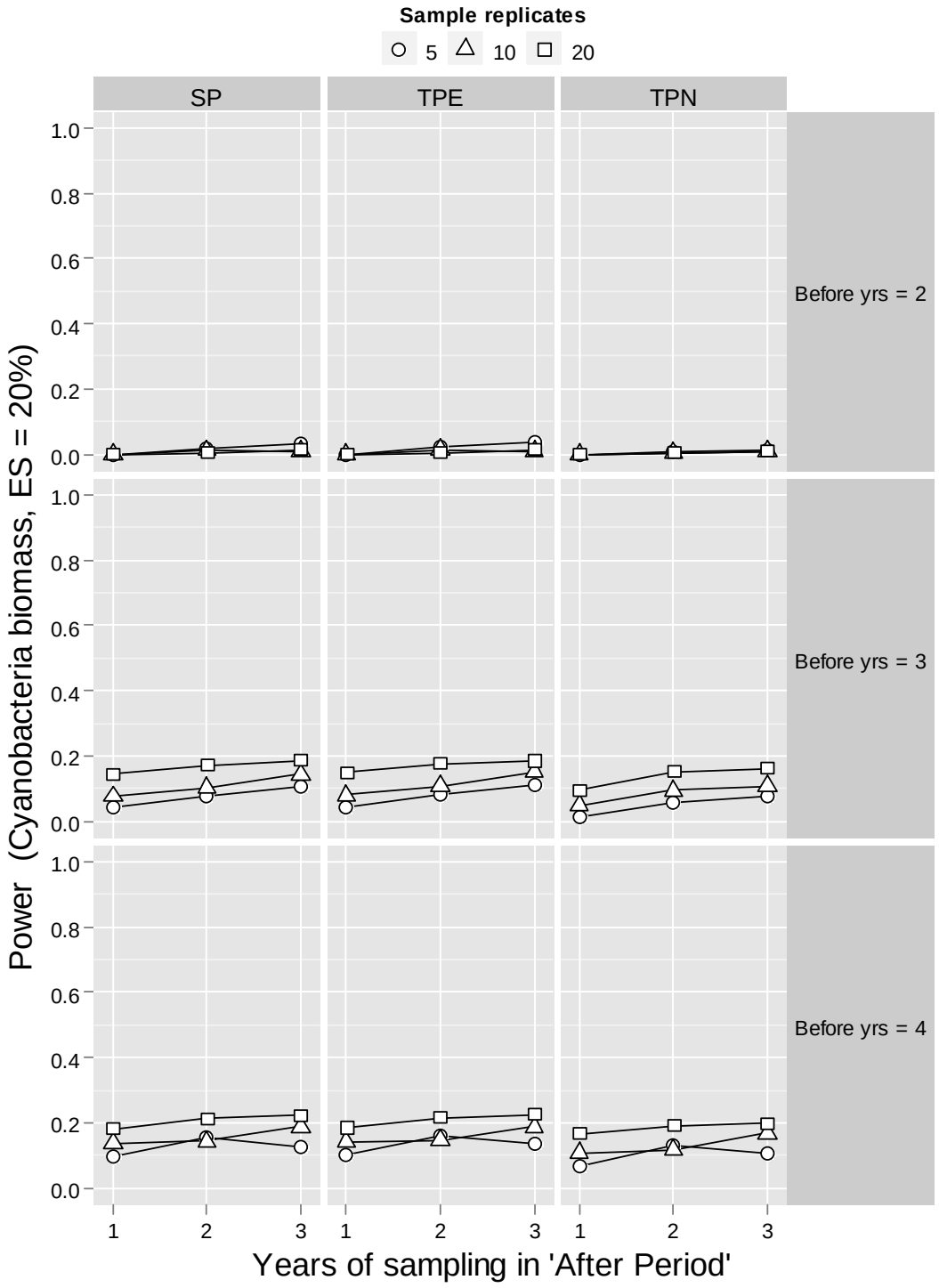


Figure 6. Cyanobacteria, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

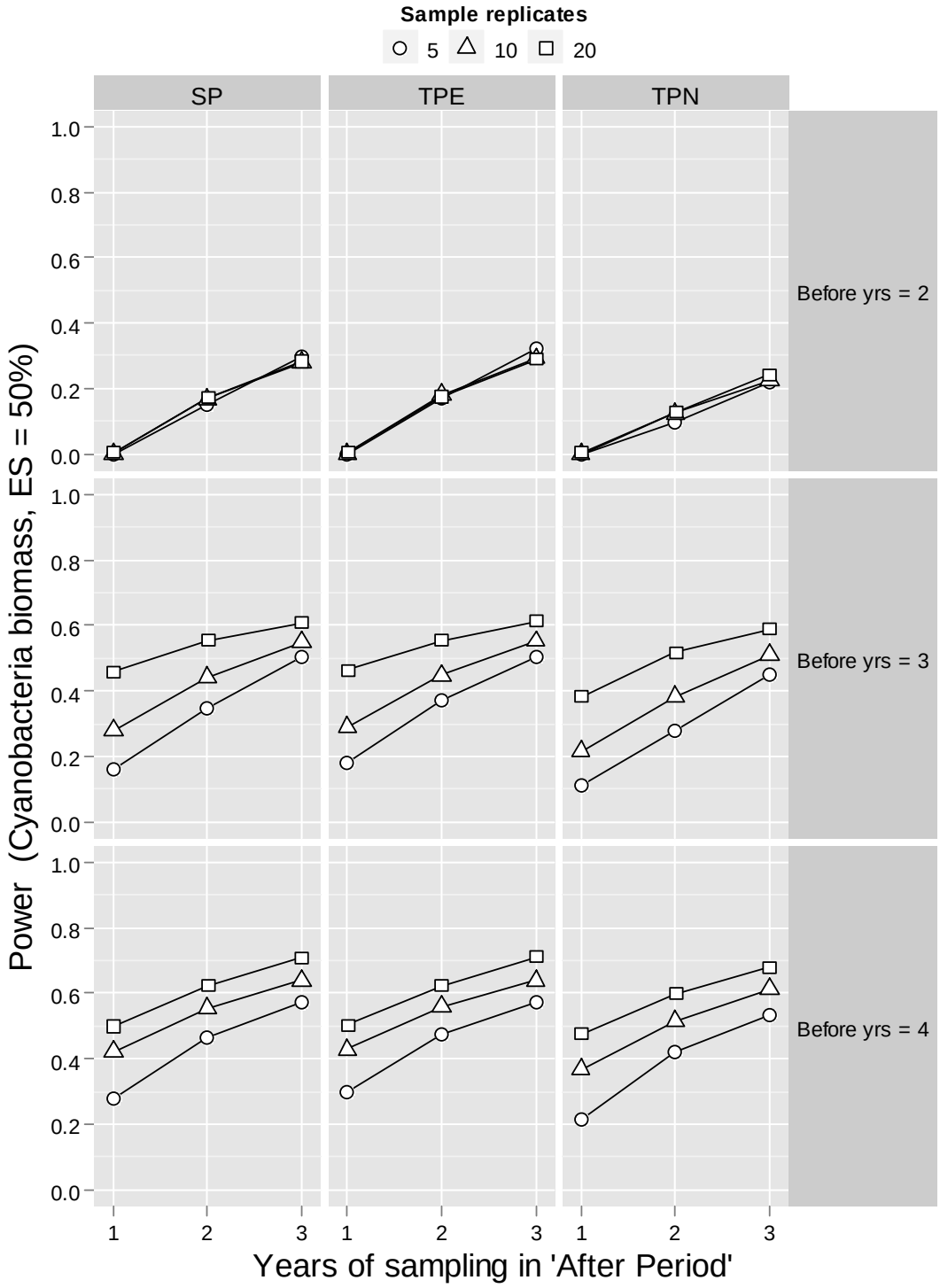


Figure 7. Diatom biomass, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

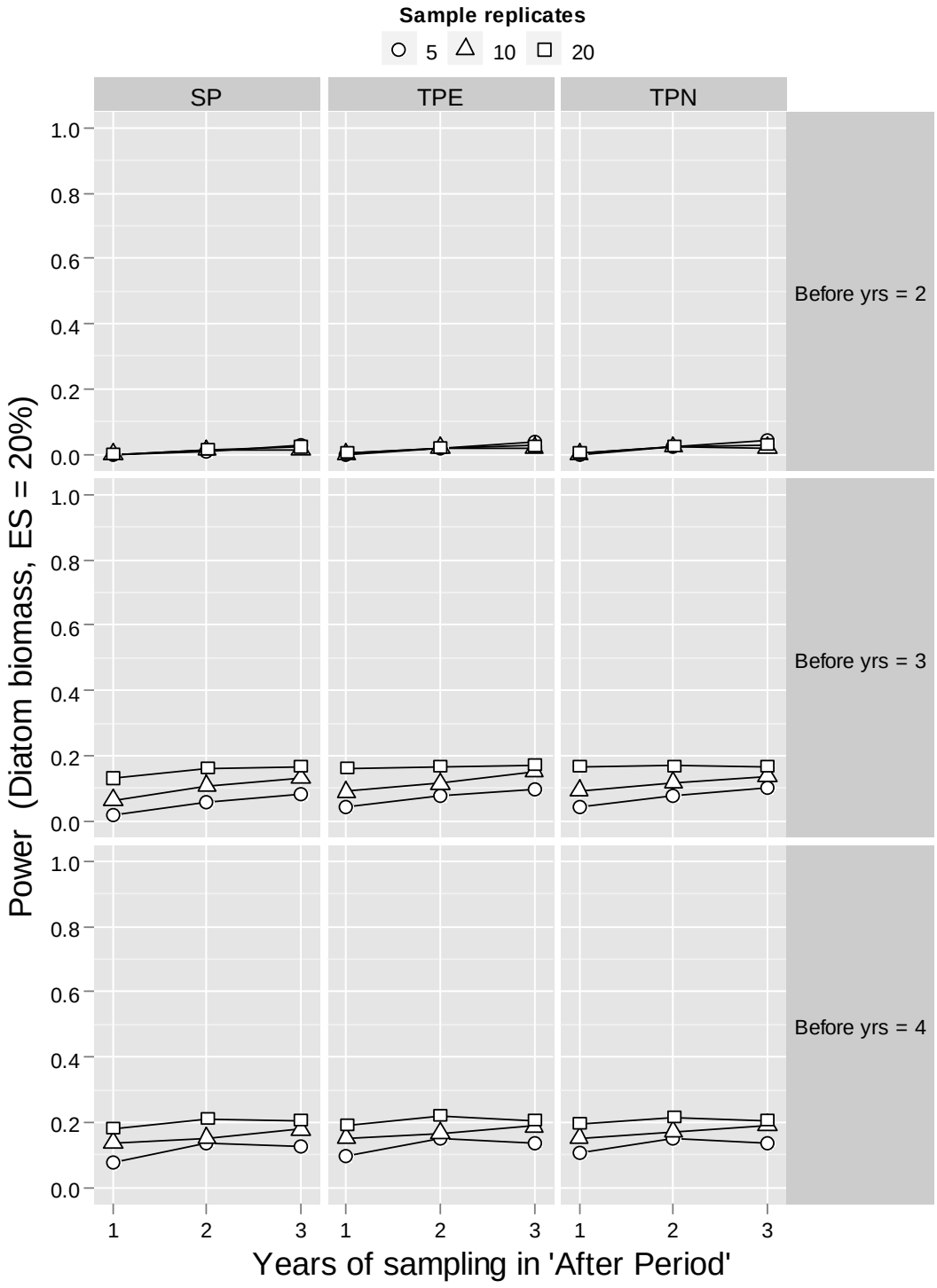


Figure 8. Diatom biomass, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

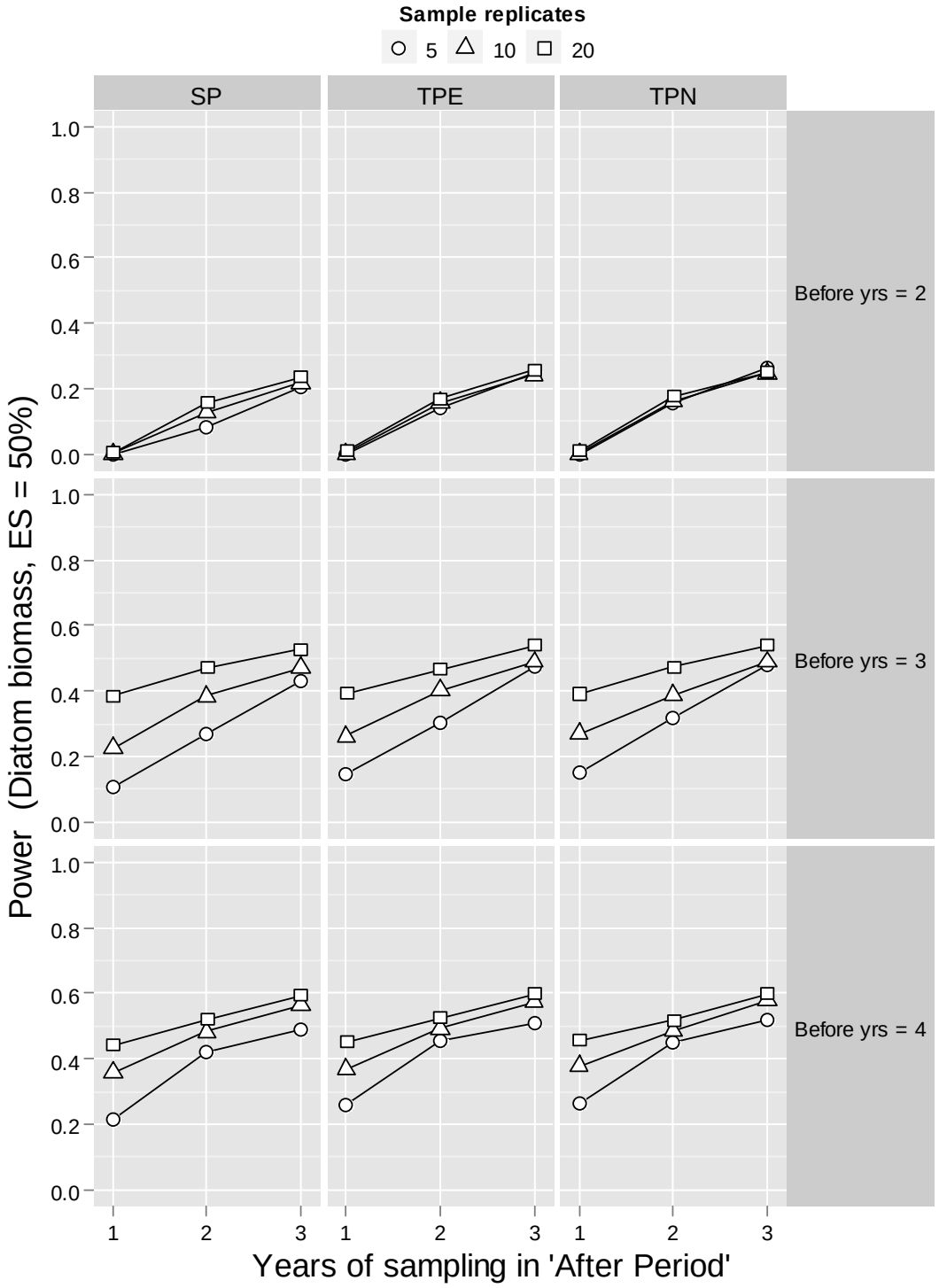


Figure 9. Total biomass, ES = 20%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).

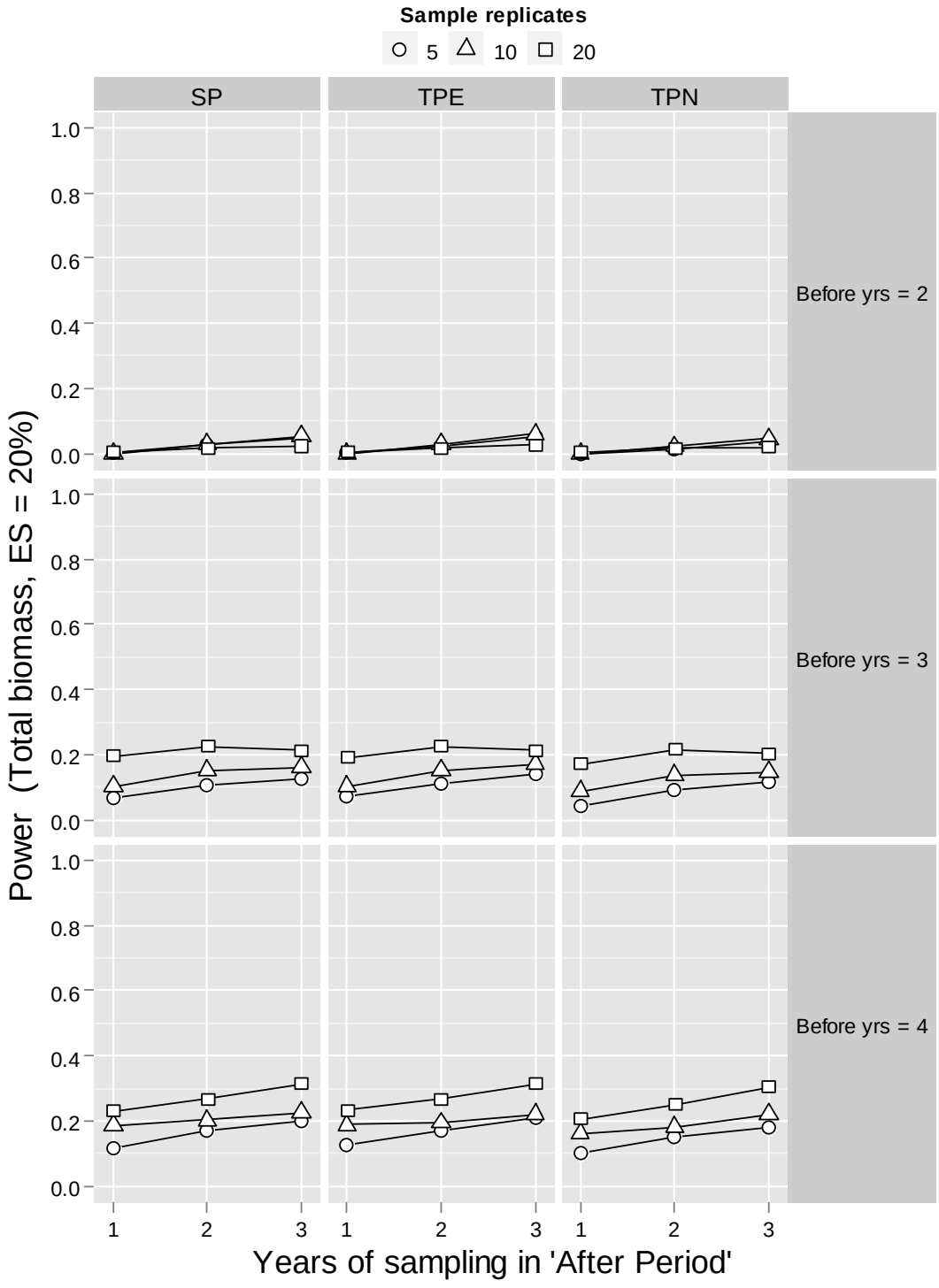


Figure 10. Total biomass, ES = 50%: Power for BACI tests (one tailed with alpha = 0.05) by station (columns) as a function of the years of sampling (before-period in rows, after-period along x-axis).