Chapter 12

Quality Assurance: Design, Precision and Management

Quality assurance (QA) is an integrated program for ensuring the reliability of monitoring and measurement data and includes quality control. Quality control (QC) refers to operational procedures for obtaining prescribed standards of performance in the monitoring and measurement process. Specific QC elements can be developed for most, if not all, project activities. All project activities, from sampling (data collection) and laboratory analysis to statistical analysis and reporting, are potential error sources (Peters 1988). Because error is cumulative and can significantly affect the results of a project, all possible efforts must be made to control it. Therefore, quality assurance is a continuous process that should be implemented throughout the entire development and operation of a program.

The purpose of an overall quality assurance project plan (QAPP), containing specific QC elements and activities, is to minimize—and when possible eliminate—the potential for error. Additionally, there are objective mechanisms for evaluating activities relative to pre-established measurement quality objectives and other project goals. The appropriateness of the investigator's methods and procedures and the quality of the data to be obtained must be ensured before the results can be accepted and used in decision making.

QA is accomplished through:

- Program design;
- Investigator training;
- Standardized data gathering and processing procedures;
- Verification of data reproducibility;
- Instrument calibration and maintenance.

As outlined below, QA requirements apply to all activities in an ecological study. More detailed guidance and examples for QA activities should be obtained from USEPA (1994c, and 1998a); more general guidance is outlined by USEPA (1993b).

12.1 Program Design

A central component of QA is overall study design which includes formulation of questions and hypotheses, experimental design, and development of analysis approaches. The classical approach by which scientists plan research consists of the following steps:

- Statement of the problem to be resolved;
- Formulation of alternative hypotheses that will explain the phenomena or, in the case of

problems that do not involve elaboration of processes, formulation of specific research questions;

- Establishment of boundaries within which to resolve the problem;
- Formulation of an experimental or study design that will falsify one or more hypotheses or answer the specific research questions;
- Establishment of uncertainty limits including setting acceptable probabilities of type I and type II errors for statistical hypothesis testing;
- Optimization of the study design including power analysis of the statistical design.

Experimental advances in basic sciences have not included the last two steps because uncertainty limits were inappropriate or unknown. Examination of experimental advances also reveals that a high degree of creativity and insight is required to formulate hypotheses and study designs; no formal planning process or "cookbook" can guarantee creativity and insight. Nevertheless, documentation of the planning process and a complete explanation of the conceptual framework help others evaluate the validity of scientific and technical achievements.

12.1.1 Formulation of a Study Design

A study design is developed to answer the specific monitoring questions developed in formulating the questions and objectives. Sampling design considerations were discussed in Chapter 5.

For quality assurance, some effort will always be required for repeated samples so that measurement error can always be estimated from a subset of sites. Repeated measurement at 10% or more of sites is common among many monitoring programs.

12.1.2 Establishment of Uncertainty Limits

The level of uncertainty associated with environmental measurements (due to natural variability, sampling error, measurement error, or other sources of uncertainty) propagates directly to the uncertainty of inferences and conclusions that can be made from the data. Establishing the limits of statistical uncertainty for conclusions also sets limits for the data to be collected (also known as Data Quality Objectives [DQOs]; Chaloud and Peck 1994). As mentioned in Chapter 5, there is a close association between sampling intensity and uncertainty. Reducing uncertainty usually results in greater costs. Assessing uncertainty, and optimizing the study design (below) require at least pilot data in hand, if not results from one year or more of monitoring.

As an example of uncertainty limits, USEPA's EMAP program established the following (Chaloud and Peck 1994):

- Estimate the status of a population of resources with 95% confidence intervals that are within 10% of the estimate;
- Determine average change in status of 20% over 10 years with 95% confidence and statistical power of 0.8.

EMAP selected 95% confidence intervals, however, there is nothing "scientific" about choosing 95% intervals over, say, 90% or 99%. The second limit above, determining

change, implies that EMAP managers were only willing to conclude a false change in status 1 time out of 20 (Type I error; false positive), but were willing to conclude a false lack of change 1 time out of 5 (Type II error, false negative).

12.1.3 Optimizing the Study Design: Evaluation of Statistical Power

A principal aspect of probability sampling is determining how many samples will be required to achieve the monitoring goals and what is the probability of making an incorrect decision based on the monitoring results. The primary tool for conducting these analyses is statistical power analysis. Evaluating statistical power is key to developing data quality criteria and performance specifications for decision making (USEPA 1996b) as well as evaluating the performance of existing monitoring programs (USEPA 1992). Power analysis provides an evaluation of the ability to detect statistically significant differences in a measured monitoring variable. The importance of this analysis can be seen by examining the possible outcomes of a statistical test. The null hypothesis (H_o) is the root of hypothesis testing. Traditionally, null hypotheses are statements of no change, no effect, or no difference. For example, the mean abundance at a test site is equal to the mean abundance of the reference sites. The alternative hypothesis (H_a) is counter to H_o, traditionally being statements of change, effect, or difference. Upon rejecting H_o, H_a would be accepted.

The two types of decision errors that could be made in hypothesis testing are depicted in Table 12-1. A Type I error (i.e., false positive) occurs when H_o is rejected although H_o is really true. A Type II error (i.e., false negative) occurs when H_o is not rejected although H_o is really false. The magnitude of a Type I error is represented by α and the magnitude of a Type II error is represented by β. Decision errors are the result of measurement and sampling design errors that were described in Section 12.1.1. A proper balance between sampling and measurement errors should be maintained because accuracy limits effective sample size and vice versa (Blalock 1979).

Comparison of Significance Level and Power

Regardless of the statistical test chosen for analyzing the data, the analyst must select the significance level of the test. That is, the analyst must determine what error level is acceptable. The probability of making a Type I error is equal to the significance level (α) of the test and is selected by the data analyst. In many cases, managers or analysts define $1-\alpha$ to be in the range of 0.90 to 0.99 (e.g., a confidence level of 90 to 99%), although there have been environmental applications where $1-\alpha$ has been set to 0.80. Selecting a 95% confidence level implies that the analyst will reject the H_o when H_o is really true, i.e., a false positive, 5% of the time.

Type II error depends on the significance level, sample size, number of replicates, variability, and which alternative hypothesis is true. The power of a test ($1-\beta$) is defined as the probability of correctly rejecting H_o when H_o is false. In general, for a fixed sample size, α and β vary inversely. Power can be increased (β can be reduced) by increasing the sample size or number of replicates. Figure 12-1 illustrates this relationship. Suppose the interest is in testing whether there is a significant difference between the means from two independent random samples. As the difference in the two sample means increases (as indicated on

Table 12-1. Errors in hypothesis testing.

Decision	State of the population (truth)	
	H_o is True	H_o is False
Accept H_o	1-α (Confidence level)	β (Type II error)
Reject H_o	α (Significance level) (Type I error)	1-β (Power)

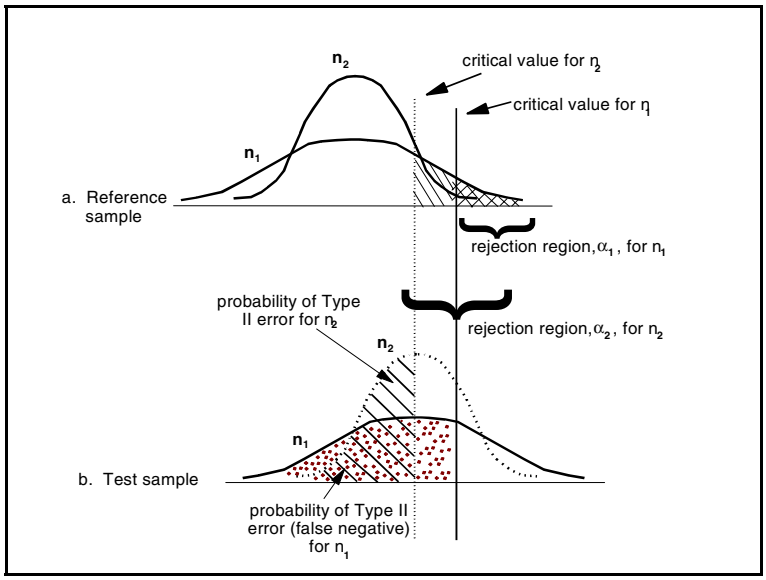

Figure 12-1
Effect of increasing sample size from n_1 to n_2 on power. The curves represent the probability distribution of the sample means from 2 samples, reference and test, and for 2 sample sizes n_1 and n_2 where $n_2 > n_1$.

the *x*-axis), the probability of rejecting H_o, the power, increases. If the real difference between the two sample means is zero, the probability of rejecting H_o is equal to the significance level, α. Figure 12-1a shows the general relationship between α and β if α is changed. Figure 12-1b shows the relationship between α and β if the sample size is increased. The tradition of 95% confidence (α = 0.05) is entirely arbitrary; there is no scientific requirement that confidence be set at 95%. Indeed, for environmental protection, power is at least as important—and possibly more important—than confidence (Peterman 1990, Fairweather 1991).

Basic Assumptions

Usually, several assumptions regarding data distribution and variability must be made to determine the sample size. Applying any of the equations described in this chapter is difficult when no historical data set exists to quantify initial estimates of proportions, standard deviations, means, or coefficients of variation. To estimate these parameters, Cochran (1963) recommends four sources:

- Existing information on the same population or a similar population;
- A two-step sample. Use the first-step sampling results to estimate the needed factors, for best design, of

the second step. Use data from both steps to estimate the final precision of the characteristic(s) sampled;

- A "pilot study" on a "convenient" or "meaningful" subsample. Use the results to estimate the needed factors. Here the results of the pilot study generally cannot be used in the calculation of the final precision because often the pilot sample is not representative of the entire population to be sampled;

- Informed judgment, or an educated guess.

For evaluating existing programs, proportions, standard deviations, means, etc. would be estimated from actual data.

Some assumptions might result in sample size estimates that are too high or too low. Depending on the sampling cost and cost for not sampling enough data, it must be decided whether to make conservative or "best-value" assumptions. Because of the fixed mobilization costs, it is probably cheaper to collect a few extra samples the first time than to realize later that additional data are needed. In most cases, the analyst should probably consider evaluating a range of assumptions regarding the impact of sample size and overall program cost. USEPA recommends that if the analyst lacks a background in statistics, he/she should consult with a trained statistician to be certain that the approach, design, and assumptions are appropriate to the task at hand.

Simple Comparison of Proportions and Means from Two Samples

The proportion (e.g., percent dominant taxon) or mean (e.g., mean number of EPT taxa) of two data sets data sets can be compared with a number of statistical tests including the parametric two-sample t-test, the nonparametric Mann-Whitney test, and two-sample test for proportions (USEPA 1996b). In this case, two independent random samples are taken and a hypothesis test is used to determine whether there has been a significant change. To compute sample sizes for comparing two proportions, p_1 and p_2, it is necessary to provide a best estimate for p_1 and p_2, as well as specifying the significance level and power (1-β). Recall that power is equal to the probability of rejecting H_o when H_o is false. Given this information, the analyst substitutes these values into the following equation (Snedecor and Cochran 1980):

Equation 12-1.

$$n_o = (Z_\alpha + Z_{2\beta})^2 \frac{(p_1 q_1 + p_2 q_2)}{(p_2 - p_1)^2}$$

where Z_α and $Z_{2\beta}$ correspond to the normal deviate. Common values of $(Z_\alpha + Z_{2\beta})^2$ are summarized in Table 12-2. To account for p_1 and p_2 being estimated, t could be substituted for Z. In lieu of an iterative calculation, Snedecor and Cochran (1980) propose the following approach: (1) compute n_o using Equation 12-1; (2) round n_o up to the next highest integer, f; and (3) multiply n_o by $(f+3)/(f+1)$ to derive the final estimate of n.

To compare the mean from two random samples to detect a change of δ; i.e., $\bar{x}_2$-$\bar{x}_1$, the following equation is used:

Equation 12-2.

$$n_o = (Z_\alpha + Z_{2\beta})^2 \frac{(s_1^2 + s_2^2)}{\delta^2}$$

Common values of $(Z_\alpha + Z_{2\beta})^2$ are summarized in Table 12-2. To account

Table 12-2. Common values of $(Z_\alpha + Z_{2\beta})^2$ for estimating sample size for use with Equations 12-1 and 12-2 (Snedecor and Cochran 1980).

Power, 1-β	α for One-sided Test			α for Two-sided Test		
	0.01	0.05	0.10	0.01	0.05	0.10
0.80	10.04	6.18	4.51	11.68	7.85	6.18
0.85	11.31	7.19	5.37	13.05	8.98	7.19
0.90	13.02	8.56	6.57	14.88	10.51	8.56
0.95	15.77	10.82	8.56	17.81	12.99	10.82
0.99	21.65	15.77	13.02	24.03	18.37	15.77

for s_1 and s_2 being estimated, Z should be replaced with t. In lieu of an iterative calculation, Snedecor and Cochran (1980) propose the following approach: (1) compute n_o using Equation 12-2; (2) round n_o up to the next highest integer, f; and (3) multiply n_o by $(f+3)/(f+1)$ to derive the final estimate of n.

A special case of Equation 12-2 arises for biocriteria, when we compare the mean of a sample to determine if the value is below some set limit, that is, if the site is impaired or below a reference threshold. The threshold is fixed by previous investigations and decisions, and is not a random variable. We ask now whether we can detect a change of δ; i.e., $C-\bar{x}_1$, where C is the biocriteria limit:

Equation 12-3.

$$n_o = (Z_\alpha + Z_{2\beta})^2 \frac{(s_1^2)}{\delta^2}$$

In Equation 12-3, Z_α is most often one-tailed, because the concern is only whether the value is below the threshold.

Sample Size Calculations for Means and Proportions

For large sample sizes or samples that are normally distributed, symmetric confidence intervals for the mean are appropriate. This is because the distribution of the sample mean will approach a normal distribution even if the data from which the mean is estimated are not normally distributed. The Student's t statistic ($t_{\alpha/2,n-1}$) is used to compute symmetric confidence intervals for the population mean, μ:

Equation 12-4.

$$\bar{x} - t_{\alpha/2,n-1}\sqrt{s^2/n} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\sqrt{s^2/n}$$

This equation is appropriate if the samples are normally distributed or the sample size is greater than 30 (Wonnacott and Wonnacott 1969), although Helsel and Hirsch (1992) suggest that highly skewed data might require more than 100 observations.

Although several approaches exist to estimate confidence levels for any percentile, many rely on assuming a normal or lognormal distribution. The approach presented here (Conover 1980) for more than 20 observations does not rely on these assumptions. Conover (1980) also provides a procedure for smaller sample sizes. To calculate the confidence interval corresponding to the median, lower quartile, or upper quartile, the following procedure is used.

1. Order the data from smallest to largest observation such that

$$x_1 \le \ldots \le x_r \le \ldots \le x_p \le \ldots x_s \le \ldots \le x_n$$

where x_p corresponds to the median; i.e., p=0.5, lower quartile; i.e., p=0.25, or upper quartile; i.e., p=0.75.

2. Compute the values of r^* and s^* as

Equation 12-5.

$$r^* = np + Z_{a/2}\ (np(1-p))^{0.5}$$

$$s^* = np + Z_{a/2}\ (np(1-p))^{0.5}$$

where $Z_{\alpha/2}$ is selected from a normal distribution table.

3. Round r^* and s^* up to the next highest integers r and s. The 1-α lower and upper confidence limits for x_p are x_r and x_s, respectively.

It can be seen from Equation 12-5 that estimation of medians or quartiles from small samples can result in large confidence intervals for the estimate. For example, the 90% confidence interval for the lower quartile of a sample of n=10 covers the first 5 observations. A sample of less than 10 observations would have a confidence interval extending below the smallest observation. This is the reasoning behind a general "rule of thumb" that estimation of reference conditions should be based on a sample of 10 or more sites, if at all possible. Figure 12-2 gives example sample size calculations for comparing proportions and population means.

12.2 Management

12.2.1 Personnel

Trained and experienced biologists should be available to provide thorough evaluations, provide support for various activities, and serve as QC checks. They should have training and experience commensurate with the needs of the program. At least one staff member should be familiar with establishing a QA framework. QA programs should document personnel responsibilities and duties and clearly delineate project organization and lines of communication (USEPA 1998a). A time line illustrating completion dates for major project milestones or other tasks can be a tremendously useful tool to track project organization and progress.

12.2.2 Resources

Laboratory facilities, adequate field equipment, supplies, and services should be in place and operationally consistent with the designed purposes of the program so that high-quality environmental data can be generated and processed in an efficient and cost-effective manner (USEPA 1992). Adequate taxonomic references and scientific literature should be available to support laboratory work, data processing, and interpretation.

12.3 Operational Quality Control

Protocols should be developed for designing a data base and for screening, archiving, and documenting data. Data screening identifies questionable data based on expected values and obvious outliers. Screening is especially important if data are gathered from a variety of sources and the original investigators and data sheets are no longer available. Figure 12-3 defines the qualitative and quantitative data characteristics that are most often used to describe data quality.

These measurement quality indicators require *a priori* consideration and definition before the data collection begins. Taken collectively, they provide a summary characterization of the data quality needed for a particular environmental decision. Duplication of approximately 10% of the total sampling effort is a common level for

Example 1—Sample size calculation for comparing proportions

To detect a difference in proportions of 0.20 with a two-sided test, α equal to 0.05, *1*-β equal to 0.90, and an estimate of p_1 and p_2 equal to 0.4 and 0.6, n_o is computed from Equation 12-1 as

$$n_o = 10.51 \frac{[(0.4)(0.6) + (0.6)(0.4)]}{(0.6 - 0.4)^2} = 126.1$$

Rounding 126.1 to the next highest integer, *f* is equal to 127, and *n* is computed as 126.1 x 130/128 or 128.1. Therefore 129 samples in each random sample, or 258 total samples, are needed to detect a difference in proportions of 0.2. Since these are proportions, the result means that the total count in the sample must be at least 129. For example, to detect the above difference in the proportion of dominant taxon (e.g., benthic macroinvertebrates or fish) of two sites, at least 129 individuals must be counted and identified in each estuary.

The example illustrates that a statistically significant difference can be easily detected in proportions if sufficient individuals are sampled. However, it is doubtful that a difference between 40% and 60% in dominant taxon is biologically meaningful.

Example 2—Sample size calculation for comparing population mean abundance

To detect a difference of 20 in mean abundance with a two-sided test. The standard deviation, *s*, was estimated as 30 for both samples based on previous studies; α was selected as 0.05; and *1*-β was selected as 0.90. Substituting these values into Equation 12-2 yields

$$n_o = 10.51 \frac{(30^2 + 30^2)}{20^2} = 47.3$$

Rounding 47.3 to the next highest integer, *f* is equal to 48, and *n* is computed as 47.3 x 51/49 or 49.2. Therefore 50 samples in each random sample, or 100 total samples, are needed to detect a difference of 20.

Figure 12-2
Example sample size calculations for comparing proportions and population means.

operational QC. Replication of samples at a randomly selected subset of field sites (usually, 10 percent of the total number is considered appropriate) is used to estimate precision, and representativeness of the samples and the methods. Splitting samples into subsamples can be used to check precision of the methodology, and reprocessing of finished samples is used to check accuracy of laboratory operations.

12.3.1 Field Operations

For the field operations aspect of an ecological study, the major QC elements are: instrument calibration and maintenance, crew training and evaluation, field equipment, sample handling, and additional effort checks. The potential errors in field operations range from personnel deficiencies to equipment problems. Field notes are integral to the documentation of

Figure 12-3
Six qualitative and quantitative data characteristics usually employed to describe data quality.

- ***Precision*** - The level of agreement among repeated measurements of the same characteristic.
- ***Accuracy*** - The level of agreement between the true and the measured value, where the divergence between the two is referred to as bias.
- ***Representativeness*** - The degree to which the collected data accurately reflect the true system or population.
- ***Completeness*** - The amount of data collected compared to the amount expected under ideal conditions.
- ***Comparability*** - The degree to which data from one source can be compared to other, similar sources.
- ***Measurability*** - The degree to which measured data exceed the detection limits of the analytical methodologies employed; often a function of the sensitivity of instrumentation.

activities and can be a potential error source if incorrect recording occurs. Training is one of the most important QC elements for field operations. Establishment and maintenance of a voucher specimen collection should be considered for biological data. Transcription errors during data entry can be reduced with double data entry. Table 12-3 gives examples of QC elements for field and laboratory activities.

12.3.2 Laboratory Operations

The QC elements in laboratory operations include sorting and verification, taxonomy, duplicate processing, archival procedures, training, and data handling. Potential error sources associated with sample processing are best controlled by staff training. Controlling taxonomic error requires well-trained staff with expertise to verify identifications. Counting error and sorting efficiency are usually the most prominent error considerations; they can be controlled by training and by duplicate processing, sorting, and verification procedures.

12.3.3 Data Analysis

Errors can occur if inappropriate statistics are used to analyze the data. Undetected errors in the data base or programming can be disastrous to interpretation. Problems in managing the data base can occur if steps are not taken to oversee the data handling, analysis, and summarization. The use of standardized computer software for data base management and data analysis can minimize errors associated with tabulation and statistical analysis. A final consideration is the possible misinterpretation of the findings. These potential errors are best controlled by qualified staff and adequate training.

12.3.4 Reporting

QC in reporting includes training, peer review, and the use of a technical editor and standard formats. The use of obscure language can often mislead the reader. Peer review and review by a technical editor are essential to the development of a sound scientific document.

Table 12-3 Example QC elements for field and laboratory activities

Project Activity	QC Element	Evaluation Mechanism
Field Sampling	Replicated samples at 10% of sites by same field crew.	Calculate relative percent difference (RPD) of index value or individual metric score
	Replicated samples at one to two of total sites by different field crew using same methods.	Calculate RPDs as above; use to evaluate consistency and bias.
Physical Habitat Assessment (Qualitative)	Ensure appropriate training and experience of operators; multiple observers.	Resume or other documentation of experience; discuss and resolve differences in interpretation.
Physical Habitat Assessment (Quantitative)	Replicated measurements at 10% of sites.	Calculate RPDs between replicate measurements; compare to preestablished precision objectives.
Laboratory: Sample Sorting	Sample residue checked for missed specimens to estimate sorting efficiency; check completed by separate lab staff.	Calculate percent recovery; compare to preestablished goals.
Laboratory: Sample Tracking	Logbook with record of all sample information.	Not applicable.
Laboratory: Taxonomic Identification	Independent identification and/or verification by specialist; ensure appropriate and current taxonomic literature available; adequate training and experience in invertebrate identifications; reference collection; exchange selected samples/specimens between taxonomists.	Calculate percent error; compare to preestablished goals.
Data Management	Proofreading; accuracy of transcription.	All transcribed data entries compared by hand to previous form—handwritten raw data, previously computer-generated tables, or data reports.
Data Analysis	Hand-check of reduced data.	For computer-assisted data reduction, approximately 10% of reduced data recalculated by hand from raw data to ensure integrity of computer algorithm.
	Appropriate statistics; training.	Review by statistician or personnel with statistical training.

Case Study: Optimization of Benthic Sampling Protocols: gear, mesh size, replicates

Ferraro et al. (1994) studied the cost-effectiveness of several alternative marine benthic sampling protocols, including sampling gear, mesh size (0.5-mm or 1.0-mm), and number of replicates (1-10), in southern California waters. Alternative sampling gear was:

- 0.1-m^2 van Veen grab
- 0.06-m^2 van Veen grab
- 0.1-m^2 van Veen grab subsampled by 1-6 core samples, 50-300-cm^2 total area subsampled.

Laboratory processing time was recorded for each sampling alternative. Twelve measures of community structure were examined. Results showed that the power of detecting differences between sites did not increase greatly for more than 4 replicates. Optimum cost-effectiveness was achieved with 5 core subsamples (250-cm^2) of 0.1-m^2 grabs, replicated 4 times at each site (Ferraro et al. 1994).

Case Study: Optimization of Benthic Sampling: Seasonal sampling, trend detection

Alden et al. (1997) examined seasonal and annual trends in estuarine benthic macroinvertebrates community measures (diversity, total abundance, biomass, % opportunities). Samples were taken seasonally (4 x per year) from 16 Chesapeake Bay sites for 9 years. Long-term trends were examined by season, and the power of detecting trends was examined for alternative sampling frequencies of 1 season, 2 seasons, or 4 seasons per year. Finally, reference and impaired sites were compared among seasons to determine if some seasons yield greater power of detection of impairment than other seasons.

Trends in indicator values were apparent and detectable in all seasons. Although 4-season sampling yielded the greatest power of trend detection, it was only marginally better than 2-season sampling and 1 season sampling. In general, summer sampling was most sensitive and yielded the greatest power, allowing detection of trends of 4%-7% change per year in abundance, diversity, and % opportunist metrics over the 9 year period. Biomass was much more variable: the minimum detectable trend was approximately 20% change per year for summer-only sampling (Alden et al. 1997).