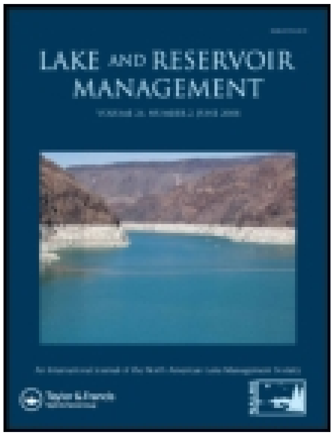

Lake and Reservoir Management

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ulrm20

An analysis of sampling programs to evaluate compliance with numerical standards: total phosphorus in Platte Lake, MI

Eric P. Smith[a] & Raymond P. Canale[b]

[a] Department of Statistics, Virginia Tech, Blacksburg, VA 24061

[b] Emeritus Professor, Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI 48109

Published online: 28 Jul 2015.

Click for updates

To cite this article: Eric P. Smith & Raymond P. Canale (2015) An analysis of sampling programs to evaluate compliance with numerical standards: total phosphorus in Platte Lake, MI, Lake and Reservoir Management, 31:3, 190-201, DOI: 10.1080/10402381.2015.1061073

To link to this article: http://dx.doi.org/10.1080/10402381.2015.1061073

Lake and Reservoir Management, 31:190–201, 2015
© Copyright by the North American Lake Management Society 2015
ISSN: 1040-2381 print / 2151-5530 online
DOI: 10.1080/10402381.2015.1061073

An analysis of sampling programs to evaluate compliance with numerical standards: total phosphorus in Platte Lake, MI

Eric P. Smith[1,*] and Raymond P. Canale[2]
[1]Department of Statistics, Virginia Tech, Blacksburg, VA 24061
[2]Emeritus Professor, Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, MI 48109

Abstract

Smith EP, Canale RP. 2015. An analysis of sampling programs to evaluate compliance with numerical standards: total phosphorus in Platte Lake, MI. Lake Reserv Manage. 31:190–201.

Platte Lake has a numerical standard for total phosphorus mandated in 2000 following judicial proceedings that requires volume weighted total phosphorus to be below 8 mg/m^3 95% of the time. Compliance has been evaluated using an extensive monitoring program that collects samples at roughly biweekly intervals. Analysis indicates that the lake has been and remains in violation of the numerical standard. This large dataset provides a unique opportunity to utilize statistical methods to evaluate compliance while controlling the Type I and Type II errors and to determine the minimum number of samples needed to detect changes in total phosphorus concentrations. Regression and analysis of variance were used to improve the sampling design. Similar techniques were used to design optimal monitoring programs for hypothetical lakes having a large range of total phosphorus concentrations and trophic conditions. Lakes with low total phosphorus concentrations require relatively few measurements during the year, but laboratory uncertainty necessitates multiple replicates. Lakes with higher total phosphorus concentrations require frequent sampling to obtain an acceptable estimate of the annual average concentration because large temporal and depth gradients are expected. Fewer replicates are indicated because laboratory measurements of higher concentrations are more reliable. A review of several national-scale studies indicates that few lakes have sufficient data to perform credible compliance evaluation or reliably detect changes in lake concentrations in response to shifts in watershed loads. Further, many laboratories have limited capability to measure the low concentrations of total phosphorus typically associated with oligotrophic lakes.

Key words: assessment, compliance, oligotrophic, standards, water quality

Platte Lake is located in the northwestern part of the Michigan Lower Peninsula (44.42.091 N, 86.06.820 W; Fig. 1). It is an oligotrophic lake with a volume of 83.5 million m^3, a mean depth of 8.2 m, and a maximum depth of 28.9 m. The Platte River, the major inflow to Platte Lake, has a mean annual discharge of ~3.5 m^3/s (USGS Gage No. 04126740). The hydrograph of the Platte River is stable because the flow is mainly groundwater originating from deep glacial outwash deposits. The mean hydraulic retention time of Platte Lake is ~0.75 years.

Algal growth in Platte Lake is limited by phosphorus, which enters the lake from point and nonpoint sources. The only significant point source of phosphorus in the watershed is the Platte River State Fish Hatchery (PRSFH) operated by the Fisheries Division of the Michigan Department of Natural Resources (Fig. 1). This facility produces coho (*Oncorhynchus kisutch*) and Chinook (*O. tshawytscha*) salmon using surface water that becomes enriched with phosphorus from fecal pellets, urine, and unconsumed feed before entering the Platte River. Platte Lake also has internal phosphorus loads from the release of phosphorus from the bottom sediments during anoxic periods and from the death and subsequent decay of coho and Chinook salmon after spawning.

Phosphorus loading from the PRSFH attained a maximum of ~1960 kg/yr in 1974, which deteriorated the water quality of the lake. Legal actions by local residents challenged this loading, and a subsequent court decision lowered the allowable load to a current level of 79 kg/yr. Further details are described by Canale et al. (2004, 2010).

[*]Corresponding author. epsmith@vt.edu

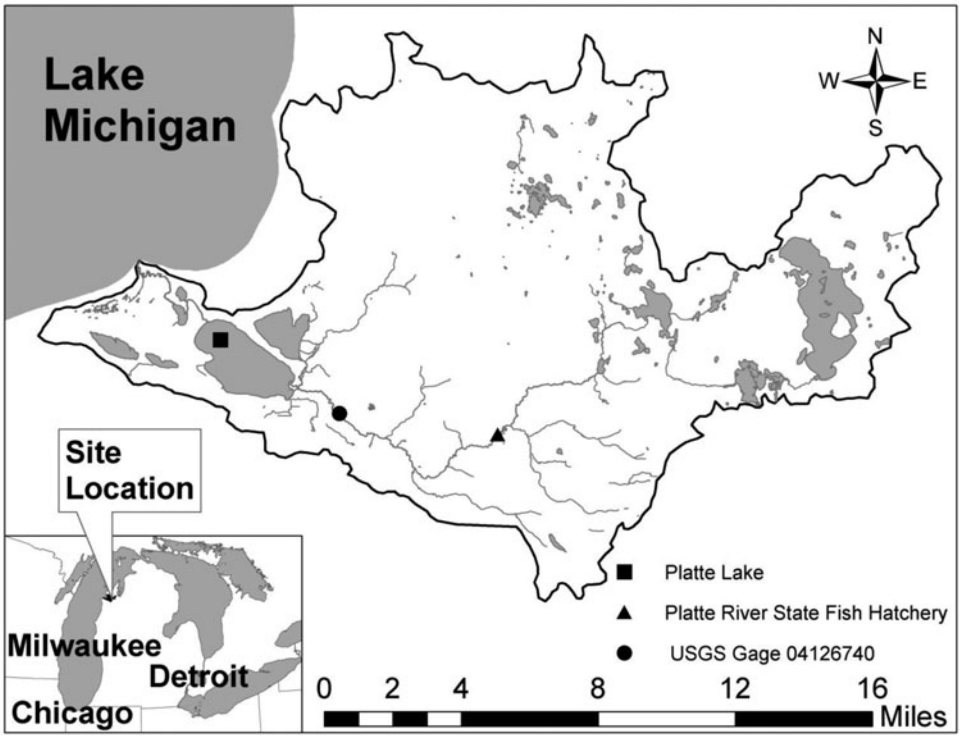

Figure 1. Platte River watershed and lake, gauging station, and hatchery locations.

The court also established a numerical phosphorus standard for Platte Lake specifying that the volume-weighted total phosphorus (TP) concentration should be maintained below 8.0 mg/m^3 95% of the time. Although not explicitly stated, the court order implied that the 95% compliance interval is 1 year. Also implicit in the standard is the assumption that the phosphorus concentration measured at a single, centrally located site in the main deep-water basin sufficiently characterizes the overall water quality of the pelagic regions of the lake. The standard does not consider shallow near-shore areas subject to wind-driven sediment resuspension or inlet tributary plumes where higher concentrations may occur prior to complete mixing with the off-shore waters.

The numerical phosphorus standard for Platte Lake was based on comparisons with the water quality of nearby lakes and an informal review of data available at the time of the 2000 court order. Since that time, an extensive monitoring program has measured the TP concentration in Platte Lake at various depths and times during the year. The main objective of this study was to perform rigorous analyses of these recent data and statistically characterize their adequacy to determine compliance with the numerical standard and to detect changes in the Platte Lake phosphorus concentration subsequent to abatement projects or changes in land use in the watershed.

The number of measurements associated with the current monitoring program for Platte Lake far exceeds typical specifications, and the cost of this intensive sampling effort cannot be sustained over the long-term. A secondary goal of this study was to determine the statistical ramifications of a monitoring program for Platte and other lakes that may have fewer measurement dates, fewer depths, and fewer or no replicates and to provide guidance on compliance monitoring. Because the laboratory measurement process is critical for accurate evaluation of compliance, several national-scale databases were used to evaluate the capability of laboratories to measure water quality.

Methods

Sampling and laboratory methods

Although no specific spatial or temporal measurement frequency was specified in the court order, water samples have been routinely collected every 2 weeks since 1990, weather and ice conditions permitting. Replicate samples are analyzed in triplicate for TP and turbidity for each sampling date and depth. Secondary data include *in situ* measurements of Secchi depth, temperature, dissolved oxygen, pH, conductivity, and oxidation–reduction potential. Other chemical and biological parameters have been measured with somewhat less frequency, including nitrate, nitrite, alkalinity, total dissolved solids, calcium, chlorophyll *a*, phytoplankton, and zooplankton.

The TP concentration was measured in the laboratory using the acid persulfate digestion-ammonium molybdate method (Eaton et al. 2005). Measurements were performed at Central Michigan University (CMU) prior to 2012 and at PRSFH after 2011. Note that the expected TP concentrations in Platte Lake (and other oligotrophic lakes) are near the detection level for most laboratory operations; therefore a number of quality control measures were implemented including the application of a spectrophotometer cuvette path length of 10 cm that increases the accuracy and precision of low concentration measurements. The laboratories performed well during independent proficiency testing in 2005, 2012, and 2014 conducted by Environment Canada (DeOliveira et al. 2013).

Description of total phosphorus measurements

The general spatial and temporal trends of the phosphorus data is described prior to more formal statistical analyses. Although all data collected after 2000 could be included, the analyses was confined to 3504 measurements collected between 2005 and 2013 where field, laboratory, and quality control procedures were the most rigorous and consistent. The analyses do not include a few outliers or incomplete datasets due to equipment malfunction or lost samples. Outliers occur, particularly in surface samples, when a measurement is affected by windrows of pollen, sloughed macrophyte material, or the shell casings of burrowing mayflies (*Hexagenia* sp.). Outliers in deeper samples usually result from strong winds that dislodge the boat anchor and resuspend sediments.

Seasonal dynamics

The median value of triplicate measurements of TP at was recorded at 8 different depths in 2009 (Fig. 2a). The first sample was taken under the ice in late January during stratified conditions. The relatively high bottom water phosphorus concentrations are an indication of sediment release under low dissolved oxygen concentrations (Fig. 2b). Note that by day 100, the ice had broken up, depleted dissolved oxygen concentrations had been restored, and the concentrations of TP were low and relatively uniform from top to bottom. The TP concentrations in the surface and mid-depth layers increased between days ~120 and 225 and then gradually decreased toward the end of the year. These changes are related to high spring flows in the Platte River and associated phosphorus loading. Bottom water TP concentrations increased during the summer in response to low dissolved oxygen concentrations that persist until fall turnover, around day 280 (Canale et al. 2010). Late summer depth gradients are also the result of settling and accumulation of algal cells and other particulate phosphorus in the bottom waters.

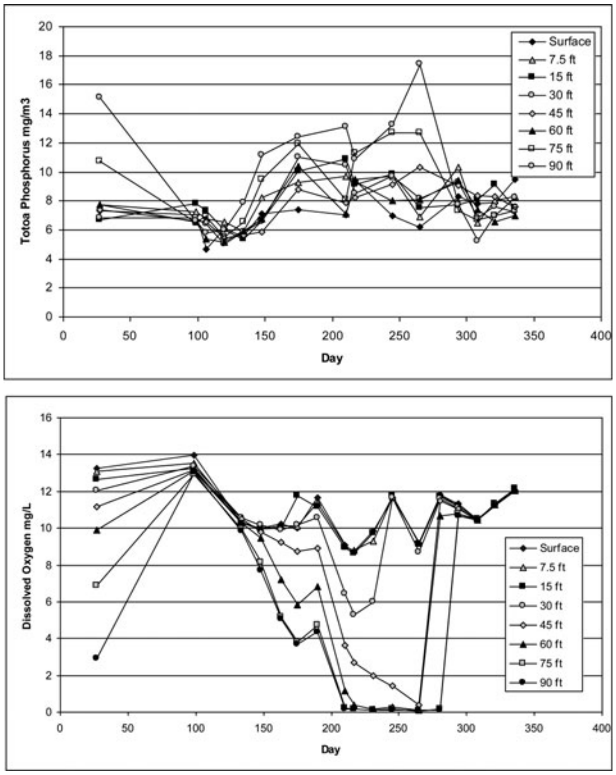

Figure 2. Platte Lake (a) total phosphorus and (b) dissolved oxygen measurements on various dates during 2009 at 8 different depths.

Depth gradients

Vertical TP concentration gradients (Fig. 3) from the 2005–2013 data for 8 depths indicate that the lowest concentrations tended to occur in the surface layer. Concentrations in the middle 5 layers were relatively uniform, and the concentrations in the bottom 2 depths were generally higher than the other layers.

Although higher concentrations occurred in the lower layers, these bottom waters represent a relatively small percent of the total Platte Lake volume. The volume fractions (λ) of the 8 layers (Fig. 4) vary considerably, with the bottom 2 layers representing only ~2.5% of the total water volume. The volume-weighted concentration (p) on any given day was calculated by multiplying the median concentration of 3 replicates measured at each depth (TP_i) by the fraction of the total volume represented by that depth and adding for all

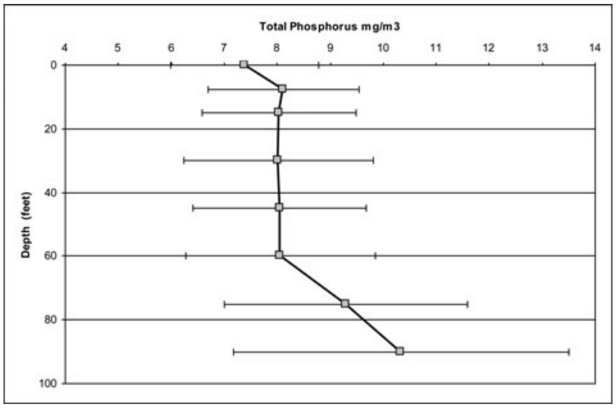

Figure 3. Measured total phosphorus concentrations in Platte Lake for 2009 to 2013 sorted by depth (mean ±1 standard deviation).

8 depths, as shown in equation 1 with $m = 8$:

$$p = \sum_{i=1}^{i=m} \lambda_i \cdot TP_i. \tag{1}$$

The median was used in the calculation rather than the mean of 3 observations to reduce the influence of occasional extreme measurements. Thus, a single volume-weighted value (p) was based on 24 separate laboratory measurements of the TP concentration. The average phosphorus concentration for all depths and all years was 8.41 mg/m^3 without consideration of volume weighting. The average concentration was reduced to 8.01 mg/m^3 after applying the weighting factors (Fig. 4). Mixed-model analysis of variance was used to compare and model depth, year, and month effects (Proc Mixed; SAS Institute Inc. 2008). The model allowed for spatial and temporal correlation in the measurements.

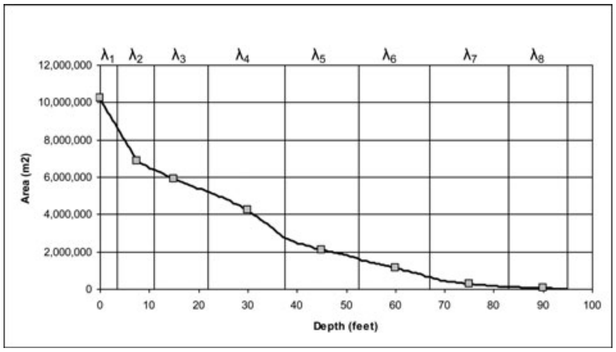

Figure 4. Area of Platte Lake as function of depth. Sample depth locations are shown as shaded squares. The vertical lines divide the lake into 8 layers with volume fractions $\lambda_1 = 0.121$, $\lambda_2 = 0.2095$, $\lambda_3 = 0.2311$, $\lambda_4 = 0.2264$, $\lambda_5 = 0.1269$, $\lambda_6 = 0.0622$, $\lambda_7 = 0.0208$, and $\lambda_8 = 0.0039$.

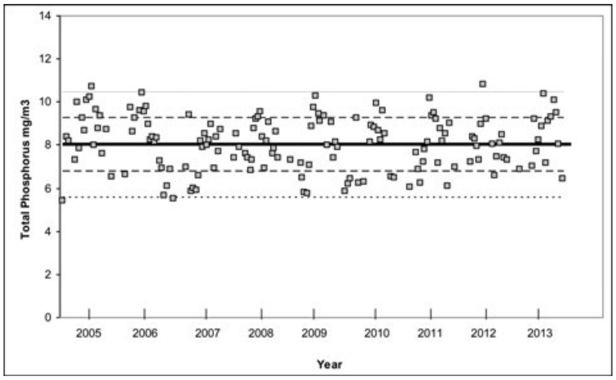

Figure 5. Long-term trend of volume-weighted total phosphorus concentrations in Platte Lake from 2005 through 2013. The dashed and dotted horizontal lines indicate the ±1 and ±2 standard deviations.

Long-term and seasonal trends

From 2005 to 2013, there were 146 measurements of the 8-layer volume-weighted TP concentration (p) (Fig. 5). Only 47% of the values were <8.0 mg/m^{3}compared to 95% required by the numerical standard; thus, the TP concentrations do not comply with the numerical standard. The standard deviation (σ) of these volume-weighted measurements was 1.23 mg/m^3. Subsequent analyses here assume that this value is a close approximation of the population standard deviation because of the large number of measurements. The horizontal lines (Fig. 5) show the mean and the ±1 and ±2 standard deviation values. Of the values, 67% are contained within 1 standard deviation, and 98% are within 2 standard deviations (for a normal population, 68% and 95% would be expected), and 53% exceed the mean, indicating a small amount of negative skewness.

The statistical analyses (discussed in following sections) were performed on the raw values as well as the log-transformed values to compensate for skewness. In all cases, the log-transformed results were similar to the nontransformed analyses and for the sake of brevity are not reported here. The monthly average value of the volume-weighted measurements increased over the year with a maximum in midsummer (Fig. 6). The general seasonal trend was consistent with the earlier discussion of the 2009 data (Fig. 2).

Sample size requirements

Measurements of the TP concentration in Platte Lake were taken to (1) determine if the lake complies with the numerical standard for TP, and (2) reliably measure changes in TP concentrations subsequent to increases or decreases of the loading from point and nonpoint sources.

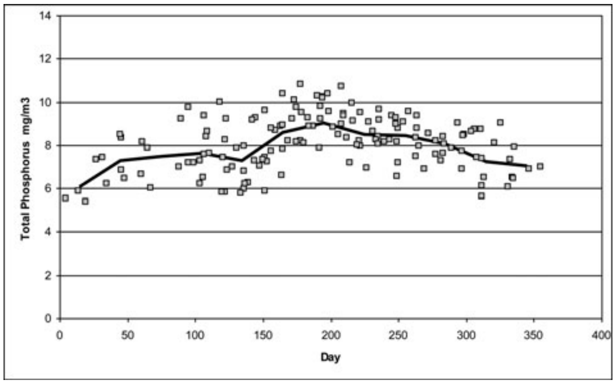

Figure 6. Seasonal variation of the volume-weighted total phosphorus concentrations in Platte Lake for data from 2005 through 2013.

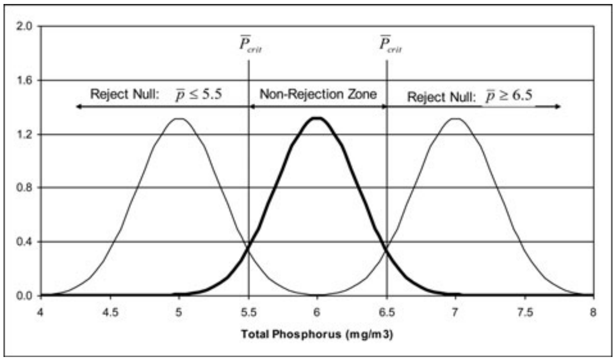

Figure 8. Compliance and alternative distributions associated with hypothesis testing for the numerical phosphorus standard for Platte Lake. The null hypothesis is represented by the distribution in the center. The alternative hypotheses are shown on the left and right of the null hypothesis. The non-rejection zone is defined by vertical lines at 5.5 and 6.5 mg/m^3. The Type I and Type II error levels were set to 0.05.

Sample size to evaluate compliance

The 95% attainment requirement specified by the court-ordered numerical standard can be expressed as an equivalent population mean value (μ) using equation 2 if the data can be approximated with a normal distribution and a reliable estimate of the population standard deviation is available as described above. This equivalent population mean is given by

$$\mu = 8.0 - z_{0.05} \cdot \sigma = 6.0 \text{ mg/m3} \tag{2}$$

for $z_{0.05} = 1.645$(95% cumulative z-score) and $\sigma = 1.23$ mg/m^3. Thus, the numerical standard will be satisfied 95% of time if the annual average TP concentration is $\le$6.0 mg/m^3 (Fig. 7).

We next determined the number of random volume-weighted measurements (p) required during the year to evaluate compliance with the desired population mean ($\mu =$ 6 mg/m^3) with a specified level of statistical confidence. The annual average of this set of measurements is formally defined here as a *sample* (with italics) and is represented as ($\bar{p}$). The number of measurements (n) actually taken during the year to calculate this annual average value is the *sample* size. The Central Limit Theorem states that the sampling distribution of multiple random *sample* means will have a normal distribution. This distribution of these *sample* means has a mean value of μ and a standard deviation of$\sigma/\sqrt{n}$.

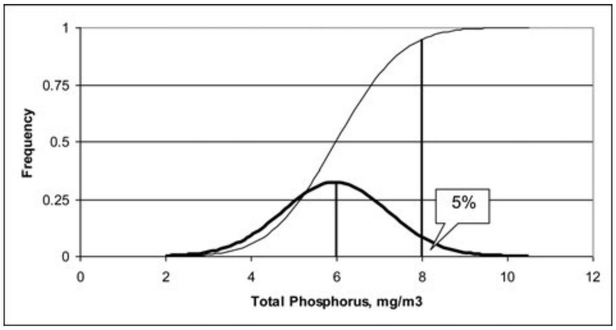

Figure 7. Normal distribution with mean of 6.0 mg/m^3 and standard deviation of 1.23 mg/m^3 equivalent to the numerical standard for Platte Lake requiring the total phosphorus concentration be <8 mg/m^3 95% of the time.

Three distributions were calculated (Fig. 8) with equal standard deviations ($\sigma = 1.23$ mg/m^3). The middle distribution is associated with the numerical standard and has a mean value of 6 mg/m^3. The upper and lower distributions are located 1 mg/m^3 above and below the numerical standard. The *sample* size that simultaneously controls the Type I and Type II errors is a function of the population standard deviation, the Type I (α) and Type II (β) significance levels, and the mean of the alternative population associated with the Type II error (μ_a). Equation 3 can be used to calculate the value of n that satisfies these requirements for a one-sided test with z-scores of z_α and z_β:

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 \cdot \sigma^2}{(\mu - \mu_a)^2}. \tag{3}$$

The center distribution and the upper and lower distributions intersect at the critical population parameter value ($\bar{P}_{crit}$), a function of the *sample* size and the population standard deviation, given by equation 4:

$$\bar{P}_{crit} = \mu \pm z_\alpha \cdot \frac{\sigma}{\sqrt{n}}. \tag{4}$$

In this example with alternative population means = 5.0 and 7.0 mg/m^3, the calculated *sample* size is 17 (rounded up from 16.4) for $\alpha = \beta = 0.05$ with the$\bar{P}_{crit}$ values equal to 5.5 and 6.5 mg/m^3. The power to detect a change for such a monitoring program is 0.95.

Alternative null hypotheses

Compliance is evaluated by comparing a *sample* ($\bar{p}$) value to the assumed hypothetical population with $\mu = 6.0$ mg/m^3, equivalent to the numerical standard. Hypothesis tests are used to quantify the comparison process. Two alternative null hypotheses can be used to test for compliance: one approach assumes that the *sample* is in compliance with the numerical standard; the other starts with the initial assumption that the *sample* violates the standard (McBride and Ellis 2001).

If the hypothesis is that Platte Lake is not in violation, then any measured *sample* mean ($\bar{p}$) less than or equal to $\bar{P}_{crit} =$ 6.5 mg/m^3 with $n = 17$ does not provide sufficient evidence to reject the null hypothesis. If the hypothesis is that Platte Lake is in violation, then the *sample* mean ($\bar{p}$) must be $\leq$5.5 mg/m^3 with $n = 17$ to provide sufficient evidence that the lake is not in violation. The non-reject regions suggest a boundary for decisions. If the *sample* mean is midway between 6.0 and 7.0 mg/m^3 or between 5.0 and 6.0 mg/m^3, then it is equally likely to have come from the alternative population, although both have low probabilities (Fig. 8). The boundary where the null is difficult to distinguish from the alternative is sometimes referred to as "guard points" (Esterby 2013), that is, points that define a region where it is "too close to call."

Sample size to detect changes

Here we distinguished between 2 populations with means of μ_1 and μ_2 using a *sample* mean ($\bar{p}_1$) of size n_1 taken from μ_1 and another *sample* mean ($\bar{p}_2$) of size n_2 taken from μ_2. The difference between the *samples* is given by$d = \bar{p}_1 - \bar{p}_2$. Several *samples* could be collected, resulting in a probability distribution for *d.* The standard error of the *d* distribution (i.e., the standard deviation of the distribution of differences) is given by equation 5, where s_1 and s$_2$ are the standard deviations of the *samples*:

$$s_{\bar{p}_1-\bar{p}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \qquad (5)$$

A hypothesis test can be performed to determine if the difference between the *sample* means is significant. The null hypothesis is that the difference between the *samples* is zero, and the alternative is that the difference is not equal to zero. The null hypothesis is rejected if the calculated value of *d* is outside the non-rejection zone, as defined by equation 6:

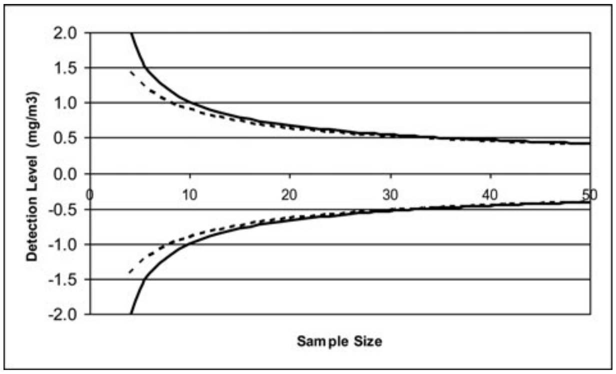

Figure 9. Detection level for 2 sample means as function of sample size with a significance level of 0.1. The solid curves were developed assuming a normal distribution. The dashed curves assume the Student-*t* distribution.

$$d = \pm t_{\alpha,df} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \qquad (6)$$

The value of $t_{\alpha,df}$ in equation 6 is the *t*-value associated with an upper level of α for the Student *t*-distribution with degrees of freedom (*df*) calculated using equation 7:

$$df = \frac{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}{\frac{1}{n_1 - 1} \cdot \frac{s_1^2}{n_1} + \frac{1}{n_2 - 1} \cdot \frac{s_2^2}{n_2}}. \qquad (7)$$

Curves were developed for *d* values for Platte Lake as a function of the *sample* size with a confidence level of $\alpha =$ 0.1 using both Student-*t* and normal distributions (Fig. 9) assuming that $n_1 = n_2$ and s$_1 = s_2 = \sigma = 1.23$ mg/m^3. We observed that the ability to detect differences with confidence increases as *n* increases. Note that if the detection level and confidence are specified, equations 6 and 7 can be used to calculate the required *sample* size. In contrast, if the *sample* size is based on budget or logistical constraints, then the trade-offs between detection level and confidence are determined by the *sample* variance. For example, a *sample* size of ~34 is required to detect an increase or decrease of 0.5 mg/m^3 of TP with $\alpha = 0.1$. If the *sample* size is reduced to 17, the confidence level associated with detection of 0.5 mg/m^3 decreases to $1 - \alpha = 0.75$. Alternatively, the 17 *sample* size can identify changes with the significance level of α maintained at 0.1, but the detection level must be increased to 0.75 mg/m^3.

An analysis of the measurement process also involves laboratory evaluation. To aid this analysis, we used data from Environment Canada (EC), which provides accredited proficiency testing (PT) services for a wide range of inorganic

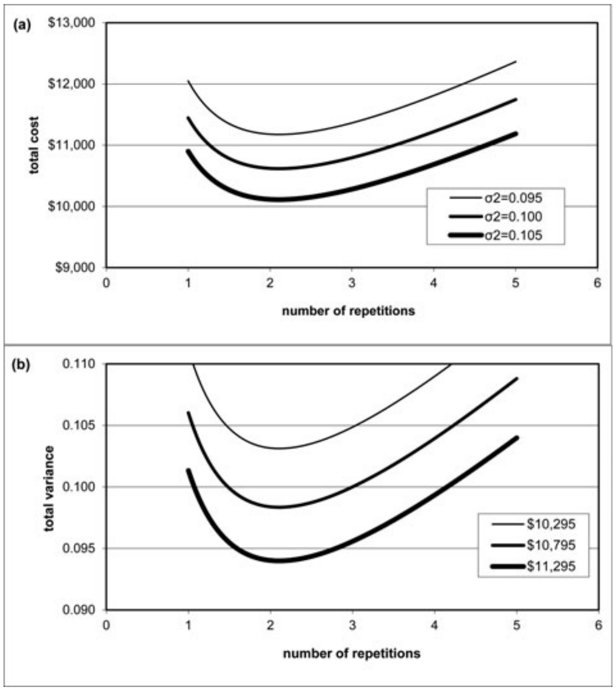

Figure 10. (a) Total annual cost of sampling program as function of the number of laboratory replications for 3 different values of the annual average variance. (b) Annual average variance as function of the number of laboratory replications for 3 different values of the total annual sampling cost.

constituents in water. These studies are designed to quantify laboratory performance and improve the quality of environmental data and provide information on the relationship between the percent coefficient of variation ($CV = 100 \cdot \sigma/\mu$) and the mean concentration of TP for various laboratories. EC proficiency tested CV values are summarized for >12,000 records where different laboratories participated in proficiency testing for measurement of TP at various concentrations. Results for the CMU and PRSFH laboratories are based on >1000 repeated measurements of TP with known concentrations. The Upstate Freshwater Institute results for CV are from Effler and O'Donnell (2010).

Results

Analysis of the Platte Lake sampling program

Eight vs. three layer volume-weighted average

The monitoring program described above has generated sufficient data to determine if fewer than 8 depths can be used to reliably estimate volume-weighted average concentrations. Inspection of the vertical concentration profile (Fig. 3) suggests that the volume-weighted average TP concentration could be adequately characterized using 3 rather than 8 layers. Water samples from 7.5, 15, 30, 45, and 60 foot depths and the 75 and 90 foot depths could be combined to create 2 composites prior to laboratory measurement. In this case, $m = 3$ where $\lambda_1 = 0.121$ to represent the volume fraction of the surface layer, $\lambda_2 = 0.854$ for the volume fraction of the middle 5 layers, and $\lambda_3 = 0.025$ for the bottom 2 layers. Linear regression that compares the 8- and 3-layer approaches results in a slope of 0.999 and an intercept of 0.02 with $R^2 = 0.979$, suggesting a 1:1 relationship. These results indicate that a future monitoring program could confidently employ 3 rather than 8 layers to determine the volume-weighted concentration.

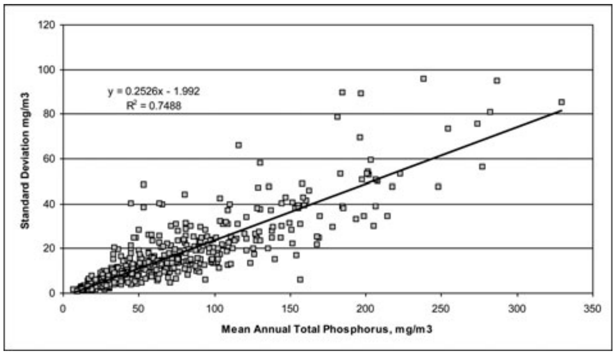

Figure 11. Water Quality Portal data for the estimated standard deviation (equation 13) and annual average total phosphorus for various lakes with 10 or more sample dates (y = standard deviation, x = mean annual total phosphorus).

Variance partitioning

Currently, triplicate measurements are employed in the monitoring program for quality control because typical concentrations in Platte Lake are near the detection level where

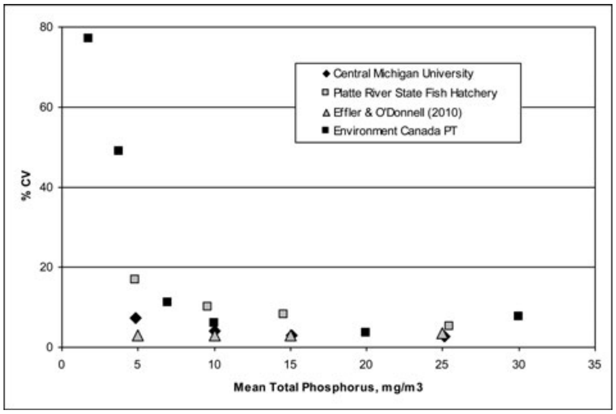

Figure 12. Coefficient of variation for measurement of a range of total phosphorus concentrations reported by various laboratories.

Table 1. Date and replication variance, optimal number of replicates (rounded up), actual number of dates, and number of equivalent dates. Calculations assume that the cost of a single sample collection (c_{dates}) is \$500 and the replicate cost (c_{reps}) is \$120 for 8 depths for 2005 to 2013. The cost of a future replicate after 2013 is \$45 for 3 depths.

Year	σ^2_{dates}	σ^2_{reps}	$\hat{n}_{reps}$	$\mathbf{n_{dates}}$	$\hat{n}_{dates}$
2005	1.51	0.274	1	18	18
2006	2.79	0.200	1	17	17
2007	1.68	0.206	1	17	17
2008	0.69	0.272	2	19	20
2009	1.87	0.269	1	15	15
2010	1.78	0.303	1	15	15
2011	1.61	0.312	1	17	18
2012	0.58	0.530	2	14	21
2013	1.37	0.706	2	14	16
future	1.50	0.600	3	17	20

colorimetric measurements are relatively imprecise (Drolc and Ros 2002). These replicates triple the laboratory cost, however, and therefore we conducted formal statistical analyses to determine how much confidence would be lost by reducing or eliminating the replicates or be gained by increasing the number of replicates.

Mixed-model ANOVA was performed on the 8-depth volume-weighted measurements for 2005–2013 to separate the variance within the 3 replicates from the variance between dates. The calculated F statistic for each year was much larger than the critical value, indicating that changes between the median of the replicates that occur between dates during the year are statistically significant despite this variation being partially obscured by variations within the 3 replicates. The partitioning of the total variance into the variance associated with the sample dates (σ^2_{dates}) and the variance of the replicates (σ^2_{reps}) was calculated (Table 1).

The variance associated with a single volume-weighted measurement (p) is given by equation 8:

$$\mathrm{var}(p) = \sigma^2_{date} + \sigma^2_{reep}. \qquad (8)$$

The variance of the *sample* ($\bar{p}$) is calculated using equation 9, where the *sample* size is equal to n_{dates} and n_{reps} is the number of replicates:

$$\mathrm{var}(\bar{p}) = \frac{\sigma^2_{date}}{n_{dates}} + \frac{\sigma^2_{rep}}{n_{dates} \cdot n_{reps}}. \qquad (9)$$

Optimal number of replicates

We determined the optimal number of replicates relative to the number of dates that minimizes the annual average variance of the samples (equation 9) for a fixed cost or minimizes the total annual cost for a fixed annual average variance. The total annual cost (C_{total}) for sample collection and laboratory measurements is determined by equation 10.

$$C_{total} = c_{dates} \cdot n_{dates} + c_{reps} \cdot n_{reps} \cdot n_{dates}. \qquad (10)$$

The formula for the optimum number of replicates ($\hat{n}_{reps}$) is derived by rearranging equation 10 to express the number of dates as a function of total cost, the cost of the dates, and the cost and number of replicates, substituting into equation 9, and minimizing with respect to the number of replicates, resulting in equation 11 (Marcuse 1949, Sokal and Rohlf 1969):

$$\hat{n}_{reps} = \sqrt{\frac{c_{dates} \cdot \sigma^2_{reps}}{c_{reps} \cdot \sigma^2_{dates}}}. \qquad (11)$$

Note that more repetitions are warranted when the variance within the replicates is high and the cost of the replicates is low; however, fewer replicates are indicated when the cost of the dates is low and the variance between the dates is high. If the cost of adding dates is similar to the cost of adding replicates, then performing multiple replicates is only warranted if the variation in the measurement process is large relative to the variation between dates. If the cost of adding sampling dates is low or the cost of sample measurement is high, then it is generally better to reduce the number of replications relative to the number of sampling dates.

The direct and indirect costs of collecting water samples from Platte Lake are larger than the costs of measuring the 8-layer volume-weighted phosphorus concentration in the laboratory. The total cost associated with sending a 2-person crew into the field to collect a sample is roughly \$500. The cost of a single in-house volume-weighted measurement is ~\$15 per laboratory analysis or \$120 for one 8-layer volume-weighted measurement. The optimal number of replicate measurements as calculated by equation 11 was roughly one for 2005 to 2011 (see Table 1). This number increased to about 2 replicates in 2012 and 2013 because the replication variance increased. If the cost of a future sample collection program is reduced to \$45 by utilizing 3 layers instead of 8, as discussed earlier, the optimal number of repetitions increases to 2.1. A suboptimal monitoring program that consists of 17 dates and 3 replications seems prudent. This program will cost slightly more (\$10,795) compared to the optimal value (\$10,113), but this additional cost is offset by a decrease in the sample variance to 0.100 from 0.105 $(\mathrm{mg/m^3})^2$ and adds a level of quality control. The total annual cost of this program varies as a function of the number of replications for 3 different values of the *sample* variance (Fig. 10a). The *sample* variance also varies as a function of the number of repetitions for 3 different fixed costs (Fig. 10b). Note that because the cost is fixed, increasing the number of replicates would require a reduction in the number of dates.

The effective sample size $\hat{n}_{dates}$ that employs only one replicate but achieves the same variance as $\hat{n}_{reps}$is given by equation 12:

$$\hat{n}_{dates} = \frac{\sigma^2_{date} + \sigma^2_{rep}}{\frac{\sigma^2_{date}}{n_{dates}} + \frac{\sigma^2_{rep}}{n_{dates} \cdot n_{reps}}}. \tag{12}$$

Calculated values of $\hat{n}_{dates}$ for various years (Table 1) show that the practice of using about 2 replicate measurements is equivalent to adding about 3 additional dates with one replicate per date when the replication variance contribution is relatively high.

Example sampling programs for other lakes

For this example, assume that a volume-weighted sample based on measurements at 3 depths is taken from a centrally located, deep-water lake site several times per year and measured for the TP concentration. The annual average of these measurements is then compared with a numerical standard to determine compliance. Five hypothetical lakes were examined: 2 oligotrophic lakes with annual average TP concentrations of 5 and 10 mg/m^3; a mesotrophic lake with an annual average concentration of 15 mg/m^3; and 2 eutrophic lakes with annual average concentrations of 25 and 50 mg/m^3. The analyses below determine the sample size needed to obtain a statistically valid estimate of the annual average concentration for comparison with the standard.

Estimates of the population standard deviation as well as the mean concentration are needed to proceed with statistical analysis and sampling program design. Unfortunately, direct calculations of TP standard deviations based on extensive monitoring data are not generally available for a wide variety of lakes; however, the standard deviation (σ) can be estimated for initial planning and sampling program design applications using equation 13 (Zhang 2007):

$$\sigma \approx \frac{\max TP - \min TP}{6}. \tag{13}$$

More than 15,000 records from the Water Quality Portal (WQP) database provided measured values for the maximum, minimum, and annual average TP concentrations as well as the number of samples taken during the year. This cooperative program integrates publicly available water quality data from the US Geological Survey National Water Information System, the Environmental Protection Agency STORET data warehouse, and the US Department of Agriculture research database (National Water Quality Monitoring Council 2006).

Table 2. Computations for the number of sample dates (rounded up) and optimal number of replicates (rounded up) for various hypothetical lakes. The population standard deviations are estimates based on Figure 11. Calculations assume a confidence interval of 90% for the mean and a Type I significance level of 0.1. The assumed cost of sample collection is $500 per date and the laboratory cost is $75 per measurement.

Parameter	Lake 1	Lake 2	Lake 3	Lake 4	Lake 5
Population mean (μ)	5	10	15	25	50
Population standard deviation (σ)	0.5	1.5	2.5	5	10
Variance(date)	0.25	2.25	6.25	25	100
Variance(rep)	0.66	0.95	1.4	1.7	3.1
Sample size (n_{dates})	11	25	31	44	44
Lab cost	$75	$75	$75	$75	$75
Date cost	$500	$500	$500	$500	$500
Optimal lab reps (n_{reps})	5	2	2	1	1

The plot of the relationship between the standard deviation calculated using equation 13 and the annual average TP concentration for 530 lakes having 10 or more measurements per year shows that as the annual average TP concentration increases, the standard deviation also increases (Fig. 11; see also Jones and Bachmann 1976, Carpenter and Brock 2006). This result is not unexpected because concentration dynamics and depth gradients generally increase in lakes with higher annual average concentrations, indicated by the annual average and estimated standard deviation (Fig. 11) for the 5 hypothetical lakes (Table 2).

Evaluation of the EC data show that the relative uncertainty in the laboratory measurements of TP increases dramatically when the concentrations are less than ~10 mg/m^3 (Fig. 12). Values for the laboratory repetition variance for the TP measurements for this example based on PRSFH results (Fig. 12) indicate that high quality lakes with low TP concentrations have relatively low seasonal variance, but these low concentrations are difficult to reliably measure in the laboratory (Table 1). In comparison, lakes with higher TP concentrations are easier to measure reliably in the laboratory, but this advantage is offset by relatively high seasonal standard deviation and variance.

To determine an adequate *sample* size, the statistical requirements associated with the calculations must be specified. We assumed for this example that the Type I significance level (α) is 0.1 and that the desired confidence interval width is equal to 10% of the mean. These are arbitrary values, and regulators may impose different specifications. The number of sample dates needed to estimate the annual average concentration with confidence is calculated using equation 14:

$$n = \left[\frac{2 \cdot z_\alpha \cdot \sigma}{Confidence\ Interval\ width} \right]^2 . \qquad (14)$$

Lakes with higher TP concentrations have higher standard deviations (Table 2) and therefore require more sample dates to satisfy the statistical specifications.

Next, we determined if replicate laboratory measurements are warranted. Equation 11 can be used to calculate the optimal number of laboratory repetitions that minimize the total annual variance of the measurements for a fixed cost. The typical cost for a commercial laboratory to measure TP is \$25, or \$75 for a volume-weighted sample (Table 2). The estimated cost to collect a single sample for this example is \$500. From this information, the calculated optimum number of replicates may be calculated (Table 2).

Note that, as expected, the lowest concentration lake requires 5 replicates (rounded up) compared to only 1 (rounded up) for the higher concentration lakes. A total of 55 measurements is required to quantify the annual average phosphorus concentrations in the oligotrophic lake compared to 44 total measurements in the eutrophic lakes. Laboratory costs are 75% of the total sampling program cost for the lowest concentration lake compared to only ∼15% for the high concentration lake.

Discussion

The goal to manage the water quality of lakes based on numerical standards and quantitative analyses is certainly worthwhile; however, such an approach must be accompanied and supported by extensive data and rigorous statistical analysis. The above analyses show that lakes with low TP concentrations require multiple laboratory replicates to ensure reliable sample measurements. These low concentrations also require a high commitment to quality assurance and control. Information on the number of laboratories and their reported quantification criteria from the WQP database (Table 3 shows that only ∼8% of the reporting laboratories can reliably measure TP concentrations <5 mg/m^3, and 82% indicated that the lowest concentration that could be reliably measured was >20 mg/m^3. The above analyses also show that lakes with high annual average TP concentrations require multiple sampling dates to ensure reliable estimation of the annual average concentration because these lakes tend to have large standard deviations over the year. About 3% of the WQP database records for the annual average TP concentration consisted of 10 or more measurements, and no records reported replicates.

Table 3. Number of laboratories (2464) and quantification levels for total phosphorus measurements for Water Quality Portal data collected from 1 January 2009 to 1 August 2014. Values are represented in mg/m^3.

Quantitation Criterion	Number
Lower reporting limit ≤5	37
Lower reporting limit ≥10	157
Method detection level ≤5	169
Method detection level >5	67
Practical quantitation limit ≥20	2034

We recommend that the monitoring specifications for compliance be linked to the constraints imposed by the regulation and standard (Barnett and O'Hagan 1997, Shabman and Smith 2003, Esterby 2013). For example, the numerical standard for Platte Lake simply stating that 95% of measurements should be <8.0 mg/m^3 is vague with regard to implementation and different agencies, or researchers might create different sampling strategies for assessment. Although the population might be defined as all potential measurements within the lake on a certain day, week, month, or year, a more useful approach would be to specify depths, horizontal locations, and sampling time intervals. Blueprints that specify the population of potential measurements should be assessed and enumerated by regulators.

Most statistical algorithms such as power analysis require a reliable numerical value for the population standard deviation. During the initial planning stage, range data may be used to estimate the population standard deviation as described earlier, but as more data become available it may be appropriate to refine this estimate. With large changes in water quality, the population standard deviation may also be expected change over time, particularly when a lake moves from one trophic status to another. For cases where abatement programs are in progress, however, concentrations are decreasing, and along with these decreases the standard deviation is also expected to decrease. Therefore, the assumption of constant variance for these types of cases has a built-in safety margin when calculating sample size.

The statistical approach used here requires random samples of the statistical population. In many lakes, however, summer and early fall concentrations may be higher than those taken in the winter and early spring, and surface concentrations may be lower than bottom concentrations (Knowlton and Jones 2006); thus, any monitoring program that favors summer or surface samples will be biased if the standard implies annual compliance. The numerical phosphorus standard for Platte Lake assumes that a single horizontal site adequately characterizes the overall water quality of Platte Lake. This assumption is clearly not adequate for elongated systems such as Lake Champlain (Smeltzer et al. 2009) or for cases where distinct phosphorus concentration differences are found in different regions or basins, such as in Lake Erie.

Enforcement of a numerical standard and conclusions regarding compliance depends on the perspective of the null hypothesis (Bross 1985, Millard 1987, Holland and Ordoukhani 1990, Smith et al. 2001). In the case of Platte Lake, if the null hypothesis assumes that the lake is not in violation, any sample consisting of 17 random volume-weighted measurements >6.5 mg/m^3 will be deemed in violation. Alternatively a sample must be <5.5 mg/m^3 to conclude that the lake is not in violation when it is initially assumed to be in violation. Samples <5.5 mg/m^3 are difficult to obtain because any contamination during sample collection, handling, and analysis almost always results in bias that tends to overestimate the actual concentration. Yet allowing a few unrepresentative high values to prematurely declare violation of the standard seems inappropriate. These circumstances place additional and unwarranted burdens when the null hypothesis assumes that the lake is not in compliance with the standard; however, these circumstances act as an additional safety factor when applying the null hypothesis that assumes the lake is in compliance.

The assumed null hypothesis is an important practical consideration for Platte Lake in terms of perceived water quality, attainability, and cost. Platte Lake is currently an oligotrophic lake with no obvious signs of serious degradation. Phosphorus is not carcinogenic, toxic, or hazardous to human health, but rather it is undesirable because it stimulates algal growth. Excessive algal production may eventually reduce water clarity and dissolved oxygen concentrations and cause taste and odor problems for drinking water supplies. All the obvious and relatively easily correctable point sources of phosphorus in the entire Platte River watershed have already been reduced to near zero. The cost of reducing the phosphorus load from the remaining nonpoint sources increases dramatically as the removal requirement increases. In the case of Platte Lake, it could be argued that the burden of proof should be on regulators to demonstrate that significant impairment exists that justifies expenditure of public funds or the placement of restrictions on private and commercial activities (see also Belsky 1984, Zander 2010). Thus, for Platte Lake we recommend that regulators view the system as satisfying the standard to evaluate compliance.

Platte Lake can be contrasted with Onondaga Lake (New York) where Effler and O'Donnell (2010) report that summer surface water TP concentrations dropped from 360 to 42 mg/m^3 between 1987 and 2009 following the reduction of phosphorus loading from a large municipal wastewater treatment plant. In cases such as these, where water quality conditions are severely degraded, it seems appropriate for the null hypothesis to assume that water quality standards are in violation. These types of eutrophic lakes with high concentrations of phosphorus (>50 mg/m^3) may have point sources of phosphorus where current technology could be used to readily reduce the loads. In these cases, the decision statistics should be conservative and the burden of proof should be on regulators and safety officials to ensure compliance. Similar logic would be appropriate when the numerical standards are applied to toxic substances or carcinogenic compounds. The exact wording of the null hypothesis for any water quality numerical standard depends on a complex matrix of biological, statistical, physical, and social considerations (many of which may be qualitative in nature). Thus, the decision regarding the null hypothesis must be established on a case by case basis and requires astute analysis, foresight, and prudent judgment.

Acknowledgments

The authors would like to thank Gary Whelan and Ed Eisch from the Michigan Department of Natural Resources and Wil Swiecki and Mike Pattison from the Platte Lake Improvement Association for their continuing support of the monitoring and laboratory program and their thoughtful comments and review of the manuscript. Dr. Scott McNaught and Jenny Estabrook from Central Michigan University measured the Lake phosphorus concentrations prior to 2012 and conducted extensive quality assurance efforts to improve reliability. Nikki Sherretz from the Platte River State Fish Hatchery measured the phosphorus concentrations after 2012. Aaron Switzer and Paul Stowe from the Platte River State Fish Hatchery collected the Lake samples. The Water Science and Technology Directorate of Environment Canada provided the TP proficiency data in Fig. 12. Laura Shumway, ORISE Research Participant at the United States Environmental Protection Agency, Office of Water, Office of Wetlands, Oceans, and Watersheds, provided the lake phosphorus data used to develop the relationship between lake annual average concentration and standard deviation in Fig. 11 and the data on laboratory quantification levels in Table 3.

References

Barnett V, O'Hagan A. 1997. Setting environmental standards: the statistical approach to handling uncertainty and variation. London (UK): Chapman and Hall.

Belsky MH. 1984. Environmental policy in the 1980s: shifting back the burden of proof. Ecol Law Quarterly 12:1–89.

Bross I. 1985. Why proof of safety is much more difficult than proof of hazard. Biometrics. 41:785–793.

Canale RP, Harrison R, Moskus P, Naperala T, Swiecki W, Whelan G. 2004. Case study: reduction of total phosphorus loads to Big Platte Lake, MI, through point source control and watershed management. Proceedings of the Water Environment Federation, Watershed, No. 4. p. 1060–1076.

Canale R, Redder T, Swiecki W, Whelan, G. 2010. Phosphorus budget and remediation plan for Big Platte Lake, Michigan. J Water Resour Plan Manage. 136:576–586.

Carpenter SR, Brock WA. 2006. Rising variance: a leading indicator of ecological transition. Ecol Lett. 9:311–318.

DeOliveira F, Simser J, Lam C, Agemian H. 2013. Environment Canada Proficiency Testing Program. Final report for rain and soft waters, major ions and nutrients in natural waters, trace elements in water, total phosphorus in water, turbidity in water, total mercury in water. EC PT Study 0103 – March 2013. Environment Canada.

Drolc A, Ros M. 2002. Evaluation of measurement uncertainty in the determination of total phosphorus using standardized spectrometric method ISO 6878. Acta Chim Slov. 49:409–423.

Eaton AD, Clesceri LS, Rice EW, Greenberg AE, Franson MH. 2005. Standard methods for the examination of water and wastewater: edition 21. Washington (DC): American Public Health Association.

Effler SW, O'Donnell SM. 2010. A long-term record of epilimnetic phosphorus patterns in recovering Onondaga Lake, New York. Fund Appl Limnol. 1771:1–18.

Esterby S. 2013. Standards, environmental. In: Encyclopedia of environmetrics, 2nd ed. El-Shaarawi AH, Piegorsch WW, editors. Chichester (UK): John Wiley and Sons. p. 2629–2637.

Holland B, Ordoukhani N. 1990. Balancing Type I and Type II error probabilities: further comments on proof of safety vs proof of hazard. Comm Stat Theory Meth. 19(10):3557–3570.

Jones JR, Bachmann RW. 1976. Prediction of phosphorus and chlorophyll levels in lakes. J Water Poll Cont Fed. 48(9):2176–2182.

Knowlton MF, Jones JJ. 2006. Natural variability in lakes and reservoirs should be recognized in setting nutrient criteria. Lake Reserv Manage. 22:161–166.

Marcuse S. 1949. Optimal allocation and variance components in nested sampling with application to chemical analysis. Biometrics. 20:189–206.

McBride GB, Ellis JC. 2001. Confidence of compliance: a Bayesian approach for percentile standards. Water Res. 35:1117–1124.

Millard SP. 1987. Proof of safety vs proof of hazard. Biometrics. 43:719–725.

National Water Quality Monitoring Council. 2006. Water quality data elements: a user guide. Technical Report No. 3. 55 p.

SAS Institute Inc. 2008. SAS/STAT 9.2 user's guide. Cary (NC): SAS Institute Inc.

Shabman L, Smith E. 2003. Implications of applying statistically based procedures for water quality assessment. J. Water Resour Plan Manage. 129(4):330–336.

Smeltzer E, Dunlap F, Simoneau M. 2009. Lake Champlain phosphorus concentrations and loading rates, 1990–2008. Lake Champlain Basin Program, Grand Isle (VT). Technical Report No. 57. 41 p.

Smith EP, Ye K, Hughes C, Shabman L. 2001. Statistical assessment of violations of water quality standards under section 303(d) of the Clean Water Act. Environ Science Tech. 35:606–612.

Sokal RR, Rohlf FJ. 1969. Biometry. The principles and practice of statistics in biological research, 3rd ed. New York (NY): WH Freeman. 887 p.

Zander J. 2010. The application of the precautionary principle in practice: comparative dimensions. London (UK): Cambridge University Press.

Zhang C. 2007. Fundamentals of environmental sampling and analysis. New York (NY): John Wiley & Sons.