Journal of Water Resources Planning and Management

VOLUME 129 / NUMBER 4 JULY / AUGUST 2003

SPECIAL ISSUE: TMDL Approach to Water Quality Management
SPECIAL EDITORS: K. H. Reckhow and Richard M. Vogel

Editorial
245 On the Need for Uncertainty Assessment in TMDL Modeling and Implementation
K. H. Reckhow

Forum
247 Engineering Water Quality Models and TMDLs
Steven C. Chapra
257 Systems Analysis, TMDLs and Watershed Approach
Douglas A. Haith

Technical Papers
261 Interplay of Science and Stakeholder Values in Neuse River Total Maximum Daily Load Process
Lynn A. Maguire
271 Integrated Approach to Total Maximum Daily Load Development for Neuse River Estuary using Bayesian Probability Network Model (Neu-BERN)
Mark E. Borsuk, Craig A. Stow, and Kenneth H. Reckhow
283 A CE-QUAL-W2 Model of Neuse Estuary for Total Maximum Daily Load Development
James D. Bowen and Jeffrey W. Hieronymus
Tim A. Wool, Steven R. Davie, and Hugo N. Rodriguez
307 Comparison of Estuarine Water Quality Models for Total Maximum Daily Load Development in Neuse River Estuary
Craig A. Stow, Chris Roessler, Mark E. Borsuk, James D. Bowen, and Kenneth H. Reckhow
315 Support of Total Maximum Daily Load Programs Using Spatially Referenced Regression Models
Gerard McMahon, Richard B. Alexander, and Song Qian
330 Implications of Applying Statistically Based Procedures for Water Quality Assessment
Leonard Shabman and Eric Smith
337  Consideration of Variability and Uncertainty in Phosphorus Total Maximum Daily Loads for Lakes  
William W. Walker Jr.

345  Coping with Uncertainty: A Case Study in Sediment Transport and Nutrient Load Analysis  
Olufemi O. Osiele, Wei Zeng, and M. B. Beck
Consideration of Variability and Uncertainty in Phosphorus Total Maximum Daily Loads for Lakes

William W. Walker Jr.

Abstract: A simplified framework for considering variability and uncertainty in developing lake phosphorus total maximum daily loads (TMDLs) is demonstrated. Explicit consideration of these factors can increase the probability that TMDL implementation will meet a defined water quality goal at an acceptable frequency. Although a lake goal is typically expressed as a seasonal or yearly average phosphorus concentration, effects of temporal variations can be captured by correlating average phosphorus concentrations with the frequency of algal blooms (defined by extreme values of chlorophyll a) or the frequency of exceeding numeric water quality standards that are directly linked to algal blooms, such as hydrogen ion concentration or transparency. The margin of safety (MOS) required to achieve the lake goal at a defined frequency and with a defined confidence level can be estimated by including stochastic terms in the phosphorus balance equation to reflect variability and uncertainty. Given limitations in the data and models typically used in developing TMDLs, the cost of the MOS, expressed in terms of percent of the total allocated load or safety factors in the design of control measures, can be large. The MOS would be expected to increase with the percent load reduction required under the TMDL, as the forecast loads become increasingly dependent on assumptions regarding the performance of best management practices or other measures for reducing loads. The magnitude and cost of the MOS can be reduced by implementing TMDLs in an iterative fashion with ongoing data collection, and model refinement to reduce uncertainties associated with forecasting the performance of phosphorus load controls and lake responses.


CE Database subject headings: Phosphorus; Water quality; Lakes.

Introduction

Federal guidelines (U.S. EPA 1999) require consideration of variability and uncertainty in the development of total maximum daily loads (TMDLs) to meet water quality standards in impaired water bodies. Considering these factors is necessary to ensure that TMDL implementation will meet objectives with reasonably high probabilities of success and in a reasonably cost-effective manner. The requirements can be met using a variety of implicit or explicit approaches. Implicit approaches embed a margin of safety (MOS) into one or more conservative assumptions in support analysis (e.g., estimation of ungauged loads or model coefficients). If the MOS is not quantified, there is some risk that the resulting load control programs would be over-designed (resulting in unnecessary regulation and expense) or underdesigned (having a low probability of meeting objectives). If the MOS is explicitly quantified, control measures will generally be overdesigned sufficiently to achieve the specified goal with a specified confidence level. This paper demonstrates an approach that explicitly quantifies the MOS in the context of a lake phosphorus TMDL analysis. While the approach is demonstrated with a relatively simple lake phosphorus loading model, it can be generalized to other types of water bodies, water quality components, and model formulations with the substitution of appropriate equations and parameter values (Borsuk et al. 2002).

The MOS can be partitioned into components that reflect variability and uncertainty. Distinctions between variability and uncertainty are often ignored and/or misunderstood. Variability refers to temporal and/or spatial variations in water quality conditions, as they relate to the management goal. The amount of variability determines the frequency at which a given numeric water quality standard will be achieved under a given loading regime. Variability is typically an inherent characteristic of the system that is insensitive to management measures. “Uncertainty” refers to random prediction error resulting from limitations in the data and models used to formulate the lake phosphorus balance, trophic response model, and/or the performance of measures to achieve the allocated loads. The level of uncertainty determines the probability of achieving the standard at a specified frequency under a given load allocation. Unlike variability, uncertainty can be reduced in many cases by collecting additional data and improving forecast models under an adaptive management framework.

Guidelines set forth by regulatory agencies may require that TMDLs be designed to achieve compliance with the relevant water quality standard at a specified frequency (e.g., <10% of values exceeding the standard). In these situations, it is important to be clear about the averaging time scale and about whether the intent is to account for variability in the system or to provide an implicit margin of safety. In the former case, the maximum excursion frequency is essentially part of the goal and a conservative design would be needed to provide a margin of safety that accounts for uncertainty (e.g., design for 5 versus 10% excursion frequency, depending upon the level of uncertainty in the TMDL derivation). In the latter case, a conservative design may not be
required, since there is already an unspecified margin of safety implicit in the target. This would imply that a final outcome with an excursion frequency higher than 10% would be “acceptable.” This approach is not recommended because the actual goal is not clearly specified. The methodology described below demonstrates how these distinctions between variability and uncertainty can be made explicitly.

**Phosphorus Goal**

The basis for establishing a phosphorus TMDL is the causal pathway linking phosphorus loading, excessive algal growth, and impairment of water uses. In some cases, excessive algal growth may lead to violations of numeric water quality standards for chlorophyll a, hydrogen ion concentration (pH), dissolved oxygen, transparency, and/or free ammonia. The U.S. EPA (2000) provides guidance for selecting an appropriate P criterion on a regional basis. If sufficient data are available, regional or lake-specific criteria can be developed based on correlations between lake phosphorus concentrations and various measures of use impairment (such as aesthetic appearance) or violations of numeric water quality standards (Heiskary and Walker 1988; Havens and Walker 2002). Depending upon lake dynamics, Lake P criteria are typically averaged over an appropriate season (annual, spring overturn, or summer) and depth interval (epilimnetic, volume-weighted-mean). Spring-overturn or summer-average epilimnetic concentrations are typically used because they are most directly correlated with algal blooms and can be predicted using relatively simple empirical phosphorus loading models of the type demonstrated below.

One consideration is that P criteria expressed as summer or yearly averages do not directly address requirements to consider seasonal variations and critical conditions under TMDL guidelines (U.S. EPA 1999). Eutrophication-related impairment of water uses and violations of water quality standards are typically episodic in nature because of the episodic nature of algal blooms. One approach to addressing this issue in deriving a phosphorus goal is to develop lake-specific or regional correlations between average phosphorus concentrations and the frequency of algal blooms or violations of water quality standards, as demonstrated in Figs. 1 and 2. These types of correlations reflect temporal and, in some cases, spatial variability in the biological response to a given average phosphorus regime. Typically, they also exhibit a threshold response pattern that provides a logical focal point for selecting a P criterion.

Correlations between average summer phosphorus concentrations and the frequency of algal blooms linked to taste-and-odor episodes were used as a partial basis for setting a phosphorus goal of 25 ppb for Vadnais Lake, Minn. (Fig. 1, Walker et al. 1989; Walker 2000a). Correlations between the frequency of spatially distributed samples with a pH exceeding 9.0 (applicable water quality standard) and lake-average P concentrations were used as a partial basis for developing a TMDL for Upper Klamath Lake, Oregon (Fig. 2, Walker 2001a). The marked pH response to algal growth in this system reflects an extremely low-buffering capacity. Since the pH excursion frequencies are computed from individual samples collected on different dates within each year and at different locations and depths, the correlation captures both spatial and temporal variability in the system.

**Total Maximum Daily Load Equation—Side One**

A variety of modeling approaches may be taken to represent the relationship between external phosphorus loads and in-lake concentrations. In the example demonstrated below, the TMDL is derived from a steady-state mass balance that equates the long-term-average external P load to the lake assimilative capacity, or the maximum external load that is consistent with meeting the defined lake concentration goal. The assimilative capacity is equal to the sum of the flushing and net retention terms of the lake phosphorus budget when Lake P concentration equals the defined target:

\[
TMDL = QP* + UAP*
\]

where TMDL=total maximum daily load (kg/year); \(Q=\)long-term-average lake outflow (hm³/year); \(P*=\)Lake P target (mg/m³); \(U=\)effective settling velocity (m/year); and \(A=\)lake surface area (m²).

Time scales longer than one day are typically relevant in formulating lake phosphorus balances. As formulated here, the TMDL refers to long-term-average (multiyear) load. Phosphorus TMDLs for Lake Okeechobee, Fla. (FDEP 2001) and Upper Klamath Lake, Oregon (ODEQE 2002) are both expressed as long-term-average loads. The “maximum” descriptor refers to the maximum long-term-average load that is consistent with meeting the lake goal, not to the extreme value of a time series. The “daily” descriptor in TMDL is included for consistency with terminology in federal guidelines (U.S. EPA 1999). While alternative time scales and interpretations the “TMDL” term might be invoked, the formulation used here is most readily applicable for use with lake phosphorus loading models. The assumption is made that, regardless of how “TMDL” is interpreted, alternative definitions of TMDL would be acceptable, as long as the analysis produces a load allocation that is consistent with achieving water quality standards.

A variety of empirical models can be used to represent the net retention term (Vollenweider 1976; NALMS 1990; Walker 1999). In this example, retention is assumed to be proportional to concentration and lake surface area. The lake is assumed to be well-mixed horizontally, so that the outflow concentration and average lake concentrations are approximately equal. Preliminary estimates of settling velocity can be derived from the literature or regional lake data (Vollenweider 1969; Chapra 1975) and subsequently calibrated to lake-specific data. The selected averaging scheme for the Lake P criterion may influence the calibrated settling velocity and average outflow rate used in the mass balance.

If sufficient data and modeling resources are available, more complex deterministic models can be used to simulate internal lake dynamics on finer spatial and temporal scales. It is difficult to generally prescribe a modeling approach, since that choice ultimately depends upon the judgment of the analyst in view of the information and problem at hand. The credibility of the analysis may be enhanced if the model is able to explain temporal (Walker and Havens 2003) and/or spatial variations (Walker 1999) in a deterministic sense, even if the goal is expressed as a long-term or seasonal average. Deterministic simulation of year-to-year variations may be desirable in situations where there is a long-term trend in the data (Walker 2000b, 2001b). Similarly, yearly simulations may also provide a basis for model testing (e.g., comparing observed data with predictions in years that are not used for calibration purposes). Extension of the simple one-layer model to include sediment and/or hypolimnetic compartments may be appropriate in situations where estimates of internal recycling and/or response times are of interest (Walker 2000b, 2001a). Caution should be exercised, however, in extending model complexity beyond the level that is supported by the site-specific data available for calibration and testing.
The other side of the TMDL equation typically partitions the long-term-average external load as follows (U.S. EPA 1999):

\[
\text{TMDL} = \Sigma \text{LAs} + \Sigma \text{WLAs} + \text{Background} + \text{MOS}
\]

(2)

where \(\Sigma \text{LAs}\) = sum of load allocations (nonpoint sources, above background) (kg/year); \(\Sigma \text{WLAs}\) = sum of waste load allocations (point sources) (kg/year); Background = background load (=natural sources) (kg/year); and MOS = margin of safety (kg/year).

Separate consideration of “internal load” is usually inappropriate because releases from bottom sediments represent recycling of phosphorus that originally entered from the watershed and would therefore be reflected in the calibrated net retention term of Eq. (1). Explicit modeling of a sediment compartment (i.e., a two-box model) may be appropriate if internal recycling is an important factor.

The LAs are assumed to reflect nonpoint sources in excess of the background load, i.e., the anthropogenic portion of the total nonpoint load. For example, a 100 kg/year total \(P\) load from an urban watershed would be reflected in two terms of the equation (say, 10 kg/year in the background term representing the expected load with an undeveloped watershed and 90 kg/year in the LA term representing the increase in load above background resulting from development of the watershed).

The background load would equal the total load to the lake if the entire watershed were undeveloped. Consideration of the background load as a separate term in the equation is useful for characterizing anthropogenic impacts and settling realistic goals. For example, it is generally not practical to implement a TMDL that is below the background load. Reaching this conclusion may indicate that the assumed lake goal is unrealistic.

The load allocation terms (LAs and WLAs) represent expected average loads that would occur under the TMDL. Discharge permit limits (typically expressed as daily or monthly maximum values) would be set to be consistent with discharging average loads represented by the WLAs while taking typical variability in effluent quantity and quality into account. In order to operate in compliance with its discharge permit, the average load from a given facility would generally be below the permit level. If discharge permit levels were equated directly to the WLAs (without considering effluent variability), it would be appropriate to consider the difference between the expected average and permitted maximum loads as part of the MOS.

---

**Fig. 1.** Bloom frequency versus total \(P\)—Vadnais Chain of Lakes, Minn. Plots—observed and predicted percent of samples with chlorophyll \(a\) values exceeding each of three bloom criteria; symbols—observed yearly values from two lakes; solid lines—predicted frequencies using listed equations; dashed lines—25 mg/m\(^3\) goal adopted for Vadnais Lake to control taste and odor problems (Walker 2000a).
Marginal of Safety

Management measures would be designed so that the expected long-term-average external load (sum of allocated point and nonpoint loads) would equal the lake assimilative capacity (TMDL) less the margin of safety (MOS). In order to draw the distinction between variability and uncertainty, it is useful to divide the MOS into two components

\[ \text{MOS} = \text{MOV} + \text{MOU} \]  

(3)

where MOV = margin of variability (kg/year); and MOU = margin of uncertainty (kg/year).

For a given TMDL, as determined by the lake assimilative capacity, increasing the MOV will reduce the expected load to the lake (sum of allocated point and nonpoint loads) and increase the “compliance rate,” or frequency of meeting a given numeric goal. Increasing the MOU will increase the “confidence level,” or probability of meeting the goal at the desired frequency. The compliance rate and confidence level assumed in formulating the TMDL may be a lake-specific policy decision and/or determined by regulations. As demonstrated below, these factors are at least as important as the selection of a numeric goal \((P^*)\) in determining the load allocation that is consistent with a given TMDL.

The MOS, MOV, and MOU consistent with a compliance rate \((\beta)\) and confidence level \((\alpha)\) can be estimated by attaching stochastic terms to the TMDL mass balance. Random year-to-year variations in Lake P concentrations are drawn from a lognormal distribution with a coefficient of variation \(S_V\). If sufficient time series data are available, an alternative approach would be to model year-to-year variations deterministically. Uncertainty in the predicted Lake P concentration under a given loading regime is modeled by attaching another lognormal deviate to the lake phosphorus balance with a coefficient of variation \(S_U\).

As estimated from Eq. (1), the TMDL represents the long-term-average load consistent with a compliance rate of 50% and confidence level of 50%. To meet the specified lake target \((P^*)\) at the specified compliance rate \((\beta)\) with a confidence level \((\alpha)\), the allocated long-term-average load \((L_A)\) would have to be reduced as follows:

\[ L_A = (Q + KA)P^*F_VF_U = \text{TMDL} F_VF_U \]  

(4)

\[ F_V = \exp(-Z_{\beta}S_V) \]  

(5)

\[ F_U = \exp(-Z_{\alpha}S_U) \]  

(6)

where \(L_A\) = allocated long-term-average load = TMDL - MOS = TMDL\(F_VF_U\); \(F_V\) = factor accounting year-to-year variability in Lake P concentration; \(S_V\) = year-to-year coefficient of variation (CV) of Lake P concentration; \(F_U\) = factor accounting for uncertainty in the predicted average Lake P concentration; \(S_U\) = model error CV for predicted average Lake P concentration; \(Z_{\beta}\) = standard normal variate with upper tail probability \(\beta\); \(Z_{\alpha}\) = standard normal variate with upper tail probability \(\alpha\); \(\beta = \) assumed compliance rate = fraction of years with Lake \(P < P^*\); and \(\alpha = \) assumed confidence level = probability Lake \(P < P^*\) at specified \(\beta\).

\(S_V\) estimates derived from variance component analyses of large lake and reservoir datasets typically range from 0.1 to 0.2 (Knowlton et al. 1984; Smeltzer et al. 1989). A lake-specific estimate can be derived if long-term monitoring data are available. By combining the above equations, the MOS, MOU, and MOV can be explicitly quantified as follows:

\[ L_A = \text{TMDL} - \text{MOS} = \text{TMDL} F_VF_U \]  

(7)

\[ \text{MOS} = \text{TMDL}(1 - F_VF_U) \]  

(8)

\[ \text{MOU} = \text{MOS}(1 - F_U)/(2 - F_U - F_V) \]  

(9)

\[ \text{MOV} = \text{MOS} - \text{MOU} \]  

(10)

Uncertainty in Lake P Forecast

The uncertainty in forecasting the long-term-average Lake P concentration resulting from a given load allocation is represented in the parameter \(S_U\). This uncertainty reflects potential errors in forecasting the performance of load control measures implemented to achieve the required load allocation, as well as poten-
The magnitude of $S_L$ would be expected to increase with the percent load reduction required under the TMDL, as the forecast loads become increasingly dependent on assumptions regarding BMP performance. For illustration purposes, it is useful to consider a simplified case in which the entire load reduction will be accomplished in a sequence of detention ponds (approximating plug flow) with second-order phosphorus removal kinetics (Walker 1987). The outflow load ($L_o$) is predicted from the inflow load ($L_i$) by the following equation:

$$L_o = L_i (1-R) = L_i (1 + KP_i T)$$  \hspace{1cm} (12)

where $R$=load reduction, as fraction of inflow load; $P_i$=inflow concentration (mg/m$^3$); $K$=second-order phosphorus removal rate (1/day/mg/m$^3$); and $T$=hydraulic residence time (days).

For a given inflow load (in this case, reflecting existing conditions), variance in the outflow load is related to uncertainty in the assumed removal rate ($K$). From a first-order error analysis, the relative standard error of the predicted outflow load is

$$S_L = RS_K$$  \hspace{1cm} (13)

where $S_K$=error CV of $P$ removal rate. Approximate parameter estimates ($K=0.0003$ l/day/mg/m$^3$ and $S_K=0.4$) are based upon performance data from 24 runoff detention ponds (Walker 1987).

Fig. 3 shows the CV of forecasted Lake $P$ concentration as a function of percent load reduction for two sets of error coefficients.
Fig. 4. TMDL sensitivity to compliance rate and confidence level. Baseline=existing conditions; A-D=load allocations for alternative assumptions; other model parameters as listed in Table 1.

The detention pond model is used here as an example. We would expect qualitatively similar patterns for other control measures (e.g., onsite BMPs), as well. For example, the potential error in forecasting small incremental reductions resulting from targeting obvious sources in the watershed (“hot spots”) would tend to be lower than the potential error in forecasting reductions resulting from implementation of BMPs on a watershed scale after obvious sources have already been controlled.

Example

As formulated above, the pond detention time \( T \) represents the decision variable in the TMDL. The objective is to find the value of \( T \) that satisfies the TMDL objective (i.e., meets Lake \( P \) target with the specified compliance rate and confidence level). Since \( T \) is proportional to volume for a given flow, it can be considered a rough surrogate for size and cost. The equations are nonlinear and a numerical solution can be derived in a spreadsheet with circular references, as detailed in Table 1.

Fig. 4 shows load allocation sensitivity to the assumed compliance rate and confidence level for the sample case defined in Table 1. The error CVs are set to represent a situation where the TMDL is being developed with relatively limited site-specific data for calibrating the lake model and evaluating BMP performance. The TMDL is independent of these assumptions because it is based exclusively on the lake assimilative capacity [Eq. (1)]. The allocated load accounts for a smaller portion of the TMDL when uncertainty and variability are considered. In the extreme case \( D \) when both factors are considered, the MOS accounts for 41% of the TMDL. The required load reduction is 70%, as compared with 50% when uncertainty and variability are ignored (Case A). Corresponding pond detention times are 80 and 33 days, respectively. The cost of overdesigning the pond by 46 days (or 139%) to provide the MOS might be considerable. The detention times in this simplified example exceed typical designs (7–30 days, Walker 1987) because the entire lake inflow is being treated in a single pond. This reduces the inflow concentration and load reduction at a given detention time below those expected in typical control programs with ponds located in critical source areas with higher inflow concentrations, as opposed to treating the entire inflow.

Phased Approach

Fig. 5 demonstrates that implementing the TMDL in an iterative fashion with incremental load reductions and ongoing data collec-
Fig. 5. Phased approach to TMDL implementation. Phase 1—initial allocation ignoring margin of safety; alternative—allocation including margin of safety based upon limited data; Phase 2—final allocation considering margin of safety after implementation of Phase 1 controls and after model refinements based upon additional lake monitoring and BMP evaluation.

<table>
<thead>
<tr>
<th>Confidence Level</th>
<th>50%</th>
<th>90%</th>
<th>50%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliance Rate</td>
<td>50%</td>
<td>80%</td>
<td>50%</td>
<td>80%</td>
</tr>
<tr>
<td>Lake Model Error CV</td>
<td>20%</td>
<td>20%</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>Pond Removal Rate CV</td>
<td>40%</td>
<td>40%</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>Allocated Load (mg/m²·yr)</td>
<td>500</td>
<td>232</td>
<td>500</td>
<td>340</td>
</tr>
<tr>
<td>Incremental Load Reduction</td>
<td>50%</td>
<td>70%</td>
<td>0%</td>
<td>20%</td>
</tr>
<tr>
<td>Cum. Load Reduction</td>
<td>50%</td>
<td>77%</td>
<td>50%</td>
<td>66%</td>
</tr>
<tr>
<td>Margin of Safety (mg/m²·yr)</td>
<td>0</td>
<td>205</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>Margin of Safety (%)</td>
<td>0%</td>
<td>41%</td>
<td>0%</td>
<td>20%</td>
</tr>
<tr>
<td>Pond Resid Time (days)</td>
<td>33</td>
<td>80</td>
<td>33</td>
<td>50</td>
</tr>
<tr>
<td>Pond Overdesign (days)</td>
<td>0</td>
<td>46</td>
<td>0</td>
<td>17</td>
</tr>
</tbody>
</table>

Conclusions

A simplified model for formal consideration of variability and uncertainty in developing a lake phosphorus TMDL is demonstrated above. Using empirical frequency response models (e.g., Figs. 1 and 2) to set lake goals provides a means of considering spatial and temporal variability while using a relatively simple, steady-state mass balance model.

There are many site-specific factors to consider in developing a TMDL (e.g., urgency of water quality problem, cost, land availability, public opinion, limitations in control technology, etc.). Large safety factors associated with the MOS may not be unusual in the context of other public works projects (e.g., bridges, buildings, etc.). On the other hand, overdesign may not be possible or practical in many situations.

Caution is advised in setting an unrealistically high-confidence level and/or compliance rates as TMDL goals. When evaluated relative to water use impairment, public health, or risk to aquatic life, there may already be a substantial margin of safety in the water quality standard or criterion that drives the TMDL. Requiring high margins of safety may hinder the progress of lake restoration by increasing costs, reducing credibility, and stimulating controversy.

An incremental or “adaptive” approach to achieving the desired compliance rate and confidence level through successive TMDLs may be appropriate, as recommended in a recent study by the National Research Council (2001). Accomplishing incremental load reductions while acquiring and analyzing new data over time can increase the probability of meeting the lake objective with each iteration of the process, as illustrated by the simple example presented above. As demonstrated in the long-term effort
to achieve water quality standards in the Everglades (Walker 1995; SFWMD 2002), a phased approach is applicable in situations where the load allocation is not immediately achievable (with or without an MOS) because of limitations in control technology.

References


