Ecology, 87(6), 2006, pp. 1472–1477

A BAYESIAN APPROACH TO RETRANSFORMATION BIAS IN TRANSFORMED REGRESSION

Craig A. Stow,[1,3] Kenneth H. Reckhow,[2] and Song S. Qian[2]

[1]*Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina 29208 USA*
[2]*Nicholas School of the Environment and Earth Sciences, Duke University, Durham, North Carolina 27712 USA*

Abstract. Ecological data analysis often involves fitting linear or nonlinear equations to data after transforming either the response variable, the right side of the equation, or both, so that the standard suite of regression assumptions are more closely met. However, inference is usually done in the natural metric and it is well known that retransforming back to the original metric provides a biased estimator for the mean of the response variable. For the normal linear model, fit under a log-transformation, correction factors are available to reduce this bias, but these factors may not be generally applicable to all model forms or other transformations. We demonstrate that this problem is handled in a straightforward manner using a Bayesian approach, which is general for linear and nonlinear models and other transformations and model error structures. The Bayesian framework provides a predictive distribution for the response variable so that inference can be made at the mean, or over the entire distribution to incorporate the predictive uncertainty.

Key words: allometric relationship; Bayesian analysis; lognormal model; log-transformed regression; Poisson regression; retransformation.

Introduction

It is common in simple linear regression analysis to log-transform the response and sometimes the predictor variable to estimate model parameter values. Log-transformation can accomplish several things:

1) It can linearize the relationship between the response and predictor variable.

2) It can stabilize the model error variance so that the assumption of a constant variance is more closely met.

3) It can remove serial correlation among the residuals.

4) It can make the assumption of conditional normality more realistic for response variables that are bounded, such as concentration data which cannot be less than zero.

Typically, however, prediction and interpretation in the natural metric are of primary interest and the simple linear model that has been estimated under a log-transformation as

$$\log Y = \hat{\alpha} + \hat{\beta} \log X \tag{1}$$

is often retransformed (exponentiated) as

$$Y = aX^{\hat{\beta}} \tag{2}$$

where log Y is the response variable, log X is the predictor variable, $\hat{\alpha}$ and $\hat{\beta}$ are the intercept and slope parameter estimates [where a "hat" ($\hat{}$) symbol denotes estimate], respectively, and $a = \exp(\hat{\alpha})$. Previous authors have pointed out that Eq. 2 is a biased estimator for the mean of Y (Sprugel 1983, Koch and Smillie 1986, Newman 1993) with a bias that is generally downward, though prediction at X values that are well outside of the calibration data set range can be positively biased (Cohn et al. 1989). This bias results, largely, from ignoring the model error term, ε, when exponentiating Eq. 1. This becomes clearer when Eq. 1 is written with the error term explicitly included as

$$\log Y = \hat{\alpha} + \hat{\beta} \log X + \varepsilon. \tag{3}$$

Then exponentiation results in

$$Y = aX^{\hat{\beta}} e^{\varepsilon}. \tag{4}$$

The problem occurs because, under standard regression assumptions, ε represents a normal distribution with mean = 0 and variance = σ^2, and the mean of $e^{\varepsilon} \neq 1$, as might be expected. When ε is exponentiated, the entire distribution, not just the mean, is exponentiated. Exponentiation of a normal distribution with mean = μ and variance = σ^2 results in a lognormal distribution with mean = $\exp(\mu + 0.5\sigma^2)$. Consequently, in Eq. 4, e^{ε} represents a multiplicative error term that is lognormally distributed with a mean = $\exp(0 + 0.5\sigma^2) = \exp(0.5\sigma^2)$. Because the mean is an average of all possible values weighted by their relative probabilities, it is not invariant to nonlinear transformations, such as exponentiation. Thus, the mean of a lognormal distribution is not obtained by exponentiating the mean of the underlying normal distribution (Crow and Shimizu 1988).

Manuscript received 23 September 2005; revised 19 December 2005; accepted 21 December 2005. Corresponding Editor: A. M. Ellison.
[3] E-mail: cstow@sc.edu

A common approach to reduce the bias in estimating Y is to multiply Eq. 1 by

$$\exp(0.5s^2) \tag{5}$$

where s^2 is a sample-based estimate for σ^2. This method reduces the bias that results from ignoring the model error term, but is generally recognized as having a slight positive bias (Smith 1993). Cohn et al. (1989) refer to the estimator represented by the correction factor in Eq. 5 as a "quasi maximum likelihood estimator" (QMLE) and also present a "minimum-variance-unbiased-estimator" (MVUE), based on a derivation by Bradu and Mundlak (1970). However, the MVUE is relatively complex to program and, because the bias of the QMLE is generally small, the MVUE has not been widely employed. Recently, the United States Geological Survey has incorporated the MVUE into a software package specifically for the problem of estimating pollutant loads in rivers and streams (Runkel et al. 2004).

For most linear regression problems either the QMLE or the MVUE provide an adequate solution to the problem of retransformation bias. However, these estimators have not been shown to be applicable for the analogous problem in a nonlinear context, or for models with non-normal error structures such as generalized linear models (McCullagh and Nelder 1989). With nonlinear models, it is still often useful to log-transform either the response variable, the right side of the equation, or both to stabilize the model error variance and/or accommodate response variables that are bounded at zero. In nonlinear models, the parameter and predictive distributions are likely to have non-standard forms, making the properties of generally applicable correction factors difficult to derive in a classical statistical framework. Additionally, depending on the model form, it is possible that the degree of bias may differ throughout the parameter and sample space. Thus, a more general framework is useful to accommodate this problem. We describe the Bayesian approach to the problem of retransformation bias. The Bayesian framework accommodates this problem quite naturally because inference and prediction are based on the posterior parameter distribution. Retransformation back to the natural metric is done by exponentiating the entire posterior distribution, resolving the bias problem in a straightforward manner. The Bayesian approach is applicable with log-transformed linear and nonlinear models, models fit using other variable transformations (Miller 1984), and models with alternative error structures.

Bayesian inference begins with Bayes' theorem:

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_\theta \pi(\theta)f(y|\theta)\,d\theta} \tag{6}$$

where $\pi(\theta\,|\,y)$ is the posterior probability of θ (the probability of the parameter vector, θ, after observing the new data, y), $\pi(\theta)$ is the prior probability of θ, (the probability of θ before observing y), and $f(y\,|\,\theta)$ is the likelihood function, which incorporates the statistical relationships as well as the mechanistic or process relationships among the predictor and response variables. Predictions for unobserved or future ys (denoted $\tilde{y}$) are assessed over the entire posterior parameter distribution as

$$\pi(\tilde{y}|y,\theta) = \int_\theta f(\tilde{y}|\theta)\pi(\theta|y)\;\; d\theta \tag{7}$$

referred to as predictive distribution.

For the log-transformed, simple linear model (Eq. 3), under the assumption that ε is distributed normally, the likelihood function is

$$f(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(\log Y - \alpha - \beta \log X)^2}{-2\sigma^2}\right] \tag{8}$$

where n is the number of observations, and y denotes observations of the response variable, Y, and the predictor variable, X. More generally, the likelihood function, for a normally distributed error term, can be expressed as

$$f(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{[\log Y - g(x,\beta)]^2}{-2\sigma^2}\right\} \tag{9}$$

where $g(x,\beta)$ is the functional form of the model. In some applications, such as our nonlinear example, $g(x,\beta)$ may also be log transformed.

Examples

We demonstrate the Bayesian approach using three examples: a simple linear regression model relating log total phosphorus concentration (kg/m^3) to log(river flow) (m^3/d), a nonlinear lake phosphorus model that predicts in-lake phosphorus concentration (mg/L) from average influent phosphorus concentration (mg/L) and water residence time (yr), and a Poisson regression (generalized linear model) relating species number to area.. We programmed these examples into WinBUGS, a free, downloadable software, for Bayesian analysis (Gilks et al. 1994). WinBUGS uses Markov chain Monte Carlo methods (McMC) to generate samples from the posterior parameter distribution. For normal, linear models numerical approximation is actually unnecessary; analytical solutions are available using appropriate prior distributions. However, WinBUGS is straightforward to program (see Supplement), allows the use of prior distributions for the normal linear model that do not result in closed form solutions (if desired), and it can accommodate model error structures in addition to the standard normal form.

Simple linear example

Our simple linear regression model is based on 26 measurements of total phosphorus concentration and concurrent daily average flow measurements collected

by the North Carolina Department of Environment and Natural Resources, and the U.S. Geological Survey, respectively, from the Neuse River in North Carolina, USA. We fit Eq. 3 with log(P concentration) and log(flow) as the response and predictor variables, respectively, using SAS 9.1 (SAS 2002) and WinBUGS. For the Bayesian estimation we used non-informative prior distributions for α, β, and σ^2.

Nonlinear example

Our example nonlinear model is a simple "input–output" or Vollenweider model (Vollenweider 1969). Various forms of this model have been fit to regional lake data bases (Canfield and Bachman 1981, Reckhow 1988) and incorporated into Eutromod (Reckhow et al. 1992, Hession et al. 1996), a spreadsheet-based lake eutrophication model. We consider the following form:

$$\log P = \log \frac{P_{in}}{1 + k\tau_w} + \varepsilon \tag{10}$$

where P is the annual average in-lake phosphorus concentration, P_{in} is the average influent concentration, τ_w is water residence time, k is an empirical constant, and ε is the normally distributed model error with mean $= 0$ and variance $= \sigma^2$. We fit this model to an example data set of 29 Florida lakes (Stow and Reckhow 1996), using nonlinear least squares (nls) as implemented in SAS software using Proc NLIN and WinBUGS. The WinBUGS estimation used non-informative priors for the model parameters and 20 000 iterations (after a 1000 iteration "burn in").

Poisson (generalized linear model) example

Our linear and nonlinear examples are both based on the assumption of a normal, additive, error structure. To illustrate an alternative error structure we show results from a Poisson regression using number of number of butterfly species as the response variable and forest patch size as the predictor variable, data of Lovejoy et al. (1984) as presented by Ramsey and Schafer (2002). Poisson regression can be applicable when the response variable represents counts that are regarded as a linear function of one or more predictor variables. This model has the form

$$N_s \sim \text{Poisson}(\text{mu}) \tag{11}$$

$$\log(\text{mu}) = \beta_0 + \beta_1 \log(\text{size}) \tag{12}$$

where N_s = species number and size = forest fragment size. We fit this model using Proc Genmod in SAS software, and in WinBUGS with non-informative priors for β_0 and β_1.

RESULTS

Simple linear example

The resultant simple linear model relating log concentration to log(flow) (Fig. 1) has the following form:

$$\log(\text{TP}) = -0.60 - 0.56(\text{flow}) + \varepsilon \quad \varepsilon \sim \mathcal{N}(0, 0.286). \tag{13}$$

The corresponding retransformation correction factor is $\exp(0.286/2) = 1.15$. Before retransformation the 95% predictive interval is evenly spaced about the regression line (Fig. 1). After retransformation the upper bound of the 95% predictive interval is further from the retransformed regression line than the lower bound, reflecting the asymmetry of the exponentiated error term (Fig. 1). The retransformed regression line is less than the bias-corrected mean, while the Bayesian mean is slightly greater than the bias-corrected mean. The Bayesian mean is slightly greater than the bias-corrected mean because the Bayesian predictive distribution incorporates both the model error variance and the posterior parameter variance.

Nonlinear example

The resultant parameter values are similar using each method. Using nls, the optimum value for k was 2.38 with a 95% confidence interval of 1.50–3.26. With WinBUGS, k had a mean of 2.38, a median of 2.36, and a 95% credible region of 1.59–3.31. The model error variance (mean squared error) obtained with nls was 0.269, while WinBUGS indicated a mean of 0.290, a median of 0.276, and a 95% credible region of 0.170–0.495 for the error variance parameter. Using the correction factor applicable for log-transformed linear models, $\exp(0.5s^2)$, results in a value of 1.14, based on nls mean squared error estimate.

To examine various regions of sample space we predicted P at the nine combinations of P_{in} and τ_w resulting from using the minimum ($P_{\text{in}} = 0.035$ mg/L, $\tau_w = 0.027$ yr), median ($P_{\text{in}} = 0.297$ mg/L, $\tau_w = 0.419$ yr), and maximum ($P_{\text{in}} = 1.923$ mg/L, $\tau_w = 3.4$ yr) observed values of each. Mean posterior P predictions are consistently higher than results obtained by exponentiating nls results (Fig. 2). The correction factor, obtained by taking the ratio of the P posterior mean: nls estimate differs within the sample space, always exceeding 1.14 mg/L, the value obtained using the linear model correction factor. This last result is not necessarily general for all nonlinear models, but occurs because the variance of the Bayesian predictive distribution incorporates both the model error variance and the posterior parameter variance, and the posterior parameter variance may differ within the sample space.

Poisson example

Parameter estimates were essentially identical using classical and Bayesian approaches; the classical approach yielded estimates of 3.68 for β_0 and 0.18 for β_1 with standard errors of 0.05 and 0.14, respectively, and similarly the Bayesian posterior means were 3.68 for β_0 and 0.18 for β_1 with respective standard deviations of 0.05 and 0.14. A comparison of predicted values in the natural metric indicates that the Bayesian predictive

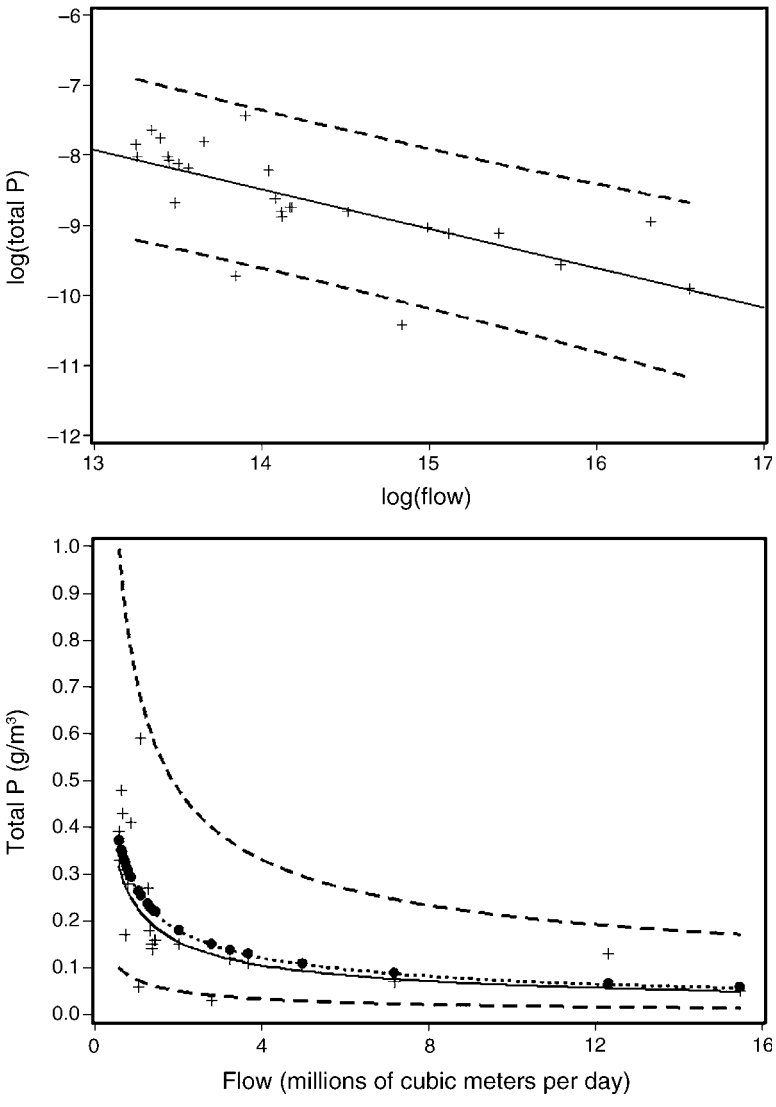

FIG. 1. Simple linear model relating log total phosphorus concentration in lakes to log(flow) before (top) and after (bottom) retransformation. In both panels, observations are denoted as "+" and dashed lines represent 95% prediction intervals. The solid line in the top panel represents the mean before retransformation; in the bottom panel, the solid line represents the model retransformed without bias correction (Eq. 2). The dotted line in the bottom panel represents the bias-corrected estimate (Eq. 5). Circles in the bottom panel represent means of Bayesian predictive distribution (Eq. 7) for each observation.

mean slightly exceeds the retransformed value throughout the observed sample space (Fig. 3).

CONCLUSION

Our main point in this brief report was to illustrate that the Bayesian approach provides a general framework for handling the problem of retransformation bias which occurs when models are fit under transformation then retransformed back to the original metric. This bias occurs because the mean of a distribution is not invariant to many nonlinear transformations. We emphasize that the Bayesian approach should not be regarded as an attempt to correct for bias. The mean of the Bayesian predictive distribution will incorporate both model error and posterior parameter variance, thus it is not exactly analogous to classical bias-corrected mean, which is based only on the model error variance, though the results will often be similar (Fig. 1). Bias is a long-term relative frequency concept applicable to classical estimators, but is not generally regarded as a relevant Bayesian concept because Bayesian inference is conditioned on the observed data (Barnett 1982). We also point out that a biased estimator can give results that are less than, greater than, or equal to the "true" but unknown mean in any given sample; bias is a long-term property of an estimator, it is not applicable to a particular estimate. The real utility of the Bayesian approach is that it provides an entire predictive

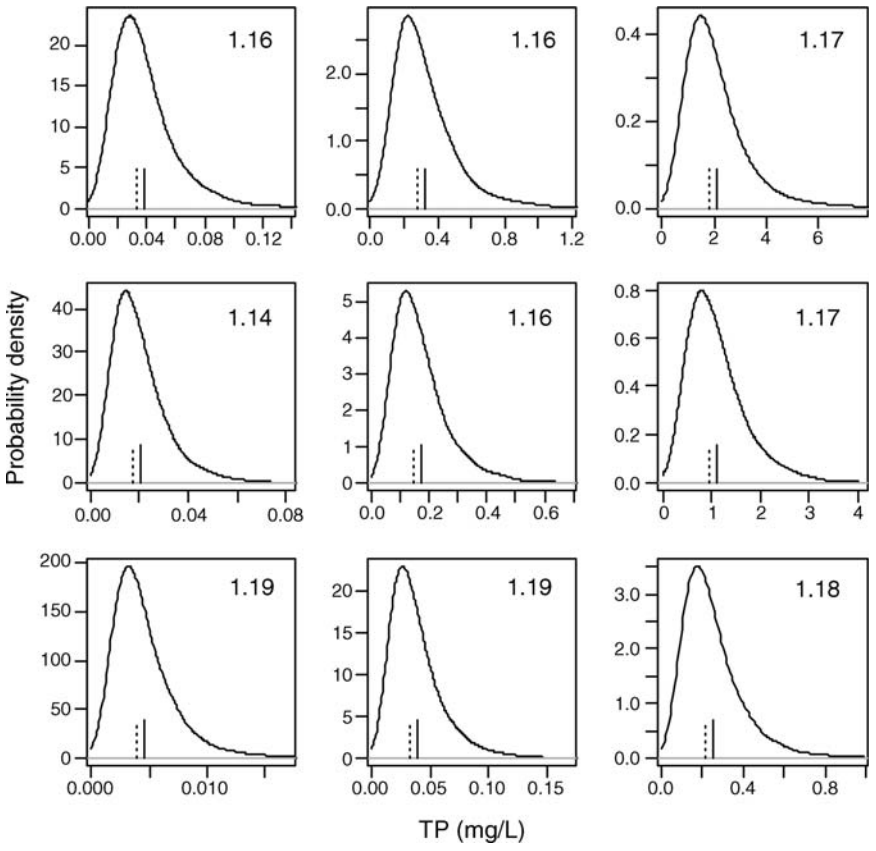

FIG. 2. Predictive distributions for total phosphorus (TP) using representative combinations of τ_w (water residence time) and P_{in} (average influent concentration of P). Top row $\tau_w = 0.027$, middle row $\tau_w = 0.419$, bottom row $\tau_w = 3.4$; first column $P_{in} = 0.035$, middle column $P_{in} = 0.297$, and third column $P_{in} = 1.923$. The predictive mean is denoted by a solid vertical line; the exponentiated value is denoted by dotted vertical line. The ratio of predictive mean : exponentiated value is indicated by the number in the upper right quadrant.

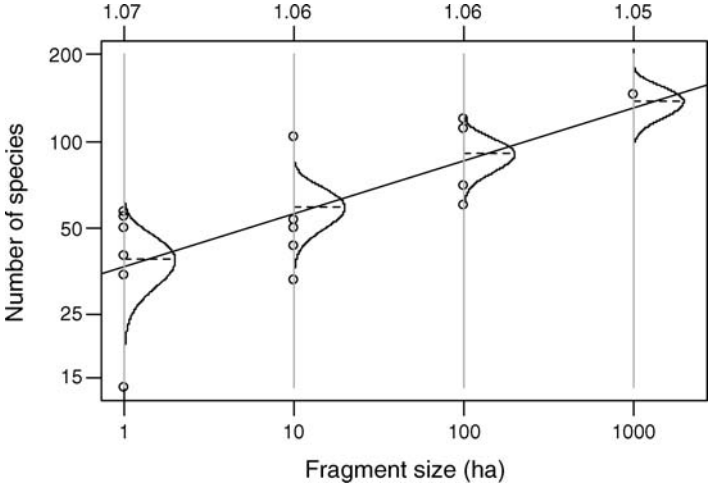

FIG. 3. Results of Poisson regression relating species number to forest fragment size. The straight line depicts retransformed model predictions; Bayesian predictive distributions are shown at fragment size values of 1, 10, 100, and 1000 ha. The solid vertical line through each distribution depicts the predictive distribution mean. Circles depict individual observations. Values at the top of the graph are the ratio of Bayesian predictive mean : retransformed value.

distribution over which inference can be made, instead of using only a point estimate such as the mean. The predictive distribution incorporates an estimate of prediction uncertainty which can be used for risk evaluation in decision-making. While the implementation of Bayesian approaches was historically limited by analytically intractable results, fast computers and cleverly written algorithms now make precise numerical approximations straightforward to generate.

Acknowledgments

This work was partially supported by EPA STAR Grant #R830883.

Literature Cited

Barnett, V. 1982. Comparative statistical inference. Second edition. John Wiley and Sons, New York, New York, USA.

Bradu, D., and Y. Mundlak. 1970. Estimation in lognormal linear models. Journal of the American Statistical Association **65**:198–211.

Canfield, D. E., and R. W. Bachmann. 1981. Prediction of total phosphorus concentrations, chlorophyll *a*, and secchi depths in natural and artificial lakes. Canadian Journal of Fisheries and Aquatic Sciences **38**:414–423.

Cohn, T. A., L. L. DeLong, E. J. Gilroy, R. M. Hirsch, and D. K. Wells. 1989. Estimating constituent loads. Water Resources Research **25**:937–942.

Crow, E. L., and K. Shimizu. 1988. Lognormal distributions, theory and applications. Marcel Dekker, New York, New York, USA.

Gilks, W. R., A. Thomas, and D. J. Spiegelhalter. 1994. A language and program for complex Bayesian modelling. Statistician **43**:169–177.

Hession, W. C., D. E. Storm, C. T. Haan, S. L. Burks, and M. D. Matlock. 1996. A watershed-level ecological risk assessment methodology. Water Resources Bulletin **32**:1039–1054.

Koch, R. W., and G. M. Smillie. 1986. Bias in hydrologic prediction using log-transformed regression models. Water Resources Bulletin **22**:717–723.

Lovejoy, T. E., J. M. Rankin, R. O. Bierregaard, K. S. Brown, L. H. Emmons, and M. E. Van der Woot. 1984. Ecosystem decay of Amazon Forest remnants. Pages 295–325 *in* M. H. Nitecki, editor. Extinctions. University of Chicago Press, Chicago, Illinois, USA.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Second edition. CRC Press, Boca Raton, Florida, USA.

Miller, D. M. 1984. Reducing transformation bias in curve fitting. American Statistician, **38**:124–126.

Newman, M. C. 1993. Regression analysis of log-transformed data: statistical bias and its correction. Environmental Toxicology and Chemistry **12**:1129–1133.

Ramsey, F. L., and D. W. Schafer. 2002. The statistical sleuth. Second edition. Duxbury, Pacific Grove, California, USA.

Reckhow, K. H. 1988. Empirical models for trophic state in southeastern United States lakes and reservoirs. Water Resources Bulletin **24**:723–734.

Reckhow, K. H., S. Coffey, M. H. Henning, K. Smith, and R. Banting. 1992. Eutromod: technical guidance and spreadsheet models for nutrient loading and lake eutrophication. Draft Report, School of the Environment, Duke University, Durham, North Carolina, USA.

Runkel, R. L., C. G. Crawford, and T. A. Cohn. 2004. Load estimator (LOADEST): a FORTRAN program for estimating constituent loads in streams and rivers. U.S. Geological Survey Techniques and Methods Book 4, Chapter A5. U. G. Geological Survey, Reston, Virginia, USA.

SAS Institute. 2002. SAS version 9.1. SAS Institute, Cary, North Carolina, USA.

Smith, R. J. 1993. Logarithmic transformation bias in allometry. American Journal of Physical Anthropology **90**: 215–228.

Sprugel, D. G. 1983. Correcting for bias in log-transformed allometric equations. Ecology **64**:209–210.

Stow, C. A., and K. H. Reckhow. 1996. Estimator bias in a lake phosphorus model with observation error. Water Resources Research **32**:165–170.

Vollenweider, R. A. 1969. Possibilities and limits of elementary models concerning the budget of substances in lakes. Archiv für Hydrobiologie **66**:1–36.

SUPPLEMENT

WinBUGS code for the examples shown in the main article (*Ecological Archives* E087-086-S1).