A COUPLED TRANSFER FUNCTION MODEL AND COMPOSITE SAMPLING STRATEGY FOR EFFICIENT MASS LOAD ESTIMATION

David R. Fox
Australian Centre for Environmetrics
University of Melbourne, Australia
david.fox@unimelb.edu.au

Abstract

Increasingly, natural resource management and environmental regulatory agencies are couching water quality objectives in terms of total contaminant load over some defined period of time. This in turn has refocussed attention on the related issues of sampling and estimation methods for mass load determination. While much good work has already been reported on these topics, it is evident that generic advice on sampling design and statistical estimation methodology is unlikely to be forthcoming. In view of this, water quality monitoring programs often lack statistical rigor with sampling designs (understandably) driven largely by considerations of cost, logistics, and expediency. Compounding the ad hoc nature of many sampling programs is the plethora of computational approaches for estimating a total mass load. When applied to relatively sparse concentration data, these different computational approaches can yield wildly different load estimates. A critical missing element in the discussion of load estimation procedures to date is the coupling of sample design and statistical estimation procedure. While considerable flexibility exists in the choice of these monitoring components, they are neither totally independent nor arbitrary considerations. In this paper we show how parameter estimation for a second-order transfer function model of daily concentration demands a composite sampling approach for data collection and analysis. In this way, the sampling design and the load estimation procedure are coupled thus avoiding the ambiguity of multiple load estimates when different estimating equations are applied to the same data. The technique is demonstrated with the estimation of a total phosphorus load in an irrigation drain in the Gippsland region of Victoria, Australia. The results of this analysis suggest that conventional estimates of load based on monthly concentration

data underestimate the true load by about 30%. Using the method described in this paper reduced this error to 0.5%.

1. INTRODUCTION

The problem of estimating mass loads (of sediments and/or nutrients) is not new (Cooper and Watts 2002, Preston et al. 1989, Richards 1998, Cohn et al 1989, Thomas 1985, Degens and Donohue 2002, Moosmann et al 2005, Chu and Sanders 2003 and others). Indeed, considerable attention has been paid to the dual problems of (i) identification of an appropriate sampling strategy (Richards 1998, Littlewood 1995, Thomas 1985, 1986, Thomas and Lewis 1993, 1995); and (ii) choice of an estimating equation for total load (Cooper and Watts 2002, Clarke 1990, Aulenbach and Hooper 2005, Preston et al. 1989, Letcher et al 1999) Although these two objectives are inter-related, they are often treated as arbitrary and/or independent. Much of the attention given to the sampling issue has focussed on the *frequency* of sampling (eg. sub-daily, daily, monthly, episodic etc.) while numerous papers have appeared that have compared the (statistical) performance of various load estimation techniques (eg. mean-based estimation, regression estimators, ratio estimators, and others) (Clarke 1990, Vogel et al. 2005, Cooper and Watts 2002). While these have been useful contributions, the difficulty it seems, is that there is no universally 'optimal' approach for mass load sampling and estimation – different circumstances will dictate different approaches. The problem is further compounded by the paucity of general recommendations that enunciate the linkages between circumstances and approaches, thus leaving the practitioner with a bewildering array of sampling strategies and estimation techniques. It has been our experience that most mass load sampling programs are *ad hoc* with data collection considerations heavily influenced by non-statistical issues such as personal preference, institutional norms, and expediency.

The key inputs for any mass load estimation are *concentration* (C_t) and *flow/discharge* (Q_t) at time *t*. The instantaneous *flux rate* (F_t) is the product of concentration and discharge (equation 1).

$$F_t = C_t \cdot Q_t \tag{1}$$

The total load or mass transported in the interval [0,T] is obtained by integrating the instantaneous flux rate:

$$Load = \int_0^T C_t \cdot Q_t \, dt \tag{2}$$

In practice, equation (2) is approximated by the summation

$$L = K \sum_{i=1}^{N} C_i \cdot Q_i \tag{3}$$

where C_i and Q_i are measurements of concentration and flow respectively and *K* is a constant.

Equation (3) is the simplest estimator and accords with intuition as it is essentially the discrete analog of equation (2). However, numerous other estimating equations have been proposed (see for example Letcher et al. 1999). The simplest sampling strategy is *systematic sampling* whereby a water sample is obtained once every *k* time periods. The main advantage of this approach is a logistical one since it is easy to implement and lends itself to automation. The disadvantage is that the intensity of sampling bears no relation to the hydrology of the water body being sampled. It is well known that for most

catchments, significant amounts of material are transported during 'peak' flow events (eg. storms). In recognition of this phenomenon, more sophisticated sampling strategies have been devised which aim to capture these high-flow/high-load events. These strategies are either *deterministic* whereby sampling occurs whenever the flow (stage height) exceeds some threshold, or *probabilistic* in which case a statistical algorithm is used to bias the sampling events towards high-flow events (Thomas 1985). In either case, a common problem is that resource constraints are such that typically only about 12 - 30 water quality samples (ie. C_i values) can be obtained. Autonomous samplers can provide flow data at finely resolved time steps. In contrast, concentration data are obtained relatively infrequently due to time, cost, and logistical considerations. Thus, it is invariably the case that flows and concentrations are *not* measured contemporaneously. The analyst is then faced with the issue of estimating, for example, an annual load using daily flow information and monthly concentrations. Common strategies involved linearly interpolating the monthly concentrations and resampling to a daily time base, or alternatively, to assume the 'spot' monthly concentration reading is indicative of the average for the month and to apply this to the total discharge over the month. Neither of these approaches is satisfactory as large errors arise (Ferguson 1986, Littlewood 1995).

In this paper, we address the data paucity and estimation issues simultaneously by using a transfer function model (Littlewood 1995, Lemke 1991) whose parameters are estimated from *monthly* water quality data. Having fitted the model and estimated other key parameters such as the variance of the random error or shock component, simulated *daily* time-series for the water quality parameter of interest can be constructed. The simulated concentrations are then matched with *actual* flows and a straightforward application of

equation (3) provides an estimate of load for the period of interest. A critical modification to the (assumed) monthly monitoring for water quality is required. Instead of taking a *single* sample once a month for analysis, the procedure outlined here requires that an *average* concentration be obtained from a *composite* monthly sample. For example, if flows are recorded on a daily basis, then a daily water sample must be obtained and stored. At the end of the month, *equal volumes* from each of the 30, say, water samples are combined. A single water quality determination is then performed on this composite sample. The rationale for the composite sampling is that it provides an estimate of the *average daily concentration* for that month. As will be seen in subsequent sections of this paper, this is a critical requirement for the development of ensuing modeling and estimation methodologies. It is important to note that the proposed methodology will be invalidated if this sampling strategy is not employed.

2. A TRANSFER MODEL FOR (DAILY) CONCENTRATION

Many modelling approaches have been used in an attempt to reconstruct (daily) time-series of nutrient concentrations (Michalak and Kitanidis 2005). These range from simple regression models where (log) concentration is assumed to be a linear function of (log) flow to more complex ARIMA models which attempt to capture some of the autocorrelation structure usually evident in both flow and concentration data. In this paper, we extend the transfer function approach adopted by Littlewood (1995). Littlewood assumed the concentration at time *i* to be a function of the flow at time *i* and the concentration at time $i-1$ (equation 4).

$$C_i = \frac{b_0}{\left(1 + B \cdot a_1\right)} \cdot Q_i \tag{4}$$

In equation 4, b_0 and a_1 are model parameters and B is the backward shift operator ie. $B \cdot X_k = X_{k-1}$.

For the watersheds we have investigated, there is evidence to suggest that the (*log*) concentration at time *i* is strongly related to the (*log*) flow at time *i* and the (*log*) concentration in the immediately two preceding time periods. In the remainder of this document, unless otherwise specified, C_i and Q_i will denote the natural logarithms of concentration and flow respectively. Our basic model is thus:

$$C_i = \frac{\alpha_0 + \alpha_1 Q_i}{\left(1 + B\beta_1 + B^2 \beta_2\right)} + \varepsilon_i \tag{5}$$

where the αs and βs are model parameters; $B^2 \cdot X_k = X_{k-2}$; and ε_i is a zero-mean random error or 'shock' component with variance σ_ε^2.

3. BASIC RESULTS

In this section we derive a number of fundamental results for the second-order transfer model that will underpin the subsequent sampling and estimation strategy. We commence with a generic, second-order transfer model (equation 6).

$$Y_j = \frac{\alpha_0 + \alpha_1 X_j}{\left(1 + B\beta_1 + B^2 \beta_2\right)} + \varepsilon_j \tag{6}$$

In deriving first and second-order moments, it is assumed *Y* and *X* have finite means $(\mu_Y$ and $\mu_X)$ and variances $(\sigma_Y^2$ and $\sigma_X^2)$ and we further assume that $\varepsilon_j \sim N(0, \sigma_\varepsilon^2) \quad \forall\ j$.

3.1 Mean of Y

Taking expectations of both sides of equation 6 gives

$$E[Y_j] = E[\alpha_0 + \alpha_1 X_j + \beta_1 Y_{j-1} + \beta_2 Y_{j-2} + \varepsilon_j]$$

$$= \alpha_0 + \alpha_1 \mu_X + \beta_1 \mu_Y + \beta_2 \mu_Y$$

and therefore

$$\mu_Y = \frac{\alpha_0 + \alpha_1 \mu_X}{1 - \beta_1 - \beta_2} \tag{7}$$

3.2 Variance of Y_j

By definition,

$$\sigma_Y^2 = Var[Y_j] = E\left[(Y_j - \mu_Y)^2\right] = E[Y_j^2] - \mu_Y^2$$

After some algebraic manipulation (see Appendix) we obtain

$$\sigma_Y^2 = \frac{\left\{2\alpha_1^2(\beta_2^2 - \beta_2 - \beta_1^2)\phi_2 + 2\alpha_1^2\beta_1\phi_1 + \alpha_1^2(\beta_2 - 1)\phi_0 + (2\beta_1^2 - 2\beta_2^2 + 2\beta_1 + \beta_2 + 1)\alpha_1^2\mu_X^2 + (\beta_2 - 1)\sigma_\varepsilon^2\right\}}{\left[(1+\beta_2)(\beta_1 + \beta_2 - 1)(\beta_1 - \beta_2 + 1)\right]} \tag{8}$$

where $\phi_k = \rho_X^{(k)}\sigma_X^2 + \mu_X^2$ and $\rho_X^{(k)}$ is the k^{th} order autocorrelation of *X*.

3.3 Covariance of Y_j and Y_{j-1}

The first-order covariance between successive values of the dependent variable is

$$Cov\left[Y_j, Y_{j-1}\right] = E\left[\left(Y_j - \mu_Y\right)\left(Y_{j-1} - \mu_Y\right)\right]$$

$$= E\left[Y_j Y_{j-1}\right] - \mu_Y^2$$

In order to evaluate this last expectation, some intermediate results are required. These are presented below and derivations are provided in the appendix.

$$E\left[X_j, Y_{j-1}\right] = \alpha_0\mu_X + \alpha_1\phi_1 + \beta_1\left[\alpha_0\mu_X + \alpha_1\phi_2 + \mu_X\mu_Y\left(\beta_1 + \beta_2\right)\right] + \beta_2\mu_X\mu_Y \tag{9}$$

$$E\left[X_j, Y_{j-2}\right] = \alpha_0\mu_X + \alpha_1\phi_2 + \mu_X\mu_Y\left(\beta_1 + \beta_2\right) \tag{10}$$

$$\begin{aligned} E\left[X_j, Y_j\right] = {} & \alpha_0\mu_X\left(1 + \beta_1 + \beta_1^2 + \beta_2\right) + \alpha_1\phi_0 + \alpha_1\beta_1\phi_1 \\ & + \phi_2\left(\alpha_1\beta_1^2 + \alpha_1\beta_2\right) + \mu_X\mu_Y\left[\left(\beta_1 + \beta_2\right)\left(\beta_1^2 + \beta_2\right) + \beta_1\beta_2\right] \end{aligned} \tag{11}$$

$$E\left[Y_j, Y_{j-1}\right] = \frac{\alpha_0\mu_Y + \alpha_1 E\left[X_j, Y_{j-1}\right] + \beta_1\left(\sigma_Y^2 + \mu_Y^2\right)}{\left(1 - \beta_2\right)} \tag{12}$$

$$E\left[Y_j, Y_{j-2}\right] = \alpha_0\mu_Y + \alpha_1 E\left[X_j, Y_{j-2}\right] + \beta_1 E\left[Y_j, Y_{j-1}\right] + \beta_2\left(\sigma_Y^2 + \mu_Y^2\right) \tag{13}$$

Therefore

$$Cov\left[Y_j, Y_{j-1}\right] =$$

$$\frac{1}{\left[\left(\beta_1 + \beta_2 - 1\right)\left(\beta_1 + \beta_1\beta_2 - \beta_2^2 + 1\right)\right]}\left\{\begin{aligned} & \alpha_1^2\left[\beta_1^3 + \beta_1^2 + \left(2\beta_2 - \beta_2^2 + 2\right)\beta_1 - \beta_2^2 + 1\right]\mu_X^2 \\ & + \alpha_1^2\left(\beta_1\beta_2^2 - 2\beta_1\beta_2 - \beta_1^3 - \beta_1\right)\phi_2 + \alpha_1^2\left(\beta_2^2 - \beta_1^2 - 1\right) + \alpha_1^2\beta_1\phi_0 - \beta_1\sigma_\varepsilon^2 \end{aligned}\right\} \tag{14}$$

4. MODEL ESTIMATION

As previously stated, a fundamental requirement of the proposed monitoring strategy is the use of composite sampling to obtain monthly 'average' concentration data. Thus, we wish to use the monthly means $\overline{Y}_k \quad k = 1, \ldots, m$ to estimate the parameters of the transfer model (equation 6). This is done by non-linear least-squares. Hence the parameter vector

$\Theta = \begin{bmatrix} \alpha_0 & \alpha_1 & \beta_0 & \beta_1 \end{bmatrix}^T$ is estimated such that $\sum_{i=1}^{m} \left(\overline{Y}_i - \hat{\overline{Y}}_i \right)^2$ is minimized, where $\hat{\overline{Y}}_i$ is the i^{th} estimated monthly mean concentration, i.e.

$$\hat{\overline{Y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{\hat{\alpha}_0 + \hat{\alpha}_1 X_j}{\left(1 + B\hat{\beta}_1 + B^2 \hat{\beta}_2\right)} \qquad j = 1, \ldots, n_i; \;\; i = 1, \ldots m. \tag{15}$$

In order to use the fitted model, the *actual* data values, Y_1 and Y_2 need to be specified. The remainder of the sequence $Y_3, \ldots, Y_N$ (where the total number of data values $N = \sum_{i=1}^{m} n_i$) is then obtained recursively using equation 6. In essence this fitted sequence models the *mean response* and <u>not</u> individual daily (log) concentrations. Large excursions above and below the mean response will arise when the variance of the random 'shock' component in equation 6 is relatively large. In practice, this is likely to be the case. The size of this discrepancy will also be magnified during the process of transforming back to the original scale (ie. exponentiation of the Ys).

One way of investigating the variance in daily concentration data is to simulate daily concentration series with the random shock component included. The advantage of this approach is that the modeled data will, to a reasonable degree, exhibit the 'correct' autocorrelation and cross-correlation (with *log*-flows) structure. However, before the simulation method can be implemented, it will be necessary to obtain an estimate of σ_ε^2. A method for estimating σ_ε^2 is developed in the next section.

4.1 Estimating the variance of the random shock component

We use a method-of-moments approach whereby the sample variance between the monthly sample means is equated to its theoretical expectation. Thus, in order to implement this approach, we require an expression for the variance of the monthly sample means.

Variance of $\overline{Y}$

Concentration data will be in the form of monthly means $\left\{\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_m\right\}$. Let n be the number of days in a month and $n' = n - 2$. From equation 6 we have:

$$n'\overline{Y} = \sum_{i=3}^{n}\left\{\alpha_0 + \alpha_1 X_i + \beta_1 Y_{i-1} + \beta_2 Y_{i-2} + \varepsilon_i\right\}$$

$$= n'\alpha_0 + \alpha_1\sum_{i=3}^{n} X_i + \beta_1\sum_{i=2}^{n-1} Y_i + \beta_2\sum_{i=1}^{n-2} Y_i + \sum_{i=3}^{n}\varepsilon_i$$

$$= n'\alpha_0 + n'\alpha_1\overline{X} + \beta_1\left(n'\overline{Y} + Y_2 - Y_n\right) + \beta_2\left(n'\overline{Y} + Y_1 + Y_2 - Y_{n-1} - Y_n\right) + n'\overline{\varepsilon}$$

and hence

$$\left[\overline{Y} - \frac{\alpha_0}{\left(1-\beta_1-\beta_2\right)}\right] = \frac{\alpha_1\overline{X} + \frac{\beta_1}{n'}\left(Y_2 - Y_n\right) + \frac{\beta_2}{n'}\left(Y_1 + Y_2 - Y_{n-1} - Y_n\right) + \overline{\varepsilon}}{\left(1-\beta_1-\beta_2\right)} \qquad (16)$$

The left side of equation 16 can be written as the linear combination $\underline{C}^T\underline{Z}$ where

$\underline{Z}^T = \left[\bar{X} \quad Y_1 \quad Y_2 \quad Y_{n-1} \quad Y_n \quad \bar{\varepsilon} \right]$ and

$$\underline{C}^T = \frac{1}{\left(1-\beta_1-\beta_2\right)}\left[\alpha_1 \quad \frac{\beta_2}{n'} \quad \frac{\left(\beta_1+\beta_2\right)}{n'} \quad \frac{-\beta_2}{n'} \quad \frac{-\left(\beta_1+\beta_2\right)}{n'} \quad 1 \right]$$

and the variance of $\bar{Y}$ is given as $Var\left[\bar{Y}\right] = \underline{C}^T \Sigma \, \underline{C}$.

For $n = 30$ it will be reasonable to assume that autocorrelations beyond a lag of about 15 are essentially zero. This means that covariances between $\left\{Y_1, Y_n\right\}$, $\left\{Y_1, Y_{n-1}\right\}$, $\left\{Y_2, Y_n\right\}$, and $\left\{Y_2, Y_{n-1}\right\}$ can be set equal to zero. Let $\Sigma = Cov\left[\underline{Z}\right]$ have elements

$$\Sigma = \begin{bmatrix} \sigma_{\bar{X}}^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_Y^2 & Cov\left[Y_j, Y_{j-1}\right] & 0 & 0 & 0 \\ 0 & Cov\left[Y_j, Y_{j-1}\right] & \sigma_Y^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_Y^2 & Cov\left[Y_j, Y_{j-1}\right] & 0 \\ 0 & 0 & 0 & Cov\left[Y_j, Y_{j-1}\right] & \sigma_Y^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{\bar{\varepsilon}}^2 \end{bmatrix}$$

then it can be shown that the variance *between* monthly sample means is given by equation 17.

$$Var\left[\bar{Y}\right] = \frac{1}{\left(1-\beta_1-\beta_2\right)^2}\left\{ \alpha_1^2 \frac{\sigma_X^2}{n} + 2\frac{\sigma_Y^2}{n^2}\left(\beta_1^2 + 2\beta_1\beta_2 + 2\beta_2^2\right) + \frac{4\left(\beta_1\beta_2+\beta_2^2\right)}{n^2} Cov\left[Y_j, Y_{j-1}\right] + \frac{\sigma_\varepsilon^2}{n} \right\} \tag{17}$$

where $Cov\left[Y_j, Y_{j-1}\right]$ is evaluated using equation 14. Denote the *sample* variance between the *m* monthly means by $S_{\bar{Y}}^2$. An estimate of the error variance may be then obtained by

equating $S_{\bar{Y}}^2$ with equation 17 and solving for σ_ε^2. An explicit formula for σ_ε^2 in terms of other model parameters is given in the Appendix.

5. STOCHASTIC SIMULATION OF DAILY CONCENTRATION SERIES

By constructing a large number of simulated daily concentration series, the variability and hence uncertainty in the estimated annual load can be examined. The advantage of this approach is that the simulated daily concentration is *conditional* on the observed discharge and the concentrations during the preceding two days. A random 'innovation' or 'shock' component is then added to the predicted concentration to mimic the inherent variability in the natural system. This approach not only permits the construction of empirical confidence intervals for the annual load, but also provides readily available estimates of other important statistics such as percentiles and maximum and minimum. The steps involved in constructing simulated series are summarized below.

1. Estimate the parameters of the transfer function model (equation 6) as outlined in section 4 above;

2. Compute the *sample* variance $\left(S_{\bar{Y}}^2\right)$ between the set of *m* observed monthly mean (*log*) concentrations;

3. Using the parameter estimates obtained at step 1 in equations 8 and 14 together with the estimated variance of the (*log*) flow data (call this $\hat{\sigma}_X^2$), compute the theoretical variance given by equation 17;

4. Find σ_ε^2 in equation 17 so that the result from step 3 equates to the result from step 2;

5. Use the estimated transfer model (equation 6) and simulate random shock components from $\varepsilon \sim N\left(0, \sigma_\varepsilon^2\right)$ for each of the $N = mn$ days in the series;

6. Estimate the annual load using equation 3 with $K = 1$;

7. Repeat step 5 N_{sim} times where N_{sim} is the number of independent simulations;

8. Examine statistics associated with N_{sim} load estimates.

This procedure is demonstrated in the following section with application to the estimation of an annual total phosphorus (TP) load.

6. EXAMPLE – ESTIMATION OF A TOTAL PHOSPHORUS LOAD

Southern Rural Water (SRW) is responsible for the management of water resources within the McAlister irrigation district (MID) in central Gippsland, Victoria, Australia. Of primary interest to SRW is the estimation of phosphorous loads from the MID to the adjoining Gippsland Lakes (Figure 1). SRW have been conducting regular monitoring of drains, streams, and rivers within the MID for a number of years. For the main drains, SRW undertakes daily monitoring of both flow and total phosphorous. Nutrient monitoring at this intensity is extremely rare and the resulting data sets unique in their temporal coverage. We have used SRW's daily monitoring data of Central Gippsland

drain #3 (CG3) for the period March 1, 1998 to 30 September, 2004 (a total of 2,406 daily values) to trial the methods presented in this paper.

A time-series plot of the daily discharge is shown in Figure 2. The measured daily TP value together with the monthly mean is shown in Figure 3. The scatter plot of TP and discharge (Figure 4) illustrates the difficulty in attempting to model TP using simple regression procedures (ie. rating curve methods). We next apply the methods presented in section 4 to estimate the transfer function model parameters. To do this we assume that the only data available to us is the daily discharge and monthly average TP concentration. A plot of the daily TP concentration and the monthly averages is shown in Figure 5. Since the complete daily flow and concentration record is available, we are able to obtain the 'true' parameter values for the transfer model of equation 6. These will be used to compare the parameter estimates based on the monthly average concentration data.

The true parameter values were found to be

$$\Theta = \begin{bmatrix} \alpha_0 & \alpha_1 & \beta_0 & \beta_1 \end{bmatrix}^T = \begin{bmatrix} -0.23102 & 0.07221 & 0.70902 & 0.12654 \end{bmatrix}^T \text{ and } \sigma_\varepsilon = 0.568413 .$$

Using the monthly mean TP data and the estimation procedure described in section 4, we obtain the *estimated* parameter values:

$$\hat{\Theta} = \begin{bmatrix} \hat{\alpha}_0 & \hat{\alpha}_1 & \hat{\beta}_0 & \hat{\beta}_1 \end{bmatrix}^T = \begin{bmatrix} -0.13392 & 0.039470 & 0.766722 & 0.124507 \end{bmatrix}^T .$$ A

comparison of the predicted monthly mean TP concentrations based on equation 6 with parameter values given by Θ and those obtained using $\hat{\Theta}$ is shown in Figure 6. It is evident from Figure 6 that the series obtained using parameters estimated using the procedure based on the monthly mean data is in very good agreement with the series

generated using the 'true' model parameters. However, as previously remarked, both of these plots represent a mean monthly response and as such are unlikely to capture the large, transient peaks in concentrations that inevitably occur. Following the procedure outlined in section 4.1, our estimate of the *between* month variance, $S_{\bar{Y}}^2$ is 0.569736 and the estimated error or 'shock' variance is $\hat{\sigma}_\varepsilon^2 = 0.1825$. A complete list of parameter values and the method of computation is given in Table 1.

Load estimation

Applying equation 3 (with $K = 1$) to the complete record of daily discharge and concentration data we obtain a 'true' load of 58,039 tonnes. When the monthly average concentrations are applied to the total monthly discharges and these quantities summed over the period, an estimate of 42,002 tonnes is obtained. This represents a significant under-estimation (28%) and is typical of the degree and direction of error encountered in practice. For example, Clement (2001) reported errors of between 15-80% in annual nutrient load estimates. Using the method outlined in section 5, 1,000 series of *daily* concentrations, each of length 2,406 days were simulated (Figure 7). Figure 7 shows output from the transfer function model together with the actual TP series. Additionally, approximate 95% confidence limits for the daily TP concentration have been provided. A total load was computed for each of the simulated series using equation 3 (with $K = 1$). The median load for these 1,000 estimates is 58,329 tonnes which is in remarkably good agreement with the true value (0.5% over-estimation). Additional insights into the load distribution are gained from an inspection of Figure 8. It can be seen from Figure 8 that,

although our method has dramatically reduced the estimation error and that the confidence intervals for either a mean or median load encompass the true load.

7. DISCUSSION

This paper brings together and extends two important aspects of mass load estimation: sampling strategy and estimation technique. While much attention has been paid to these issues separately, relatively little has been said about their inter-dependency. Our approach has been to extend the transfer function modelling approach of Littlewood (1995) to adequately characterize daily stream concentration data. Considerations of sampling cost and logistics led us to develop a composite sampling strategy that provides a surrogate measure of the average monthly concentration based on a single water quality determination. As in statistical ANOVA models, information contained in the observed variation *between* these monthly averages enables key parameters of the transfer model to be estimated. Finally, stochastic variation is randomly generated on a daily time-step and added to the predicted mean daily concentration to generate series of daily concentration data. A total mass load is computed for each simulated series by coupling the generated daily concentrations with the observed daily flows. The mean and variance of this collection of estimated loads can be computed from which point and interval estimates for the total load are derived. An example is provided whereby the error in the estimated total load is reduced from approximately 30% to less than 1%.

8. REFERENCES

Aulenbach, B.T. and Hooper, R.P. (2005) Improving stream solute load estimation by the composite method: A comparative analysis using data from the Panola mountain research watershed. *Proceedings of the 2005 Georgia water Resources Conference*, April 25-27, 2005, University of Georgia. Kathryn J. Hatcher (ed.).

Chu, A.K. and Sanders, B.F. (2003) Data requirements for load estimation in well-mixed tidal channels. *Journal of Environmental Engineering*, **129(8)**, 765-773.

Clarke, R.T. (1990) Statistical characteristics of some estimators of sediment and nutrient loadings. *Water Resources Research*, **26(9)**, 2229-2333.

Clement, A. (2001) Improving uncertain nutrient load estimates for Lake Balaton. *Water Science and Technology*, **43(7)**, 279-286.

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

Cooper, D.M. and Watts, C.D. (2002) A comparison of river load estimation techniques: application to dissolved organic carbon. *Environmetrics*, **13**, 733-750.

Degens, B.P. and Donohue, R.D. (2002) Sampling mass loads in rivers: A review of approaches for identifying, evaluating and minimising estimation errors. Waters and Rivers Commission, Water Resource Technical Series No. WRT 25.

Ferguson, R.I. (1986) River loads underestimated by rating curves. *Water Resources Research*, **22(1)**, 74-76.

Letcher, R.A., Jakeman, A.J., McKee, L.J., Merritt, W.S., Eyre, B.D., and Baginska, B. (1999) Review of techniques to estimate catchment exports. Technical Report 99/73, October, NSW EPA, Sydney, Australia.

Littlewood, I.G. (1995) Hydrological regimes, sampling strategies and assessment of errors in mass load estimates for United Kingdom rivers. *Environment International*, **21(2)**, 211-220.

Lemke, K.A. (1991) Transfer function models of suspended sediment concentration. *Water Resources Research*, **27(3)**, 293-305.

Preston, S.D., Bierman, V.J., and Silliman, S.E. (1989) An evaluation of methods for the estimation of tributary mass load. *Water Resources Research*, **25(6)**, 1379-1389.

Michalak, A.M. and Kitanidis, P.K. (2005) A method for the interpolation of nonnegative functions with an application to contaminant load estimation. *Stochastic Environmental Research Risk Assessment*, **19**, 8-23.

Moosmann, L., Müller, B., Gächter, R., Wüest, A., Butscher, E. and Herzog, P. (2005) Trend-oriented sampling strategy and estimation of soluble reactive phosphorus loads in streams. *Water Resources Research*, **41**, W01020, doi:10.1029/2004WR003539

Richards, R.P. (1998) Estimation of pollutant loads in rivers and streams: A guidance document for NPS programs. Water Quality laboratory, Heidelberg, Ohio.

Thomas, R.B. (1985) Estimating total suspended sediment yield with probability sampling. *Water Resources Research*, **21(9)**, 1381-1388.

Thomas, R.B. and Lewis, J. (1993) A comparison of selection at list time and time-stratified sampling for estimating suspended sediment loads. *Water Resources Research*, **29(4)**, 1247-1256.

Thomas, R.B., and Lewis, J. (1995) An evaluation of flow-stratified sampling for estimating suspended sediment loads. *Journal of Hydrology*, **170**, 27-45.

Vogel, R.M., Rudolph, B.E., and Hooper, R.P. (2005) Probabilistic behavior of water-quality loads. *Journal of Environmental Engineering*, **131(7)**, 1081-1089.

APPENDIX – Derivation of equations

We define $\phi_k = \rho_X^{(k)}\sigma_X^2 + \mu_X^2$ where $\rho_X^{(k)}$ is the autocorrelation at lag *k* for *X*.

Equation 10

$$\begin{aligned} E\left[X_j Y_{j-2}\right] &= E\left[X_j\left(\alpha_0 + \alpha_1 X_{j-2} + \beta_1 Y_{j-3} + \beta_2 Y_{j-4} + \varepsilon_{j-2}\right)\right] \\ &= E\left[\alpha_0 \mu_X\right] + \alpha_1 E\left[X_j X_{j-2}\right] + \beta_1 E\left[X_j Y_{j-3}\right] + \beta_2 E\left[X_j Y_{j-4}\right] \end{aligned}$$

As we have used a second-order transfer function model, it is assumed that correlations between flow and concentration beyond lags of two time periods are negligible. Thus the last two expectations in the equation above are set equal to the product of the individual means. Furthermore, $E\left[X_j X_{j-1}\right] = Cov\left[X_j, X_{j-1}\right] + \mu_X^2 = \rho_X^{(1)}\sigma_X^2 + \mu_X^2 = \phi_1$ and similarly $E\left[X_j X_{j-2}\right] = \phi_2$. Therefore

$$E\left[X_j Y_{j-2}\right] = \alpha_0 \mu_X + \alpha_1 \phi_2 + \mu_X \mu_Y \left(\beta_1 + \beta_2\right).$$

Equation 9

$$\begin{aligned} E\left[X_j Y_{j-1}\right] &= E\left[X_j\left(\alpha_0 + \alpha_1 X_{j-1} + \beta_1 Y_{j-2} + \beta_2 Y_{j-3} + \varepsilon_{j-1}\right)\right] \\ &= E\left[\alpha_0 \mu_X\right] + \alpha_1 E\left[X_j X_{j-1}\right] + \beta_1 E\left[X_j Y_{j-2}\right] + \beta_2 E\left[X_j Y_{j-3}\right] \\ &= \alpha_0 \mu_X + \alpha_1 \phi_1 + \beta_1 E\left[X_j Y_{j-2}\right] + \beta_2 \mu_X \mu \end{aligned}$$

and substituting equation 10 for $E\left[X_j Y_{j-2}\right]$ gives

$$E\left[X_j, Y_{j-1}\right] = \alpha_0 \mu_X + \alpha_1 \phi_1 + \beta_1\left[\alpha_0 \mu_X + \alpha_1 \phi_2 + \mu_X \mu_Y \left(\beta_1 + \beta_2\right)\right] + \beta_2 \mu_X \mu_Y .$$

Equation 11

$$E\left[X_j Y_j\right] = E\left[X_j\left(\alpha_0 + \alpha_1 X_j + \beta_1 Y_{j-1} + \beta_2 Y_{j-2} + \varepsilon_j\right)\right]$$
$$= \alpha_0 \mu_X + \alpha_1 E\left[X_j^2\right] + \beta_1 E\left[X_j Y_{j-1}\right] + \beta_2 E\left[X_j Y_{j-2}\right]$$

Noting that $E\left[X_j^2\right] = \phi_0$ and substituting $E\left[X_j Y_{j-1}\right]$ and $E\left[X_j Y_{j-2}\right]$ with their previously derived expressions gives (after some algebraic manipulation)

$$E\left[X_j, Y_j\right] = \alpha_0 \mu_X \left(1 + \beta_1 + \beta_1^2 + \beta_2\right) + \alpha_1 \phi_0 + \alpha_1 \beta_1 \phi_1$$
$$+ \phi_2\left(\alpha_1 \beta_1^2 + \alpha_1 \beta_2\right) + \mu_X \mu_Y \left[\left(\beta_1 + \beta_2\right)\left(\beta_1^2 + \beta_2\right) + \beta_1 \beta_2\right]$$

Equation 12

$$E\left[Y_j Y_{j-1}\right] = E\left[Y_{j-1}\left(\alpha_0 + \alpha_1 X_j + \beta_1 Y_{j-1} + \beta_2 Y_{j-2} + \varepsilon_j\right)\right]$$
$$= E\left[\alpha_0 \mu_X\right] + \alpha_1 E\left[X_j Y_{j-1}\right] + \beta_1 E\left[Y_{j-1}^2\right] + \beta_2 E\left[Y_{j-1} Y_{j-2}\right]$$
$$= \alpha_0 \mu_Y + \alpha_1 E\left[X_j Y_{j-1}\right] + \beta_1\left(\sigma_Y^2 + \mu_Y^2\right) + \beta_2 E\left[Y_{j-1} Y_{j-2}\right]$$

Noting that $E\left[Y_j Y_{j-1}\right] = E\left[Y_{j-1} Y_{j-2}\right]$ we have

$$\left(1 - \beta_2\right) E\left[Y_j Y_{j-1}\right] = \alpha_0 \mu_Y + \alpha_1 E\left[X_j Y_{j-1}\right] + \beta_1\left(\sigma_Y^2 + \mu_Y^2\right)$$

and on replacing $E\left[X_j Y_{j-1}\right]$ with equation 9 we obtain

$$E\left[Y_j, Y_{j-1}\right] = \frac{\alpha_0 \mu_Y + \alpha_1 E\left[X_j, Y_{j-1}\right] + \beta_1\left(\sigma_Y^2 + \mu_Y^2\right)}{\left(1 - \beta_2\right)}.$$

Equation 13

$$\begin{aligned} E\left[Y_j Y_{j-2}\right] &= E\left[Y_{j-2}\left(\alpha_0 + \alpha_1 X_j + \beta_1 Y_{j-1} + \beta_2 Y_{j-2} + \varepsilon_j\right)\right] \\ &= E\left[\alpha_0 \mu_X\right] + \alpha_1 E\left[X_j Y_{j-2}\right] + \beta_1 E\left[Y_{j-1} Y_{j-2}\right] + \beta_2 E\left[Y_{j-1}^2\right] \\ &= \alpha_0 \mu_Y + \alpha_1 E\left[X_j Y_{j-2}\right] + \beta_1 E\left[Y_{j-1} Y_{j-2}\right] + \beta_2\left(\sigma_Y^2 + \mu_Y^2\right) \end{aligned}$$

Noting that $E\left[Y_{j-1} Y_{j-2}\right] = E\left[Y_j Y_{j-1}\right]$ we have

$$E\left[Y_j, Y_{j-2}\right] = \alpha_0 \mu_Y + \alpha_1 E\left[X_j, Y_{j-2}\right] + \beta_1 E\left[Y_j, Y_{j-1}\right] + \beta_2\left(\sigma_Y^2 + \mu_Y^2\right).$$

Equation 8

By definition $\sigma_Y^2 = Var\left[Y_j\right] = E\left[\left(Y_j - \mu_Y\right)^2\right] = E\left[Y_j^2\right] - \mu_Y^2$.

Now,

$$\begin{aligned} E\left[Y_j^2\right] &= E\left\{\left(\alpha_0 + \alpha_1 X_j + \beta_1 Y_{j-1} + \beta_2 Y_{j-2} + \varepsilon_j\right) Y_j\right\} \\ &= E\left\{\left[\alpha_0 Y_j\right] + \alpha_1 X_j Y_j + \beta_1 Y_j Y_{j-1} + \beta_2 Y_j Y_{j-2} + Y_j \varepsilon_j\right\} \\ &= \alpha_0 \mu_Y + \alpha_1 E\left[X_j Y_j\right] + \beta_1 E\left[Y_j Y_{j-1}\right] + \beta_2 E\left[Y_j Y_{j-2}\right] \end{aligned}$$

Replacing the expectations with equations 11, 12, and 13 respectively gives

$$\sigma_Y^2 = \frac{\left\{2\alpha_1^2\left(\beta_2^2 - \beta_2 - \beta_1^2\right)\phi_2 + 2\alpha_1^2\beta_1\phi_1 + \alpha_1^2\left(\beta_2 - 1\right)\phi_0 + \left(2\beta_1^2 - 2\beta_2^2 + 2\beta_1 + \beta_2 + 1\right)\alpha_1^2\mu_X^2 + \left(\beta_2 - 1\right)\sigma_\varepsilon^2\right\}}{\left[\left(1+\beta_2\right)\left(\beta_1 + \beta_2 - 1\right)\left(\beta_1 - \beta_2 + 1\right)\right]}$$

Explicit solution of error variance

Equations 8, 14, and 17 can be combined and solved for σ_ε^2. Doing this yields

$$\sigma_\varepsilon^2 = A_0\,\phi_0 + A_1\,\phi_1 + A_2\,\phi_2 + B\mu_X^2 + C\,\sigma_X^2 + D\,\sigma_{\bar{Y}}^2$$

where

$$A_0 = \frac{2\alpha_1^2}{M}\left[2\beta_2^2\left(\beta_2 - 1\right) - 2\beta_1\beta_2 - \beta_1^2\left(1 + \beta_2\right)\right]$$

$$A_1 = \frac{-4\alpha_1^2}{M}\left[\beta_1^3 + 2\beta_1\beta_2^2 - \beta_2^4 + \beta_1^3\beta_2 + \beta_1^2\beta_2^2 + 2\beta_2\beta_1^2 + \beta_2^2 + \beta_1\beta_2 - \beta_2^3\beta_1\right]$$

$$A_2 = \frac{4\alpha_1^2}{M}\left[\left(\beta_1 + 2\right)\beta_2^4 + \left(\beta_1^2 - 2\right)\beta_2^3 - \left(\beta_1^2 + 3\beta_1 + 3\right)\beta_1\beta_2^2 - \left(\beta_1^2 + 2\beta_1 + 2\right)\beta_1^2\beta_2 - \beta_1^4\right]$$

$$B = \frac{-2\alpha_1^2}{M}\left[\begin{array}{r} 2\left(\beta_1 + 3\right)\beta_2^4 + 2\left(\beta_1^2 + \beta_1 - 1\right)\beta_2^3 - 2\left(\beta_1^3 + 4\beta_1^2 + 5\beta_1 + 2\right)\beta_2^2 \\ -\left(2\beta_1^3 + 6\beta_1^2 + 9\beta_1 + 4\right)\beta_1\beta_2 - \left(2\beta_1^2 + 2\beta_1 + 1\right)\beta_1^2 \end{array}\right]$$

$$C = \frac{-n\,\alpha_1^2}{M}\left[\beta_2^3 - \beta_2^2 - \left(\beta_1^2 + 1\right)\beta_2 - \left(\beta_1^2 - 1\right)\right]$$

$$D = \frac{n^2}{M}\left[\beta_2^5 + \left(2\beta_1 - 3\right)\beta_2^4 - 2\left(2\beta_1 + 1\right)\beta_2^3 - 2\left(\beta_1^3 - 1\right)\beta_2^2 - \left(\beta_1^4 - 4\beta_1 + 3\right)\beta_2 - \left(\beta_1^4 - 2\beta_1^3 + 2\beta_1 - 1\right)\right]$$

$$M = \left[\left(n-4\right)\left(\beta_2 - 1\right)\beta_2^2 - \left(n\beta_1^2 - 2\beta_1^2 - 4\beta_1 + n\right)\beta_2 - \left(n-2\right)\beta_1^2 + n\right]$$

ACKNOWLEDGEMENTS

The author is grateful to Isabelle Gabas and Southern Rural Water for making available the monitoring data used in the example. This work was jointly funded by the West Gippsland Catchment Management Authority, Southern Rural Water, Goulburn Murray Water, the University of Melbourne, and CSIRO Land and Water.

FIGURES

Figure 1. Study region - Gippsland Lakes in eastern Victoria, Australia (inset).

Figure 2 Time series of daily discharge (ML) for the period March 1, 1998 to 30 September, 2004 in irrigation drain CG3.

Figure 3 Time series of daily total phosphorus (TP) concentration $\left(\mu g \mathrm{L}^{-1}\right)$ for the period March 1, 1998 to 30 September, 2004 in irrigation drain CG3.

Figure 4 Scatter diagram of daily TP concentration $\left(\mu g \mathrm{L}^{-1}\right)$ and daily discharge (ML) $\left(\mu g \mathrm{L}^{-1}\right)$ for the period March 1, 1998 to 30 September, 2004 in irrigation drain CG3.

Figure 5 Daily total phosphorus (TP) concentration $\left(\mu g \mathrm{L}^{-1}\right)$ (gray lines) and monthly average TP concentration $\left(\mu g \mathrm{L}^{-1}\right)$ (solid dots) for the period March 1, 1998 to 30 September, 2004 in irrigation drain CG3.

Figure 6 Predicted monthly mean *ln*(TP) concentration using 'true' transfer function model parameters (solid line) and estimated transfer function model parameters (broken line).

Figure 7 Simulated daily *ln*(TP) concentrations (gray lines); actual daily *ln*(TP) concentrations (crosses); and approximate 95% confidence envelope (solid black lines).

Figure 8 Statistical summary of 1,000 simulated total loads.

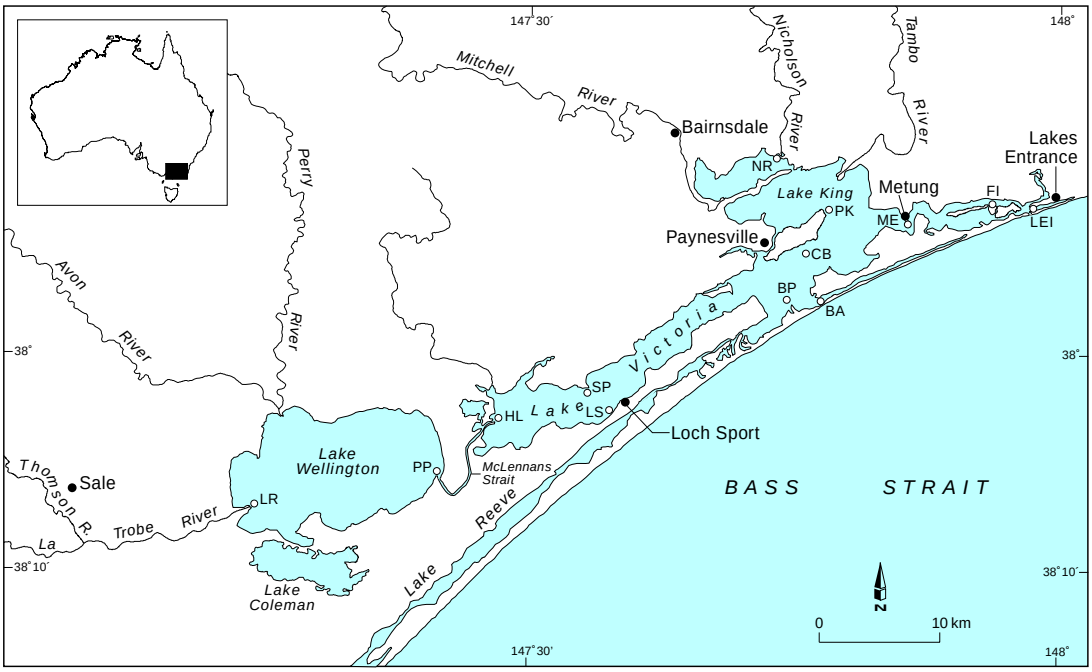

147°30´
148°
Mitchell River
Nicholson River
Tambo River
Bairnsdale
Lakes Entrance
NR
Lake King
Metung
FI
PK
ME
LEI
Paynesville
CB
BP
BA
Perry River
Avon River
38°
Victoria
SP
HL
Lake
LS
Loch Sport
Lake Wellington
PP
McLennans Strait
Sale
Thomson R.
LR
Trobe River
La
Reeve
BASS STRAIT
38°10´
Lake Coleman
Lake
N
0
10 km
147°30'
148°

Figure 1

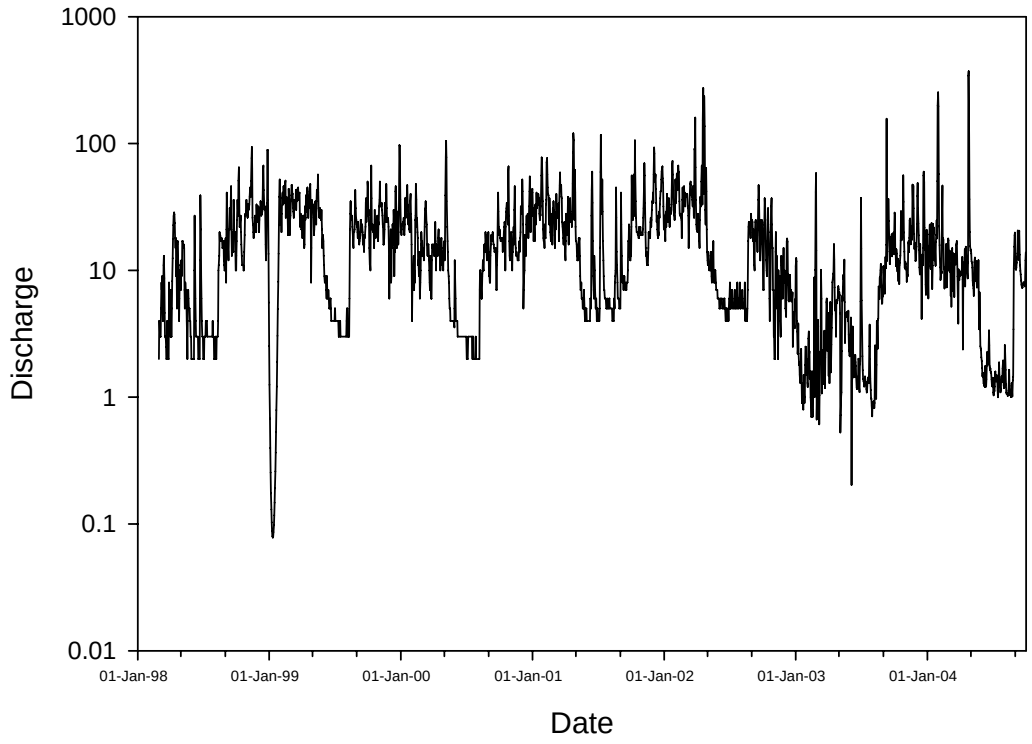

1000
100
10
1
0.1
0.01
Discharge
01-Jan-98
01-Jan-99
01-Jan-00
01-Jan-01
01-Jan-02
01-Jan-03
01-Jan-04
Date

Figure 2

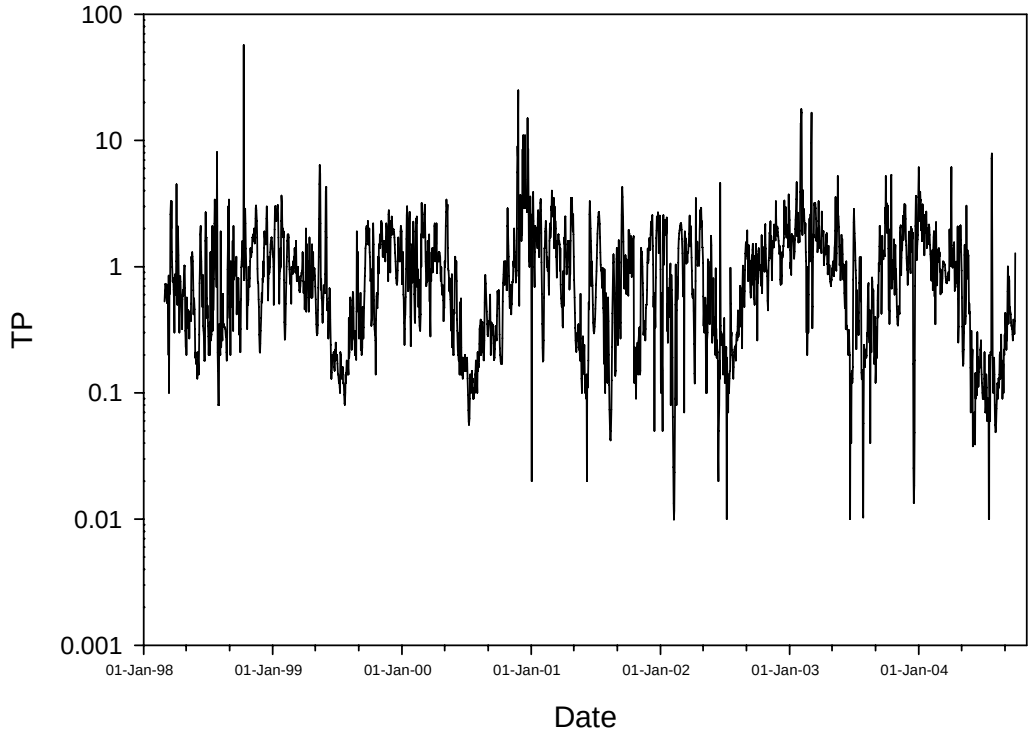

100
10
1
0.1
0.01
0.001
TP
01-Jan-98
01-Jan-99
01-Jan-00
01-Jan-01
01-Jan-02
01-Jan-03
01-Jan-04
Date

Figure 3

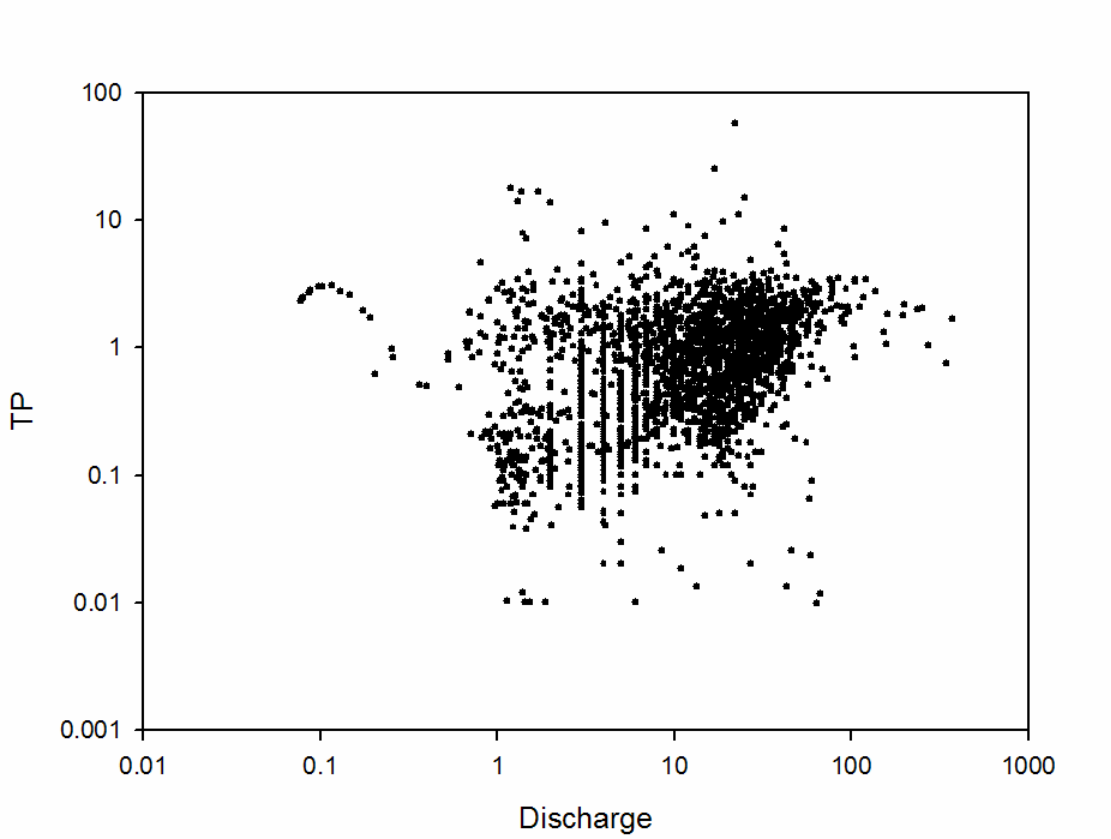


Figure 4

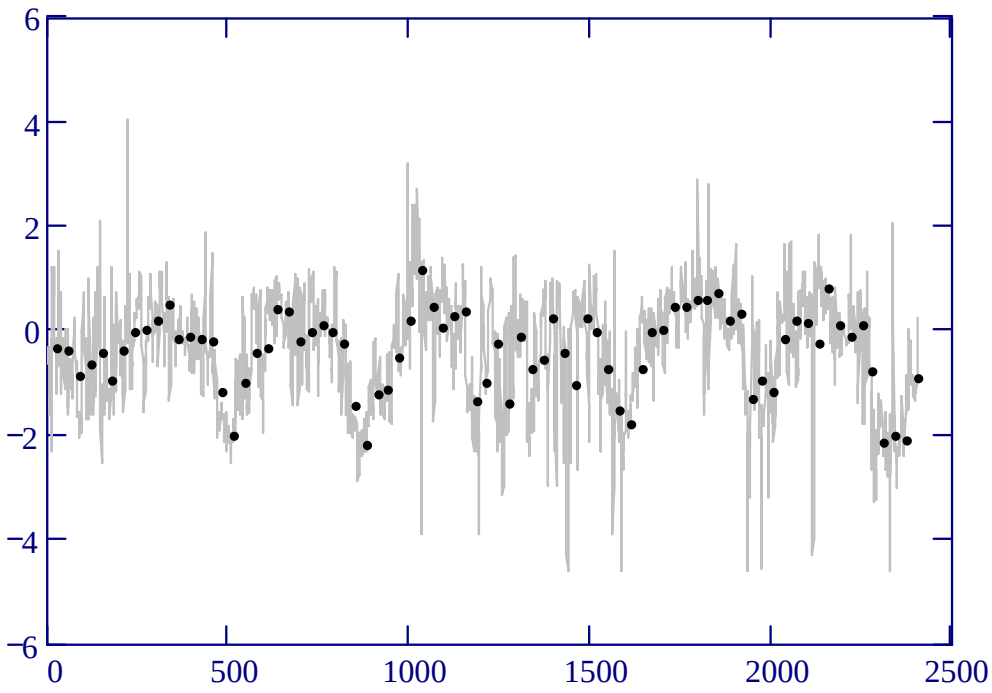


Figure 5

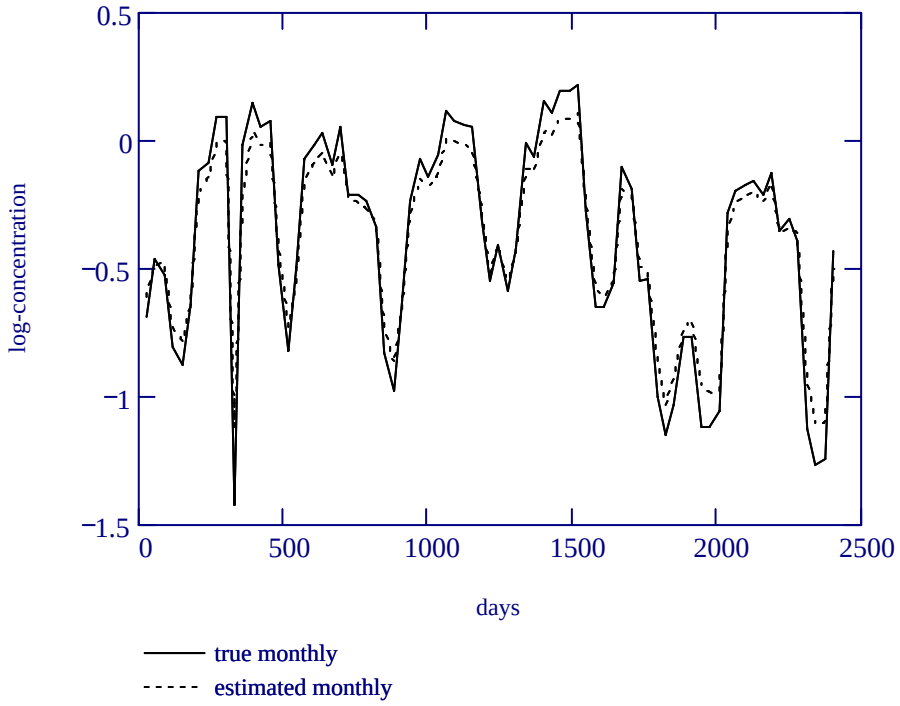

0.5
0
−0.5
−1
−1.5
log-concentration
0
500
1000
1500
2000
2500
days
true monthly
estimated monthly

Figure 6

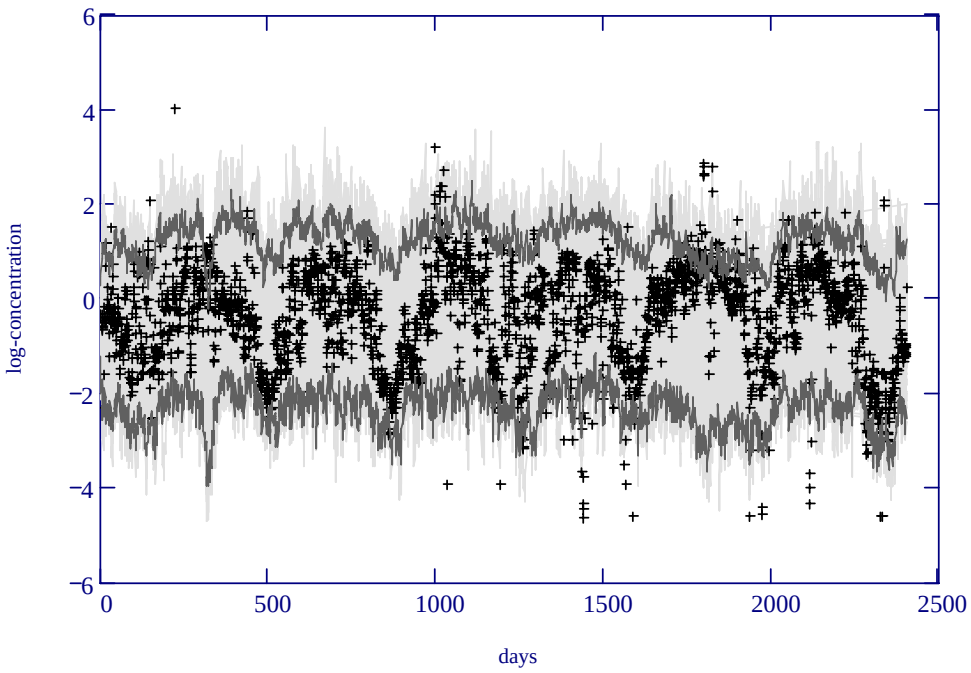
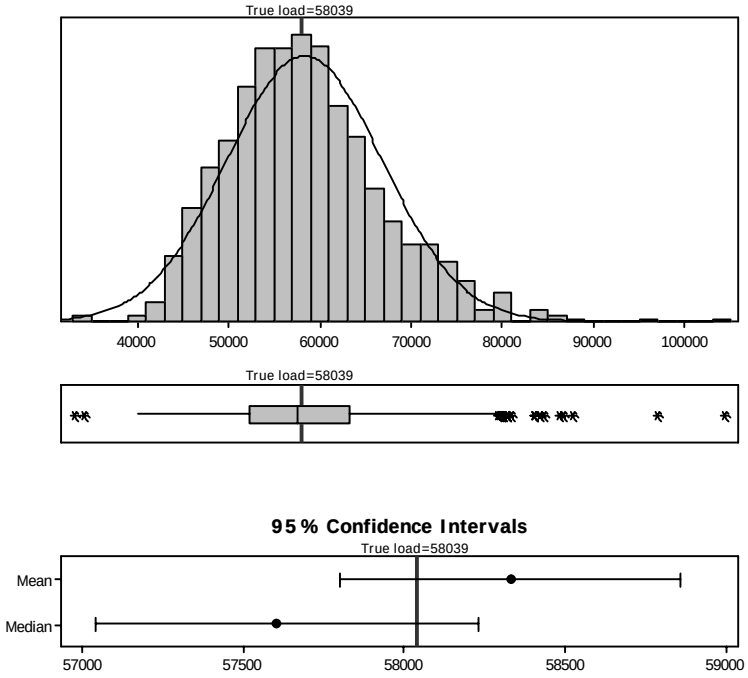


Figure 7



Anderson-Darling Normality Test	
A-Squared	4.26
P-Value <	0.005
Mean	58329
StDev	8541
Variance	72946043
Skewness	0.67809
Kurtosis	1.21739
N	1000
Minimum	33075
1st Quartile	52312
Median	57606
3rd Quartile	63196
Maximum	104389
95% Confidence Interval for Mean	
57799	58859
95% Confidence Interval for Median	
57045	58231
95% Confidence Interval for StDev	
8182	8933

Figure 8

TABLES

Table 1 Parameters for transfer function model (equation 6), estimated or actual values, and method of computation.

Parameter	Value or estimate	Method
μ_X	2.324867	Mean of *all* daily *ln*(discharge) data
μ_Y	-0.3880	Equation 7 and $\hat{\Theta}$
S_X^2	1.308395	Variance of *all* daily *ln*(discharge) data
$S_{\bar{Y}}^2$	0.569736	Sample variance of *monthly* mean concentration data
$Cov\left[Y_j, Y_{j-1}\right]$	0.799261	Equation 14 and $\hat{\Theta}$
σ_ε	0.427200	Section 4.1