18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

An extended regression approach to estimating loads and their uncertainties in Great Barrier Reef catchments

Wang, Y-G.[1], P. Kuhnert[2], B. Henderson[3] and L. Stewart[4]

[1]CSIRO Division of Mathematical and Information Sciences, Indooroopilly, Queensland
[2]CSIRO Division of Mathematical and Information Sciences, Cleveland, Queensland
[3]CSIRO Mathematical and Information Sciences, Canberra, Australian Capital Territory
[4]CSIRO Land and Water, Townsville, Queensland
Email: you-gan.wang@csiro.au

Abstract: There are numerous load estimation methods available, some of which are captured in various online tools. However, most estimators are subject to large biases statistically, and their associated uncertainties are often not reported. This makes interpretation difficult and the estimation of trends or determination of optimal sampling regimes impossible to assess.

In this paper, we first propose two indices for measuring the extent of sampling bias, and then provide steps for obtaining reliable load estimates by minimizing the biases and making use of possible predictive variables. The load estimation procedure can be summarized by the following four steps.

(i) output the flow rates at regular time intervals (e.g.10 minutes) using a time series model that captures all the peak flows;

(ii) output the predicted flow rates as in (i) at the concentration sampling times, if the corresponding flow rates are not collected;

(iii) establish a predictive model for the concentration data, which incorporates all possible predictor variables and output the predicted concentrations at the regular time intervals as in (i); and

(iv) obtain the sum of all the products of the predicted flow and the predicted concentration over the regular time intervals to represent an estimate of the load.

The key step to this approach is in the development of an appropriate predictive model for concentration. This is achieved using a generalized regression (rating-curve) approach with additional predictors that capture unique features in the flow data, namely the concept of the first flush, the location of the event on the hydrograph (e.g. rise or fall) and cumulative discounted flow. The latter may be thought of as a measure of constituent exhaustion occurring during flood events. The model also has the capacity to accommodate autocorrelation in model errors which are the result of intensive sampling during floods. Incorporating this additional information can significantly improve the predictability of concentration, and ultimately the precision with which the pollutant load is estimated. We also provide a measure of the standard error of the load estimate which incorporates model, spatial and/or temporal errors. This method also has the capacity to incorporate measurement error incurred through the sampling of flow. We illustrate this approach using the concentrations of total suspended sediment (TSS) and nitrogen oxide (NOx) and gauged flow data from the Burdekin River, a catchment delivering to the Great Barrier Reef. The sampling biases for NOx concentrations range from 2 to 10 times indicating severe biases. As we expect, the traditional average and extrapolation methods produce much higher estimates than those when bias in sampling is taken into account.

Keywords: *Biased sampling, bootstrap, Load estimation, Suspended sediment, Uncertainty*, water quality.

1. INTRODUCTION

Quantifying the amount of sediment, nutrients and pesticides (via a "load") entering into the Great Barrier Reef (GBR) is a primary focus for water quality improvement plans that aim to halt or reverse the decline in reef health over the next 5 years (The State of Queensland and Commonwealth of Australia 2003). Here, the term "load" represents the amount of material (whether sediment, nutrients or pesticides) transported past a location in a river over a specified period of time. Substantial work has been undertaken to define loads under varying conditions and assumptions and, a large body of literature has been devoted to this topic.

In this paper, we are primarily interested in quantifying the uncertainty in loads, where uncertainty is comprised into three components: measurement error, stochastic uncertainty and knowledge uncertainty. Many approaches in the literature do not explicitly incorporate uncertainty. Those that do, focus on some aspects of uncertainty but not all. For example, there are many simulation based approaches that tackle uncertainty by examining the variability amongst load methodologies (Guo et al. 2002, Etchells et al. 2005), while others develop an approximation for various loads estimation approaches. Tarras-Walhberg & Lane (2003) use Monte Carlo simulation to generate alternative log concentration values for their regression model and thus enable a family of curves to be generated. Rustomji and Wilkinson (2009) use the bootstrap technique to resample the residuals and place confidence intervals around estimates of load based on a non-linear regression approach.

We attempt to develop a general estimation procedure that provides reliable (with minimum bias) load estimates and their associated uncertainties. The prediction for both flow and concentration is performed at the same regular time-intervals to alleviate sampling biases and correlation is introduced into the modeling process to account for serial dependence. We acknowledge that there may be times when concentration is at best weakly related to discharge. What we are advocating is to consider other explanatory variables and attributes of the hydrograph because these additional variables may still be useful in explaining substantial amounts of the variability. In the worst-case scenario, where there is no predictive power, the regression model can be seen to default back to predicting an average concentration, and the load estimated becomes a form for the popular average estimator.

The structure of this paper is as follows. We will first introduce two types of bias indices measuring the extent of possible biases in the data (Section 2). Section 3 provides the estimation procedures and prediction methodology. The results from applying these methods to the Burdekin catchment are presented in Section 4. Finally in Section 5, we provide some discussion around the methodology and possible implications for future sampling.

2. TWO INDICES MEASURING SAMPLING BIASES

To illustrate bias sampling consider the following estimates of bias for concentration and flow recorded at Inkerman Bridge in the Burdekin catchment. Flow data is gauged and recorded hourly, while concentration, in particular NOx and total suspended sediment (TSS) are recorded less frequently.

Table 1: Illustration of bias in the sampling regime for the Burdekin catchment at Inkerman Bridge. (Data courtesy of Miles Furnas and Allan Mitchell, AIMS)

Year	$\bar{Q}$	$\bar{Q}_{\text{int}}$	$\bar{Q}_C$	Bias Index (R_q)	Bias Index (R_c)	n
1995/1996	67.88	67.09	373.91	1.012	5.57	19
1996/1997	276.56	27.63	1735.68	1.011	6.34	78
1997/1998	257.01	25.11	2611.64	1.007	10.24	39
1998/1999	21.09.73	218.39	574.67	1.006	2.63	70
1999/2000	435.74	43.82	2294.19	1.004	5.29	100

The table above summarises the flow data recorded across five water years, $\overline{Q}$ represents the mean recorded flow, $\overline{Q}_{\text{int}}$ represents the mean interpolated flow (hourly) and $\overline{Q}_C$ represents the mean flow evaluated where concentration was measured. The number of concentration records, *n*, is indicated by the last column of the table.

We computed the ratio of the sample average flow to the interpolated flow ($R_q = \overline{Q} / \overline{Q}_{\text{int}}$) and the ratio of the flow recorded only when concentration has been measured, relative to the interpolated flow ($R_c = \overline{Q}_C / \overline{Q}_{\text{int}}$). These indices measure the extent of the sampling biases. Table 1 shows that the bias resulting from using an interpolated flow record across all 5 water years is negligible. However, the bias incurred from using flow that is only recorded when concentration is measured are very substantial (up to 10 times). If we sample concentrations more often at high flows, the sample mean for the concentration will be over-represented and bias in load estimate may occur (for methods that use the average concentration). This highlights the importance of accounting for biases in the loads estimation procedure.

3. LOAD ESTIMATION

There are numerous methods for estimating pollutant loads (see, e.g. Phillips *et al.*, 1999; Letcher et al., 2002). In general, the total Load, *L* can now be estimated by

$$\hat{L} = K \sum_{m=1}^{M} \hat{c}_m \hat{q}_m \delta, \qquad (1)$$

where $(\hat{c}_m, \hat{q}_m)$ are the observed or imputed/estimated values for time interval, *m*. Note, the time interval δ must be small such that $(\hat{c}_m, \hat{q}_m)$ can be regarded as a constant during each interval and to avoid possible bias in $\hat{L}$. The parameter *K* here is a unit-conversion constant. It is important to note that if the time interval for the flow sampling is not a constant, bias will arise when the flow is related to the time intervals.

The fundamental question now is how to obtain $(\hat{c}_m, \hat{q}_m)$ at regular time intervals. We propose using the following steps to obtain the most reliable load estimates by minimizing the biases (step i and ii) and making use of possible predictive variables (step iii),

(i) output the flow rates at regular time intervals (e.g.10 minutes) using a time series model that captures all the peak flows,

(ii) output the predicted flow rates as in (i) at the concentration sampling times, if the corresponding flow rates are not collected,

(iii) establish a predictive model for the concentration data and output the predicted c_i concentrations at the regular time intervals as in (i), the predictive model should incorporate possible predictive variables, and

(iv) obtain the sum of all the products of the predicted flow and the predicted concentration over the regular intervals for each reporting year (e.g. water year).

Step (i) is necessary to remove possible bias in the flow data. For example, if the sample mean of the flow data may not be representative of the average water discharge, potential bias may occur in the load estimation. Nevertheless, step (i) is relatively easy to achieve using a LOESS smoother with a short span for instance. Step (ii) is necessary for matching the corresponding flow data to the concentration data. This step becomes unnecessary if the flow data are available at the same time when concentration data are collected. Step (iii) is the key component and may require substantial modelling. We will adopt an extended rating curve approach and focus on developing useful predictive variables to increase the accuracy of the concentration estimates as promoted by Thomas and Lewis (1995). It should be noted that load estimates can still be biased even if we have a well established predictive model when Step (i) and (ii) are not followed. However, the accuracy of the approach relies heavily on the strength and consistency of the relationship. We rely on the rating-curve approach and extend it in three novel aspects.

(a). Develop informative variables that capture the underlying hydrological processes, which may improve the predictive accuracy.

(b). Generalize the linear relationship to more flexible non-linear functions (polynomials and splines).

(c). Allow the model errors to be correlated.

There are several hydrological phenomena that should be considered in estimating the sediment and nutrient loads of riverine systems and the impact of these on the derived load will vary from site to site. "First flush"

is the phenomenon whereby the first significant channelised flow of the wet season is generally accompanied by relatively high sediment and nutrient concentrations (Thomas and Lewis, 1995). Precipitation in GBR catchments occurs predominantly within a well-defined, summer wet season (November to April). The run-off and interflow associated with a wet season's initial, flow-inducing precipitation event tends to pick up unconsolidated, fine sedimentary material and nutrients that have accumulated on or just below the land surface of the catchment. These materials accumulate due to natural weathering, disturbance, anthropogenic activity (e.g. land cultivation) and biomass decay during the relatively long, intervening dry period between wet seasons (Wallace et al. 2008) and are readily entrained by the event runoff. There are physical processes which result in non-unique relationships between discharge and concentration at several temporal scales. Within events, there can be hysteresis in the concentration-discharge relationship, such that this relationship varies to form a clockwise or anti-clockwise loop. Frequently, concentration is higher on the rising limb of the hydrograph, due to depletion of sediment availability during the event (Thomas and Lewis, 1995); and possibly also higher rainfall intensity and sediment transport capacity on the rising limb.

Between events in a given season or year, concentration can generally decline due to depletion of available sediment. This depletion may be caused by transport of material weathered during prior dry seasons and increase in vegetation cover through the season.

Based on the covariates discussed above we consider the following model

$$\log(c_i) ,= \beta_0 + \sum_{k=1}^{9} \beta_k x_{ki} + \varepsilon_i \qquad (2)$$

where x_{1i} is log (Q), x_{2i} is the {log(Q^2)}, x_{3i} is time in days (t), x_{4i} is $\sin(2\pi t / 365.25)$, x_{5i} is $\cos(2\pi t / 365.25)$, x_{6i} is the limb (1=rising, -1=falling, 0=normal flush), x_{7i} is the cumulative discounted flow (CDF), $(1-a)\sum_{j\geq 0} a^j \hat{Q}_{i-j} / (1-a^i)$. Here a represents the discount factor between 0 and 1 and is chosen to be 0.95 per day (and hence roughly 0.5 per fortnight). We intend to further investigate the effect of this CDF and report elsewhere. The terms $x_{8i} = x_{1i}x_{6i}$ and $x_{9i} = x_{6i}x_{7i}$ are interaction terms and allow the effect of the limb to alter according to flow and the discount factor. In matrix notation, we can write (2) as $\log(C) = X_0\beta + \varepsilon$, where β is the parameter vector and X_0 is the n by 10 design matrix. The residuals ε_i are assumed to follow a first order Autoregressive (AR(1)) process. It is important to have the ranges of (x_i, c_i) and covered in the data or spurious predictions may occur. To this end, it may be necessary to impose an upper limit on C.

This format is an example of a generalized approach. Other covariates exist and may be important in some circumstances. What we are proposing here is a generalized framework of USGS model (Cohn *et al*., 1992), and illustrating how it may be used to improve predictive power and the accuracy / precision with which we measure pollutant loads, rather than the set of covariates for all occasions. The same extensions are relevant to the functional forms, which may range between simple linear to highly flexible spline relationships.

Using matrix notation, we can write the predicted concentrations as $(\hat{c}_m)_{1\leq m\leq M} = \exp\{X_1(X_0'X_0)^{-1}X_0 z\}$, in which X_0 is the design matrix in model fitting while X_1 is the M by 10 design matrix for predicting the M concentrations at the designed time sequence, and $z = \{\log(c_i)\}_{i=1,\ldots,n}$. For regression models, $\hat{c}_i$ may be predicted by the flow data via a parametric model. In general, we have L predicted by $\exp(\hat{l} + \varepsilon)$ where $\hat{l} = \sum_m \hat{c}_m \hat{q}_m$. Because $E(C) = E\{\exp(X_1\beta + \varepsilon\} = \exp(X_1\beta + \sigma^2/2\}$ and $E(\hat{c}_m) = c_m \exp(\sigma_m{}^2/2)$, where σ^2 and σ_m^2 are the variances of ε and $\hat{z}_m$, our proposed load estimator with bias correction is given by

$$\hat{L} = T\delta\sum_{m=1}^{M} \hat{c}_m \hat{q}_m \exp\{(s^2 - s_m^2)/2\}, \qquad (3)$$

where s^2 and s_m^2 are the estimates of variance of ε and $\hat{z}_m$. Ferguson (1986) and Koch and Smillie (1986) propose at least three other ways of correcting this estimator. We will not focus on fine tuning such bias because it is relatively small considering the other types of model bias and the associated uncertainties which will be considered below. Note that $\exp(s^2/2)$ may be replaced by the smearing estimate (Duan 1983), $\sum_{i=1}^{n}\exp(\hat{\varepsilon}_i)/n$, where $(\hat{\varepsilon}_i)$ are the residuals from the regression model.

We shall be interested in establishing the predictive variance of L in which a model error ε cannot be eliminated by increasing the sample size. The model error ε is assumed to have a variance σ^2 and correlation matrix $R(\rho)$ with autocorrelation parameter ρ which measures temporal correlation. Denote the vector of the load estimates at the regular intervals as (L_m), and after some algebra,

$$\begin{aligned}\mathrm{var}(\hat{L}) &= \mathrm{trace}\{\mathrm{var}(\hat{\beta})SS^T\} \\ &\quad + \alpha_1^2\{\sum L_m^2\{1+\partial f/\partial\log(Q_m)\}^2 + \alpha_2^2\left[\sum L_m\{1+\partial f/\partial\log(Q_m)\}\right]^2,\end{aligned}$$

where $f(\hat{Q})$ is the regression model on log scale, and $\partial f/\partial\log(Q_m)=\beta_1$ for the traditional rating curve model, $S = X_1^T(L_m)_{1\le m\le M}$ is a matrix of *K* by *M* (*K* is the number of parameters), SS^T is a square matrix of *K* by *K*, and α_1 and α_2 are the coefficient of variation (CV) of the independent measurement error and spatial/temporal random effects in log(Q). We will illustrate the calculation using different error rates in the example.

4. CASE STUDY: THE BURDEKIN CATCHMENT

The Burdekin River drains an area of 130,000 km^2 and the catchment is the second largest draining to the GBR lagoon and is the largest in terms of mean gauged annual discharge. The distribution of land use within the catchment is represented by approximately 1% cropping, primarily confined to the Burdekin Delta, cattle grazing 95% and 4% other uses. Development of the catchment by European settlers began in the mid-1800s with the introduction of sheep and cattle grazing (Lewis et al. 2007) and the probable commencement of alluvial mining. It is generally accepted that post settlement activities such as these would have increased the annual average flux of sediment to the GBR lagoon (see e.g., Belperio 1979) and in recent years trace-element analysis of coral cores has provided evidence in support of that proposition (McCulloch et al. 2003; Lewis et al. 2007). We use the Burdekin as an initial case study to evaluate and explore the methods that we have developed for estimating loads and quantifying uncertainty.

Data collected at Inkerman Bridge was collected by the Australian Institute of Marine Science (AIMS) between 1987 and 2000 as part of their riverine monitoring program for the purpose of calculating annual loads. Flow data was recorded by a Natural Resources and Water (NRW) gauge located at Clare, which resides approximately 20km upstream from the sampling site where concentrations of TSS and NOx were recorded.

We fitted different models to a subset of the data (Water Year 1996/1997) as described in Section 3. The best model based on GCV is M1 which has intercept, limb, log(Flow), its quadratic term; 2 periodic terms (annual and 6 month cycles), and spline of CDF with serial correlations (correlation is estimated to be 0.75). Average and extrapolation produce very highly biased estimates due to the opportunistic sampling designs, while the rating curve models and ratio (and Beale) estimators perform similarly (Table 2) For TSS our new estimates are about 15% higher than the ratio and Beale's estimates (when there is a clear relationship indicated in Fig 1 (top left panel), while for NOx, where there is no apparent relationship between flow and concentration (see bottom left panel of Fig. 1), we obtained similar estimates to the Ratio counterparts. The proposed methodology provides a natural way of further incorporating possible informative variables and quantifying the uncertainties. In Table 2, 95% two confidence intervals (A and B) are given for two sets of hypothesized error rates of (α_1,α_2). Our framework will also enable us to study efficiency of different designs. Designs with occasional sampling at the beginning and the end of the water year (although flow is ambient) would be able to improve the prediction model.

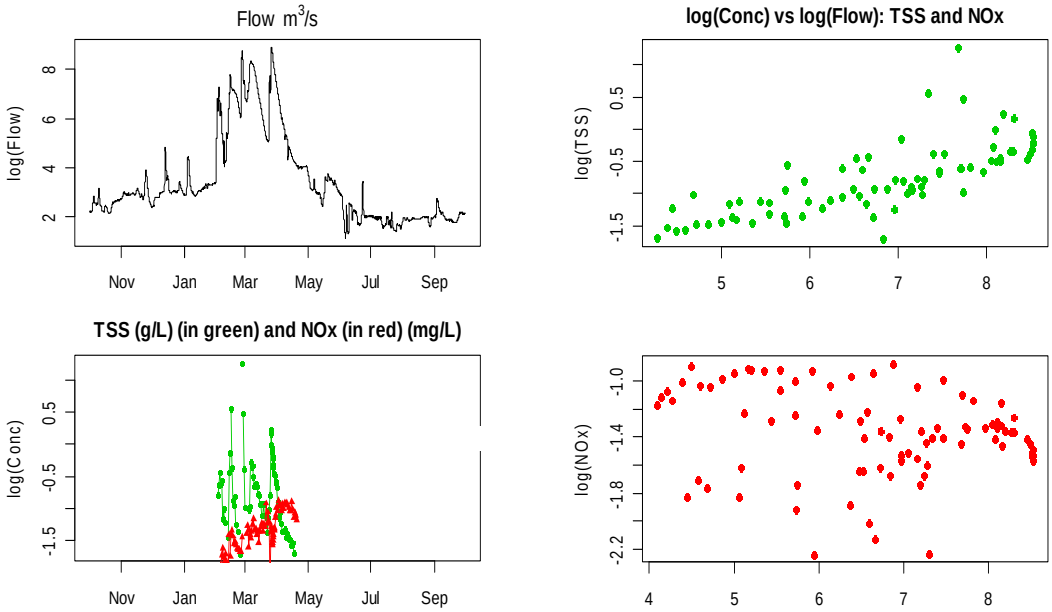

Fig. 1: Plots showing the flow and concentration data (left panels) of TSS (n=78, g/L)and NOx (n=86, mg/L). The relationships between the concentration and flow are plotted on the right panels (on log scale).

Table 2: Estimates of the total TSS and NOx load, $\hat{L}$, the corrected load, $\hat{L}_C$, and the 95% confidence intervals assuming $(\alpha_1, \alpha_2) = (0.10, 0.05)$ and $(\alpha_1, \alpha_2) = (0.30, 0.10)$.

Load	$\hat{L}$ (t)	$\hat{L}_C$	*n*	95%CI (A)	95%CI (B)	Avg	Extrp	Ratio	Beale
TSS (Mt)	6.98	7.26	78	(5.00, 10.54)	(4.36, 12.10)	27.67	37.43	5.99	6.00
NOx (Kt)	2.02	2.04	86	(1.87, 2.22)	(1.75, 2.38)	12.47	11.72	2.01	2.01

5. DISCUSSION

The relationship between flow and concentration and the nature of the sampling undertaken will dictate the type of model required. In some situations, a model with flow and concentration only will provide an adequate model, however in other situations, although there may not be an obvious relationship between flow and discharge, other variables such as the cumulative discounted flow or rising/falling limb may be important. These terms need to be thoroughly investigated in any modeling exercise undertaken.

Regression estimators can be highly biased, especially if systematic sampling is used in an event responsive system (Preston et al. 1992). Preston et al. (1989) also found that estimates of load produced via a regression approach can be less accurate than those produced by the ratio estimator if a small number of samples are collected and the relationship is not well understood. However, in these examples, other covariates were not explored and the models were based predominantly on functions of flow. We believe that the extended regression approach provides a natural way of handling the sampling bias and that other classical methods do not have such advantages. Despite this, regression-based estimates improve when adequate sampling has been undertaken over a broad range of conditions, thus providing the best estimates with low error, it is still reliable even when there is little no relationships between the concentration and flow rate. Additional work can be carried to validate the methodology further and determine what combinations of variables are useful for modelling certain types of datasets and to investigate more complicated modeling structures that include interactions to determine whether a more complex model is required. We also aim to compare the proposed method with others by decimating a dataset with intensive concentration records.

ACKNOWLEDGMENTS

Inkerman Bridge dataset was kindly provided by Miles Furnas and Alan Mitchell (AIMS). We also wish to thank Rebecca Bartley for internally reviewing this paper and two referees for their constructive comments.

REFERENCES

Belperio, A. P. (1979), The combined use of wash load and bed material load rating curves for the calculation of total load: an example from the Burdekin River, Australia. Catena 6, 317-329.

Cohn, T. A., D. L. Caulder, E. J. Gilroy, L. D. Zynjuk, and Summers, R. M. (1992), The validity of a simple statistical model for estimating fluvial constituent loads: an empirical study involving nutrient loads entering Chesapeake Bay. Water Resources Research **28**: 2353-2363.

Duan, N. (1983), Smearing estimate: A nonparametric retransformation method. Journal of the American Statistical Association, **78**: 605-610

Etchells, T.,Tan, K.S.and Fox, D.R. (2005),. Quantifying the uncertainty of nutrient load estimates in the Shepparton irrigation region. Pages 170-176 *in* MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and new Zealand, Melbourne Australia.

Ferguson, R.I. (1986), Hydraulics and hydraulic geometry. Progress in *Physical Geography,* 10, 1-31.

Guo, Y.P., Markus, M. and Demissie, M. (2002), Uncertainty of nitrate-N load computations for agricultural watersheds. *Water Resources Research,* 38(10), 1185, doi:10.1029/2001WR001149.

Koch, R.W. and Smillie (1986), Comment on ``River loads underestimated by rating curves" by R. I. Ferguson. *Water Resources Research,* 22**,** 2121-2122.

Letcher, R.A., Jakeman, A.J., Calfas, M., Linforth, M., Baginska, B. and Lawrence, I.. (2002), A comparison of catchment water quality models and direct estimation techniques. *Environmental Modelling & Software* 17, 77-85.

Lewis, S.E., Shields, G.A., Kamber, B.S. and Lough, J.M. (2007), A multi-trace element coral record of land-use changes in the Burdekin River catchment, NE, Australia. Palaeogeography, Palaeoclimatology, Palaeoecology **246**: 471-487.

McCulloch, M.T., Fallon, S., Wyndham, T., Hendy,E., Lough, J. M. and Barnes, D. (2003), Coral record of increased sediment flux to the inner Great Barrier Reef since European settlement. *Nature,* 421, 727-730.

Phillips, J.M., Webb, B.W., Walling, D.E. and Leeks, G.J.L. (1999), Estimating the suspended sediment load of rivers in the LOIS study area using infrequent samples. *Hydrological Processes*, 13,,1035-1050.

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

Preston, S.D., Bierman, J.V.J. and Silliman, S.E. (1992), Impact of flow variability on error in estimation of Tributary mass loads. *Journal of Environmental Engineering*, 118, 402-418.

Rustomji, P., and Wilkinson, S.N. (2009), Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated using rating curves. *Water Resources Research,* 44.

Tarras-Wahlberg, N.H., and Lane, S. N. (2003), Suspended sediment yield and metal contamination in a river catchment affected by El Nino events and gold mining activities: the puyango river basin, southern ecuador. *Hydrological Processes*, 17, 3101-3123.

The State of Queensland and Commonwealth of Australia. (2003), Reef water quality protection plan: For catchments adjacent to the Great Barrier Reef World Heritage area. Queensland Department of Premier and Cabinet, Brisbane, see http://www.reefplan.qld.gov.au/library/pdf/reefplan.pdf.

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

Wallace, J., Stewart, L., Hawdon, A. and Keen, R. (2008), The role of coastal floodplains in generating sediment and nutrient fluxes to the great barrier reef lagoon in Australia. In: Abstracts: Ecohydrological Processes and Sustainable Floodplain Management Opportunities and Concepts for Water Hazard Mitigation and Ecological and Socioeconomic Sustainability, 19-23 May 2008. Lodz, Poland.