WATER RESOURCES RESEARCH, VOL. 49, 8565–8578, doi:10.1002/2013WR013559, 2013

Estimating bootstrap and Bayesian prediction intervals for constituent load rating curves

Olga Vigiak[1,2] and Ulrike Bende-Michl[3,4]

Received 23 January 2013; revised 21 November 2013; accepted 27 November 2013; published 19 December 2013.

[1] Assessment of constituent loads in rivers is essential to evaluate water quality of streams and estuaries; however, uncertainty in load estimation may be large and must be considered and communicated together with estimates. In this comparative study, the usefulness of two existing methods (bootstrap and Bayesian inference) to assess uncertainty in constituent loads estimated with an improved eight-parameter rating curve is demonstrated. Bootstrap prediction intervals and Bayesian credible intervals were estimated for daily and monthly loads obtained with a rating curve applied to routine monitoring sampling data sets of nitrate (NO3-N), reactive phosphorus (RP), and total phosphorus (TP) of the Duck River, in Tasmania (Australia). Predicted loads and prediction intervals were compared to benchmark loads obtained by an independent, high frequency monitoring program. The eight-parameter rating curve resulted in better prediction of NO3-N and TP than RP loads. Both inference methods successfully generated prediction intervals. The bracketing frequency (i.e., the fraction of prediction intervals that comprised benchmark loads) of bootstrap prediction intervals was 50–65% of daily or monthly benchmark loads. Bracketing frequency of Bayesian credible intervals was consistently higher, and included 74–85% of benchmark daily loads and 80% or more of benchmark monthly loads. Both methods proved to be robust to the presence of an artificial outlier. Prediction intervals were affected by the distribution of the regression error, hence they reflected uncertainty in the regression data set and limitations in the rating curve formulation. They did not account for other sources of uncertainty, i.e., they were still conservative predictions of load uncertainty.

Citation: Vigiak, O., and U. Bende-Michl (2013), Estimating bootstrap and Bayesian prediction intervals for constituent load rating curves, *Water Resour. Res.*, *49*, 8565–8578, doi:10.1002/2013WR013559.

1. Introduction

[2] Assessment of constituent loads of rivers is necessary to benchmark water quality conditions of streams and estuaries, to measure changes in river water quality over time, to evaluate Total Maximum Daily Load (TMDL) compliance, and to calibrate catchment-scale models that inform water quality management [e.g., *Cohn et al*., 1992; *Johnes*, 2007; *Stenback et al*., 2011; *Saad et al*., 2011; *Wang et al*., 2011; *Verma et al*., 2012; *Wellen et al*., 2012].

[3] Commonly, water discharge is measured continuously at monitoring stations. Conversely, constituents are monitored periodically, typically at weekly to monthly intervals; these periodic measurements may (or more frequently may not) be complemented by additional sampling during runoff events [*Johnes*, 2007; *Bartley et al*., 2012]. Several methods exist to estimate constituent loads as a function of discharge, ranging from simple or stratified average or ratio methods, to rating curves, and interpolation methods. More recently composite methods have been developed that merge regression and interpolation techniques [*Aulenbach and Hooper*, 2006; *Verma et al*., 2012]. Each method has recognized advantages and limitations, and their appropriateness depends on constituent and sample characteristics [*Letcher et al*., 1999; *Johnes*, 2007; *Verma et al*., 2012]. Rating curves (i.e., regression of constituent concentration or load against discharge and auxiliary covariates) are recommended methods for sediment and nutrient load estimation, as the relationship between these constituent concentrations and discharge in the log-log scale often appears close to linear [*Cohn et al*., 1992; *Stenback et al*., 2011; *Saad et al*., 2011; *Wang et al*., 2011], although concave or convex shapes can also be

Additional supporting information may be found in the online version of this article.

[1]Now at Joint Research Centre of the European Commission, Institute for Environment and Sustainability, Water Resources Unit, Ispra, Varese, Italy.
[2]Formerly at Department of Primary Industry of Victoria, Rutherglen Centre, Rutherglen, Victoria, Australia.
[3]Division of Land and Water Resources, Commonwealth Scientific and Industrial Research Organisation, Canberra, Australian Capital Territory, Australia.
[4]Bureau of Meteorology, Canberra, Australian Capital Territory, Australia.

Corresponding author: O. Vigiak, Joint Research Centre of the European Commission, Institute for Environment and Sustainability, Water Resources Unit, I-21027 Ispra, Varese, Italy. (olga.vigiak@jrc.ec.europa.eu)

0043-1397/13/10.1002/2013WR013559

observed [*Horowitz*, 2003; *Crowder et al*., 2007]. *Cohn et al*. [1992] proposed a seven-parameter general rating curve to estimate constituent concentration and load as a function of discharge, season, and long-term trend, that remains among the most common methods of load estimation in the United States and worldwide [*Stenback et al*., 2011].

[4] Rating curve fitting is, however, impaired when samples do not represent the full complexity of hydrological processes affecting constituent loads during runoff events, such as hysteresis or exhaustion [e.g., *Wang et al*., 2011]. This is often the case under monitoring schemes based on periodic sampling, whose samples are taken in prevalence at low discharge, so that the constituent-discharge relationship is ill-defined at high discharges and heteroscedasticity of residuals in the rating curve is often observed [*Cox et al*., 2008]. Further, back-transformation of mean values from the log to the arithmetic space requires the application of a correction factor CF, whose formulation has been the subject of much literature [e.g., *Cohn et al*., 1992; *Cox et al*., 2008 and studies cited therein]. To reduce problems of heteroscedasticity of residuals and back-transformation errors associated with log-linear model fitting, recent literature has advocated fitting nonlinear models [e.g., *Asselman*, 2000] or generalized linear models [*Cox et al*., 2008; *Wang et al*., 2011; *Kuhnert et al*., 2012].

[5] Given the importance of constituent loads for natural resource management, the uncertainty associated with load estimation methods should be quantified [*Tan et al*., 2005; *Johnes*, 2007; *Krueger et al*., 2009; *Wang et al*., 2011; *Kuhnert et al*., 2012]. Provision of uncertainty information, e.g., in the form of prediction intervals, for loads estimated with rating curve methods is not trivial because the assumption of homoscedasticity of residuals is often violated [*Petersen-Øverleir*, 2004]. Furthermore, constituent loads are generally estimated daily, i.e., at the scale at which the rating curves are fitted to, but then aggregated and reported at monthly or annual periods, i.e., at the scale that is of interest for management; hence, uncertainty information needs to be propagated from daily to longer periods.

[6] Nonparametric methods that relax distributional assumptions in the variables and residuals may offer alternative approaches to estimate load prediction intervals. In addition, Monte Carlo procedures have been used to estimate confidence intervals of constituent loads by several methods [*Coats et al*., 2002; *Guo et al*., 2002; *Tan et al*., 2005]. In this context, the bootstrap method, which mimics the sampling scheme by resampling with replacement observations or residuals [*Efron and Tibshirani*, 1998; *Chernick*, 2008], appears appealing. *Aulenbach and Hooper* [2006] and *Ide et al*. [2011] used bootstrap to quantify predicted load's precision under several sampling schemes and to assess minimum sampling requirements to achieve precision targets. *Rustomjii and Wilkinson* [2008] applied bootstrap of residuals to predict 95% confidence intervals of annual sediment loads using a nonlinear rating curve. Clearly, there is scope in further exploring the usefulness of the bootstrap method applied to constituent rating curves.

[7] Under a different paradigm, Bayesian inference combines the use of prior information about the distribution of model parameters and the likelihood function from the observed data set to provide belief posterior distributions of parameters and variable estimates. From the posterior distributions, credible intervals at given probability levels can be calculated. Bayesian inference is highly parametric as it requires formulating assumptions on the distribution of parameters in addition to the variables. However, it offers also considerable parametric flexibility. Thus, Bayesian inference has gained momentum since the advent of computing power has made running the Markov-chain Monte Carlo (MCMC) algorithm more feasible in real time. Applications in stage-discharge rating curve estimation have already been proven useful [*Moyeed and Clarke*, 2005; *Retain and Petersen-Øverleir*, 2011]. Bayesian inference has recently been applied to constituent rating curve estimation by *Alameddine et al*. [2011] and *Kulasova et al*. [2012].

[8] The objective of this study was to demonstrate the usefulness of bootstrap and Bayesian inference in estimating prediction intervals for daily and monthly loads when using an improved eight-parameter rating curve based on *Cohn et al*. [1992] applied to routine monitoring data sets. Eleven year data sets (1999–2009) of routine monitoring of three constituents, namely nitrate (NO3-N), reactive phosphorus (RP), and total phosphorus (TP), in the Duck River, Tasmania (Australia), were used to fit the rating curves. The rating curves were applied to a prediction data set (April 2008 to April 2010) derived from daily discharge data to provide estimates of loads and associated 95% prediction intervals. The prediction intervals were compared to benchmark ("true") constituent loads, obtained from an independent, high frequency nutrient monitoring data set taken from April 2008 to April 2010 at the same location [*Bende-Michl et al*., 2013].

2. Materials and Methods

2.1. The Duck River Catchment and Data Sets

[9] The Duck River catchment is located on the temperate North West Tasmanian coast (Australia, Figure 1). The long-term average annual rainfall is approximately 1150 mm. Monthly rainfall exceeds potential evapotranspiration from April to October and is highest for the month July; from November to March monthly potential evapotranspiration exceeds monthly rainfall, and February and March are the driest months [*DPIWE*, 2003]. Daily discharge and constituent concentrations were measured at Scotchtown station (station number 14214; GDA 94 Northing: 5473785, Easting: 341557), which drains 369 km^2 of the Duck River catchment (Figure 1). Historical (1995–2010) mean daily discharge at the station is 4.9 m^3/s, ranging from 0.4 to 88.5 m^3/s. Nutrient exports in the Duck River catchment are amongst the highest of Tasmania [*DPIWE*, 2003], and nutrient concentrations frequently exceed *ANZECC* [2000] trigger values for lowland rivers in Tasmania (e.g., 0.19 mg/L NO3-N, 0.02 mg/L DRP, and 0.05 mg/L TP).

[10] A detailed description of the study area can be found in *Bende-Michl et al*. [2013] and *Verburg et al*. [2012]. Briefly, the catchment is underlain by sedimentary mud and siltstones in the eastern and southern hills, by basaltic material in the north-eastern hills, and by marine sediments in the western hills. The lowland valley plains consist of alluvial deposits such as gravel, sand, silts and Quaternary

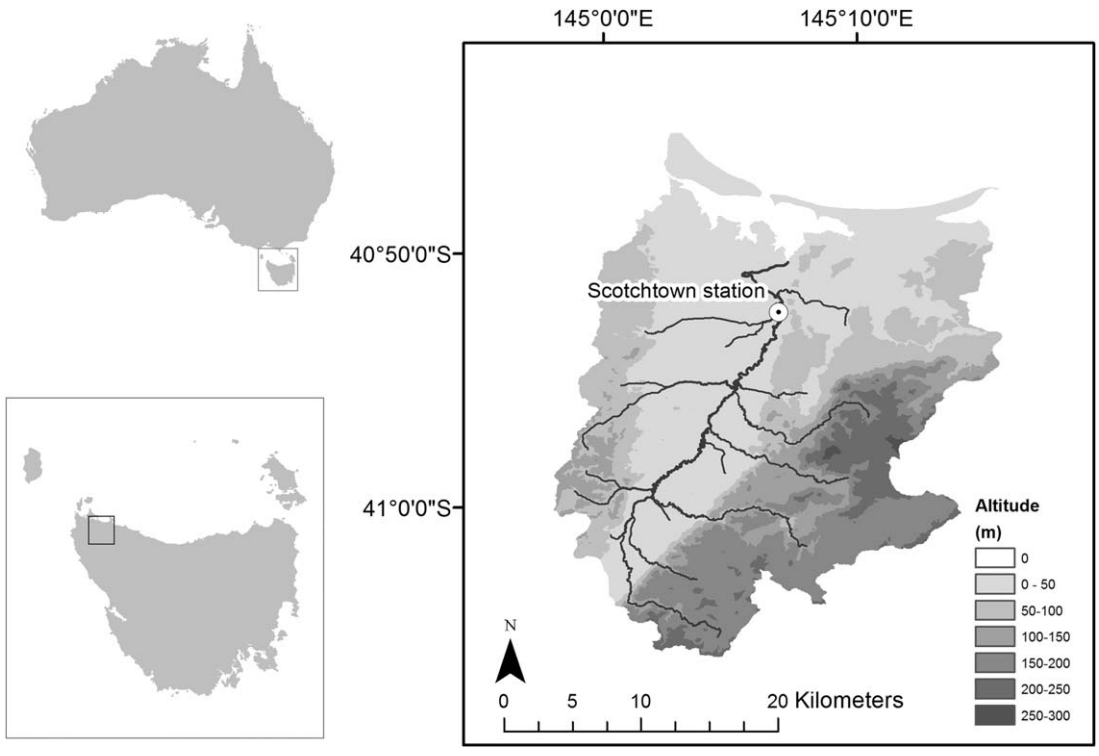

Figure 1. Location and topography of the Duck River catchment.

coastal dunes and are prone to rapid saturation and water logging [*Richley*, 1978]. Major land use types comprise intensive dairy production in the middle flat parts of the catchment, with extensive grazing, forestry, and some cropping in the hills. No sewage treatment plants are located in the study area; however, a dairy factory is located about 14 km above the sampling point.

[11] Discharge and historical water quality data at Scotchtown station for 1999–2009 were obtained from Water Information System Tasmania web service [*DPIWE of Tasmania*, 2011] (Table 1). Constituent (NO3-N, RP, and TP) concentration data were compiled from monthly monitoring, State of Rivers, and flood sampling programs. This data set was used for fitting the rating curves (data set I) and included 110 samples of NO3-N and RP, 116 samples of TP (Table 1), and daily discharge for the same period. The scatter plots of constituent concentrations against log-transformed daily discharge Q (m^3/s, Figure 2a) showed that NO3-N distribution deviated quite substantially from linear, with a change in the concentration-discharge relationship at about 1 m^3/s discharge (i.e., -0.9 in the x axis of Figure 2a where discharge is centered on the regression data set geometric mean value of 2.46 m^3/s), above which NO3-N concentration changed little with discharge. RP and TP concentrations showed a more linear relationship with discharge in the log-log space.

[12] From April 2008 to April 2010, a high frequency nutrient monitoring study was conducted at the same location [*Bende-Michl et al*., 2013; *Verburg et al*., 2012]. A detailed description of instrument selection [*Bende-Michl and Hairsine*, 2010] as well as the setup, maintenance, and quality control procedures of this monitoring program can be found in *Bende-Michl et al*. [2013]. NO3-N was measured in situ every 5 min by UV/Vis spectroscopy (s::can

Table 1. Statistical Summary of Regression Data Set I (Routine Sampling 1999–2009) and Evaluation Data Set E (High Frequency Monitoring From April 2008 to April 2010) of the Duck River, Tasmania (Australia)

	Daily Discharge Q (m3/s)		NO3-N (mg/L)		RP (mg/L)		TP (mg/L)	
	I	E	I	E	I	E	I	E
Size	110	743	110	553	110	652	116	566
Minimum	0.41	0.41	0.072	0.013	0.003	0.002	0.018	0.010
10th percentile	0.74	0.57	0.205	0.181	0.008	0.023	0.037	0.089
25th percentile	0.90	0.83	0.291	0.346	0.014	0.048	0.058	0.153
Median	2.10	2.19	0.424	0.571	0.025	0.068	0.097	0.241
Mean	4.67	5.12	0.446	0.616	0.083	0.114	0.224	0.384
75th percentile	5.70	6.50	0.551	0.855	0.052	0.110	0.192	0.414
90th percentile	11.85	12.73	0.700	1.090	0.216	0.251	0.667	0.832
Maximum	88.50	58.69	1.120	2.653	0.918	1.477	2.460	3.890

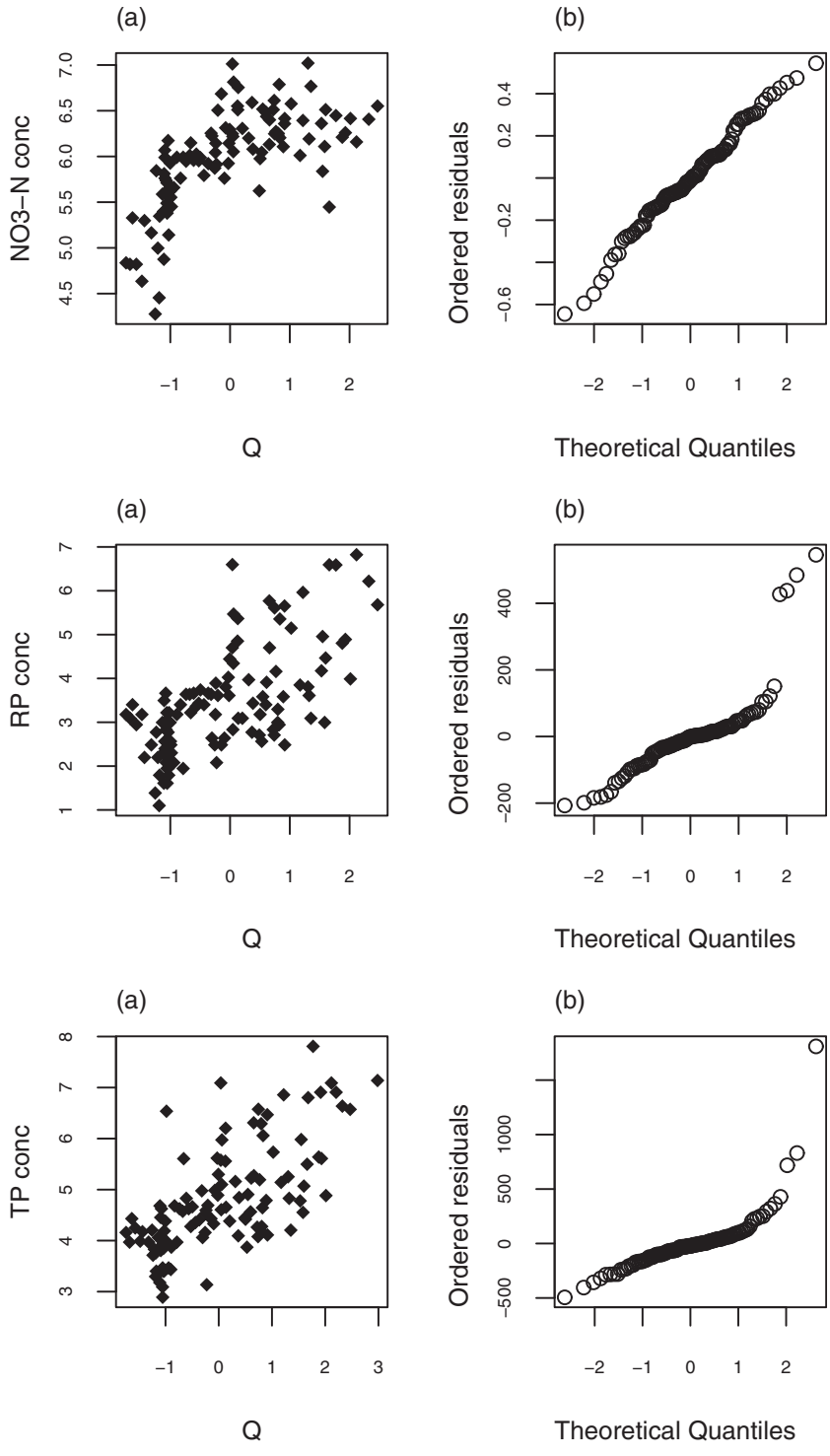

Figure 2. (a) Log-log distribution of constituent concentrations (mg/L) against daily discharge Q (m^3/s) in the regression data sets (daily discharge in the *x* axis is centered on its geometric mean); and (b) norm-quantile plots of rating curve (equation (1)) residuals.

spectrolyser measuring at 2 nm intervals between the 180 and 720 nm wavelengths) and analyzed based on the selective adsorption of electromagnetic radiation in the ultraviolet wavelength region (180–390 nm) [*Van den Broeke et al*., 2006]. To measure RP and TP concentrations, river samples were pumped to the instruments, filtered through a 20 μm filtration system, and measured with two automated bankside wet chemistry analyzers (SysteaMicromac C). Samples were analyzed by colorimetric analysis, following the standard molybdate blue methods with UV persulphate oxidation determined at 880 nm wavelength. TP analysis included a persulphate pretreatment digestion step [*Gray et al*., 2006]. RP was measured every 38 min, whereas TP was measured at hourly intervals. Analysis results were

stored in a database management system and checked for data quality and integrity through strict data quality control protocol [*Bende-Michl et al*., 2013]. The filtration for RP samples at 20 μm instead of the 0.45 μm employed in routine monitoring means that a slightly larger fraction of particulate phosphorus was potentially included in the high frequency RP concentrations compared to RP routine monitoring data set. Similarly, coarse P contents (>20 μm) could have been potentially missed in the high frequency TP concentration data set. However, analysis showed that high frequency monitoring P concentrations matched well routine sampling P concentrations, i.e., the two concentrations were commensurable [*Bende-Michl et al*., 2013].

[13] From these high frequency concentration and discharge monitoring, time-weighted nutrient loads at daily intervals were obtained, which represented the benchmark daily loads of the evaluation data set E. The data set comprised 553 NO3-N, 652 DRP, and 566 TP daily loads (Table 1). Due to technical issues, some gaps in daily load series were present. To obtain benchmark monthly loads, the sum of daily loads of each month was divided by the ratio of discharge for which load was available to the total discharge of the month. If this ratio was <0.7, the month load was excluded from the monthly evaluation data set.

2.2. Rating Curve Formulation

[14] The general rating curve proposed by *Cohn et al*. [1992] was slightly modified in this study by adding a covariate that accounted for constituent exhaustion proposed by *Wang et al*. [2011]:

$$\ln c_i = \beta_0 + \sum_{k=1}^{7} \beta_k x_{k,i} + \varepsilon_i \quad (1)$$

where ln indicates the natural logarithm; c_i is the daily constituent concentration (in mg/L) of day i; the intercept β_0 basically acts as scaling factor; x_k are covariates that together account for discharge, hysteresis, seasonality, exhaustion, and long-term trend (Table 2) [*Cohn et al*., 1992; *Wang et al*., 2011]; and ε_i is the error (residual), which is assumed to be independent, identically, and normally distributed with mean 0 and constant variance. Covariates included in the rating curve were selected because they could all be derived from daily discharge time series; they were continuous and were assumed to exert a linear effect in the log-log space. The covariates were centered on their mean value to improve numerical stability.

[15] The rating curves were fitted to the regression data sets I (composed of $c_i - x_i$ observation vectors). Given the distribution of NO3-N concentration in the log-log space (Figure 2a), NO3-N rating curve was stratified using 1 m^3/s as threshold to separate low from medium to high flow conditions. This reduced the size of the regression data set to 39 data entries for low flow conditions, and 71 data entries for medium-high flow. equation (1) was fitted (and applied) independently for low and medium-high flow. However, covariate x_2 (quadratic term of daily discharge, Table 2) was excluded because it was highly correlated to x_1 (daily discharge), and its exclusion improved rating curve fitting. Conversely, RP and TP rating curves were defined as in equation (1) with no need of further optimization. For RP and TP, some heteroscedasticity of the residuals was detected, e.g., in the normal quantile-quantile plots of residuals (Figure 2b; diagnostic plots of least-square regression fitting and residuals plot of each constituents are shown in the supporting information). Therefore, in these cases least-square derived confidence and prediction intervals would not be correct.

2.3. Prediction Intervals

[16] The rating curves were applied to a prediction data set J that was derived from daily discharge time series of Scotchtown station from April 2008 to April 2010. Daily loads were predicted as:

$$L_j = wQ_j \exp(\hat{c}_j + e) \quad (2)$$

where L_j is the constituent load (kg/d) of day j of the prediction data set J, w is a unit conversion factor, Q_j is the daily discharge (m^3/s), $\hat{c}_j$ is the constituent concentration (mg/L) predicted using equation (1); and e is an error term based on the estimated variance of ε in equation (1). Given that we estimated prediction loads and not load means, no correction factor CF term was applied in the back-transformation.

2.3.1. Bootstrap Prediction Intervals

[17] To estimate prediction intervals of daily and monthly loads, random-X resampling bootstrap was applied to the regression data set I, i.e., the $c_i - x_i$ observation vectors were resampled with replacement. Random-X

Table 2. Description and Physical Interpretation of Covariates Used in the Study to Estimate Constituent Concentrations[a]

	Mathematical Expression	Explanation	Physical Interpretation
$x_{1,i}$	$\ln Q_i$	ln = natural logarithm; Q_i = discharge of day i, in ML/d	Discharge
$x_{2,i}$	$(\ln Q_i)^2$	Quadratic term of x_1	Hysteresis
$x_{3,i}$	$\sin(2\pi M_i/12)$	M_i = month of day i	Seasonality
$x_{4,i}$	$\cos(2\pi M_i/12)$		
$x_{5,i}$	$\frac{\sum_{z=1}^{i} d^{i+1-z} Q_z}{\sum_{z=1}^{i} d^{i+1-z}}$	A discount factor of daily discharge Q up to day i; the parameter d regulates the influence of past discharge on day i constituent load. It was set to 0.95 after *Wang et al*. [2011], meaning that the past fortnight discharge most influences constituent concentration of day i	Constituent exhaustion
$x_{6,i}$	T_i	T_i = year of day i	Long-term trend
$x_{7,i}$	T_i^2	Quadratic term of x_6	

[a]Covariates $x_1 - x_4$ and $x_6 - x_7$ were proposed by *Cohn et al*. [1992]; covariate x_5 was proposed by *Wang et al*. [2011].

resampling was preferred since it was less affected by violation of model assumptions than fixed-X resampling, i.e., resampling of residuals [*Efron and Tibshirani*, 1998; *Petersen-Øverleir*, 2004; *Chernick*, 2008]. The regression data set I was resampled with replacement B times (set to 2000) to produce an I* regression data set (where the asterisk indicates a bootstrap replicate) of the same size of I. The rating curve (equation (1)) was fitted to each I* and daily loads were predicted with equation (2) for the prediction data set J and summed at monthly periods. The error term e in equation (2) was set as a random resample of the regression residuals ε^* of the I* regression data set. Finally, 95% bootstrap prediction intervals (95CI) for daily and monthly loads were predicted with the bias-corrected and accelerated (BCa) percentile method [*Efron and Tibshirani*, 1998], using the 2.5 and 97.5 percentile of the B bootstrap realizations. Although slightly less efficient to compute, the BCa method is more reliable than the simple percentile method when bootstrap realization distributions are not Gaussian [*Efron and Tibshirani*, 1998].

2.3.2. Bayesian Credible Intervals

[18] In this study, noninformative priors for Bayesian inference were applied to all the parameters. In equation (1), prior distributions for all rating curve parameters were set as uniform in the interval ($\sim$U [−10; 10]) after testing that the range was sufficiently large for application to all coefficients and constituents. (A sensitivity analysis on the influence of prior settings of coefficients is provided in the supporting information). The distribution of the concentration ln(c) was set as normal with mean equal to the mean logarithm concentration of the regression data set and large variance (N[$\overline{c_i}$, σ^2]). The error term e in equation (2) was set as N[0, σ^2]. Finally, the precision parameter τ ($1/\sigma^2$) was set as a gamma distribution Ga [0.001,0.001].

[19] The MCMC runs were set with a burn-in period of 10,000 runs, whose results were discarded after visually checking that burn-in length was sufficient. To reduce correlation in the posterior distributions only one result in every 50 runs of the following 1,00,000 runs was retained (thinning). The 2.5 and 97.5 percentile of the posterior distribution were used to identify the 95% credible interval ($95C_RI$).

2.4. Comparison of Predictions With Benchmark Loads

[20] Daily and monthly predicted loads of data set J were compared to the independent benchmark load data set E via visual inspection of scatter plots as well as selected summary statistics. These included the adjusted R^2 (adjR^2) as a measure of variance explained by the model adjusted for the number of explanatory variables in the model, the median prediction error (bias b) as a measure of predicted load accuracy, the relative Root-Mean-Square Error (RRMSE) as a measure of accuracy and precision of predicted loads, and *Lin*'s [1989, 2000] concordance coefficient r_c as a measure of agreement.

[21] Prediction intervals were compared for precision and observed accuracy [*Jin et al*., 2010; *Engeland et al*., 2010; *Li et al*., 2010]. The precision was measured with the average relative interval length (*ARIL*, [*Jin et al*., 2010]):

$$\mathrm{ARIL} = \frac{1}{M}\sum_{m=1}^{M}\frac{Limit_{up,m} - Limit_{low,m}}{O_m} \quad (3)$$

where for the M benchmark loads, $Limit_{up,m}$ indicated the upper boundary of the 95CI or $95C_RI$ for mth benchmark load; $Limit_{low,m}$ indicated its lower boundary; and O_m indicated the benchmark load. A low *ARIL* indicates precise intervals. The observed accuracy of the prediction intervals was assessed with the bracketing frequency, i.e., the fraction P_{95} (%) of benchmark loads that were bracketed by the 95% prediction intervals [*Li et al*., 2010]. A high P_{95} indicates high accuracy. It is important to note that the bracketing frequency P_{95} cannot be expected to be as high as the theoretical coverage (95%), because the latter is the coverage that would occur in the long term, when multiple independent samples were to be collected. In our practical case study, the bracketing frequency P_{95} refers instead only to a single independent sample.

[22] All statistical analyses were conducted using R [*R Core Team*, 2012]; Bayesian analysis was done through OpenBugs [*Lunn et al*., 2009], controlled from R using the BRugs package [*Ligges et al*., 2012].

3. Results

3.1. Daily Loads

[23] Table 3 summarizes the agreement between daily load predictions obtained with the rating curve fitted by least-square regression (in this case, *Duan*'s [1983] smear correction factor CF was applied for back-transformation), bootstrap and Bayesian median loads with benchmark loads. Prediction intervals and benchmark loads for the prediction period are shown in Figure 3.

[24] Predicted NO3-N daily loads showed very good agreement with benchmark loads (Table 3), but failed to predict the very low loads (<5 kg/day) observed in February 2009, and the relatively low loads in October 2008 and October 2009 (Figures 3 and 4). The lack of predictive

Table 3. Agreement of Daily Load (kg/day) Predictions by Ordinary Least-Square, Bootstrap, and Bayesian Fitting of Rating Curve With Benchmark Loads[a]

	Adj R^2	b	RRMSE	r_c	*ARIL*	P_{95}
NO3-N (n = 553)						
Least-square[b]	0.89	9	0.66	0.94	NA	NA
Bootstrap	0.89	9	0.67	0.94	1.64	0.52
Bayesian	0.89	10	0.66	0.94	1.69	0.72
RP (n = 652)						
Least-square[b]	0.41	−27	1.68	0.47	NA	NA
Bootstrap	0.41	−28	1.68	0.43	13.59	0.61
Bayesian	0.41	−28	1.68	0.43	10.97	0.87
TP (n = 566)						
Least-square[b]	0.75	−29	1.15	0.82	NA	NA
Bootstrap	0.76	−31	1.14	0.83	9.87	0.63
Bayesian	0.76	−33	1.14	0.82	2.94	0.84

[a]AdjR^2 = adjusted R^2 of benchmark loads versus predictions; b = median bias; RRMSE = relative root mean square error; r_c = Lin's concordance coefficient; *ARIL* = average relative interval length of prediction intervals; P_{95} = fraction of benchmark loads bracketed by the prediction intervals; n = sample size.

[b]Log-linear least-square regression (equations (1) and (2)), for which confidence intervals cannot be inferred due to violations in regression assumptions.

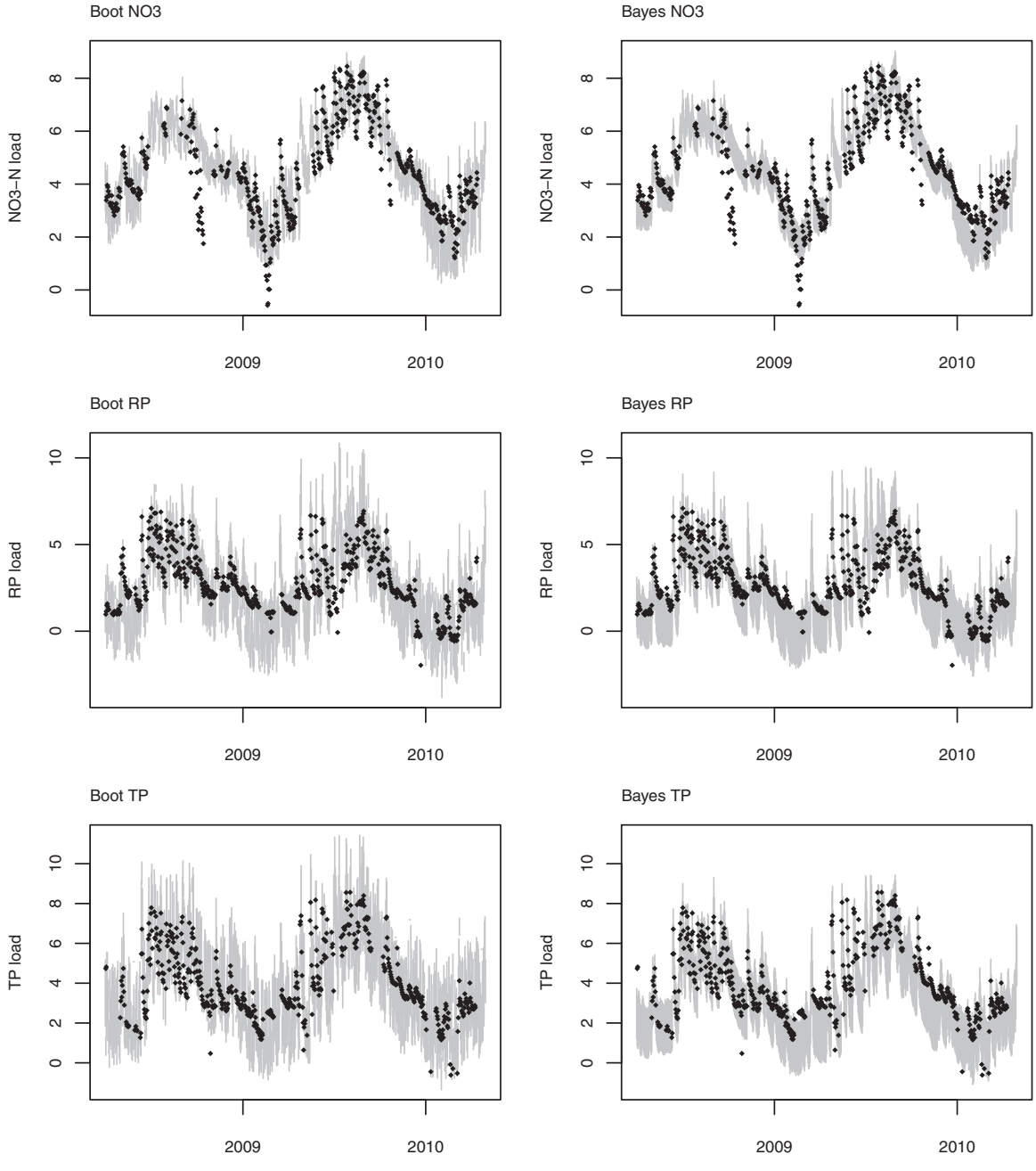

Figure 3. Comparison of daily benchmark loads (black points) with (left) bootstrap prediction intervals and (right) Bayesian credible intervals for three constituents (NO3-N, RP, and TP) at Duck River, Tasmania, in April 2008–2010. Loads are expressed in the log-scale (ln kg/day).

capability at low flow in February 2009 is likely the result of the limited range and sample size of NO3-N concentrations in the regression data set compared to the benchmark data set (Table 1). The relatively low loads (<100 kg/day) observed at the end of the wet season (October 2008 and 2009) are a combination of medium discharge but low NO3-N concentration, which are interpreted as a response to limit in supply of NO3-N sources [*Bende-Michl et al*., 2013]. This is a seasonal exhaustion effect that the rating curve failed to characterize properly, mainly because the very low concentrations of NO3-N measured in these critical periods by the high frequency monitoring data set were absent from the regression data set. Thus, the seasonal covariates (x_3 and x_4) could not capture this process appropriately. On the other hand, covariate x_5 was set to consider only short term (i.e., the previous 15 days) exhaustion. A second, longer term exhaustion set for a longer lag period (by choosing different values for d in Table 2) [e.g., *Kuhnert et al*., 2012] could have been added to the rating curve to characterize this effect, possibly in substitution of the seasonal covariates (x_3 and x_4). However, the exhaustion covariate (x_5) already showed considerable correlation with discharge and the second term of discharge (covariates x_1 and x_2), thus the addition of another covariate that

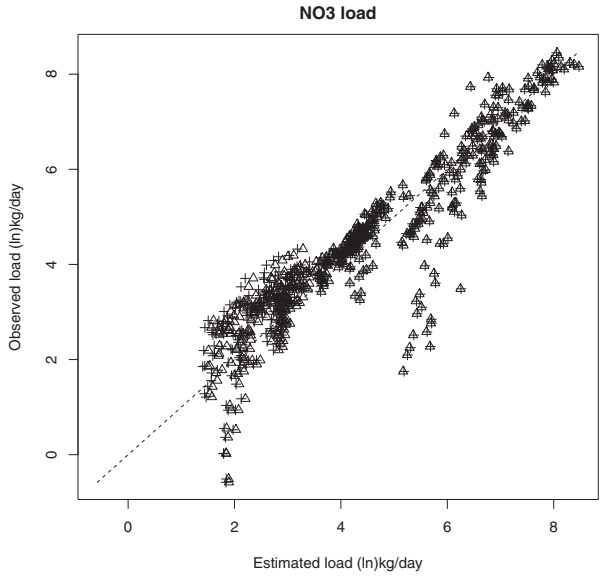

Figure 4. Benchmark NO3-N daily loads (log-scale, ln kg/day) against predicted daily loads by bootstrap (Δ) or Bayesian inference (+).

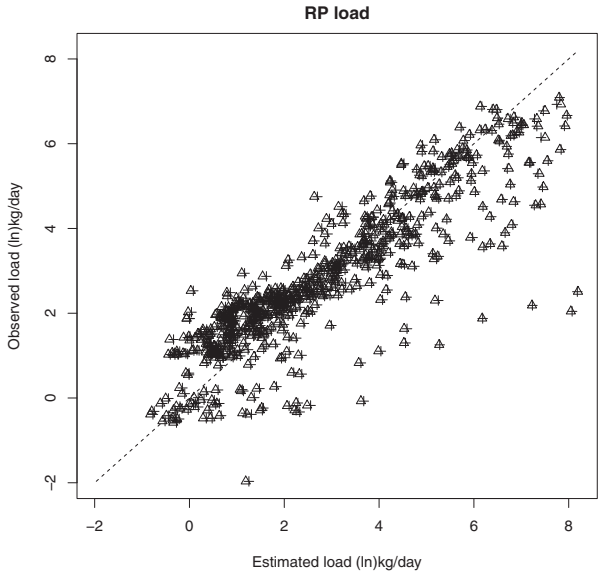

Figure 5. Benchmark RP daily loads (log-scale, ln kg/day) against predicted daily loads by bootstrap (Δ) or Bayesian inference (+).

depended on discharge could introduce numerical instability. In any case, the lack of low concentration observations in the critical period in the regression data set precluded the possibility of capturing this process. As a consequence, both bootstrap 95CI and Bayesian $95C_RI$ failed to include these low loads (Figure 3). Bootstrap prediction intervals were narrower, but of lower bracketing frequency than Bayesian credible intervals (P_{95}, Table 3).

[25] The rating curve yielded poor agreement of RP predictive loads with benchmark loads. It tended to overpredict high loads and under-predict low loads (Table 3 and Figure 5). The poor performance of the rating curve can be mostly ascribed to limitations in the regression data set, which had fewer data entries at the upper end of concentration range compared to the high frequency monitoring data set (Table 1), also a high scatter can be observed at low flow (Figure 2b). Differences in RP measurement protocols in the routine and high frequency monitoring sampling programs, e.g., in sampling techniques, filtering size applied, and laboratory analysis, could be another source of error, albeit this was considered marginal [*Bende-Michl et al*., 2013]. More likely, the regression data set may have been insufficient to define the rating curve properly and the rating curve formulation failed to account for all the processes affecting RP load in the Duck River. The high frequency data set showed that RP was subject to rapid "flushing" despite relative low flows at the beginning of the wet season [*Bende-Michl et al*., 2013]; this process was not included in the rating curve. Improvements in model formulation could therefore potentially be achieved by using alternative covariates to account for flushing using a generalized linear mixed model [*Wang et al*., 2011; *Kuhnert et al*., 2012]. As a result of the higher uncertainty in the rating curve fitting, prediction intervals were large; bootstrap prediction intervals were of similar range, yet with lower bracketing frequency, than the Bayesian credible intervals (Figure 3 and Table 3).

[26] Predictions of TP loads were in good agreement with benchmark loads (Table 3 and Figures 3 and 6), but with higher uncertainty than for NO3-N. Prediction intervals were large: bootstrap prediction intervals were more than twice the Bayesian credible intervals (Table 3, *ARIL*), yet again resulted in lower bracketing frequency than the Bayesian credible intervals. In all three cases, the Bayesian credible interval bracketing frequency was about 20% higher than bootstrap prediction intervals.

3.2. Monthly Loads

[27] Aggregation of daily loads to monthly periods may result in better load predictions if errors in daily loads are

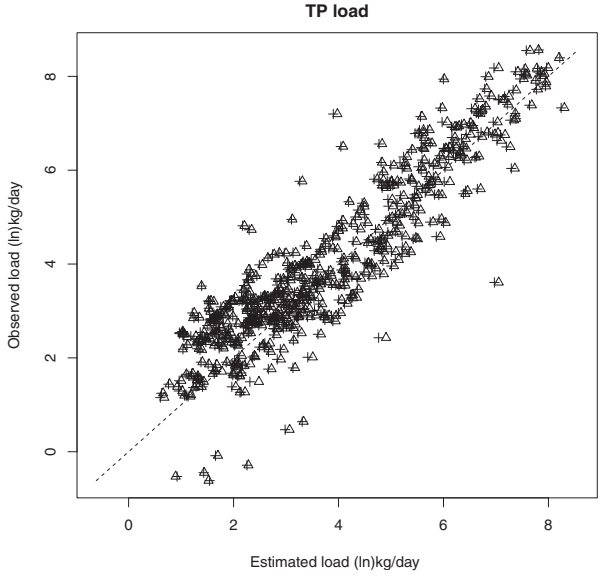

Figure 6. Benchmark TP daily loads (log-scale, ln kg/day) against predicted daily loads by bootstrap (Δ) or Bayesian inference (+).

compensated for, i.e., if aggregation smoothes out the noise rather than the signal [*Horowitz*, 2003; *Crowder et al.*, 2007]. In Figure 7 monthly prediction intervals are compared to benchmark loads. Generally, the monthly load predictions (in terms of summary statistics) as well as prediction intervals (in terms of precision and observed accuracy) improved for monthly aggregated data (Table 4).

[28] It should be noted that comparison of predictions with benchmark load is weaker at the monthly scale than at the daily scale. First, the number of evaluation points dropped considerably going from daily to monthly. Second, and possibly more importantly, benchmark loads were corrected for time series gaps in proportion to discharge of missing load days. Monthly benchmark loads therefore cannot be considered strictly "true" loads, but tended to better correspond to predicted loads where the constituent relationship with discharge was stronger.

[29] Nevertheless, these results showed that assessment of prediction intervals could be easily extended to aggregated periods. Bayesian credible intervals were larger than bootstrap prediction intervals in all three cases. The bracketing frequency of Bayesian credible intervals increased to

Table 4. Agreement of Monthly Load (kg/month) Predictions by Ordinary Least-Square, Bootstrap, and Bayesian Fitting of Rating Curve With Benchmark Loads[a]

	AdjR^2	*b*	RRMSE	r_c	*ARIL*	P_{95}
NO3-N (n = 15)						
Least-square[b]	0.99	−49	0.16	0.99	NA	NA
Bootstrap	0.99	−77	0.16	0.99	0.64	0.53
Bayesian	0.99	−69	0.16	0.99	1.16	0.80
RP (n = 22)						
Least-square[b]	0.34	−641	0.99	0.41	NA	NA
Bootstrap	0.34	−670	1.00	0.36	7.15	0.63
Bayesian	0.34	−661	1.00	0.37	8.03	0.95
TP (n = 17)						
Least-square[b]	0.90	1032	0.51	0.84	NA	NA
Bootstrap	0.91	459	0.49	0.85	1.85	0.53
Bayesian	0.92	864	0.47	0.84	2.39	1.00

[a]AdjR^2 = adjusted R^2 of benchmark loads versus predictions; *b* = median bias; RRMSE = relative root mean square error; r_c = Lin's concordance coefficient; *ARIL* = average relative interval length of prediction intervals; P_{95} = fraction of benchmark loads bracketed by the prediction intervals; *n* = sample size.

[b]Log-linear least-square regression (equations (1) and (2)), for which confidence intervals cannot be inferred due to violations in regression assumptions.

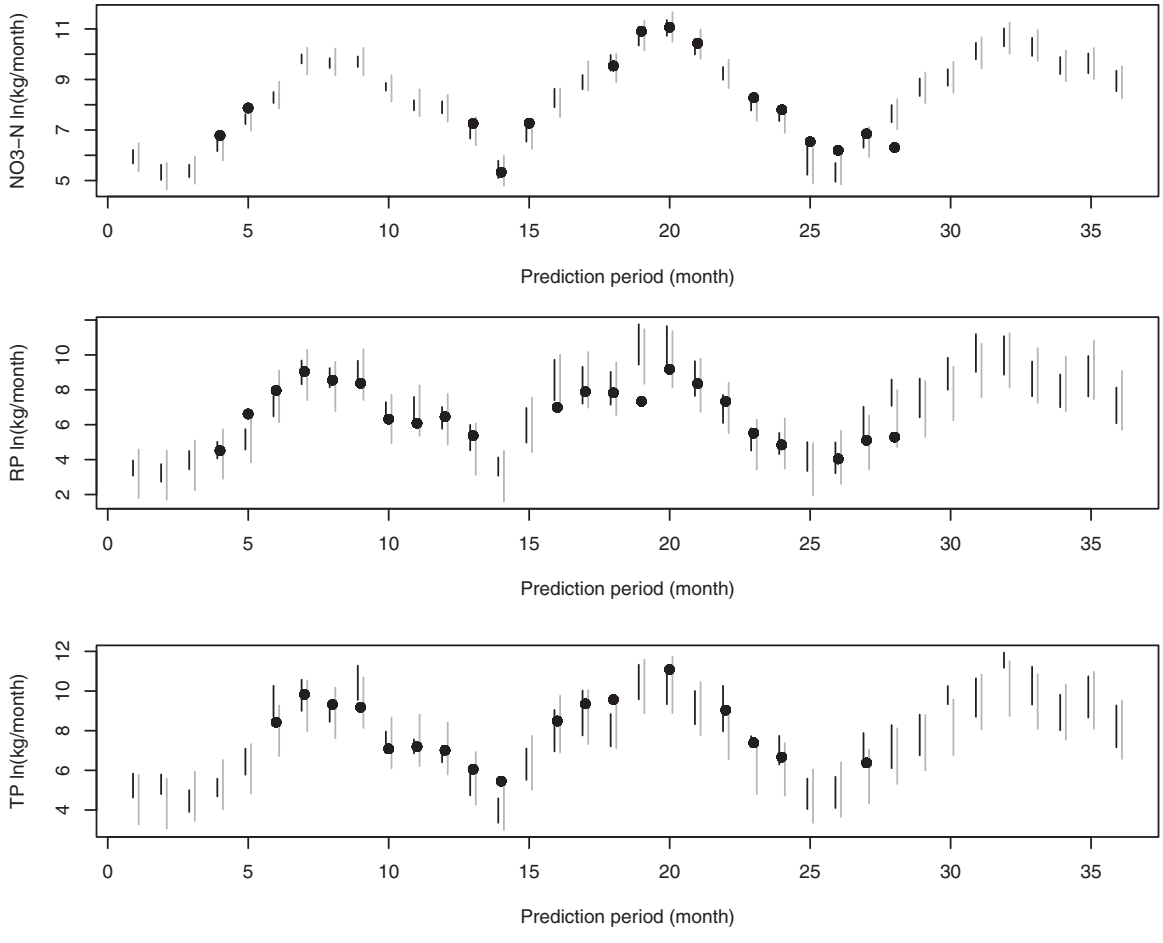

Figure 7. Monthly load prediction intervals for the period April 2008 to March 2010: bootstrap confidence intervals (black lines), Bayesian credible intervals (gray line), and benchmark loads (black points) are shown in logarithmic scale. (top) NO3-N; (middle) RP; and (bottom) TP.

80% or above, and was higher than bootstrap prediction intervals. Conversely, the bracketing frequency of bootstrap prediction intervals did not improve at monthly time steps.

3.3. Testing Method Robustness

[30] A more thorough testing of the inference methods requires an evaluation of their robustness, for example in the presence of an outlier in the regression data set I [*Abrahart et al*., 2011]. This is important, as laboratory or transcription errors may frequently occur, while their rectification may be impaired, for example if the databases are published online and sources are difficult to contact [*Abrahart et al*., 2011]. Outliers in constituent data sets are difficult to detect because monitored concentrations are usually not normally distributed, and removal of outliers is always a controversial practice. The RP data set in the Duck River offers a good example of a data set that was ill-defined at high concentrations.

[31] In order to test the robustness of the three methods in the presence of outliers, the original TP regression data set was artificially altered by changing the data entry of day 28/7/1999 from 0.54 mg/L (at 5.45 m^3/s during the recession curve of a runoff event) to 5.4 mg/L. The artificial outlier is admittedly an extreme case: it has high leverage (diagnostic plots and residual plots are in the supporting information), it is higher than the maximum TP concentrations recorded in the regression and evaluation data sets (Table 1), and it is high for similar environments [*DPIWE*, 2003; *Holz*, 2010]). However, had this entry occurred in the original data set, it would have been difficult to identify it as an outlier statistically: a *Shapiro and Wilks* [1965] test for normality of the TP concentration (with or without the outlier) would reject the hypothesis of log-normal distribution at probability level $p = 0.001$, therefore, outlier tests such as *Grubbs*'s [1950] test could not be performed.

[32] Notwithstanding this extreme case, the introduction of the artificial outlier did not impact on load predictions substantially. As an example, Figure 8 shows the distribution of two covariate coefficients, i.e., the coefficient β_1 for x_1 (discharge) and the coefficient β_5 for x_5 (exhaustion), under the original case (black line) and in the artificially

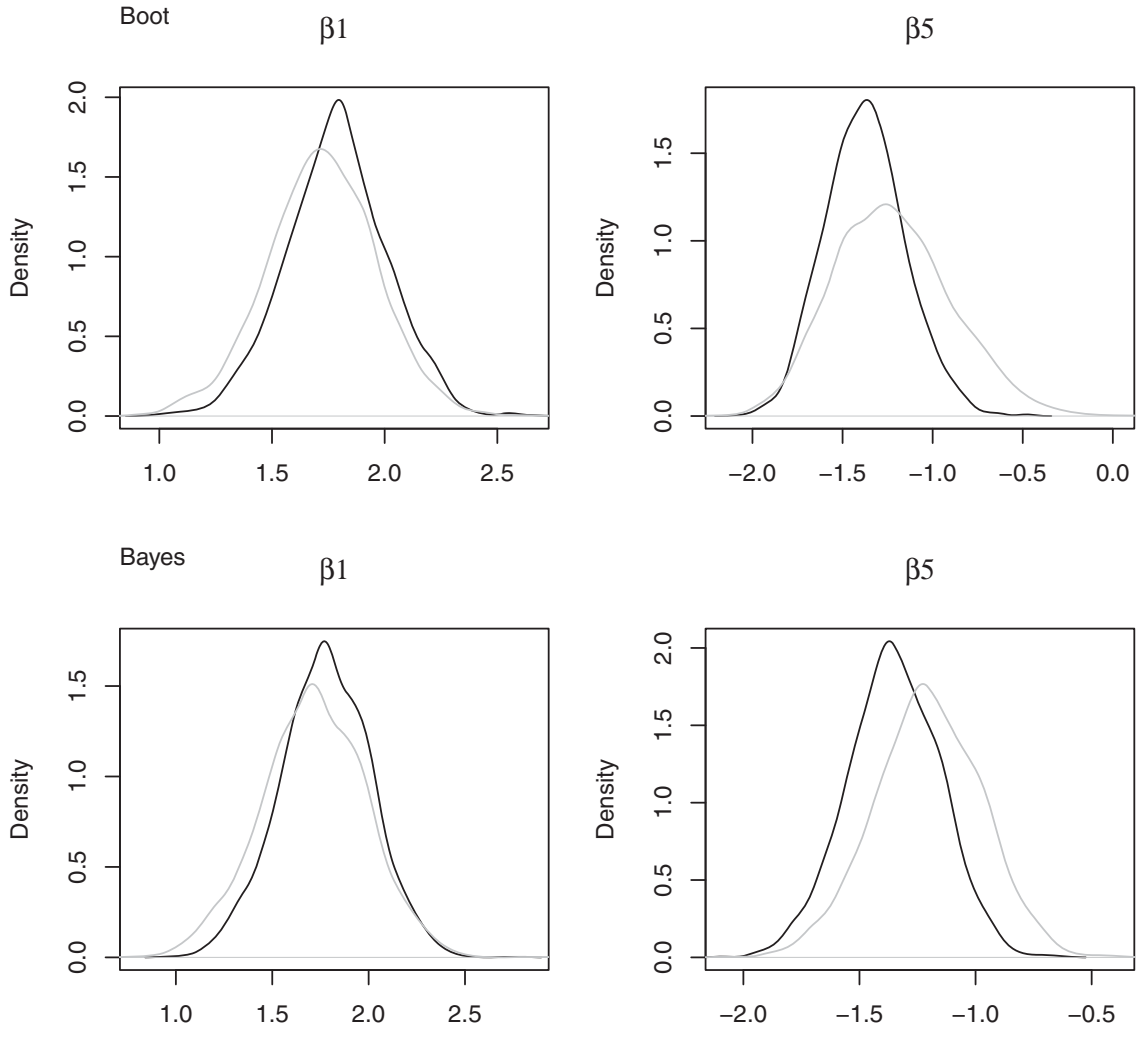

Figure 8. Distributions of β_1 and β_5 parameters for the TP original regression data set (black lines) and in the presence of an artificially inserted outlier (gray lines); top = bootstrap; and bottom = Bayesian posteriori distributions (Bayes).

altered data set (gray line) (the distribution of all coefficients is shown in the supporting information). The parameter distribution changed slightly, showing an increase in uncertainty, particularly for covariate x_5, but not substantially.

[33] Predictions were not much affected by these changes; contrary to expectations, the introduction of the outlier improved predictions of benchmark loads (Table 5). The introduction of an artificial outlier resulted in much larger bootstrap prediction intervals compared to the original data set (Table 5), yet the bracketing frequency of daily loads did not change. Conversely, the introduction of the outlier resulted in Bayesian credible intervals that were slightly larger but of similar bracketing frequency than for the original data set.

4. Discussion

[34] The choice of adopting bootstrap or Bayesian methods to assess prediction intervals may reside largely on philosophical and conceptual grounds, with bootstrap being a frequentist, nonparametric approach, as opposed to the Bayesian paradigm, which makes inference on the basis of prior knowledge and requires definition of prior parameter distributions (albeit these can be noninformative). However, both frequentist and Bayesian methods are used in environmental sciences. Confidence intervals and credible intervals have been compared in environmental studies, and can be used in conjunction to assess uncertainty of very complex systems [*Lu et al*., 2012]. This paper does not dwell in the theoretical and conceptual backgrounds and merits of both methods; *Lu et al*. [2012] offers a good theoretical treatise of the merits of regression confidence intervals (albeit not bootstrap) and Bayesian credible intervals for some models of relevance for environmental systems.

[35] Rather, we limit ourselves to consider the practical outcomes of prediction intervals by both methods in an applied case study recognizing that our results are limited to a single independent "benchmark" data set. The practitioner that is called to assess prediction intervals for rating curve loads would be interested in conveying intervals that are as accurate and precise as possible. Prediction intervals by least-square regression cannot be usually formulated for rating curve methods, given the violation of assumption in homoscedasticity of residuals. The addition of Monte-Carlo methods (either by bootstrap or Bayesian inference) allowed estimating prediction intervals at daily and monthly periods, but did not improve load accuracy compared to simple log-linear least-square regression with a smear CF (Tables (3–5)).

[36] Bootstrap prediction intervals were narrower than Bayesian credible intervals in the case of NO3-N but larger in the case of RP and TP. The bracketing frequency of bootstrap prediction intervals was about 40–65% (Tables 3 and 4). Bracketing frequency of benchmark loads by Bayesian credible intervals was consistently higher than bootstrap prediction intervals. Although credible intervals are belief intervals and their interpretation does not claim an exact probability of estimation [e.g., *Lu et al*., 2012], it is nonetheless reassuring for the practitioner that posterior credible intervals bracketed 74% or more of benchmark loads.

[37] The large differences in bootstrap intervals and Bayesian credible intervals were unexpected, in theory the two methods should have converged to similar results. A major cause for differences in the intervals lays in the definition of the error term e in equation (2), which was set as a resample of residuals ε with repetition in bootstrap while it was normally distributed with 0 mean and variance equal to variance of residuals ε in the Bayesian analysis. The relationships between sample size of the regression data set, the distribution of regression residuals ε, and the precision and accuracy of prediction intervals was explored using a simplified rating curve for the case of TP (see supporting information). ııııThe prediction intervals were sensitive to sample size and distribution of residuals ε; particularly in the case of bootstrap intervals. At small sample sizes, random resampling of residuals meant that a single large residual could be resample frequently, resulting in very large prediction intervals. Conversely, in the Bayesian case, one large residual caused large variance of residuals and error term; however, at small sample sizes prior settings exerted more control on the posterior distributions than the data itself, so the impact of sample size was reduced compared to the bootstrap method. As the sample size increased, the error term variance decreased, and the precision of prediction intervals improved for both methods, i.e., the *ARIL* decreased. Also, differences between the methods became smaller and the prediction intervals by the two methods tended to converge. The central theorem would suggest that with sample sizes even larger than those available in the Duck River case study, the distribution of residuals ε in the two methods would converge, and differences in prediction intervals would disappear. This could not be tested with our data sets; future research using for example synthetic data sets could be used to explore this further. Bracketing frequency of prediction intervals was sensitive to the error term e too; with smaller error and more precise prediction intervals, the bracketing frequency reduced. Thus, precision and accuracy of prediction interval reflected the distribution of residuals ε, hence limits in model

Table 5. Agreement of TP Daily (kg/day) and Monthly (kg/month) Load Predictions by Ordinary Least-Square, Bootstrap, and Bayesian Fitting of Rating Curve With Benchmark Loads in the Presence of an Artificially Inserted Outlier[a]

	AdjR^2	b	RRMSE	r_c	*ARIL*	P_{95}
TP (n = 566)						
Least-square[b]	0.78	−25	1.07	0.88	NA	NA
Bootstrap	0.79	−32	1.07	0.89	31.49	0.65
Bayesian	0.79	−32	1.07	0.87	2.80	0.93
TP (n = 17)						
Least-square[b]	0.93	697	0.43	0.96	NA	NA
Bootstrap	0.95	412	0.37	0.97	5.34	0.59
Bayesian	0.94	522	0.39	0.93	3.07	1.00

[a]AdjR^2 = adjusted R^2 of benchmark loads versus predictions; b = median bias; RRMSE = relative root mean square error; r_c = Lin's concordance coefficient; *ARIL* = average relative interval length of prediction intervals; P_{95} = fraction of benchmark loads bracketed by the prediction intervals; n = sample size.

[b]Log-linear least-square regression (equations (1) and (2)), for which confidence intervals cannot be inferred due to violations in regression assumptions.

formulation and regression data sets. This experiment suggested that at least in this case study, the definition of the error term e as given in the Bayesian method was more appropriate than the one set in the bootstrap method. Nevertheless, we considered the error settings coherent with each method, and did not further tested alternative technical choices.

[38] Improvements of load predictions could be achieved with better model formulation. The appropriateness of the rating curve application depends on the constituent characteristics as well as on the active hydrological processes. For example, several studies found that the application of rating curve was more justified for constituents that have an important particulate components dominated by surface processes than for dissolved chemicals, which may be dominated by subsurface flow dynamics [*Coats et al*., 2002; *Guo et al*., 2002; *Johnes*, 2007; *Ide et al*., 2011; *Stenback et al*., 2011]. The Duck River results showed that the rating curve proposed in this study was not effective in capturing hydrological processes that were active at the beginning of the wet season, like "flushing" in the case of RP, or at its end, like the exhaustion of constituent supply observed in the case of NO3-N. Additional covariates [e.g., such as those proposed by *Wang et al*., 2011; *Kuhnert et al*., 2012] should be considered to account for this seasonal effects, possibly in substitution of the seasonality covariates (x_3 and x_4). Generalized linear mixed model formulations could potentially help including discrete covariates and accounting for autocorrelation in the data set [*Aulenbach and Hooper*, 2006; *Verma et al*., 2012; *Kuhnert et al*., 2012]. Such formulations could be accommodated within the bootstrap and Bayesian inference methods tested in this study.

[39] Improvements in the load predictions could also be achieved by enlarging the regression data sets. Particularly, the full range of flow conditions should be covered to include seasonal changes in nutrient supply and transport, and thus more observation close to the maxima and minima [e.g., *Saad et al*., 2011; *Bende-Michl et al*., 2013]. For example, in the Duck River, monitoring the first events of the season would provide useful information on such extremes. Large data set could also support flow stratification and better separate flow conditions characterized by different hydrological processes. This is important where low flow conditions are likely to become more frequent, for example during prolonged droughts that could lead to building up of constituent supply in the landscape, available for flushing at the onset of a wetter period. Targeted monitoring programs can help improving the understanding of hydrological processes [*Bende-Michl et al*., 2013], and guide the formulation of appropriate rating curve methods.

[40] Prediction intervals as applied in this study only reflected uncertainty due to the estimation method given the data sets. Other sources of uncertainty were not accounted for. Important sources of uncertainty that should be considered within a larger uncertainty framework are uncertainties in discharge [*Wang et al*., *2011*; *Hamilton and Moore*, 2012; *Tomkins*, 2012], and in water quality concentration data [*Harmel et al*., 2009; *McMillan et al*., 2012]. Uncertainty in discharge estimated with stage-discharge rating curves has been estimated to be around 10–20% at medium flow, but can be considerably larger at low flow and at high flow, and large variability exists depending on site conditions, discharge estimation method, and aggregation period [*Harmel et al*., 2009; *Hamilton and Moore*, 2012; *Tomkins*, 2012]. Uncertainty in concentration measurements due sampling and analytical procedures may be even larger [*Harmel et al*., 2009; *McMillan et al*., 2012]. The impact of such uncertainties on loads estimated with the constituent rating curve could be evaluated in a larger framework, e.g., using RMSE propagation methods [*Harmel et al*., 2009], or a Bayesian framework [e.g., *Kulasova et al*., 2012].

5. Conclusions

[41] Provision of 95% prediction intervals of constituent loads is important information that should be conveyed together with load estimates [*Tan et al*., 2005; *Krueger et al*., 2009; *Wang et al*., 2011]. In this study, bootstrap and Bayesian methods were used to overcome the obstacles associated with least-square regression, and to generate rating curve loads and prediction intervals for daily and longer periods.

[42] Application of an improved eight-parameter rating curve that included terms for discharge, hysteresis, seasonality, exhaustion, and long-term trend to routine monitoring data sets of the Duck River yielded good load predictions for NO3-N (stratified for low and medium-high flow) and TP when compared to an independent data set of loads obtained by high frequency monitoring. Poor predictions of some very low benchmark loads of NO3-N and RP loads showed that the rating curve proposed in this study was not effective in capturing hydrological processes that were active at the beginning of the wet season, like "flushing" in the case of RP, or at its end, like the exhaustion of constituent supply observed in the case of NO3-N. Such shortcomings could also be ascribed to limitations in the regression data set including size and range compared to the independent, high frequency monitoring sample used as benchmark, confirming that estimations could be improved with better sampling of constituent concentrations, although the sampling strategy may differ for constituents and environments [e.g., *Johnes*, 2007; *Ide et al*., 2011; *Verma et al*., 2012].

[43] The precision and accuracy of prediction intervals were sensitive to the distribution of residuals in the regression data set. The 95% bootstrap prediction intervals included about 40–65% of daily and monthly benchmark loads. Bayesian credible intervals consistently achieved a higher bracketing frequency of benchmark loads, about 74–85% of daily loads and 80% or more of monthly loads. From a practical perspective, thus, Bayesian inference can be recommended to formulate credible intervals to account for uncertainty in constituent load estimation by rating curve, particularly in small sample size conditions such as those tested in this case study.

[44] Prediction intervals as estimated in this study were sensitive to the distribution of the regression error, hence they reflected uncertainty in the regression data set and limitations in the rating curve formulation. They did not account for other sources of uncertainty, i.e., they were still conservative predictions of load uncertainty.

[45] **Acknowledgments.** D. Fox (Environmetrics Australia) and S. Norng (DPI Victoria) provided mentoring in the Bayesian analysis. Bayesian diagnostic plots were produced with an R script of M. Wand (UTS). The high frequency nutrient monitoring data study was funded by the

Commonwealth Environment Research Facilities (CERF) Programme, through the Australian Department of the Environment, Water, Heritage, and the Arts (DEWHA), and by CSIRO Water for a Healthy Country Flagship. Anna M. Roberts, David Nash, Murray Hannah (DEPI Victoria), Kirsten Verburg, and Quanxi Shao (CSIRO), as well as the Journal anonymous reviewers offered very useful comments to the manuscript.

References

Abrahart, R. J., N. J. Mount, N. A. Ghani, N. J. Clifford, and C. W. Dawson (2011), DAMP: A protocol for contextualising goodness-of-fit statistics in sediment discharge data-driven modelling, *J. Hydrol.*, *409*, 596–611.

Alameddine, I., S. S. Qian, and K. H. Reckhow (2011), A Bayesian changepoint-threshold model to examine the effect of TMDL implementation on the flow-nitrogen concentration relationship in the Neuse River basin, *Water Res.*, *45*, 51–62.

ANZECC (2000), Australian Water Quality Guidelines for Fresh and Marine Waters, Agric. and Resour. Manage. Counc. of Aust. and N. Z., Kingston, Australia.

Asselman, N. E. M. (2000), Fitting and interpretation of sediment rating curves, *J. Hydrol.*, *234*, 228–248.

Aulenbach, B. T., and R. P. Hooper (2006), The composite method: An improved method for stream-water solute load estimation, *Hydrol. Processes*, *20*, 3029–3047.

Bartley, R., W. J. Speirs, T. W. Ellis, and D. K. Waters (2012), A review of sediment and nutrient concentration data from Australia for use in catchment water quality models, *Mar. Pollut. Bull.*, *65*, 101–116.

Bende-Michl, U., and P. B. Hairsine (2010), A systematic approach to choosing an automated nutrient analyser for river monitoring, *J. Environ. Monit.*, *12*, 127–134, doi:10.1039/b910156j.

Bende-Michl, U., K. Verburg, and H. P. Cresswell (2013), High frequency nutrient monitoring to infer seasonal patterns in catchment source availability, mobilisation and delivery, *Environ. Monit. Assess.*, *185*(11), 9191–9219, doi:10.1007/s10661-013-3246-8.

Chernick, M. R. (2008), *Bootstrap Methods A Guide for Pratictioners and Researchers, Wiley Ser. Probab. Stat.*, 2nd ed., John Wiley, Hoboken, N. J.

Coats, R., F. Liu, and C. R. Goldman (2002), A Monte Carlo test of load calculation methods, Lake Tahoe Basin, California, Nevada, *J. Am. Water Resour. Assoc.*, *38*, 719–730.

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

Cox, N. J., J. Warburton, A. Armstrong, and V. J. Holliday (2008), Fitting concentration and load rating curves with generalized linear models, *Earth Surf. Processes Landforms*, *33*, 25–39.

Crowder, D. W., M. Demissie, and M. Markus (2007), The accuracy of sediment loads when log-transformation produces nonlinear sediment load-discharge relationships, *J. Hydrol.*, *336*, 250–268.

DPIWE (2003), State of the river report for the Duck river catchment, in *Tech. Rep. WAP 03/08*, Water Assess. and Plann. Branch, Hobart.

DPIWE (2011), Water information system of Tasmania, The Department of Primary Industries, parks, Water and Environment, Hobart, Australia. [Available at http://water.dpiw.tas.gov.au/wist/ui, last accessed 16 Nov. 2011.]

Duan, N. (1983), Smearing estimate: A nonparametric retransformation method, *J. Am. Stat. Assoc.*, *78*, 605–610.

Efron, B., and R. Tibshirani (1998), *An Introduction to the Bootstrap, Monogr. Stat. Appl. Probab.*, vol. 57, Chapman and Hall, New York.

Engeland, K., B. Renard, I. Steinsland, and S. Kolberg (2010), Evaluation of statistical models for forecast errors from the HBV model, *J. Hydrol.*, *384*, 142–155.

Gray, S., G. Hanrahan, I. McKelvie, A. Tappin, F. Tse, and P. Worsfold (2006), Flow analysis techniques for spatial and temporal measurement of nutrients in aquatic systems, *Environ. Chem.*, *3*(1), 3–18, doi:10.1071/EN05059.

Grubbs, F. E. (1950), Sample criteria for testing outlying observations, *Ann. Math. Stat.*, *21*, 27–58.

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

Hamilton, A. S., and R. D. Moore (2012), Quantifying uncertainty in streamflow records, *Can. Water Resour. J.*, *37*, 3–21.

Harmel, R. D., D. R. Smith, K. W. King, and R. M. Slade (2009), Estimating storm discharge and water quality data uncertainty: A software tool for monitoring and modeling applications, *Environ. Modell. Software*, *24*, 832–842.

Holz, G. K. (2010), Sources and processes of contaminant loss from an intensively grazed catchment inferred from patterns in discharge and concentration of thirteen analytes using high intensity sampling, *J. Hydrol.*, *383*, 194–208.

Horowitz, A. J., (2003), An evaluation of sediment rating curves for estimating suspended sediment concentrations for subsequent flux calculations, *Hydrol. Processes*, *17*, 3387–3409.

Ide, J., M. Chiwa, N. Higashi, R. Maruno, Y. Mori, and K. Otsuki (2011), Determining storm sampling requirements for improving precision of annual load estimates of nutrients from a small forested watershed, *Environ. Monit. Assess.*, *184*(8), 4747–4762, doi:10.1007/s10661-011-2299-9.

Jin, X., C.-Y. Xu, Q. Zhang, and V. P. Singh (2010), Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model, *J. Hydrol.*, *383*, 147–155.

Johnes, P. J. (2007), Uncertainties in annual riverine phosphorus load estimation: Impact of load estimation methodology, sampling frequency, baseflow index and catchment population density, *J. Hydrol.*, *332*, 241–258.

Krueger, T., J. N. Quinton, J. Freer, C. J. A. Macleaod, G. S. Bilotta, R. E. Brazier, P. Butler, and P. M. Haygarth (2009), Uncertainties in data and models to describe event dynamics of agricultural sediment and phosphorus transfer, *J. Environ. Qual.*, *38*, 1137–1148.

Kuhnert, P. M., B. L. Henderson, S. E. Lewis, Z. T. Bainbridge, S. N. Wilkinson, and J. E. Brodie (2012), Quantifying total suspended sediment export from the Burdekin River catchment using the loads regression estimator tool, *Water Resour. Res.*, *48*, W04533, doi:10.1029/2011WR011080.

Kulasova, A., P. J. Smith,K. J. Beven, S. D. Blazkova, and H. Hlavacek (2012), A method of computing uncertain nitrogen and phosphorus loads in a small stream from an agricultural catchment using continuous monitoring data, *J. Hydrol.*, *458–459*, 1–8.

Letcher, R. A., J. A. Jakeman, W. S. Merritt, L. J. McKee, B. D. Eyre, and B. Baginska, (1999), Review of techniques to estimate catchment exports, *EPA Tech. Rep. 99/73*, 139 pp., Environ. Protect. Auth. of N. S. W., Sydney.

Li, L., J. Xia, C.-Y. Xu, and V. P. Singh (2010), Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models., *J. Hydrol.*, *390*210–221.

Ligges, U., S. Sturtz, A. Gelman, G. Gorjanc, and C. Jackson (2012), *R interface to the OpenBUGS MCMC software*, v 0.7–5, Vienna, Austria. [Available at www.r-project.org, last accessed 26 Mar 2012.]

Lin, L. I.-K. (1989), A concordance correlation coefficient to evaluate reproducibility, *Biometrics*, *45*, 255–268.

Lin, L. I.-K. (2000), A note on the concordance correlation coefficient, *Biometrics*, *56*, 324–325.

Lu, D., M. Ye, and M. C. Hill (2012), Analysis of regression confidence intervals and Bayesian credible intervals for uncertainty quantification, *Water Resour. Res.*, *48*, W09521, doi:10.1029/2001WR011289.

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best (2009), The BUGS project: Evolution, critique, and future directions (with discussion), *Stat. Med.*, *28*, 3049–3082. [Available at http://www.openbugs.info/w/, last accessed 26.03.2012.]

McMillan, H., T. Krueger, and J. Freer (2012), Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality, *Hydrol. Processes*, *26*, 4078–4111.

Moyeed, R. A., and R. T. Clarke (2005), The use of Bayesian methods for fitting rating curves, with case studies, *Adv. Water Res.*, *28*, 807–818.

Petersen-Øverleir, A. (2004), Accounting for heteroscedasticity in rating curve estimates, *J. Hydrol.*, *292*, 173–181.

R Core Team (2012), R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. Version 2.14.2. [Available at http://www.r-project.org/foundation/, last accessed 26 Mar. 2012.]

Retain, T., and A. Petersen-Øverleir (2011), Dynamic rating curve assessment in unstable rivers using Ornstein-Uhlenbeck processes, *Water Resour. Res.*, *47*, W02524, doi:10.1029/2010WR009504.

Richley, R. (1978), *Land Systems of Tasmania: Region 3*, Dep. of Agric., Hobart.

Rustomji, P., and S. N. Wilkinson (2008), Applying bootstrap resampling to quantify uncertainty in fluvial suspended sediment loads estimated

using rating curves, *Water Resour. Res.*, *44*, W09434, doi:10.1029/2007WR006088.

Saad, D. A., G. E. Schwarz, D. M. Robertson and N. L. Booh (2011), A multi-agency nutrient dataset used to estimate loads, improving monitoring design, and calibrate regional nutrient SPARROW models, *J. Am. Water Resour. Assoc.*, *47*, 933–949.

Shapiro, S. S., and M. B. Wilks (1965), An analysis of variance test for normality (complete samples), *Biometrika*, *52*, 591–611.

Stenback, G. A., W. G. Crumpton, K. E. Schilling, and M. J. Helmers (2011), Rating curve estimation of nutrient loads in Iowa rivers, *J. Hydrol.*, *396*, 158–169.

Tan, K. S., D. R. Fox, and T. Etchells (2005), *GUMLEAF: Generator for Uncertainty Measures and Load Estimates Using Alternative Formulae: User Guide and Reference Manual*, p. 36, Aust. Cent. Environ., Univ. of Melbourne, Parkville, Australia.

Tomkins, K. M. (2012), Uncertainty in streamflow rating curves: Methods, controls, and consequences, *Hydrol. Processes*, doi:10.1002/hyp.9567

Van den Broeke, J., G. Langergraber, and A. Weingartner (2006) Online and in situ UV/vis spectroscopy for multi-parameter measurements: A brief review, *Spectrosc. Eur.*, *18*(4), 15–18.

Verburg, K., H. Cresswell, U. Bende-Michl, J. Gibson, and P. Hairsine (2012), Spatial diagnosis of catchment water quality: Using multiple lines of evidence, in *Landscape Logic: Integrating Science for Landscape Management*, edited by T. Lefroy et al., pp. 83–102, CSIRO Publ., Collingwood, Vic.

Verma, S., M. Momcilo, and R. A. Cooke (2012), Development of error correction techniques for nitrate-N load estimation methods, *J. Hydrol.*, *432–433*, 12–25.

Wang, Y.-G., P. Kunhert, and B. Henderson (2011), Load estimation with uncertainties from opportunistic sampling data: A semi-parametric approach, *J. Hydrol.*, *396*, 148–157.

Wellen, C., G. B. Arhonditsis, T. Labencki, and D. Boyd (2012), A Bayesian methodological framework for accommodating interannual variability of nutrient loading with the SPARROW model, *Water Resour. Res.*, *48*, W10505, doi:10.1029/2012WR011821.