The application of high temporal resolution data in river catchment modelling and management strategies

by Crockford, L., O'Riordain, S., Taylor, D., Melland, A.R., Shortle, G. and Jordan, P.

DOI: doi:10.1007/s10661-017-6174-1

Crockford, L., O'Riordain, S., Taylor, D., Melland, A.R., Shortle, G. and Jordan, P. 2017. The application of high temporal resolution data in river catchment modelling and management strategies. *Environmental Monitoring and Assessment*, 189:451

Environmental Monitoring and Assessment

The application of high temporal resolution data in river catchment modelling and management strategies

--Manuscript Draft--

Manuscript Number:	EMAS-D-17-00632R1
Full Title:	The application of high temporal resolution data in river catchment modelling and management strategies
Article Type:	Original Research
Corresponding Author:	Lucy Crockford, Ph.D Harper Adams University UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Harper Adams University
Corresponding Author's Secondary Institution:	
First Author:	Lucy Crockford, Ph.D
First Author Secondary Information:	
Order of Authors:	Lucy Crockford, Ph.D
	Sean O'Riordain, Ph.D
	David Taylor, Ph.D
	Alice R Melland, Ph.D
	Ger Shortle, B.Sc
	Philip Jordan, Ph.D
Order of Authors Secondary Information:	
Funding Information:	Teagasc (Walsh Fellowship) — Dr. Lucy Crockford
	Science Foundation Ireland (10/CE/1855) — Dr Sean O'Riordain

The application of high temporal resolution data in river catchment modelling and management strategies

Crockford, L. (ORCID: 0000-0001-8336-4149) [a,b,c*], O'Riordain, S.[d], Taylor, D.[e], Melland, A. R.[f], Shortle, G.[a], Jordan, P. [a, g].
[a] The Agricultural Catchments Programme, Teagasc, Johnstown Castle, Wexford, Ireland
[b] Geography, School of Natural Sciences, Trinity College Dublin, Ireland
[c] Crop and Environment Sciences, Harper Adams University, Shropshire, UK
[d] Statistics, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[e] Department of Geography, National University of Singapore, Singapore
[f] National Centre for Engineering in Agriculture, University of Southern Queensland, Australia
[g] School of Environmental Sciences, University of Ulster, Coleraine, Northern Ireland, United Kingdom

*Corresponding author
Crop and Environment Sciences
Harper Adams University
Edgmond
Shropshire
TF10 8NB
UK
+44 (0)1952 815476

lcrockford@harper-adams.ac.uk

Acknowledgements

The authors would like to thank the Teagasc Agricultural Catchments Programme (ACP – funded by the Department of Agriculture, Food and Marine, Ireland) scientists, technologists, technicians and advisors for support received throughout this project. The authors also acknowledge John Haslett of the School of Computer Science and Statistics, Trinity College Dublin, for model conceptualization and statistical support, and two anonymous referees for their very helpful comments. The lead author was supported by a Walsh Fellowship through Teagasc and the second author was funded, in part, by Science Foundation Ireland, grant 10/CE/1855 to Lero – the Irish Software Engineering Research Centre (www.lero.ie). Concepts of LAM testing with high resolution P data were developed with help from Dr. Rachel Cassidy, Agri-Food Biosciences Institute, Belfast. The research is based on a confidential dataset collected by a government funded research programme and, as a result, is unavailable for public access. However, summary statistics of all model outputs are provided in the online resource for transparency.

Abstract

Modelling changes in river water quality, and by extension developing river management strategies, has historically been reliant on empirical data collected at relatively low temporal resolutions. With access to data collected at higher temporal resolutions, this study investigated how these new dataset types could be employed to assess the precision and accuracy of two phosphorus (P) load apportionment models (LAMs) developed on lower resolution empirical data.

Predictions were made of point and diffuse sources of P across ten different sampling scenarios. Sampling resolution ranged from hourly to monthly through the use of 2000 newly created datasets from high frequency P and discharge data collected from a eutrophic river draining a 9.48 km^2 catchment. Outputs from the two LAMs were found to differ significantly in the P load apportionment (51.4% versus 4.6% from point sources) with reducing precision and increasing bias as sampling frequency decreased. Residual analysis identified a large deviation from observed data at high flows. This deviation affected the apportionment of P from diffuse sources in particular.

The study demonstrated the potential problems in developing empirical models such as LAMs based on temporally relatively poorly-resolved data (the level of resolution that is available for the majority of catchments). When these models are applied *ad hoc* and outside an expert modelling framework using extant datasets of lower resolution, interpretations of their outputs could potentially reduce the effectiveness of management decisions aimed at improving water quality.

Keywords

Agriculture; modelling; phosphorus; water quality; pollution; high frequency data

Introduction

Cultural eutrophication, due to high concentrations of phosphorus (P) and nitrogen (N), currently presents a major and widespread challenge for water managers (Williams and Kimball 2013; Fonseca et al. 2014; Binzer et al. 2016). River water quality, globally, is greatly influenced by anthropogenic pollution from point and diffuse sources (Sharpley et al. 2013), with the former having severe impacts during periods of low dilution in rivers (e.g. during spring and summer in temperate climates; Withers et al. 2012; Withers et al. 2014; Begum et al. 2016) when diffuse sources are relatively inactive. Diffuse sources of nutrients in a catchment may be activated during, for example, subsequent periods of heavy rain and resultant increased surface and sub-surface flows (Sharpley and Wang 2014; Begum et al. 2016).

Despite the implementation of point and diffuse source reduction strategies in many countries, water quality in some catchments has remained poor, with only 33% of those reviewed by Verdonschot et al. (2013) showing evidence of improved water quality. Agricultural intensification (Jansons et al. 2002; Moreno-Ostos et al. 2007) and/or significant sewage or industrial effluent (Jarvie et al. 2006) are factors largely identified as slowing recovery in rivers. This is noted, for example, in European Union (EU) countries that have implemented Water Framework Directive (WFD; OJEC 2000) policies, particularly where river managers are attempting to improve water quality that in the first cycle failed the WFD objectives to improve or protect good status (e.g. de Vries and de Boer 2010). Remediation strategies may be improved by determining the contribution of each nutrient source, as this allows a better targeting of mitigation measures (Verdonschot et al. 2013). Often, this source determination is provided by models and especially when there is a paucity of empirical data (Bowes et al. 2008; Yang and Wang 2010). However, technological advances and reductions in the cost of equipment have resulted in the availability of high temporal resolution datasets (Melland et al. 2012; Bieroza and Heathwaite 2015; Campbell et al. 2015; Perks et al. 2015; Rode et al. 2016). Such high temporal resolution datasets have been used to test the precision and accuracy of models developed using data collected at much lower frequencies, for example at daily or monthly intervals (see Cassidy and Jordan 2011 and Skeffington et al. 2015).

For P, empirical load apportionment models (LAMs; Bowes et al. 2008; Greene et al. 2011) have been used to allocate relative contributions from different P sources using stream chemistry and flow data only (Bowes et al. 2008; Bowes et al. 2009; Bowes et al. 2010; Chen et al. 2015). When used with extant data, these models provide a cost effective, labour efficient means of estimating the P load in rivers apportioned to either diffuse or point sources (Schoumans et al. 2009) making them attractive tools for catchment management and pollution risk assessment. However, some limitations have been identified with LAMs. These limitations comprise i) the requirement of P concentrations at high flows to adequately describe the diffuse signal during storm events; ii) the necessity of a short sampling time step; and iii) the assumption that once the model has been fitted to flows at a low sampling frequency, the LAM will adequately describe all flows available at a higher frequency resolution (Bowes et al. 2008).

High temporal resolution data could thus be used to quantify the potential impact of these limitations on catchment risk assessment. In addition, they could provide the basis for analysis of the variability in model outputs in an individual catchment, which is an important step in the modelling process (Robson 2014; Chaudhary and Hantush 2017), using a range of data collection scenarios similarly to previous studies (Cassidy and Jordan 2011; Skeffington et al. 2015).

In the current study, the effects of sampling frequency and timing on the outputs of two different LAMs, used as “off the shelf” risk assessment tools, were investigated. The two LAMs are described in Bowes et al. (2008) and Greene et al. (2011) and are hereafter referred to as, respectively, BM and GM. The precision and accuracy of the two LAMs are assessed within their applicability as tools in catchment management. Determinations of accuracy and precision were based on a comparison of model outputs between models, across sampling scenarios and with data collected at a high temporal resolution from an agricultural catchment in the Republic of Ireland that had experienced nutrient pollution problems in the recent past (Melland et al. 2012).

Study area

The study catchment (Figure 1, 9.48 km^2, 53° 49’ 15”N, 6° 24’ 16” W) is identified as Arable B in detail elsewhere (Jordan et al. 2012; Mellander et al. 2012; Mellander et al. 2014) and drains into the Dee and Glyde rivers and eventually to the Irish Sea off the eastern coast of Ireland. Ranging between 225 and 28 mASL, land use in the catchment comprises roughly equal proportions of arable land and grassland with a livestock density of 1.36 LU ha^{-1}. Moderately drained gleyic brown earth and groundwater gley soils overlay calcareous greywacke and mudstone bedrock, which is often highly fractured and may provide fast pathways for groundwater flow. Surface and near-surface flow, associated with acute, storm dependent P transfer from diffuse sources (e.g. Jordan et al. 2007), is considered a predominant contributor to river flow, while chronic P pollution from rural point sources is relatively important at base flow (Melland et al. 2012; Murphy et al. 2015). The rural population density for the catchment is 14 houses km^{-2} (Melland et al. 2012), with wastewater treated in septic tank systems, generating a potential human P load of 372 kg year^{-1}, or 39.2 kg year^{-1} km^{-2}, based on data provided in Jordan et al. (2012). The proportion of this P lost from septic tank systems and entering the stream network is, however, likely to vary due to the range of working and defective effluent treatment stages and temporary breakdowns that occur from time to time (e.g. Withers et al. 2014). Nevertheless, based on Carvalho et al. (2005), the minimum P load exported from fully working systems under these input conditions is estimated as 63 kg year^{-1}. There are no other urban or industrial point sources but farmyards, where waste management may be poor, pose an additional unquantified risk (Murphy et al. 2015).

Materials and Methods

High Temporal Resolution Data Collection

Total reactive phosphorus (TRP – unfiltered and undigested) concentrations are used in water quality assessments in Ireland (SI 272 2012) and so were used in this analysis. Concentrations of TRP were measured sub-hourly by a bankside P analyser (Phosphax-Sigmatax, HACH, Germany; operational range 0.010 – 5.000 mg L^{-1}) following Eisenreich et al. (1975). This equipment has been used extensively in catchment research projects throughout Ireland and the UK (e.g. Wade et al. 2012; Mellander et al. 2014; Outram et al. 2014; Campbell et al. 2015). Concentrations of TRP in each sample were determined on a molybdate-antimony blue complex (DIN EN 38405 D11 – updated to DIN EN ISO 6878) that was auto calibrated against a standard concentration (2 mg L^{-1}). A pressure transducer (Orpheus-mini, OTT, Germany and ADC and C31, OTT, Germany) monitored river stage height equated to flow (Q, m^3 s^{-1}) using sub-hourly rated records of water level at a Corbett non-standard flat-v weir. Data were transferred to a WISKI 7 database management system for quality control, processing and archiving.

Data management and collection scenarios

Total reactive P and Q data collected over a three year period (1st April 2010 – 31st March 2013) were examined for outliers and any apparent anomalies were discarded following consultation with data managers. Hourly averages were calculated from sub-hourly data (on average three TRP datapoints per hour and six Q datapoints per hour) for ease of forward processing and to remove any bias caused by equipment and sampling anomalies. The data were then resampled to reflect different sampling scenarios of frequency and timing using functions programmed in R (R Core Team 2014).

New descriptors were derived from the date and times of observations, to include day of the week, week of the year, weekend (Saturday or Sunday) and working hours (8.00-18.00, Monday to Friday). Once the data were primed for resampling, subsets were sampled from the original dataset for each sampling scenario into new combinations (C; Table 1). Sampling scenarios were designed to reflect realistic sampling frequencies of daily, three times per week, weekly and monthly. Daily datasets (C1a-C1d) were also repeated to observe the effect of restricting sampling to during working hours, the hourly change in P apportionment and the difference between night and day. Of the 2000 monthly datasets, only 999 were fitted with the BM, due to non-convergence after 5 days of analysis for 1001 datasets. However, this still provided a statistically sufficient number of datasets for model performance analysis.

The Load Apportionment Models

The two LAMs used in the current study (BM and GM) are able to estimate the relative contributions of P from point and diffuse sources based on measurements of P concentration at particular river flow rates (Q). The BM (Eqn. 1) used two functions; the first constrained (B < 1) to represent a reduction in P concentration as Q increases (point sources) and the second constrained (D > 1) to show an increase in P concentration as Q increases (diffuse sources):

$$\mathrm{P} = A.\,Q^{B-1} + C.\,Q^{D-1} \qquad \text{(Eqn. 1)}$$

Where A, B, C, and D are time in-variant model coefficients; Q is flow; and P is the P concentration. B constrained to < 1, D constrained to >1.

The GM (Eqn. 2) comprised three functions, with no constraints, in a polynomial nonlinear regression. First, a complete inverse proportional relationship between P concentration and Q (point sources); second, a linear relationship between P concentration and Q (diffuse sources); and third, a quadratic relationship between P concentration and Q, to account for hysteresis caused by source depletion in the dataset, i.e. when diffuse sources have become exhausted in the catchment but the flow continues to increase.

$$P = aQ^{-1} + bQ + cQ^2 \quad \text{(Eqn. 2)}$$

Where a, b and c are time-invariant model coefficients; Q is flow and P is the phosphorus concentration

Coefficients from each model were then manipulated to provide four model outputs using hydrological data: i) flows at which point sources no longer dominate load (Q_e), ii) percentage of flows dominated by point sources, iii) TRP cumulative load over three years and, iv) point load apportionment. Details on the method of calculation of each of these outputs are available in the online resource. However, the main differences between the algorithms are that BM allows variation in the inverse proportionality between Q and P, which is to account for P lost to sediments (Bowes et al. 2008), while GM focuses more on source depletion and the linear relationship between Q and P from diffuse sources.

The modelling process

All datasets were analysed using the R programming package "phoslam" (developed by the authors of this study), which provided the best fit for each model by the least squares method (stats::nls). The code was assessed as fit for purpose by the developers of the two LAMs being assessed (pers. comm. M. Bowes and S. Greene) in independent blind tests on dummy datasets. Standard errors for the load apportionment of hourly data were calculated based on 500 replicates using bootstrapping (Efron 1979). Standard errors for the resampled datasets (as described in section 2.2) for the four model outputs were calculated based on 2000 resampled datasets.

As part of the modelling process, the Akaike Information Criterion (AIC; k = 2; Akaike 1974) was calculated for each dataset output to provide a basis for model selection and, in this case, model comparison. The AIC is used to quantify model fit and takes into account the complexity of a model. AIC was used as the model selection criterion ahead of R^2 because of the non-linearity of and power functions employed by the two LAMs (Spiess and Neumeyer 2010).

Analysis of model outputs

Full high resolution hydrograph

Each modelled line was compared visually to the original observed dataset, with calculated residuals (Equation 3) also analysed visually for change over increasing Q.

Residual = $(\hat{y} - y_t)$ (Eqn. 3)

Where $\hat{y}_t$ is the estimated value for TRP load using model coefficients at time *t*; y_t is the observed value for TRP load at time *t*.

The residual as a percentage of observed load was calculated to show the degree of error (see online resource, Table 1SI). In this case, the range of percentage error obtained for increments of Q illustrated the applicability of the modelled line to observed loads across the range of flows, for all sampling strategies.

Between models and sampling strategies

The mean, standard deviation, skewness and kurtosis for model outputs for each sampling strategy were calculated as part of the evaluation. Accuracy and reproducibility of model outputs between sampling strategies were also analysed using four tests which are outlined below and described in detail in Table 1SI.

1. Using Direct Value Comparison (Bennett et al. 2013), the estimated total cumulative TRP load for each fitted model was compared with the observed total cumulative TRP load based on hourly mean data.
2. Root mean square error (RMSE) values for three of the 2000 resampled datasets (pertaining to maximum, minimum and median modelled total cumulative TRP load) for each sampling strategy were calculated using Eqn. 4.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \quad \text{(Eqn. 4)}$$

Where $\hat{y}_t$ is the estimated value for TRP load using model coefficients at time *t*; y_t is the observed value for TRP load at time *t*; and *n* is the sample number.

3. The outputs for each of the resampled datasets were then checked for normality using the Anderson-Darling test (Anderson and Darling 1952) using nortest::ad.test in R. Histograms, showing the frequency of outputs for each parameter, were produced along with box plots showing the range of coefficient values for each sampling strategy. To improve the normality of the datasets, 75% of the outputs of each sampling strategy were re-sampled 30 times, and the means of these new datasets provided a new dataset (n=30) which was normally distributed ($p < 0.01$).
4. Using these new normally distributed datasets, the difference between sampling strategies were identified by ANOVA and the Tukey honestly significant difference (TukeyHSD) test, with no limitation on degrees of freedom. The differences between LAMs were determined based on the original outputs from each combination dataset using unpaired t-tests in Prism 5.0, as the degrees of freedom were < 3000, with Welch correction for unequal variances (Welch 1947).

Results

Dataset construction

Quality controlled mean hourly data (calculated from high frequency sub-hourly data) from the installed instrumentation provided 24867 paired data-points for TRP concentration and Q for three years 1st April 2010 – 31st March 2013, out of a possible 26304 data-points (Figure 2). This represented 96%, 93% and 94% completeness for the periods, respectively, 2010 to 2011, 2011 to 2012 and 2012 to 2013. The estimated baseflow index (BFI; ratio of baseflow to total flow) of 0.66, determined by Local Minimum Method (Pettyjohn and Henning, 1979), is considered moderate for this given size of catchment and highlights the previous poor-moderate drainage class described for this catchment (Melland et al. 2012). Table 1 provides the descriptive statistics of the sample numbers for the newly constructed datasets.

High temporal resolution dataset

Over the three year period, the total observed cumulative TRP load, based on hourly averages of concentration and flow, was 1380.35 kg or 145.6 kg km^{-2}. This total provided the value for Direct Value Comparison.

According to the BM and based on hourly data (Figure 3a), the total cumulative TRP load for the three years 2010-2013 was 1390.49 kg (RMSE of 0.23 mg s^{-1}). Owing to the availability of high temporal resolution data, the estimation of TRP loads at high flows by the GM was found to be a negative number (due to the third function for hysteresis; Figure 3b).

Using these TRP load estimations, the GM calculated total cumulative load as 1380.35 kg with a RMSE of 0.36 mg s^{-1}. Despite producing an accurate total cumulative P load, a negative value for P load is not logical, and also affected the calculation of load apportionment at lower sampling frequencies. To mitigate this, any estimation of negative concentration at high flow (above ~99%ile) was converted to concentration by point sources only (i.e. omitting the diffuse and hysteresis functions in the model; Figure 4). In this case, this increased the estimated load to 1438.20 kg (RMSE of 0.22 mg s^{-1}). As the model estimates P load, and point sources are believed to be continuous irrespective of flow, this modification to the model was justified as the point source load must still be accounted for while the diffuse load is absent. Any reference to the GM from this point forward is in relation to this modified model at high flows, unless stated otherwise.

Sampling scenarios

Tables 2SI and 3SI (Online Resource) show the outputs for all combinations of new datasets, according to, respectively, the BM and GM models. One estimate of total cumulative TRP load by the BM using C2b (sampling three days per week) was deemed an outlier at 1.66×10^{22} kg and removed. The extremely large range of outputs for total cumulative TRP load estimation by C4 precluded the calculation of a standard deviation. Thirty two C4 datasets had problems with convergence (coefficient C was calculated to equal 0) and could not provide a value for Qe and so were omitted. The coefficients obtained for each model varied widely (Figures 5i and 5ii), with variance in those coefficients that describe diffuse sources (BM: C, D and GM: b, c) increasing substantially as sampling frequency decreased. This could be indicative of the difficulty in accurately defining contributions from diffuse sources with a reduced sampling frequency.

High variability in the prediction of the contribution from diffuse sources was also evident due to the increased range of values for all model outputs within each combination dataset as sampling frequency decreased. As the sampling frequency reduced to weekly, implausible values for total cumulative TRP load were produced. This was particularly the case when the monthly datasets were used, with values as high as 9.6×10^{209} kg (BM) and 37 391.8 kg (GM) being predicted. High variability may also be attributed to the seasonal nature of the study site, where coefficients attempt to describe the TRP-Q relationship over two different halves of the year (as identified by Jordan et al. 2012).

All model outputs showed increasing ranges, and bias (skewness values outside -1 to +1, kurtosis ≠ 0) as sampling frequency decreased (Figures 6 and 7, Figures 1SI and 2SI). The outputs achieved by the BM using the monthly sampling frequency (C4) were particularly variable, with 838 out of 999 datasets overestimating the total load by between 1% and 7.0 x 10^{208} %. Marked differences in output between the two LAMs were evident, particularly in estimates of cumulative load and load apportionment to point sources (Figures 6 and 7).

Timing of sampling appears to have had little effect on the estimated percentage of flows dominated by point sources, or P load apportionment for both models. For example, the means for each of the daily sampling model outputs were similar (Qe values of 0.47, 0.46, 0.45 and 0.48 $m^3 s^{-1}$ and 0.044, 0.043, 0.045 and 0.042 $m^3 s^{-1}$ for, respectively, the BM and GM (Tables 2SI and 3SI)). Estimated total loads, however, showed a large divergence between means within a sampling frequency (Tables 2SI and 3SI), particularly for the BM. Thus, although coefficients may have provided a similar answer for Qe and percentage of flows dominated by point sources, they still impacted the precision of total load estimation and the resulting source apportionment.

Statistical analysis of model outputs

Between models

Differences between estimations of contribution from point and diffuse sources to overall TRP load between the two LAMs were particularly evident. The BM output based on the hourly data indicated that 51.4% (95% CI: 48.4% – 54.8%) of the TRP load came from point sources, compared with only 4.2% (95% CI: 4.1% – 4.6%) according to the GM. Similar divergence in other model outputs was evident, with the percentage of flows dominated by point sources (i.e. number of flows below Qe for each model) ranging from 94.8% (Qe: 0.416 $m^3 s^{-1}$) for BM to 37.7% (Qe: 0.049 $m^3 s^{-1}$) for the GM. Significant ($p<0.001$) differences between models were also evident in the mean Qe values, percentage of time flows were dominated by point sources, point apportionment, and estimated total cumulative TRP load.

Between sampling strategies

The newly created datasets, from the means of 75% random samples of model outputs within each combination, were found to be normally distributed ($p<0.01$), except for total cumulative TRP load estimation by the BM using C3 and C4 and Qe using BM for C4 (Table 4SI). Due to the high variation in loads estimated using BM using sampling strategies C3 and C4, and the non-normality of the Qe values using C4, C3 and C4 were excluded from the Tukey HSD test for between sampling strategies (GM values were included). Within each model, the means and standard deviations (from the means of the newly created resampled datasets) between sampling strategies were, for the most part, significantly different from each other ($p<0.05$) for all four model outputs. The means of some combination datasets had non-significant differences (Table 5SI). None of the datasets had significantly similar means for all four model outputs.

Residuals analysis

RMSE scores (Table 2) were, as expected, high for the model parameters resulting in maximum modelled total cumulative TRP loads, while minimum modelled total cumulative TRP loads resulted in the lowest RMSE scores. Residuals stayed quite close to zero until Q reached ~ 1 $m^3 s^{-1}$, when the rate of increase in residual error rose significantly as sampling frequency decreased (Figure 8).

Estimations of TRP load at high and low flows were wide-ranging for both BM and GM, reflected by residuals as a percentage of observed load (Figure 9 and Table 6SI; error range 200-4000%). Both residual errors and percentage residual errors at high flows were many magnitudes higher than at low flows (Figure 8 and Table 6SI), particularly for the BM. Additionally, loads modelled at lower Q values had large errors when examined as a percentage of observed load.

The initial analysis of outputs from the data used indicated that, although the GM provided a more precise range of values for point apportionment and total load estimation, the BM had a consistently better averaged AIC value (i.e., a better fit).

Discussion

Applying high temporal resolution data in model assessment

The availability of high temporal resolution river water quality data has enabled a comparative assessment of two relatively widely referred to LAMs, over a range of sampling scenarios. Both LAMs generated very good approximations of estimated total cumulative TRP loads (1390.49 kg (BM) and1438.20 kg (GM)) when compared with observed values. However, they were statistically very different ($p < 0.01$) for all model outputs. The differences were largely attributed to the differing model construction (see online resource or Bowes et al. 2008 and Greene et al. 2011). When considering model selection, the models were shown to have differing strengths when used *ad hoc* with extant data. The BM appeared to have the better fit with observed values while the GM, by comparison, generated narrower ranges of model parameters. The most important output, point source contribution, was particularly polarised. This was largely due to the dissimilar algorithms used and the way Qe values are calculated due to the structure of algorithm employed.

The river has been observed to have a high concentration of P throughout summer low flows (Jordan et al. 2012). Therefore, point sources are expected to dominate (Jarvie et al. 2010) and to contribute a much higher proportion of P than estimated by the GM. Choosing models based on expectation rather than performance can, however, lead to incorrect conclusions on which model is describing the change in concentration with flow more accurately – mainly because of model abstraction and idealization (Chakravartty 2010). In this case, the change of concentration with flow based on hourly mean data was known, but future users of LAMs in general may be reliant on lower frequency sampling to determine model parameters. This highlights the need to include an indication of variability, as a measure of confidence, in model outputs.

Using a high temporal resolution dataset, the large errors by percentage at both low and high flows were clearly apparent, with the error at low flow observations probably caused by the highly variable TRP concentration data. This variability was inadequately modelled by a single line and may be further affected by increased scale and varying base-flow indices (Johnes 2007). Further model development to account for these factors could allow for improved model performance using, as a minimum, daily-resolved data as an input. Similarly, studies of the effects of quickflow as the predominant contributor to streamflow, using high temporal resolution sampling, may provide additional hydrological understanding required for future improvements in model development. However, incorporating the impact of these various factors may lead to a more process-based model (as used by Romagnoli et al. 2017), thus rendering somewhat redundant the concept of empirical models as an easy solution to complex hydrological problems, such as load apportionment.

Residual errors were found to be particularly high at high flows (Figure 8), and as sampling frequency decreased (Table 2). Previous users of the BM found variable estimation of P concentration at low flows (McDowell et al. 2011; Trevisan et al. 2012). Future model development could be improved by using suitable artificial data sequences (Bennett et al. 2013) that may identify the optimum limits of a particular model. This highlights the utility of generating high resolution time series water quality data for, *inter alia*, model testing.

Sampling strategy design

Although a higher sampling frequency will potentially provide more precise outputs when modelling environmental data, balancing resources and uncertainty must be considered when designing a sampling regime (Schoumans et al. 2009). In this study, as the sampling frequency increased, the residual error was reduced and the range of load estimation and other model outputs narrowed. However, using a monthly sampling frequency, nearly all of the resampled datasets overestimated the total cumulative TRP load ~ in some cases by several orders of magnitude. As the LAMs have been designed to represent trends, i.e., changes in P over changes in flow, it follows that monthly data would not be of a high enough resolution to quantify this relationship adequately (Kristensen and Bøgestrand 1996). The WFD implementation has generally resulted in a hierarchical design for sampling frequency (Petit 2010), with EU member states putting more resources into failing catchments to identify the driving factors of eutrophication (Priestly 2015). Consequently, in most other catchments where sampling frequency remains low (usually monthly due to sampling budget constraints) LAMs there are unlikely to prove effective as management tools.

While some studies have looked at the effects on model outputs of reducing sampling frequency (Jarvie et al. 2010; Cassidy and Jordan 2011; Wade et al. 2012; Bieroza et al. 2014), few have investigated specific timing, either during the day or during the week. Dissolved oxygen saturation over a 24 hour period has shown a distinct diurnal cycle (Wade et al. 2012). Yet sampling regimes may be implemented to collect a river sample within a 3 hour window of a particular day of the week, and few are collected outside of normal working hours. Model outputs in this study were statistically significantly different, depending on what day of the week a water sample was collected (estimates of point apportionment could differ by 30% depending on the days of the week sampling took place). Similarly, daily night-time and daytime modelled total cumulative TRP loads were statistically different (p<0.001). Hence sampling that takes place at regular intervals with a relatively low frequency may be missing processes and sources of P occurring at specific times of day and/or on specific days. Some studies suggest that weekly sampling combined with storm sampling will provide the best range of concentration with flow data for use in modelling. However, even using this method, McDowell et al. (2011) could only achieve a dataset that covered 60% of their site's flow duration curve after 6 years of sampling.

Implications for catchment management strategies using load models

This study and others (Cassidy and Jordan 2011; Chen et al. 2013) have illustrated the challenges associated with the accurate prediction of P loads at high flows. Johnes (2007) found total P annual load was progressively overestimated by each model tested as sampling decreased from daily to monthly. However, Wade et al. (2012), on the much larger River Thames, UK, saw little improvement in annual total P load estimation using a simple nutrient load estimation algorithm when the frequency of sampling was reduced. In the current study, sampling three times per week resulted in only slightly higher RMSE and a small reduction in uncertainty in cumulative load estimation (Table 2).

This poses a number of problems when considering effective management strategies to improve river water quality. For example, improvements in sewage treatment are likely to be viewed as the optimal management response to model outputs identifying point sources as the predominant contributor of P load in a river. However, high uncertainties associated with the model outputs may render improvements in sewage treatment futile. Similarly, model outputs suggestive of a strong diffuse source contribution of P load could lead to inappropriate and ultimately ineffective measures applied to farming practices in the area. Point sources can be particularly important during the late spring and early summer, i.e. during much of the ecologically critical growing season (Jarvie et al. 2013; Jarvie et al. 2014). Consequently the ability to model TRP load at low flows adequately, thereby reducing the risk of incorrectly attributing P loads to either point or diffuse sources, is of vital importance to the effective management of river eutrophication. This is especially so in mixed landuse, mixed P source catchments where those sources have different hydrological dependencies (Jordan et al. 2007).

The success of river restoration measures is dependent on the implementation of adequate post restoration monitoring (Feld et al. 2011). Here the length of monitoring is important but, as shown in this study, also sampling frequency. As another possible application of LAMs could be to identify a change in P load apportionment, and/or reduction in annual cumulative P load, following implementation of remediation measures (e.g., Greene et al. 2011), this would also be constrained by the uncertainty observed in these model outputs when tested using high resolution data.

The decrease from a possible 365 to 150 observations per year to only 50 (i.e. weekly) was shown to significantly reduce the precision of each of the models for all four model outputs. However, daily sampling appeared to provide some parsimony and a trade-off between sample temporal resolution and model requirements (for total cumulative load estimation). Even with a daily sampling scenario, the provision of observed cumulative TRP load from high temporal resolution data highlighted the potential for poor prediction of TRP load (290% load overestimation for C1a).

While this study focuses only on two models that were devised for physiographically different catchments, the overall finding relating to the effects of sampling frequency and timing on model outputs has much wider implications. This is particularly the case given that both models have been relatively widely applied. Use of the BM and GM in the current study revealed divergent outputs based on varying input data from the same catchment. At extreme ends, one model suggested the contribution from point sources was low (magnitude-centric) while the other estimated nearly all flows in the year to be dominated by point sources (duration-centric). Source apportionment, even using samples collected at a daily interval, resulted in high prediction variability (particularly for the BM) and presented a problem for modelling in rivers that historically have been sampled on a monthly basis. Nevertheless, the estimated rural-point source TRP load of 63 kg year^{-1} using the method by Carvalho et al. (2005) - approximately 14% of total observed cumulative load) is similar to that approximated by the GM. However, there is an equal element of uncertainty with this source estimation due to unresolved point source origins, condition and risk (Melland et al. 2012; Murphy et al. 2015).

Withers et al. (2009) highlighted the oversimplification of nutrient modelling using LAMs, evident in the use of a single modelled line to describe the clustering of points at low flows. Similarly Neal et al. (2010) discussed the requirement of a larger number of variables (which may vary spatially as well) to model P transport in rivers fully. The complexity of P-Q relationships has also been recognized in high-resolution datasets by Bowes et al. (2015). Lower resolution data were found to mask important P processes leading to a need for more complex model assumptions and may require a completely new analytical method (Chaudhary and Hantush, 2017). It is clear that the use of LAMs needs to be developed on catchment specific data but development, and therefore model predictions as shown here, may be constrained by the quality and resolution of the input data.

Conclusions

One of the benefits of using higher resolution environmental data is the ability to assess the limitations of existing empirical models that are often employed for river catchment management. Sampling frequency has been identified as an important factor in model performance previously and this study developed a method to quantify this effect on two commonly-used LAMs. This was particularly pertinent as point sources have become recognised as influential in eutrophic episodes because of their dominance during the ecologically critical period of spring/summer.

Three clear outcomes from this study were:

1. Interrogation of high frequency data allowed the assessment of the precision of models over a range of sampling frequencies and timings
2. Accuracy of model outputs may be improved by partitioning the data collected seasonally
3. The main difference between the two LAMs was in the apportionment of TRP to point sources. This has important implications for their use in catchment management.

Regarding outcome 1, variation in modelled total cumulative TRP load across sampling scenarios showed that daily sampling appeared to show some compromise between resource and model requirements at the scale of the study – this requires further investigation. Errors were particularly evident at extremes of the flow curve and therefore could be reduced by targeted sampling campaigns. Although, timing of sampling also affected the accuracy of model outputs. The results presented here highlight the need for robust statistical testing and provision of confidence intervals for output data. This will ensure selection of the most appropriate model and that confidence can be attached to the implementation of measures aimed at improving river water quality.

Outcome 2 highlights the seasonal nature of the data used and that processes are different due to land management practices and weather patterns in temperate climates. Future models should use data that have been seasonally partitioned to determine if accuracy may be improved.

Outcome 3 revealed the disparity often displayed using different models on identical datasets. In this study, output from one model suggested that improvement in water quality would be best achieved through measures that target diffuse sources in the catchment (magnitude-centric) whereas the other model pointed firmly towards point sources being an important factor in poor water quality (duration-centric). Thus as tools for river management, empirical models, such as LAMs, need to be considered within focused and expert based modelling frameworks. Moreover, the LAM used is likely to require calibrating to accommodate local catchment characteristics. Ignoring these factors is likely to lead to widely varying results and challengeable decisions as shown by the application of high temporal resolution data in this study.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE T Automat Contr*, 19, 716-723. doi: 10.1109/TAC.1974.1100705

Anderson, T.W., & Darling, D. A. (1952). Asympotic theory of certain goodness of fit criteria based on stochastic processes. *Ann Math Stat*, 23, 193-212. doi: projecteuclid.org/euclid.aoms/1177729437

Begum, S., Adnan, M., McClean, C. J., Cresser, M. S. (2016). A critical re-evaluation of controls on spatial and seaonsal variations in nitrate concentrations in river waters throughout the River Derwent catchment in North Yorkshire, UK. *Environ Monit Assess*, 188, 305. doi: 10.1007/s10661-016-5305-4

Bennett, N. D., Croke, B. F. W., Guariso, G., Guillaume, J. H. A., Hamilton, S. H., Jakeman, A. J., et al. (2013). Characterising performance of environmental models. *Environ Modell Softw*, 40, 1-20. doi: 10.1016/j.envsoft.2012.09.011

Bieroza, M. Z., Heathwaite, A. L., Mullinger, N. J., & Keenan, P. O. (2014). Understanding nutrient biogeochemistry in agricultural catchments: the challenge of appropriate monitoring frequencies. *Environ Sc: Processes Impacts*, 16, 1676-1691. doi: 10.1039/c4em00100a

Bieroza, M. Z., & Heathwaite, A. L. (2015). Seasonal variation in phosphorus concentration-discharge hysteresis inferred from high-frequency in situ monitoring. *J Hydrol*, 524, 333-347. doi: 10.1016/j.jhydrol.2015.02.036

Binzer, A. Guil, C., Rall, B. C. Brose, U. (2016). Interactive effects of warming, eutrophication and size structure; impacts on biodiversity and food-web structure. *Global Change Biology*, 22(1), 220-227.

Bowes, M. J., Jarvie, H. P., Halliday, S. J., Skeffington, R. A., Wade, A. J., Loewenthal, M., et al. (2015). Characterising phosphorus and nitrate inputs to a rural river using high-frequency concentration-flow relationships. *Sci Total Environ*, 511, 608-620. doi: 10.1016/j.scitotenv.2014.12.086

Bowes, M. J., Neal, C., Jarvie, H. P., Smith, J. T., & Davies, H. N. (2010) Predicting phosphorus concentrations in British rivers resulting from the introduction of improved phosphorus removal from sewage effluent. *Sci Total Environ*, 408, 4239-4250. doi: 10.1016/j.scitotenv.2010.05.016

Bowes, M. J., Smith, J. T., Jarvie, H. P., & Neal, C. (2008). Modelling of phosphorus inputs to rivers from diffuse and point sources. *Sci Total Environ*, 395, 125-138. doi: 10.1016/j.scitotenv.2008.01.054

Bowes, M. J., Smith, J. T., Jarvie, H. P., Neal, C., & Barden, R. (2009). Changes in point and diffuse source phosphorus inputs to the River Frome (Dorset, UK) from 1966 to 2006. *Sci Total Environ*, 407, 1954-1966. doi: 10.1016/j.scitotenv.2008.11.026

Campbell, J. M., Jordan, P., & Arnscheidt, J. (2015). Using high-resolution phosphorus data to investigate mitigation measures in headwater river catchments. *Hydrol Earth Syst Sc*, 19, 453-464. doi: 10.5194/hess-19-453-2015

Carvalho, L., Maberly, S. C., May, L., Reynolds, C., Hughes, M., Brazier, R., et al. (2005). Risk Assessment Methodology for Determining Nutrient Impacts in Surface Freshwater Bodies. Bristol: Environment Agency.

Cassidy, R., & Jordan, P. (2011). Limitations of instantaneous water quality sampling in surface-water catchments: Comparison with near-continuous phosphorus time-series data. *J Hydrol*, 405, 182-193. doi: 10.1016/j.jhydrol.2011.05.020

Chakravartty, A. (2010). Perspectivism, inconsistent models, and constrastive explanation. *Stud Hist Philos Sci*, 41, 405-412. doi: 10.1016/j.shpsa.2010.10.007

Chaudhary, A., & Hantush, M. M. (2017). Bayesian Monte Carlo and maximum likelihood approach for uncertainty estimation and risk management: Application to lake oxygen recovery model. *Water Research*, 108, 301-311. doi: 10.1016/j.watres.2016.11.012

Chen, D., Dahlgren, R. A., & Lu, J. (2013). A modified load apportionment model for identifying point and diffuse source nutrient inputs to rivers from stream monitoring data. *J Hydrol*, 501, 25-34. doi: 10.1016/j.jhydrol.2013.07.034

Chen, D. J., Hu, M. P., Guo, Y., & Dahlgren, R. A. (2015). Reconstructing historical changes in phosphorus inputs to rivers from point and nonpoint sources in a rapidly developing watershed in eastern China, 1980-2010. *Sci Total Environ*, 533, 196-204. doi: 10.1016/j.scitotenv.2015.06.079

Core Team, R. (2014). R, a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing

Efron, B. (1979). Bootstrap Methods: Another look at the Jacknife. *Ann Stat*, 7, 1-26. doi: 10.1214/aos/1176344552

Eisenreich, S. J., Bannerman, R. T., & Armstrong, D. E. (1975). Simplified phosphorus analysis technique. *Environ Lett*, 9, 45-53.

Feld, C. K.., Birk, S., Bradley, D. C., Hering, D., Kail, J., Marzin, A., et al. (2011). From Natural to Degraded Rivers and Back Again: A Test of Restoration Ecology Theory and Practice. In G. Woodward. (Ed.), *Advances in Ecological Research 44* (pp 120-209). Amsterdam: Elsevier,.

Fonseca, B. M., de Mendonça-Galvão, L., Padovesi-Fonseca, C., Monteiro de Abreu, L., & Fernandes, A. C. M. (2014). Nutrient baselines of Cerrado low-order streams: comparing natural and impacted sites in Central Brazil. *Environ Monit Assess*, 186, 19-33. doi: 10.1007/s10661-013-3351-8

Greene, S., Taylor, D., McElarney, Y. R., Foy, R. H., & Jordan, P. (2011). An evaluation of catchment-scale phosphorus mitigation using load apportionment modelling. *Sci Total Environ*, 409, 2211-2221. doi: 10.1016/j.scitotenv.2011.02.016

Jansons, V., Vagstad, N., Sudars, R., Deelstra, J., Dzalbe, I., & Kirsteina, D. (2002). Nutrient losses from point and diffuse agricultural sources in Latvia. *Landbauforsch Volk* 52, 9-17.

Jarvie, H. P., Neal, C., & Withers, P. J. A. (2006). Sewage-effluent phosphorus: a greater risk to river eutrophication than agricultural phosphorus? *Sci Total Environ*, 360, 246-253. doi: 10.1016/j.scitotenv.2005.08.038

Jarvie, H. P., Sharpley, A. N., Spears, B., Buda, A. R., May, L., & Kleinman, P. J. A. (2014). Water quality remediation faces unprecedented challenges from "Legacy Phosphorus". *Environ Sci Technol*, 47, 8997-8998. doi: 10.1021/es403160ha

Jarvie, H. P., Sharpley, A. N., Withers, P. J. A., Scott, J. T., Haggard, B. E., & Neal, C. (2013). Phosphorus mitigation to control river eutrophication: murky waters, inconvenient truths, and "postnormal" science. *J Environ Qual*, 42, 295-304. doi: 10.2134/jeq2012.0085

Jarvie, H. P., Withers, P. J. A., Bowes, M. J., Palmer-Felgate, E. J., Harper, D. M., Wasiak, K., et al. (2010). Streamwater phosphorus and nitrogen across a gradient in rural-agricultural land use intensity. *Agric Ecosyst Environ*, 135, 238-252. doi: 10.1016/j.agee.2009.10.002

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. doi: 10.1016/j.jhydrol.2006.07.006

Jordan, P., Arnscheidt, A., McGrogan, H., & McCormick, S. (2007). Characterising phosphorus transfers in rural catchments using a continuous bank-side analyser. *Hydrol Earth Syst Sc*, 11, 372-381.

Jordan, P., Melland, A., Mellander, P. E., Shortle, G., & Wall, D. (2012). The seasonality of phosporus transfers from land to water: Implications for trophic impacts and policy evaluation. *Sci Total Environ*, 434, 101-109. doi: 10.1016/j.scitotenv.2011.12.070

Kristensen, P., & Bøgestrand, J. (1996). Surface water quality monitoring. Denmark: National Environmental Research Institute.

McDowell, R. W., Snelder, T., Littlejohn, R., Hickey, M., Cox, N., & Booker, D. J. (2011). State and potential management to improve water quality in an agricultural catchment relative to a natural baseline. *Agric Ecosyst Environ*, 144, 188-200. doi: 10.1016/j.agee.2011.07.009

Melland, A. R., Mellander, P. E., Murphy, P. N. C., Wall, D., Mechan, S., Shine, O., et al. (2012). Stream water quality in intensive cereal cropping catchments with regulated nutrient management. *Environ Sci Policy*, 24, 58-70. doi: 10.106/j.envsci.2012.06.006

Mellander, P. E., Jordan, P., Wall, D., Melland, A. R., Meehan, R., Kelly, C., et al. (2012). Delivery and impact bypass in a karst aquifer with high phosphorus source and pathway potential. *Water Res*, 46, 2225-2236. doi: 10.1016/j.watres.2012.01.048

Mellander, P. E., Melland, A. R., Murphy, P. N. C., Wall, D. P., Shortle, G., & Jordan, P. (2014). Coupling of surface water and groundwater nitrate-N dynamics in two permeable agricultural catchments. *Journal of Agricultural Science*, 152, S107-S124. doi: 10.1017/S0021859614000021

Moreno-Ostos, E., da Silva, S. L. R., de Vicente, I., & Cruz-Pizarro, L. (2007). Interannual and between-site variability in the occurrence of clear water phases in two shallow Mediterranean lakes. *Aquat Ecol*, 41, 285-297. doi: 10.1007/s10452-006-9072-0

Murphy, S., Jordan, P., Mellander, P. E., & O'Flaherty, V. (2015). Quantifying faecal indicator organism hydrological transfer pathways and phases in agricultural catchments. *Sci Total Environ*, 520, 286-299. doi: 10.1016/j.scitotenv.2015.02.017

Neal, C., Jarvie, H. P., Williams, R., Love, A., Neal, M., Wickham, H., et al. (2010). Declines in phosphorus concentration in the upper River Thames (UK): Links to sewage effluent cleanup and extended end-member mixing analysis. *Sci Total Environ*, 408, 1315-1330. doi: 10.1016/j.scitotenv.2009.10.055

Noges, P., Noges, T., Ghiani, M., Sena, F., Fresner, R., Friedl, M., et al. (2011). Increased nutrient loading and rapid changes in phytoplankton expected with climate change in stratified South European lakes: sensitivity of lakes with different trophic state and catchment properties. *Hydrobiologia*, 667, 255-270. doi: 10.1007/s10750-011-0649-9

OJEC (2000) Council Directive 2000/60/EC establishing a framework for Community action in the field of water policy. Luxembourg: European Parliament and the Council of the European Union.

Outram, F. N., Lloyd, C. E. M., Jonczyk, J., Benskin, C. M. H., Grant, F., Perks, M. T., et al. (2014). High-frequency monitoring of nitrogen and phosphorus response in three rural catchments to the end of the 2011-2012 drought in England. *Hydrol Earth Syst Sc*, 18, 3429-3448. doi: 10.5194/hess-18-3429-2014

Perks, M. T., Owen, G. J., Benskin, C. M. H., Jonczyk, J., Deasy, C., Burke, S., et al. (2015). Dominant mechanisms for the delivery of fine sediment and phosphorus to fluvial networks draining grassland dominated headwater catchments. *Sci Total Environ*, 523, 178-190. doi: 10.1016/j.scitotenv.2015.03.008

Petit, K. (2010). Progress in monitoring river quality. France: Office nationale de l'eau et des milieux aquatiques (ONEMA).

Pettyjohn, W. A. and Henning, R.(1979). Preliminary estimate of groundwater recharge rates, related streamflow and water quality in Ohio. *Ohio State. University Water Resources Center*, Project Completion Report Number 552, 323.

Priestly, S. (2015). Water Framework Directive: achieving good status of water bodies, Briefing Paper Number CBP 7246. House of Commons Library, London.

R Core Team (2014). R, a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Robson, B. (2014). State of the art in modelling of phosphorus in aquatic systems: Review, criticisms and commentary. *Environ Modell Softw*, 61, 339-359. doi: 10.1016/j.envsoft.2014.01.012

Rode, M., Wade, A. J., Cohen, M. J., Hensley, R. T., Bowes, M. J., Kirchner, J. W., et al. (2016). Sensors in the stream: the high-frequency wave of the present. *Environmental Science and Technology*, 50(19), 10297-10307. doi: 10.1021/acs.est.6b02155

Romagnoli, M., Portapila, M., Rigalli, A., Maydana, G., Burgués, M., & García C. M. (2017). Assessment of the SWAT model to simulate a watershed with limited available data in the Pampas region, Argentina. *Sci Total Envirot*, 596-597, 437-450. doi: 10.1016/j.scitotenv.2017.01.041

Schoumans, O. F., Silgram, M., Groenendijk, P., Bouraoui, F., Andersen, H. E., Kronvang, B., et al. (2009).

Description of nine nutrient loss models: capabilities and suitability based on their characteristics. *J Environ Monitor*, 11, 506-514. doi: 10.1016/j.agee.2011.07.00910.1039/b823239c

Sharpley, A., Jarvie, H. P., Buda, A., May, L., Spears, B., & Kleinman, P. (2013). Phosphorus legacy: overcoming the effects of past management practices to mitigate future water quality impairment. *J Environ Qual*, 42, 1308-1326. doi: 10.2134/jeq2013.03.0098

Sharpley, A., & Wang, X. (2014). Managing agricultural phosphorus for water quality: Lessons from the USA and China. *Journal of Environmental Sciences*, 26, 1770-1782. doi: 10.1016/j.jes.2014.06.024

SI 272 (2012) European Communities Environmental Objectives (Surface Waters) (Amendment) Regulations. Dublin: Stationery Office.

Skeffington, R. A., Halliday, S. J., Wade, A. J., Bowes, M. J., & Loewenthal, M. (2015). Using high-frequency water quality data to assess sampling strategies for the EU Water Framework Directive. *Hydrol Earth Syst Sc*, 19, 2491-2504. doi: 10.5194/hess-19-2491-2015

Spiess, A. N., & Neumeyer, N. (2010). An evaluation of R^2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. *BMC Pharmacology*, 10, 6. doi: 10.1186/1471-2210-10-6

Trevisan, D., Quétin, P., Barbet, D., & Dorioz, J. M. (2012). POPEYE: A river-load oriented model to evaluate the efficiency of environmental policy measures for reducing phosphorus losses. *J Hydrol*, 450-451, 254-266. doi: 10.1016/j.hydrol.2012.05.001

Verdonschot, P. F. M., Spears, B. M., Feld, C. K., Brucet, S., Keizer-Vlek, H., Borja, A., et al. (2013). A comparative review of recovery processes in rivers, lakes, estuarine and coastal waters. *Hydrobiologia*, 704, 453-474. doi: 10.1007/s10750-012-1294-7

Wade, A. J., Palmer-Felgate, E. J., Halliday, S. J., Skeffington, R. A., Loewenthal, M., Jarvie, H. P., et al. (2012). Hydrochemical processes in lowland rivers: insights from in situ, high-resolution monitoring. *Hydrol Earth Syst Sc*, 16, 4323-4342. doi: 10.5194/hess-16-4323-2012

Welch, B. L. (1947). The generalization of students problem when several different population variances are involved. *Biometrika*, 34:28-35. doi: 10.2307/2332510

Williams, A. A., & Kimball, M. E. (2013). Evaluation of long-term trends in hydrographic and nutrient parameters in a southeast US coastal river. *Environ Monit Assess*, 185, 10495-10509. doi: 10.1007/s10661-013-3347-4

Withers, P. J. A., Jarvie, H. P., Hodgkinson, R. A., Palmer-Felgate, E. J., Bates, A., Neal, M., et al. (2009). Characterization of phosphorus sources in rural watersheds. *Environ Qual*, 38, 1998-2011. doi: 10.2134/jeq2008.0096

Withers, P. J. A., Jordan, P., May, L., Jarvie, H. P., & Deal, N. (2014). Do septic tank systems pose a hidden threat to water quality? *Frontiers in Ecology and the Environment*, 12, 123-130. doi: 10.1890/130131

Withers, P. J. A., May, L., Jarvie, H. P., Jordan, P., Doody, D., Foy, R. H., et al. (2012). Nutrient emissions to water from septic tank systems in rural catchments: uncertainties and implications for policy. *Environ Sci Policy*, 24, 71-82. doi: 10.1016/j.envsci.2012.07.023

Yang, Y. S., & Wang, L. (2010). A Review of Modelling Tools for Implementation of the EU Water Framework

Directive in Handling Diffuse Water Pollution. *Water Resources Management*, 24(9), 1819-1843. doi: 10.1007/s11269-009-9526-y

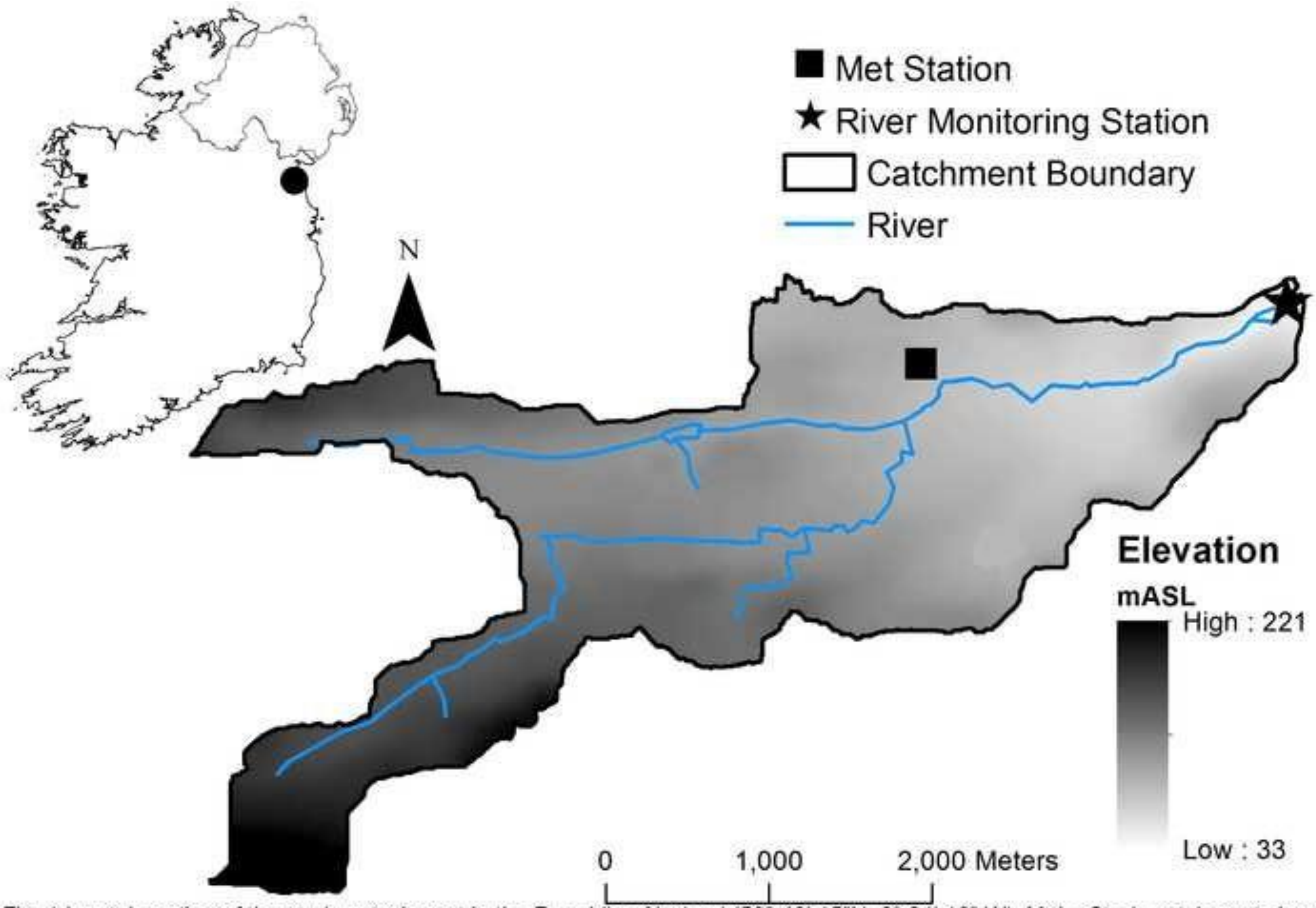

Fig. 1 Inset: Location of the study catcvhment in the Republic of Ireland (53° 49' 15"N, 6° 24' 16" W), Main: Study catchment showing topography, river-channel and monitoring stations (ArcGIS 10.2.2).

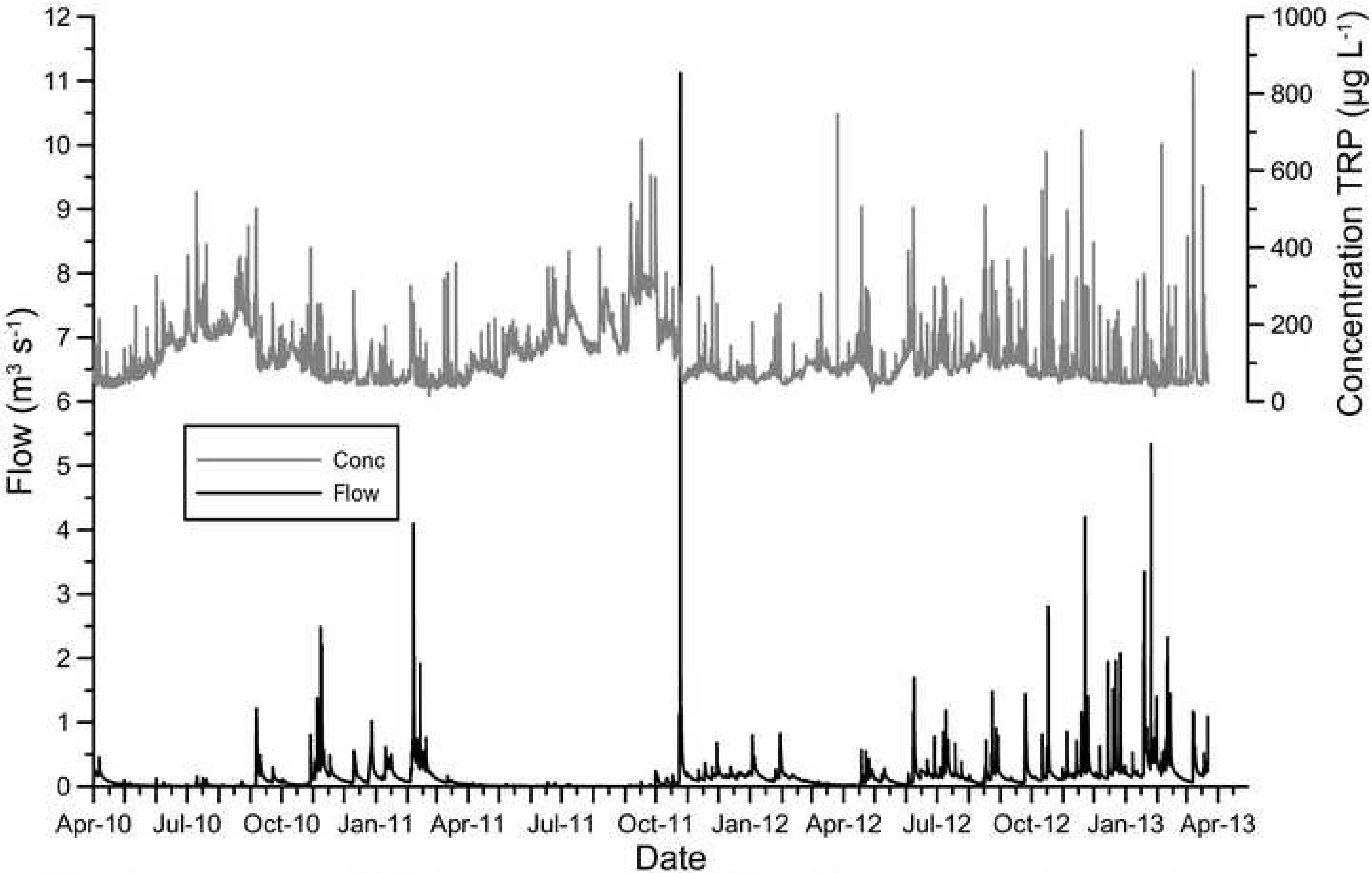

Fig. 2 Time-series of TRP concentration and Flow for the period 1st April 2010 - 31st March 2013 (Grapher 9.0).

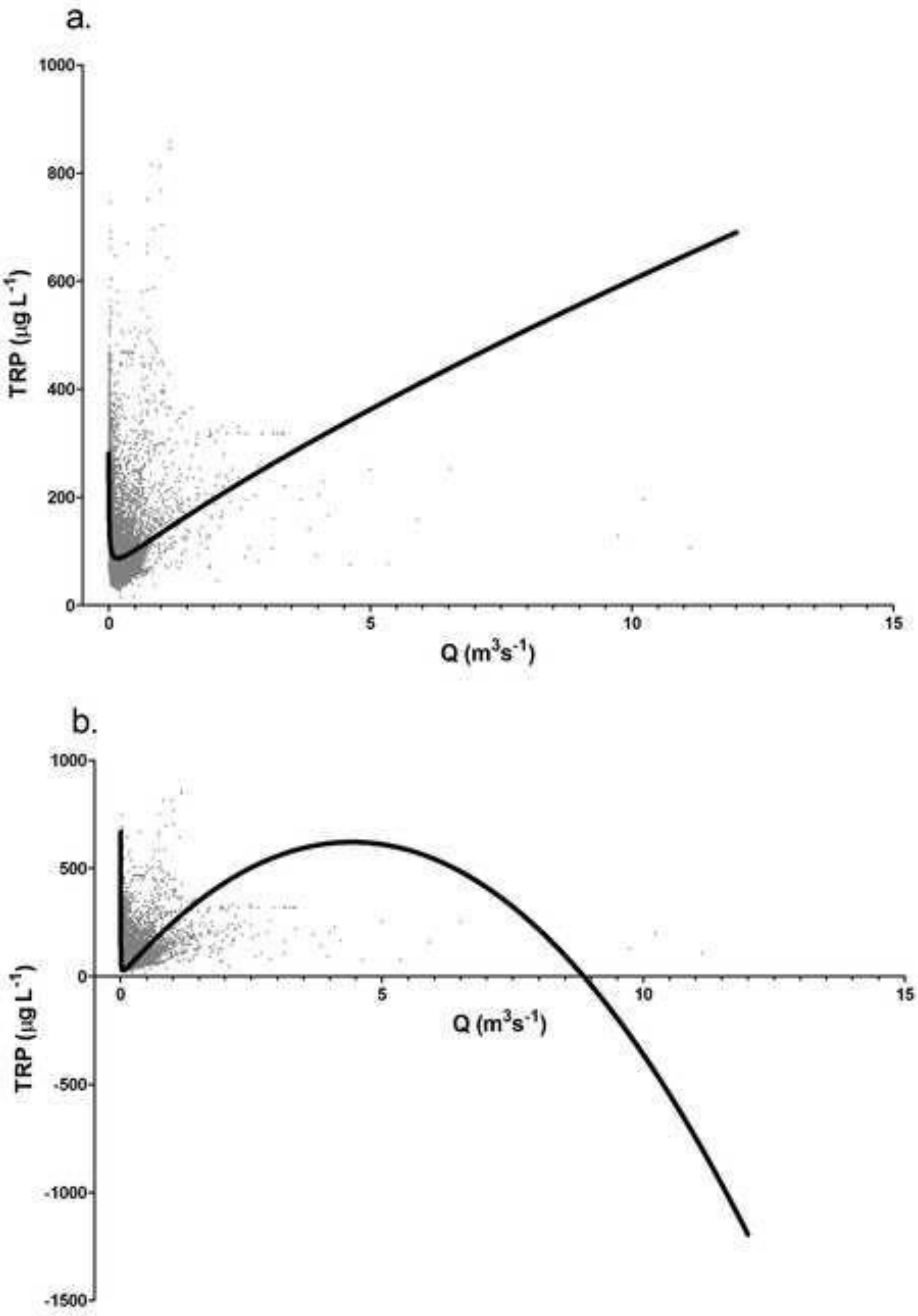

Fig. 3 a) BM modelled line with full total reactive phosphorus (TRP) and Q dataset shown in grey, b) GM modelled line before adjustment with full dataset shown in grey (Prism 5.0).

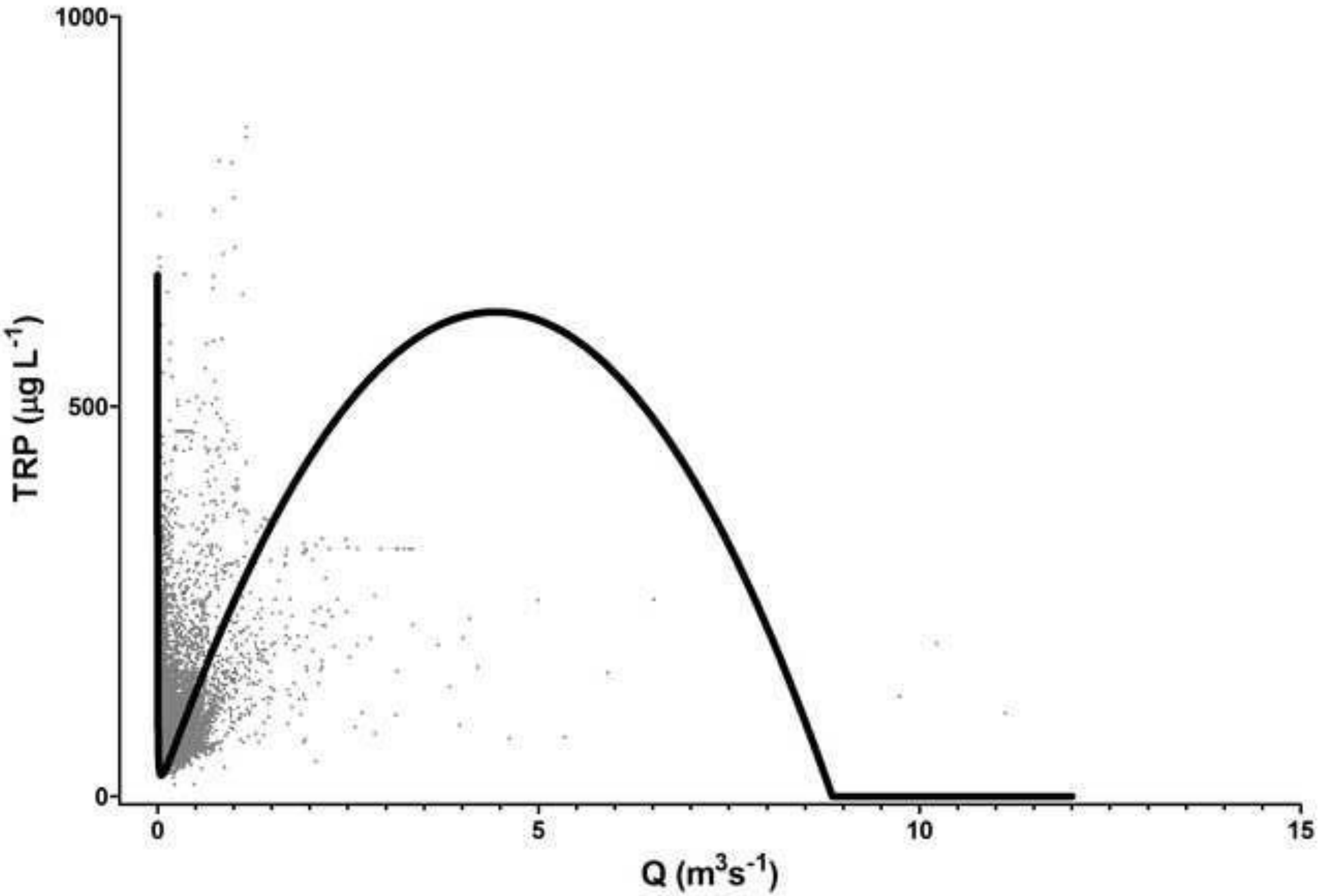

Fig. 4 GM modelled line with diffuse total reactive phosphorus load set to zero at high (~99th %ile) and full total reactive phosphorus (TRP) and Q dataset shown in grey (Prism 5.0).

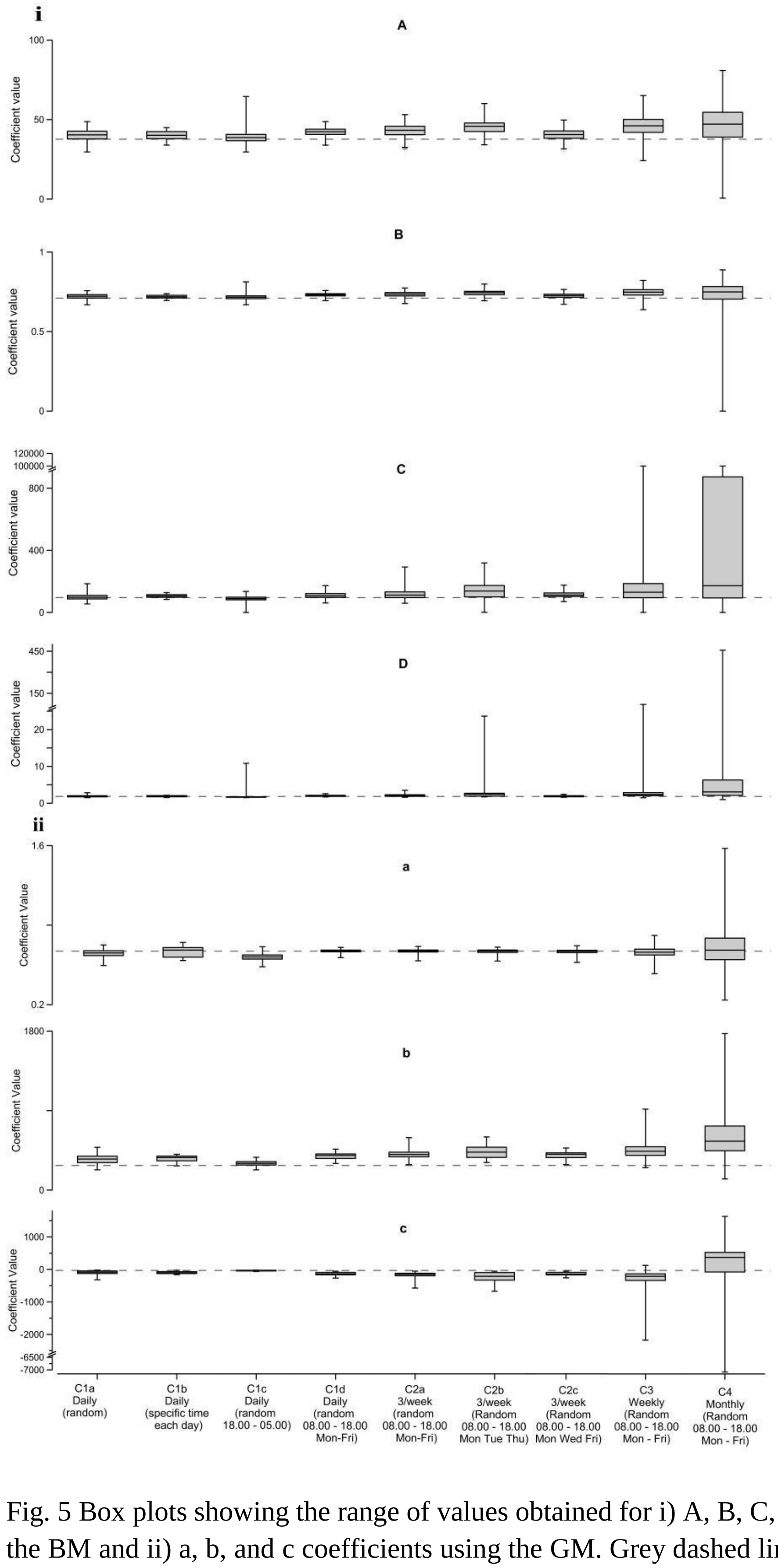

Fig. 5 Box plots showing the range of values obtained for i) A, B, C, and D coefficients using the BM and ii) a, b, and c coefficients using the GM. Grey dashed line denotes the coefficient value of the best fit line using hourly data (Grapher 9.0).

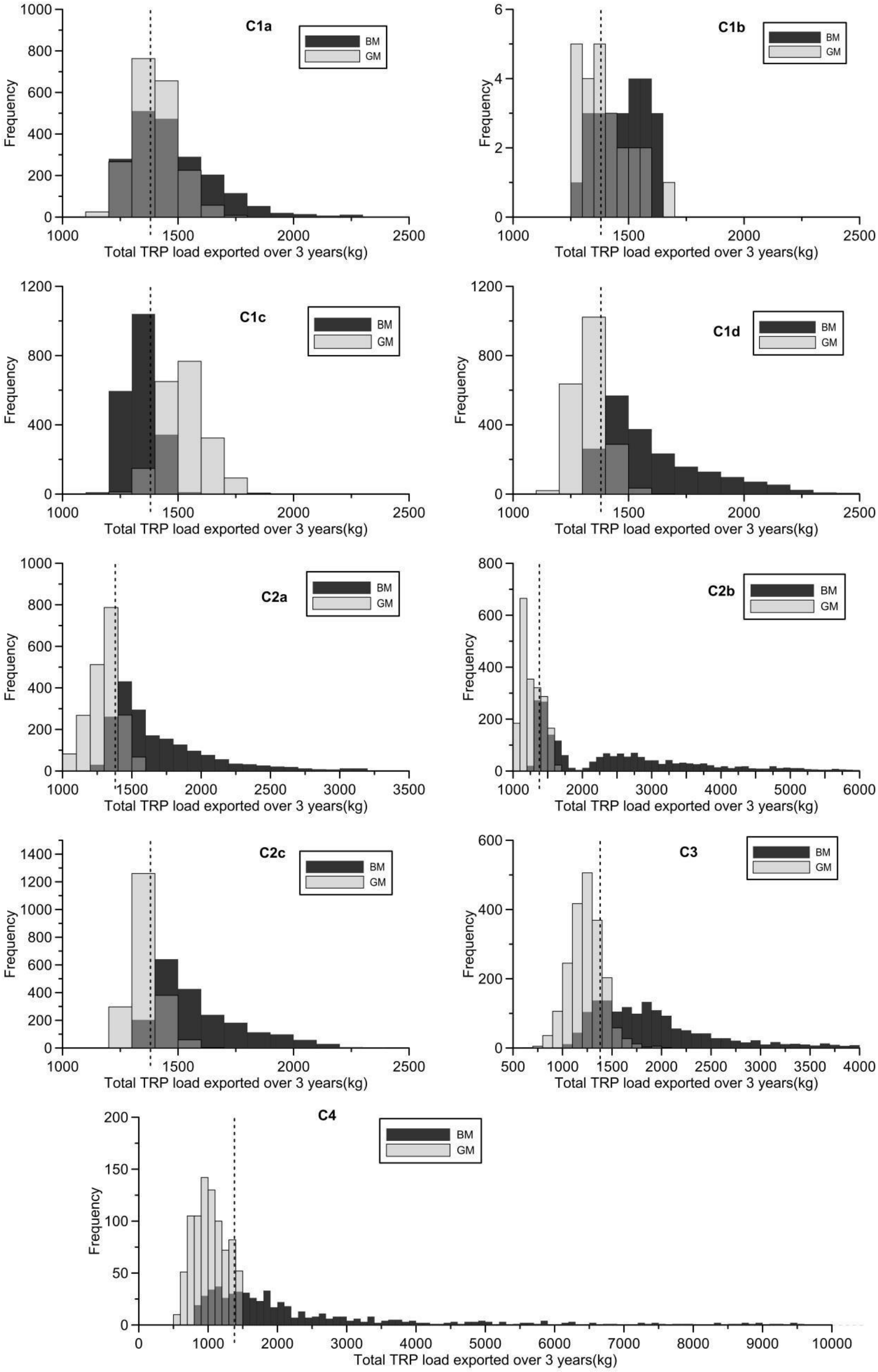

Fig. 6 Frequency distribution of the modelled total cumulative total reactive phosphorus (TRP) load for each sampling strategy, based on resampling 2000 times, C1a) daily (random), C1b) daily (specific time each day; n= 24). C1c) daily (random, 18.00-05.00), C1d) daily (random, Mon-Fri, 08.00-18.00), C2a) three times per week (random, Mon-Fri, 08.00-18.00), C2b) three times per week (random, Mon Tue Thu, 08.00-18.00), C2c) three times per week (random, Mon Wed Fri, 08.00-18.00), C3) weekly (random. Mon-Fri, 08.00-18.00), and 999 times C4) monthly (random, Mon-Fri. 08.00-18.00). Bin size was 100 for both LAMs (BM and GM), dashed line denotes total observed TRP load. Grey shading indicates histogram overlap (Grapher 9.0).

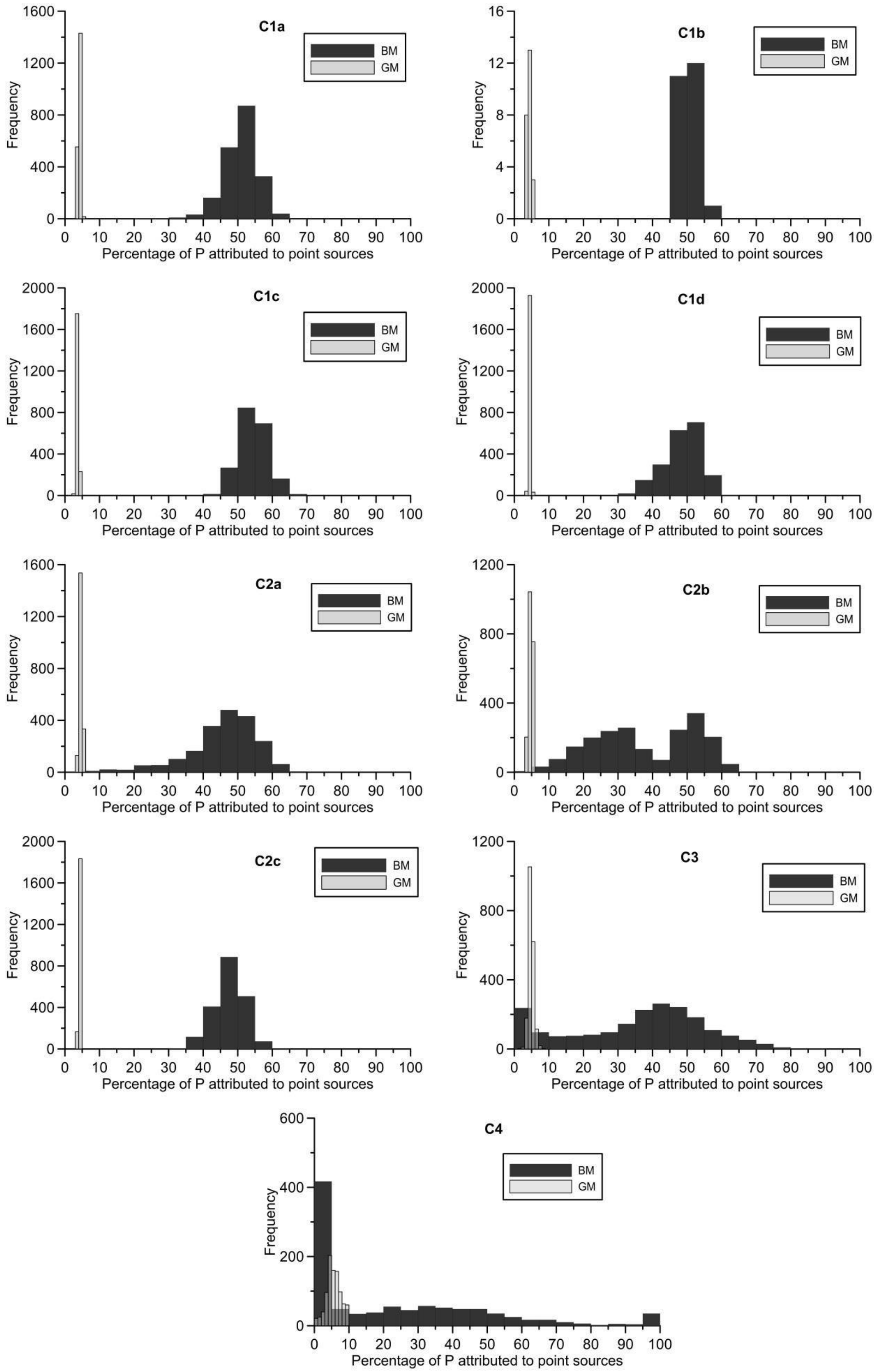

Fig. 7 Frequency distribution of the P load apportionment to point sources for each sampling strategy, based on resampling 2000 times, C1a) daily (random), C1b) daily (specific time each day; n= 24). C1c) daily (random, 18.00-05.00), C1d) daily (random, Mon-Fri, 08.00-18.00), C2a) three times per week (random, Mon-Fri, 08.00-18.00), C2b) three times per week (random, Mon Tue Thu, 08.00-18.00), C2c) three times per week (random, Mon Wed Fri, 08.00-18.00), C3) weekly (random, Mon-Fri, 08.00-18.00), and 999 times C4) monthly (random, Mon-Fri, 08.00-18.00). Bin size was 100 for both LAMs (BM and GM), dashed line denotes total observed TRP load. Grey shading indicates histogram overlap (Grapher 9.0).

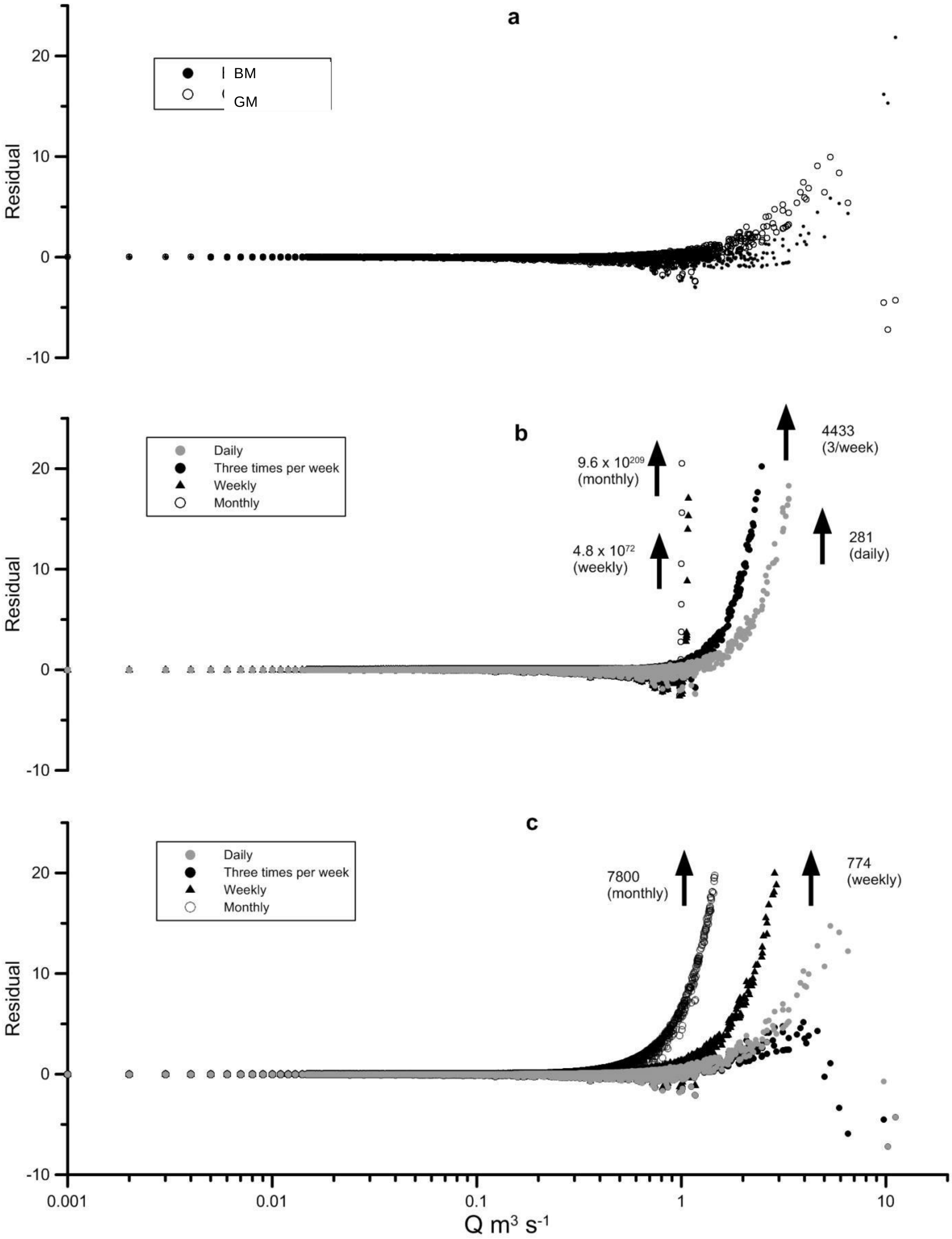

Fig. 8 Residuals between modelled line for each sampling strategy (using datasets with max total reactive phosphorus (TRP) modelled load) and observed TRP load, a) BM versus GM using hourly dataset, b) BM only and c) GM only. Values next to arrows show maximum residual obtained for sampling combination (Grapher 9.0).

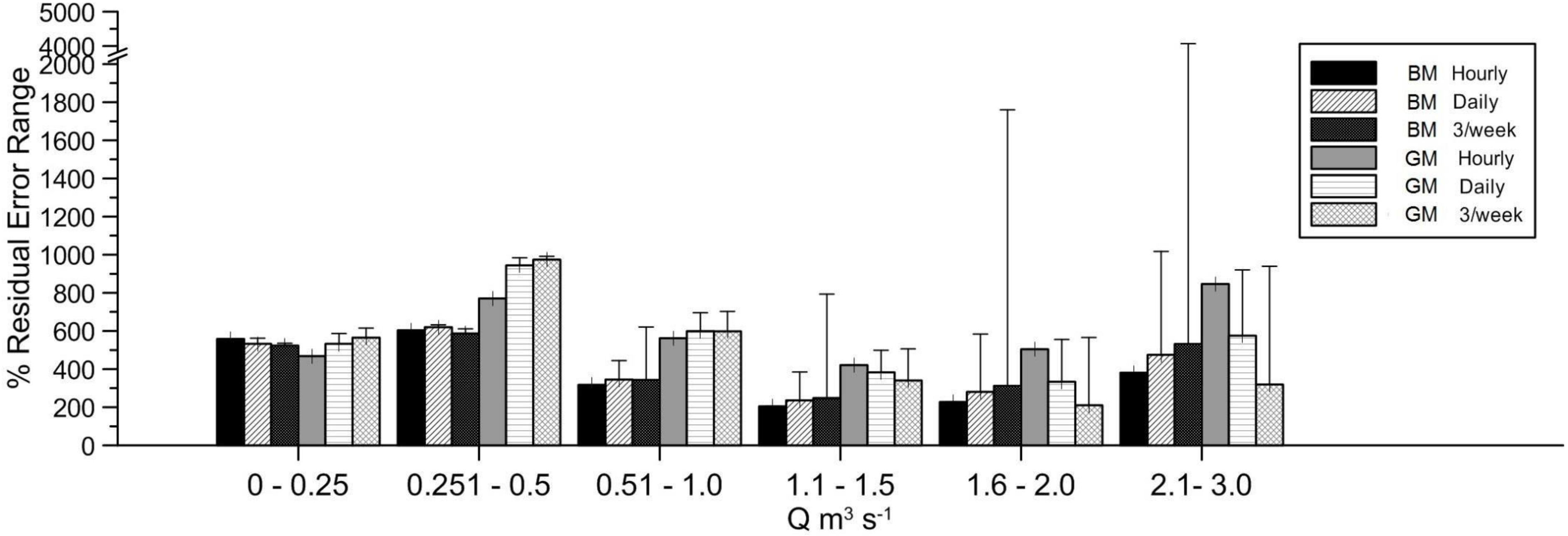

Fig. 9 Range of residuals as a percentage of observed total reactive phosphorus loads as Q increases. Error bars indicate the largest range between min and max values of % residual error for the particular model and sampling strategy (Grapher 9.0).

Table 1 Description and summary statistics of resampled datasets for each sampling combination

C	Sampling Scenario	No. of Datasets	No. of Datapoints per Dataset				
			Mean	SD	Median	Min	Max
C1a	Daily – Random in 24 hours, 7 days per week	2000	1069	0.00	1069	1069	1069
C1b	Daily (same hour each day)	24	1036	4.45	1036	1029	1044
C1c	Daily (Night) - Random 18.00 - 05.00	2000	1068	0.00	1068	1068	1068
C1d	Daily (Day) – Random Mon - Fri, 08.00-18.00	2000	769	0.00	769	769	769
C2a	Three days per week - random (Mon-Fri 08.00-18.00)	2000	461	1.44	462	459	463
C2b	Three days per week - Mon Tue Thu 08.00-18.00	2000	462	0.00	462	462	462
C2c	Three days per week - Mon Wed Fri 08.00-18.00	2000	460	0.00	460	460	460
C3	Weekly – Random, Mon-Fri 08.00-18.00	2000	157	0.00	157	157	157
C4	Monthly – Random, Mon-Fri 08.00-18.00	999	36	0.00	36	36	36

Table 2 Root mean square error (RMSE) for each combination dataset, using coefficients from individual datasets modelling the highest, lowest and median total cumulative total reactive phosphorus (TRP) load by the BM, when compared with observed load using high frequency Q data.

Sampling Strategy	Meta Data	Model	RMSE (mg s^{-1})		
			Highest Estimated TRP Load	Lowest Estimated TRP Load	Median Estimated TRP Load
C1a	Daily – Random	BM	5.90	0.12	0.33
		GM	0.31	0.14	0.14
C1b	Specific time each day	BM	0.47	0.12	0.30
		GM	0.30	0.14	0.13
C1c	Daily – Random (18.00-05.00)	BM	11.16	0.12	0.14
		GM	0.36	0.20	0.17
C1d	Daily – Random (Mon-Fri; 08.00-18.00)	BM	2.71	0.17	0.42
		GM	0.18	0.16	0.14
C2a	Three days per week - Random (Mon-Fri; 08.00-18.00)	BM	39.76	0.14	0.62
		GM	0.19	0.18	0.14
C2b	Three days per week (Mon, Tue, Thu; 08.00-18.00)	BM	62.55	0.21	1.91
		GM	0.19	0.18	0.16
C2c	Three days per week (Mon, Wed, Fri; 08.00-18.00)	BM	2.22	0.15	0.38
		GM	0.19	0.15	0.15
C3	Weekly – Random (Mon-Fri; 08.00-18.00)	BM	3.04E+70	0.14	0.96
		GM	7.37	0.19	0.16
C4	Monthly – Random (Mon-Fri; 08.00-18.00)	BM	1.36E+144	0.16	0.31
		GM	72.97	0.19	0.17