


A method for uncertainty constraint of catchment discharge and phosphorus load estimates

Michael J. Hollaway¹  | Keith J. Beven¹ | Clare Mc W. H. Benskin¹ | Adrian L. Collins² | Robert Evans³ | Peter D. Falloon⁴ | Kirsty J. Forber¹ | Kevin M. Hiscock⁵ | Ron Kahana⁴ | Christopher J. A. Macleod⁶ | Mary C. Ockenden¹ | Martha L. Villamizar⁷ | Catherine Wearing¹ | Paul J. A. Withers¹ | Jian G. Zhou⁸ | Nicholas J. Barber⁹ | Philip M. Haygarth¹

¹Lancaster Environment Centre, Lancaster University, Lancaster, UK

²North Wyke, Rothamsted Research, Okehampton, UK

³Global Sustainability Institute, Anglia Ruskin University, Cambridge, UK

⁴Met Office Hadley Centre, Exeter, UK

⁵School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich, UK

⁶James Hutton Institute, Aberdeen, UK

⁷School of Engineering, Liverpool University, Liverpool, UK

⁸School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester, UK

⁹Geography Department, Durham University, Durham, UK

Correspondence

Michael J. Hollaway, Lancaster Environment Centre, Lancaster University, Bailrigg, Lancaster LA1 4YQ, UK.
Email: m.hollaway@lancaster.ac.uk

Funding information

Natural Environment Research Council (NERC), Grant/Award Numbers: NE/K002406/1, NE/K002430/1 and NE/K002392/1; Department for Environment, Food and Rural Affairs (Defra), Grant/Award Numbers: LM0304, WQ0212, WQ0211 and WQ0210; Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme, Grant/Award Number: GA01101

Abstract

River discharge and nutrient measurements are subject to aleatory and epistemic uncertainties. In this study, we present a novel method for estimating these uncertainties in colocated discharge and phosphorus (P) measurements. The “voting point”-based method constrains the derived stage-discharge rating curve both on the fit to available gaugings and to the catchment water balance. This helps reduce the uncertainty beyond the range of available gaugings and during out of bank situations. In the example presented here, for the top 5% of flows, uncertainties are shown to be 139% using a traditional power law fit, compared with 40% when using our updated “voting point” method. Furthermore, the method is extended to in situ and lab analysed nutrient concentration data pairings, with lower uncertainties (81%) shown for high concentrations (top 5%) than when a traditional regression is applied (102%). Overall, for both discharge and nutrient data, the method presented goes some way to accounting for epistemic uncertainties associated with nonstationary physical characteristics of the monitoring site.

KEYWORDS

discharge, epistemic and aleatory uncertainty, nutrient load, rating curve, voting point

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2018 John Wiley & Sons, Ltd.

1 | INTRODUCTION

Effective evaluation of process-based water quality models requires an understanding of the uncertainties in the observational data used in calibration processes, as estimates of catchment discharge and nutrient loads are affected by significant uncertainties (Beven, Buytaert, & Smith, 2012; Coxon et al., 2015; Harmel, Cooper, Slade, Haney, & Arnold, 2006; Harmel, Smith, King, & Slade, 2009; Johnes, 2007; McMillan, Krueger, & Freer, 2012; Westerberg, Guerrero, Seibert, Beven, & Halldin, 2011). In some cases, the uncertainties may be such that for some events the observational data may not be useful for model calibration and evaluation (Beven & Smith, 2015; Beven, Smith, & Wood, 2011; Beven & Westerberg, 2011).

Continuous river discharge measurements are often obtained by observing the river water level (stage), and then using a fitted curve (rating curve) to convert these to an estimated discharge. A rating curve is a model of the relationship between observed stage and discharge for a gauging site. As a result, uncertainties in the resultant discharge data can come from errors in river stage measurement, errors in the gauged discharges, lack of gauged data over parts of the curve (particularly the higher end), uncertainties that arise from the fitting of the rating curve itself (e.g., structural error in the model used), and changes in the rating curve over time (McMillan & Westerberg, 2015). When considering water quality data, nutrient loads are calculated for a specific period (typically daily) using river discharge along with measurements of concentrations of the nutrient of interest (e.g., phosphorus [P]). Therefore, uncertainties in load estimations arise from uncertainties in the discharge estimates and in the sampling and measurement of determinand concentrations in addition to their aggregation to the temporal and spatial scales of interest (McMillan et al., 2012).

In hydrology, all important uncertainties can be considered to be epistemic in nature: that is, they arise from a lack of knowledge of the key underlying processes (e.g., Beven, 2016; Nearing et al., 2016). However, we may choose to treat some uncertainties as aleatory (i.e., they arise from simple random variability). Typically, measurement errors in variables such as stage or nutrient concentrations are treated as aleatory variables. In contrast, epistemic uncertainties can include changes to the river cross section, vegetation growth, the effect of sampling and analysis protocols on concentration measurements, and the choice of a functional form for the rating curve; all of which can affect subsequent estimates of discharges and nutrient loads.

Many previous studies have attempted to estimate the uncertainties in both discharge and water quality measurements using a wide range of techniques (Harmel et al., 2009; Johnes, 2007; Moatar & Meybeck, 2005; Webb et al., 1997). For discharge, it is common to fit a statistical regression model to the available stage and discharge gaugings, which allows a statistical estimate of uncertainty in the rating curve. Simple power law or polynomial functions have often been applied, or multisegment functions where the rating curve appears to show a complex shape (e.g., Herschy, 1999). There are, however, alternatives, including drawing on fuzzy regression and fuzzy set concepts (Blazkova & Beven, 2009; Krueger et al., 2010; Pappenberger et al., 2006) and nonparametric regression techniques (e.g., LOWESS, Cleveland, 1979; Coxon et al., 2015), which have been employed on stage-discharge measurements and water quality

variables to estimate the uncertainties in discharge and nutrient concentrations and load calculations (Lloyd, Freer, Johnes, Coxon, & Collins, 2016). A further method, focused on uncertainty in the rating curve, has been suggested by McMillan and Westerberg (2015). Their voting point method allowed for situations where channel form and velocities might change over time so that many candidate rating curves might be plausible.

In this study, we extend the voting point method to use water balance data to constrain rating curve uncertainties and also apply the voting point method in the estimation of continuous nutrient concentrations and loads. Further to this, we also place our results in context with other uncertainty techniques by comparing with estimates from using more traditional methods (e.g., fitting a power law function to the available observations).

2 | METHODS

2.1 | Study area

Newby Beck is a small headwater subcatchment located in the River Eden basin in the North West of England, in the United Kingdom. The catchment is approximately 12.5 km² in size with an average elevation of 234 m above sea level. The discharge measurement site for this catchment is a rated section of channel, with water level data collected (with a Schlumberger Water Services (SWS) Mini-Diver) at 15-min intervals. As shown in Figure 1, the cross-sectional area of the rated section of the channel changes significantly at higher water levels, which could contribute to the epistemic uncertainties associated with the discharge produced by the rating curve.

Fourteen discharge measurements were available to develop a site specific rating curve. In addition, a high frequency bankside monitoring station was situated at the outlet, which recorded nitrate (NO₃), total P (TP), and total reactive P (TRP) at 30-min intervals (Outram et al., 2014). The TP and TRP measurements were conducted using a Hach Lange combined Sigmatax sampling module and Phosphax Sigma analyser (Perks et al., 2015). Rainfall over the catchment was recorded at 15-

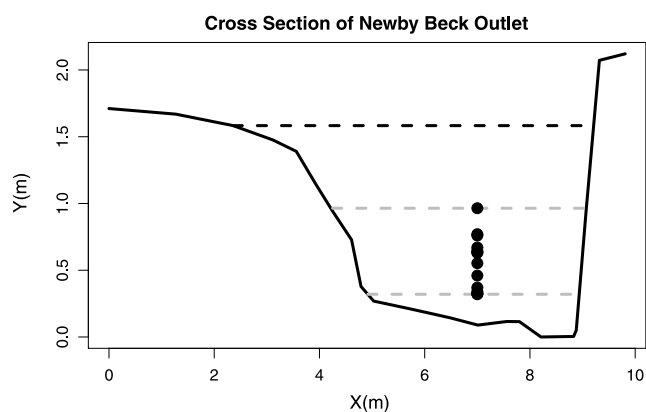


FIGURE 1 Cross section geometry of Newby Beck outlet. The black dots show the heights of each of the 14 available gaugings. The grey dashed lines show the channel cross section at the top and bottom of the gauged range, and the black dashed line highlights the change in channel cross section at high flows well beyond the gauged range

min intervals by three tipping bucket rain gauges, and daily rainfall data were obtained from a gauge in the centre of the catchment from the Met Office Integrated Data Archive System network (Met Office, 2012). Other meteorological data were provided by an automatic weather station, which was located towards the centre of the catchment. For this study, the data were analysed over three hydrological years (2011–2012, 2012–2013, and 2013–2014).

2.2 | Constraining uncertainty at high flows using water balance information

As applied by McMillan and Westerberg (2015) rating curves are typically fitted using power laws or segmented power laws. Typically, when presented with limited numbers of river gaugings (in this application only 14 in-bank flow measurements were available), the majority of the data falls at the lower flows, and therefore, extrapolation is required beyond the end of the gauged range. With the power law, this extrapolation often introduces large uncertainties, particularly for overbank flow. At the Newby Beck outlet, very few of the gaugings are at the higher flow values (Figure 1). To constrain the extrapolation beyond the gauged range in such cases, we have incorporated the Velocity Area Rating Extension (VARE) model of Ewen, Geris, O'Donnell, Mayes, and O'Connell (2010). The advantage of the VARE method is that local knowledge of the gauging site, such as river cross-sectional area (Figure 1) and water balance estimates, can be used to constrain this extrapolation beyond the gauged range by imposing a maximum velocity that can be achieved in the river channel. In VARE, a sigmoidal function (G , Equation 2) that varies between two limits (the maximum and minimum stream velocities, v_{\max} and v_{\min}) is used to determine the average velocity in the stream for a given stage measurement:

$$X = \text{MIN}\left(1, \frac{y - y_{\min}}{y_{\max} - y_{\min}}\right), \quad (1)$$

$$G = \frac{1}{2} \left[1 + \frac{\tanh(2\alpha X^\beta - \alpha)}{\tanh(\alpha)} \right], \quad (2)$$

$$v = v_{\min} + (v_{\max} - v_{\min}) \cdot G \quad (3)$$

where y is the measured stage, y_{\min} is the minimum stage, y_{\max} is the maximum stage, α and β are parameters that control the sigmoidal function, and v is the VARE calculated velocity. The velocity can then be used with the cross-sectional area of the stream at stage y to determine the discharge (Ewen et al., 2010). In this case, we assume that the velocity is zero at the bottom of the channel (therefore giving us the values of y_{\min} and v_{\min}), and thus need to calibrate the remaining four parameters of the VARE model (α , β , y_{\max} , and v_{\max}). Furthermore, the VARE model can be calibrated over a long period (in this case, three hydrological years), such that a water balance is approximately satisfied, allowing for uncertainty in both rainfall and actual evapotranspiration estimates (see below). To explore the rating curve uncertainty, a Monte Carlo analysis was run using the VARE rating curve model; 2,000 parameter sets of the four VARE parameters were sampled using random uniform sampling and evaluated using an extended voting point method.

2.3 | The extended voting point method

In the voting point method of McMillan and Westerberg (2015), randomly generated rating curves are assessed using an informal likelihood measure based on how close each curve falls to each of the available discharge–water level pairs. A logistic function was used to account for error in the gauging measurements, although they suggest this can be replaced with a function of the modeller's choosing. In this study, we replace the logistic function with a triangular fuzzy weighting measure, which uses Equations 4 and 5 to calculate a normalized score ($Score(g)$ in Equation [4]) and weight ($W(g)$ in Equation [5]) at each of the 14 available gauging points (see Figure 2).

$$Score(g) = \begin{cases} (\hat{Y}_g - y_g) / (y_g - y_{\min,g}) & \hat{Y}_g < y_g \\ (\hat{Y}_g - y_g) / (y_{\max,g} - y_g) & \hat{Y}_g \geq y_g \end{cases} \quad (4)$$

$$W(g) \begin{cases} [(Score(g) - L_{lwr}) / \text{abs}(L_{lwr})]^N & L_{lwr} \leq Score(g) < 0 \\ [(L_{upr} - Score(g)) / \text{abs}(L_{upr})]^N & 0 \leq Score(g) < L_{upr} \\ 0 & Score(g) \notin (L_{lwr}, L_{upr}) \end{cases} \quad (5)$$

Here, \hat{Y}_g is the rating curve estimated value of discharge; y_g is the gauged discharge value; $y_{\min,g}$ is the lower limit of error (see below); and $y_{\max,g}$ is the upper limit of error for a given gauging point. This therefore gives a score of zero for a value at the best estimate of the observed value, -1 at the lower limit and $+1$ at the upper limit. If the normalized score lies between -1 and $+1$ for a given gauging, a triangular fuzzy weight ($W(g)$, Figure 2) is calculated for the gauging point g in Equation 5. Here, L_{lwr} and L_{upr} are the lower and upper limits of the normalized scores required to consider the simulated values acceptable across all the gauged points (in this case -1 and 1); and N is a shaping factor (set to 1 in this case).

The method requires that the limits $y_{\min,g}$ and $y_{\max,g}$ can be specified for each gauging point. This information is not usually available for single gauging points but typical uncertainties of $\pm 10\%$ for in-bank flows have been determined, for example, by Schmidt and Yen (2008); Krueger et al. (2010); McMillan et al. (2012). A rating curve model was then considered behavioural based on the constraint that it returned a normalized score of between 0 and 1 for at least one gauging point. Any behavioural parameter sets from the 2,000 sampled are assigned an overall voting point likelihood weighting (L_{vp}) based on Equation 6:

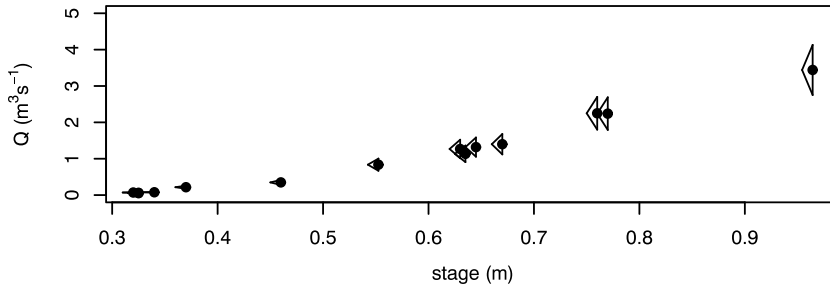
$$L_{vp} \propto w_{vp} \sum_{g=1}^n W(g), \quad (6)$$

where $W(g)$ is the weight at gauging g , and w_{vp} is the voting point weighting based on the number of gaugings the candidate curve passes through. The voting point weighting is calculated as follows:

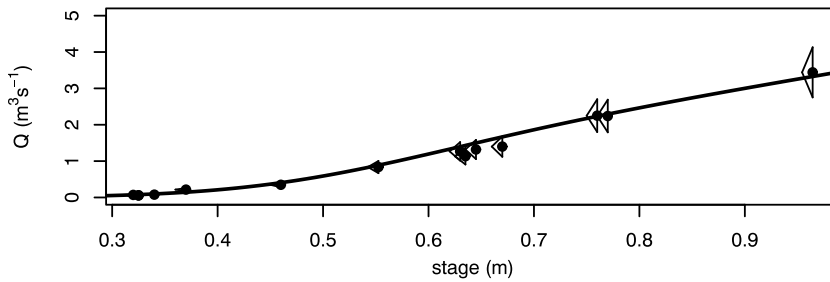
$$w_{vp} = \left(\frac{\max(hfit) - \min(hfit)}{\max(h) - \min(h)} \right) \cdot \left(\frac{\max(qfit) - \min(qfit)}{\max(q) - \min(q)} \right), \quad (7)$$

where h and q are the gauged stage and flow values; $hfit$ and $qfit$ are the subsets that are intersected by the candidate curve. As discussed by McMillan and Westerberg (2015), w_{vp} is a weighting based on the

(a) Gaugings and triangular fuzzy errors



(b) Weight sample curve by sum of hits and water balance



(c) Continue Monte-Carlo sampling of candidate curves

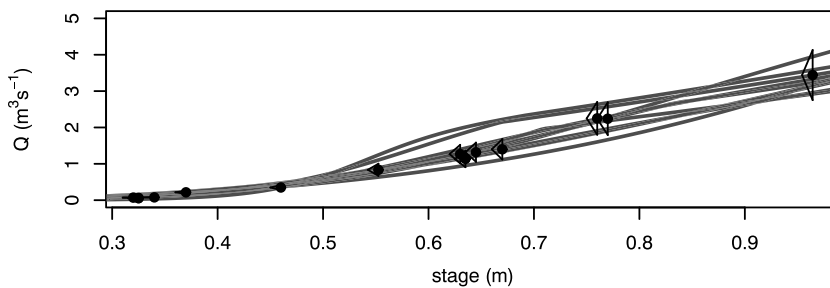


FIGURE 2 Schematic of the voting point method for discharge. (a) Errors on discharge measurements are estimated as $\pm 10\%$ of the gauged value, and defined using triangular fuzzy weighting (see text). (b) A candidate Velocity Area Rating Extension sigmoidal rating curve is sampled, and the number of gauging points the curve passes through is counted. The number of hits, along with a test of water balance satisfaction (not shown), allows the voting point likelihood of that curve to be calculated. (c) Further, candidate curves are selected using Monte Carlo sampling until a predetermined number of curves have been sampled. The 95% prediction uncertainty (95PPU) bounds on the resultant discharge time series are defined on the basis of the number of behavioural curves and their associated voting point weightings

space (area) of gauging points that the candidate curve spans. This is to avoid situations where the distribution of gaugings is highly skewed (in this case, towards lower stages), which can lead to divergence from the gauging points, particularly at the top and bottom ends of the stage range.

In addition to this, a further constraint was imposed on each of the 2,000 candidate curves, in that the modelled mass balance was required to fall within a particular tolerance of the observed value as calculated from the rainfall and weather station data. The water balance was determined by comparing the total discharge estimated using the candidate rating curve to the total rainfall minus the estimated evapotranspiration (estimated using the FAO Penman-Monteith equation (Allen, Pereira, Raes, & Smith, 1998) for a crop representative of improved grassland, from data measured by the automatic weather station) during the calibration period (2011–2012, 2012–2013, and 2013–2014 hydrological years). It is assumed that the change in storage over this time is negligible relative to other uncertainties. To allow for errors in the water balance calculation arising from both the estimates of rainfall over the catchment area, the evapotranspiration estimate and change in catchment storage, rating curves were accepted if they gave estimated discharges within 10% of the water balance estimate. A normalized score ($Score_{mb}$) was calculated using Equation 4; allowing $\pm 10\%$ tolerance

on the water balance) and if this fell between -1 and 1 , a likelihood weight (L_{mb}) was calculated as follows:

$$L_{mb} \begin{cases} [(Score_{mb} - MB_{lwr}) / abs(MB_{lwr})]^N & MB_{lwr} \leq Score_{mb} < 0 \\ [(MB_{upr} - Score_{mb}) / abs(MB_{upr})]^N & 0 \leq Score_{mb} < MB_{upr} \\ 0 & Score_{mb} \notin (MB_{lwr}, MB_{upr}) \end{cases} \quad (8)$$

If a candidate rating curve was classed as behavioural (likelihoods > 0) using both the mass balance and the voting point criteria described above, an overall likelihood weighting for each behavioural candidate curve was calculated as follows:

$$L_{VARE} = \frac{L_{vp} \cdot L_{mb}}{C}, \quad (9)$$

where L_{VARE} is the overall weighting, L_{vp} is the likelihood measure calculated for the voting point fit to gaugings, L_{mb} is the likelihood measure for the mass balance criteria, and C is a scaling factor such that the sum of likelihoods scales to unity in each case. If either evaluation measure returned a likelihood of zero (L_{vp} or L_{mb}), then according to Equation 9, the overall likelihood (L_{VARE}) is also zero and the candidate curve is classed as nonbehavioural and plays no further role in the analysis. Once a set of behavioural models has been obtained, prediction quantiles can be formulated at a given point on the curve (i) as follows:

$$P(\hat{Z}_i < z_i) = \sum_{j=1}^{j=N} L[M(\Theta_j) | \hat{Z}_{i,j} < z_i]. \quad (10)$$

Here, P is the prediction quantile for \hat{Z}_i (the simulated value of variable Z at point i using candidate curve $M(\Theta_j)$) being less than z ; L is the likelihood weighting associated with candidate curve $M(\Theta_j)$; Θ_j is the j th parameter set; and N is the number of candidates accepted as behavioural. We then define the 95 percent prediction uncertainty (PPU) limits on the estimated discharge from the 2.5 and 97.5 quantiles derived from Equation 10. The upper and lower limits of uncertainty on the resultant discharge time series were based around the 2.5 and 97.5 percentiles, respectively (95PPU limits). The 50th percentile (median) was defined as the best estimate of the observed discharge.

2.4 | Extension of voting point method to nutrient data

Uncertainty in calculated nutrient loads results from both the discharge uncertainty, the concentrations measurements themselves, and the cross-sectional and temporal variability. To evaluate the uncertainty in nutrient concentrations from the bank-side analyser, in situ measurements from the instrument were paired with laboratory analysed spot and ISCO samples taken at a corresponding time. An empirical power law relationship was then fitted using Monte Carlo analysis to identify behavioural parameter sets. A sample of 2,000 parameter sets (of the power law) was tested, and the modified voting point method was used to assign likelihood weightings to each proposed parameter set. The laboratory analysed sample was assumed to be the best estimate of the true concentration, and the deviation between the in situ measurement and this value was used to define the unit normalized limits for calculation of evaluation scores. In this case, the evaluation scores were calculated using the approach outlined in Equation 4), and the overall weighting of each candidate curve was calculated on the basis of the

number of measurement pairs intersected (following a similar approach to Equations 5–6). The method will be demonstrated for the case of TP concentrations and loads.

The unique combinations of the behavioural discharge and TP concentration time series from the voting point analysis were then used to determine TP loads using Equation 11:

$$Load = \sum_{i=1}^n Q_i C_i, \quad (11)$$

where Q_i is the discharge at time i , C_i is the concentration, and n is the number of measurement time steps in a day. Any day with missing data was excluded from the model evaluation. As with discharge, prediction quantiles were calculated at each time step, with the combined final likelihood weight of each behavioural model determined as follows:

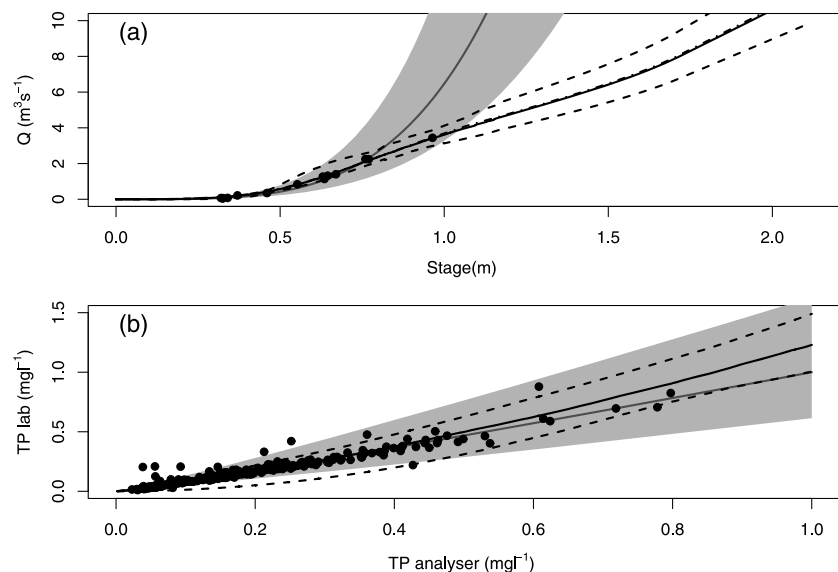
$$L_{load} = \frac{L_{VARE} \cdot L_{conc}}{C}, \quad (12)$$

where L_{load} is the overall likelihood of each TP load time series, L_{VARE} is the likelihood weighting of each behavioural parameter set from the rating curve uncertainty analysis, and L_{conc} is the likelihood weighting from the concentration uncertainty analysis. C is a scaling factor, such that the sum of likelihoods scale to unity in each case. As with discharge, the upper and lower limits of uncertainty on the resultant load time series were based around the 2.5 and 97.5 percentiles, respectively (95PPU limits). The 50th percentile (median) was defined as the best estimate of the observed in-stream load.

3 | RESULTS

Figure 3 shows the uncertainty limits calculated for discharge (Figure 3a) and TP concentrations (Figure 3b). Overall, the uncertainty interval (based on 95% prediction quantiles) on the discharge measurements was, on average, 70% throughout the duration of the

FIGURE 3 (a) Rating curve at the Newby Beck outlet as estimated using the Velocity Area Rating Extension method. Solid line shows curve with best fit to gaugings, large dashed lines show 95% prediction bounds, and black dots show the gaugings. The dashed and dotted line shows the official rating curve. The solid dark grey line shows a standard power law fitted with regression and the grey shading shows the 95% prediction intervals from the regression analysis. (b) Rating between total phosphorus (TP) concentration as measured using the bank-side analyser and corresponding samples analysed in the lab. The solid line shows the best fit to the lab analysed data, and the dashed lines show the 95% prediction bounds. The black dots show the pairs of TP concentrations from the analyser and the lab. The solid dark grey line shows a standard power law fitted with regression and the grey shading shows the 95% prediction intervals from the regression analysis



calibration period, with a range of 21–140%. The higher relative uncertainty intervals were seen in the low flow periods (here defined as the lowest 5% of discharges, which equates to values $<0.032 \text{ m}^3 \text{ s}^{-1}$), where they were on average 128%. However, this equated to a mean absolute uncertainty interval of $0.032 \text{ m}^3 \text{ s}^{-1}$. In contrast, the high flow periods (here defined as the highest 5% of discharges which equates to values greater than $1.22 \text{ m}^3 \text{ s}^{-1}$) had much smaller relative uncertainty intervals, on average 40%. This range is much larger compared with those determined during a recent study on 500 UK catchments (Coxon et al., 2015), which showed that the majority of catchments had 20–40% relative uncertainty intervals, though the maximum uncertainty of 140% determined for Newby Beck here is much lower than the maximum value of 397% quoted by Coxon et al. (2015).

Figure 3a also shows a comparison with a rating curve generated for this catchment using the traditional power law (fitted using regression). The power law (solid grey line in Figure 3a) gives much higher values at the high end of the rating than when the water balance constraint is imposed for the VARE method (solid black line in Figure 3a).

Furthermore, outside the range of the available gaugings, the uncertainty (95% prediction intervals from the regression) in the rating curve (grey shading) is much larger than the curve generated from the voting point method (large dashed black lines). The power law regression gives 157% uncertainty on discharge for the high flows (top 5%), compared with 40% when using the VARE voting method. For the low flow (bottom 5%), both methods produce similar uncertainties, with the power law regression showing slightly higher average uncertainties at 139%, compared with 128% from the VARE voting method.

The uncertainty intervals (based on 95% prediction quantiles from the fitted empirical power law), generated from the comparison between the continuous bank-side analyser data and the lab analysed samples, showed a similar pattern with the lowest 5% of concentrations (those below 0.0049 mg L^{-1}) showing the highest relative uncertainty intervals (on average 231%). For the higher concentrations (the top 5%; 0.179 mg L^{-1}), the intervals were smaller, at around 81%. The TP concentration and discharge uncertainties are reflected in the TP load calculations, which see a relative interval of on average 292% for the lowest loads (bottom 5%) and 74% for the highest loads (top

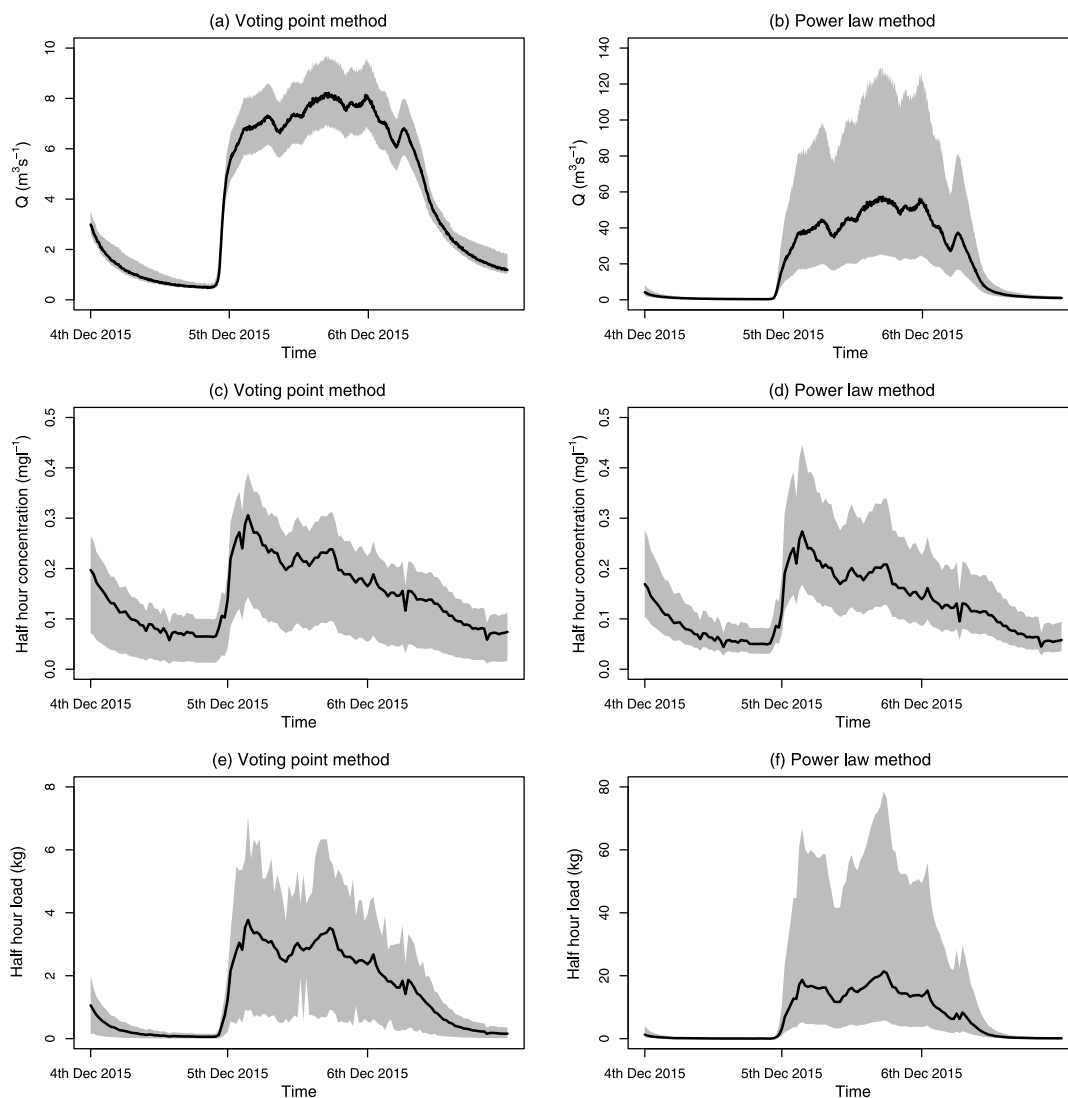


FIGURE 4 Time series of stream discharge, half-hour total phosphorus concentration and half-hour TP load during Storm Desmond (December 5–6, 2015) for the voting point method (a, c, and e) and the power law method (b, d, and f). The black line shows the median value, and the grey shading shows the 95% uncertainty limits derived for both methods. Note the difference in scale on the y axis for the power law method

5%). Overall, these intervals are larger than those reported by McMillan et al. (2012), who provided a summary of uncertainties in water quality data showing relative errors of up to 150% on TP loads and concentrations. However, recent work by Lloyd et al. (2016), employing the use of a bank-side analyser similar to that used at Newby Beck, resulted in uncertainties of up to 83% on the estimation of TP loads when compared with laboratory analysed data.

Figure 3b shows the relationship between the bank-side analyser and laboratory TP concentrations as predicted using a power law fitted using standard regression (solid grey line shows fit, and grey shading shows 95% prediction intervals). As with the discharge, the uncertainty intervals at the higher concentrations (top 5%) were much greater using the regression (102%) than with the voting point method (81%). For the lower concentrations (bottom 5%), however, the regression tended to show much lower uncertainties at 103% compared with the voting point method (231%). Note that because none of the rating curves using either the functional form (Equation 2) or power law can have negative values, these large uncertainty values indicate that the distribution of estimated values at any particular flow or load must be skewed.

4 | DISCUSSION AND CONCLUSIONS

This scientific briefing presents a new approach to the estimation of uncertainty in rating curves applied to discharge and water quality measurements. This method builds upon a modified voting point method (McMillan & Westerberg, 2015) combined with the VARE model of Ewen et al. (2010). This helps constrain the maximum discharge, particularly in situations where the river is likely to go out of bank. This is demonstrated in Figure 4, which shows a comparison of the river discharge from VARE and the power law methods (Figure 4 a,b) during Storm Desmond (December 5–6, 2015), where there was widespread flooding and out of bank flow. Using VARE, the discharge peaks at $8.2 \text{ m}^3 \text{ s}^{-1}$ with an uncertainty range of $7.0\text{--}9.7 \text{ m}^3 \text{ s}^{-1}$ (Figure 4a). If the power law method is used, we are well beyond the gauged range (Figure 3a). The maximum discharge during Desmond was $57.4 \text{ m}^3 \text{ s}^{-1}$ with a much larger uncertainty range of $25.4\text{--}129.5 \text{ m}^3 \text{ s}^{-1}$ (Figure 4b). Therefore, the use of the VARE and voting method allows the modeller to constrain the uncertainty using local knowledge of the catchment.

Furthermore, VARE allows the hydrologist to impose a (uncertain) mass balance constraint on the evaluation of candidate rating curves using available weather data over a long period (three hydrological years in this application to Newby Beck). This, therefore, ensures that the rating curve model is consistent with the catchment water balance (see Beven and Smith (2015), for example, where this is not the case in another catchment). However, it is acknowledged here that the uncertainty in the mass balance calculation is dependent on the accuracy in the available weather data and consequent precipitation and evapotranspiration estimates on which to perform the analysis.

The advantage of the VARE method in the voting point framework is that the weighting imposed on the overall likelihood of a candidate model can be stronger towards either the fit to the gaugings or the mass balance (e.g., a multiplier can be added to each likelihood in

Equation 9 when calculating the overall likelihood for a candidate curve, L_{VARE}). The weighting towards either constraint can be split evenly or allowed to give preference to one of the measures depending on the model user, knowledge of the catchment, the available data to calculate mass balance, and the nature of the application the model user wants to use the model for.

As we have demonstrated in this work (Figures 3 and 4), the downside of using the power law method to fit rating curves is often the lack of available gaugings during high flow periods. Therefore, when extrapolating the curve beyond the gauged range, there is the potential for overestimation at the higher end of the curve (Figure 3). In effect the power law does not take account of the rapid change in cross-sectional area and consequent decrease in average velocities that often arises in overbank flows. Hydraulic modelling can go some way to reducing such errors in the extrapolation of the rating curve, but then requires specific assumptions about changes in roughness coefficients or conveyance. In our case, the VARE approach avoids this by imposing hydrological consistency through the uncertain mass balance constraint. This reduces the uncertainty when extrapolating the curve beyond the gauged range, as shown in Figure 3.

There are other epistemic uncertainties that can lead to nonstationarities in rating curves that are not always obvious. For example, during a flood there can be changes in the physical cross section of the channel due to erosion or sediment build up (Lang, Pobanz, Renard, Renouf, & Sauquet, 2010). This can alter the stage-discharge relationship from any single calibrated curve. Using the voting point method in combination with the VARE approach aims to reduce this uncertainty by assuming that each of our 14 gaugings are representative of a given rating curve at the time of measurement. Therefore, our condition of any candidate curve only needing to hit one gauging to be classed as behavioural aims to account for any potential variation in the rating curve with time.

We also present an extension of the voting point method to account for uncertainties in our P observations and the translation of these errors through to the estimates of daily P loads. As most water quality models typically work on a mass balance basis, the focus is on uncertainties in the observed load data. As load data are calculated using the combination of discharge and concentration, the errors in both measurements must be accounted for.

Therefore, the error in the load measurement (for this particular dataset) will be a combination of rating curve uncertainty, procedural and instrument error in the measurement of nutrient concentrations (in this case P), and cross-sectional variation. Previous methods to estimate load uncertainty (Johnes, 2007; Lloyd et al., 2016) provide some estimation of this combination of errors. However, the discharge error is based on the aforementioned power law rating curve fitted using methods such as LOWESS. Therefore, these methods are susceptible to the issues of extrapolation beyond the range of the gauging data. Our application of the VARE method to estimate the discharge component of the load calculation accounts for this issue as discussed above.

For the concentration errors, we have employed similar methods to those used previously, whereby the bank-side analyser data are compared with those generated in a lab, to check for inconsistencies

in the measurements. However, the previous methods tend to quantify the relationship between these data using a regression analysis or LOWESS that requires a fit to all data pairings. As with discharge data, epistemic errors in nutrient data can arise due to changes in the monitoring equipment, such as instrument drift in the bank-side analyser data over time. Therefore, the relationship between laboratory data (which is often generated infrequently, such as with gauging data) and the in situ data may shift. Therefore, to account for these epistemic errors, we utilized the voting point method to estimate the uncertainty in our bank-side analyser data, assuming the lab data were the best estimate of the true measurement.

Overall, the uncertainties in concentrations at Newby Beck (~231% for the lowest 5% of concentrations and ~81% for the highest 5% of concentrations) and loads (relative interval of on average 292% for the lowest loads [bottom 5%] and 74% for the highest loads [top 5%]) were similar to those reported by previous studies (Lloyd et al., 2016; McMillan et al., 2012). However, we tend to show higher relative uncertainties towards the lower end of the range.

Again, when the stream went out of bank during Storm Desmond (Figure 4), the application of the extended voting point method led to more constrained uncertainties on TP load. The maximum half hourly TP load using the voting point method was 3.7 kg with an uncertainty range of 0.9–7.1 kg (Figure 4e). With the power law method, the estimated load was much higher at 21.4 kg with an uncertainty range of 5.8–78.6 kg (Figure 4f). As shown in Figure 4c,d, both the voting point method and the power law method produce similar uncertainty estimates on the TP concentrations, with the voting point method tending to show slightly higher uncertainties towards the lower concentrations (as abovementioned). Therefore, during Storm Desmond, the higher levels of uncertainty exhibited for TP loads when using the power law method are most likely as a result of the large errors shown at the higher end of the rating curve. Our combined VARE and voting point method approach significantly constrains this load uncertainty (Figure 4e) and consequently the estimate of the total load from the catchment integrated over time because of the importance of the high flow events in P export.

As the computational cost of running this procedure is relatively cheap, and as more gauging information or additional data regarding the characteristics of the catchment become available, the rating curve information or the empirical relationship between the lab and in situ P measurements can be updated easily. This will allow further constraints on the estimation of uncertainties in the discharge, nutrient concentrations, and estimated loads. These uncertainties can then be used as limits of acceptability in the evaluation of water quality models as demonstrated by Holloway et al. (2018).

ACKNOWLEDGMENTS

This study was funded by the Natural Environment Research Council (NERC) as part of the NUTCAT 2050 project, Grants NE/K002392/1, NE/K002430/1, and NE/K002406/1. This work was supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra. The authors are grateful to the Eden Demonstration Test Catchment (Eden DTC) research platform for provision of the field

data (Department for Environment, Food and Rural Affairs (Defra), Projects WQ0210, WQ0211, WQ0212, and LM0304). The contribution of ALC was funded via BBSRC grant award I03303 - Soil to Nutrition-sustainable intensification-optimisation at multiple scales. The data used in this study are openly available from the Lancaster University data archive (<http://dx.doi.org/10.17635/lancaster/researchdata/227>). The DTC data are available from the Eden DTC consortium until the data archive is transferred to Defra (Department for Environment, Food & Rural Affairs) as the holding body.

ORCID

Michael J. Holloway  <http://orcid.org/0000-0003-0386-2696>

REFERENCES

- Allen, R. G., Pereira, L. S., Raes, D., Smith, M. (1998). *Crop evapotranspiration—Guidelines for computing crop water requirements*. FAO - Food and Agriculture Organization of the United Nations.
- Beven, K. (2016). Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal*, 61, 1652–1665. <https://doi.org/10.1080/02626667.2015.1031761>
- Beven, K., Buytaert, W., & Smith, L. A. (2012). On virtual observatories and modelled realities (or why discharge must be treated as a virtual variable). *Hydrological Processes*, 26, 1905–1908. <https://doi.org/10.1002/hyp.9261>
- Beven, K., & Smith, P. (2015). Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *Journal of Hydrologic Engineering*, 20, 15. [https://doi.org/10.1061/\(asce\)he.1943-5584.0000991](https://doi.org/10.1061/(asce)he.1943-5584.0000991)
- Beven, K., Smith, P. J., & Wood, A. (2011). On the colour and spin of epistemic error (and what we might do about it). *Hydrology and Earth System Sciences*, 15, 3123–3133. <https://doi.org/10.5194/hess-15-3123-2011>
- Beven, K., & Westerberg, I. (2011). On red herrings and real herrings: Disinformation and information in hydrological inference. *Hydrological Processes*, 25, 1676–1680. <https://doi.org/10.1002/hyp.7963>
- Blazkova, S., & Beven, K. (2009). A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resources Research*, 45(12). <https://doi.org/10.1029/2007wr006726>
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829–836. <https://doi.org/10.1080/01621459.1979.10481038>
- Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., & Smith, P. J. (2015). A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. *Water Resources Research*, 51, 5531–5546. <https://doi.org/10.1002/2014wr016532>
- Ewen, J., Geris, J., O'Donnell, G., Mayes, Q., & O'Connell, E. (2010). Multiscale experimentation, monitoring and analysis of long-term land use changes and flood risk—SC060092: Final science report. Newcastle University.
- Harmel, R. D., Cooper, R. J., Slade, R. M., Haney, R. L., & Arnold, J. G. (2006). Cumulative uncertainty in measured streamflow and water quality data for small watersheds. *Transactions of the ASABE*, 49, 689–701.
- Harmel, R. D., Smith, D. R., King, K. W., & Slade, R. M. (2009). Estimating storm discharge and water quality data uncertainty: A software tool for monitoring and modeling applications. *Environmental Modelling and Software*, 24, 832–842. <https://doi.org/10.1016/j.envsoft.2008.12.006>
- Herschy, R. W. (1999). *Hydrometry principles and practices* (2nd ed.). New York: Wiley Blackwell.

- Hollaway, M. J., Beven, K. J., Benskin, C. M. H., Collins, A. L., Evans, R., Falloon, P. D., ... Haygarth, P. M. (2018). The challenges of modelling phosphorus in a headwater catchment: Applying a 'limits of acceptability' uncertainty framework to a water quality model. *Journal of Hydrology*. <https://doi.org/10.1016/j.jhydrol.2018.01.063>
- Johnes, P. J. (2007). Uncertainties in annual riverine phosphorus load estimation: Impact of load estimation methodology, sampling frequency, baseflow index and catchment population density. *Journal of Hydrology*, 332, 241–258. <https://doi.org/10.1016/j.jhydrol.2006.07.006>
- Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J. A., Bilotta, G. S., Brazier, R. E., ... Haygarth, P. M. (2010). Ensemble evaluation of hydrological model hypotheses. *Water Resources Research*, 46. <https://doi.org/10.1029/2009WR007845>
- Lang, M., Pobanz, K., Renard, B., Renouf, E., & Sauquet, E. (2010). Extrapolation of rating curves by hydraulic modelling, with application to flood frequency analysis. *Hydrological Sciences Journal*, 55, 883–898. <https://doi.org/10.1080/02626667.2010.504186>
- Lloyd, C. E. M., Freer, J. E., Johnes, P. J., Coxon, G., & Collins, A. L. (2016). Discharge and nutrient uncertainty: Implications for nutrient flux estimation in small streams. *Hydrological Processes*, 30, 135–152. <https://doi.org/10.1002/hyp.10574>
- McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, 26, 4078–4111. <https://doi.org/10.1002/hyp.9384>
- McMillan, H. K., & Westerberg, I. K. (2015). Rating curve estimation under epistemic uncertainty. *Hydrological Processes*, 29, 1873–1882. <https://doi.org/10.1002/hyp.10419>
- Met Office. (2012). Met Office Integrated Data Archive System (MIDAS) land and marine surface stations data (1853-current). NCAS British Atmospheric Data Centre (ed.). <https://catalogue.ceda.ac.uk/uuid/220a65615218d5c9cc9e4785a3234bd0>
- Moatar, F., & Meybeck, M. (2005). Compared performances of different algorithms for estimating annual nutrient loads discharged by the eutrophic River Loire. *Hydrological Processes*, 19, 429–444. <https://doi.org/10.1002/hyp.5541>
- Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W., & Weijis, S. V. (2016). A philosophical basis for hydrological uncertainty. *Hydrological Sciences Journal*, 61, 1666–1678. <https://doi.org/10.1080/02626667.2016.1183009>
- Outram, F. N., Lloyd, C. E. M., Jonczyk, J., Benskin, C. M. H., Grant, F., Perks, M. T., ... Lovett, A. L. (2014). High-frequency monitoring of nitrogen and phosphorus response in three rural catchments to the end of the 2011–2012 drought in England. *Hydrology and Earth System Sciences*, 18, 3429–3448. <https://doi.org/10.5194/hess-18-3429-2014>
- Pappenberger, F., Matgen, P., Beven, K. J., Henry, J.-B., Pfister, L., & de, P. F. (2006). Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources*, 29, 1430–1449. <https://doi.org/10.1016/j.advwatres.2005.11.012>
- Perks, M. T., Owen, G. J., Benskin, C. M. H., Jonczyk, J., Deasy, C., Burke, S., ... Haygarth, P. M. (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. <https://doi.org/10.1016/j.scitotenv.2015.03.008>
- Schmidt, A. R., & Yen, B. C. (2008). Theoretical development of stage-discharge ratings for subcritical open-channel flows. *Journal of Hydraulic Engineering*, 134, 1245–1256. [https://doi.org/10.1061/\(ASCE\)0733-9429\(2008\)134:9\(1245\)](https://doi.org/10.1061/(ASCE)0733-9429(2008)134:9(1245))
- Webb, B. W., Phillips, J. M., Walling, D. E., Littlewood, I. G., Watts, C. D., & Leeks, G. J. L. (1997). Load estimation methodologies for British rivers and their relevance to the LOIS RACS (R) programme. *Science of the Total Environment*, 194–195, 379–389. [https://doi.org/10.1016/S0048-9697\(96\)05377-6](https://doi.org/10.1016/S0048-9697(96)05377-6)
- Westerberg, I., Guerrero, J. L., Seibert, J., Beven, K. J., & Halldin, S. (2011). Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrological Processes*, 25, 603–613. <https://doi.org/10.1002/hyp.7848>

How to cite this article: Hollaway MJ, Beven KJ, Benskin CMWH, et al. A method for uncertainty constraint of catchment discharge and phosphorus load estimates. *Hydrological Processes*. 2018;32:2779–2787. <https://doi.org/10.1002/hyp.13217>