HYDROLOGICAL PROCESSES
Hydrol. Process. **24**, 1270–1284 (2010)
Published online 2 March 2010 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/hyp.7587

Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions

Hilary McMillan,[1]* Jim Freer,[2] Florian Pappenberger,[3] Tobias Krueger[4] and Martyn Clark[1]

[1] *National Institute of Water and Atmospheric Research Ltd, Christchurch, New Zealand*
[2] *School of Geographical Sciences, University of Bristol, Bristol, UK*
[3] *European Centre for Medium-Range Weather Forecasts, Reading, UK*
[4] *University of East Anglia, Norwich, UK*

Abstract:

In order to quantify total error affecting hydrological models and predictions, we must explicitly recognize errors in input data, model structure, model parameters and validation data. This paper tackles the last of these: errors in discharge measurements used to calibrate a rainfall-runoff model, caused by stage–discharge rating-curve uncertainty. This uncertainty may be due to several combined sources, including errors in stage and velocity measurements during individual gaugings, assumptions regarding a particular form of stage–discharge relationship, extrapolation of the stage–discharge relationship beyond the maximum gauging, and cross-section change due to vegetation growth and/or bed movement. A methodology is presented to systematically assess and quantify the uncertainty in discharge measurements due to all of these sources. For a given stage measurement, a complete PDF of true discharge is estimated. Consequently, new model calibration techniques can be introduced to explicitly account for the discharge error distribution. The method is demonstrated for a gravel-bed river in New Zealand, where all the above uncertainty sources can be identified, including significant uncertainty in cross-section form due to scour and re-deposition of sediment. Results show that rigorous consideration of uncertainty in flow data results in significant improvement of the model's ability to predict the observed flow. Copyright © 2010 John Wiley & Sons, Ltd.

KEY WORDS River flow; model calibration; rating curve; uncertainty; stage–discharge; gravel-bed river

Received 23 June 2009; Accepted 16 November 2009

INTRODUCTION

Conceptual hydrological models are important tools for understanding and predicting catchment responses to measured or modelled climate and land-use scenarios. However, the necessary gross simplifications which occur when translating a complex perceptual model of catchment behaviour into a conceptual model lead to recognized model structural omissions and model parameters which cannot be directly related to measured physical properties (Beven, 2006). Calibration methods must therefore be used to identify model parameters, based on measured data. The most commonly used calibration methods, based on minimization of squared errors, make the implicit assumption that the only source of error is a Gaussian 'measurement error'. In truth, there are many different sources of error including uncertainties in input data (e.g. precipitation, temperature), calibration/validation data (e.g. streamflow), model structure and parameters. Where the incidence and distribution of each of these error sources is not explicitly recognized (a difficult task in very many cases; Beven *et al*., 2008), the calibration process may yield biased parameter estimates (e.g. Kavetski *et al*., 2006a,b;Vrugt *et al*., 2008; Thyer *et al*., 2009). This in turn leads to biased model predictions, and a loss of the potential opportunity to learn more about model error sources and methods to mitigate these.

Our aim in designing model calibration techniques must therefore be to properly account for each uncertainty source and appropriately quantify or parameterize the resulting error distribution, which may be non-stationary in time (e.g. see Freer *et al*., 2004). This paper takes one step towards that goal, by presenting a methodology to explicitly quantify one of these uncertainty sources, namely errors in the computed discharge series used to calibrate a rainfall-runoff model, caused by uncertainty in the rating curve used to transform continuously measured stage data into discharge. This uncertainty in turn derives from a combination of sources, including errors in stage and velocity measurements during gaugings, assumption of a particular form of stage–discharge relationship, extrapolation of the stage–discharge relationship beyond the maximum gauging and cross-section change due to vegetation growth or bed movement. This paper demonstrates how, using knowledge of each of these factors, a complete PDF of true discharge may be estimated for a given measured stage value.

Several previous studies have investigated methods of including uncertainties in the stage–discharge relationship, from the first to suggest a statistical framework for those uncertainties (Venetis, 1970) to many modern studies (Petersen-Øverleir, 2004; Moyeed and

* Correspondence to: Hilary McMillan, National Institute of Water and Atmospheric Research Ltd, Christchurch, New Zealand.
E-mail: h.mcmillan@niwa.co.nz

Clark, 2005; Pappenberger *et al*., 2006; Reitan and Petersen-Øverleir, 2006; 2009; Di Baldassarre and Montanari, 2009; Krueger *et al*, 2009; Liu *et al*., 2009). These all rely on fitting a single set of gaugings (i.e. measured stage/discharge points) using a single rating curve of specified form, and investigate the uncertainty in the parameters of that rating curve. For example, Pappenberger *et al*. (2006) use eight data points to fit a power-law curve (Manning equation formulation) and hence determine an 'envelope curve': upper and lower acceptable limits on discharge prediction. Krueger (2009) fits stage–discharge relationships to two weirs at experimental field sites, with the power-law form and bed level defined by the appropriate weir equation, and again models are scored as having 'perfect fit' within the resulting envelope curve, with linear decline in performance measure outside this.

Other studies have considered the possible effects of uncertainty in the stage–discharge relationship on calibration of, and predictions from, rainfall-runoff models. Aronica *et al*. (2006) calibrated a conceptual linear–nonlinear rainfall-runoff model using upper and lower bounds for multipliers of a rating curve and demonstrated the resulting change in prediction limits. Montanari (2004) simulated uncertainty in the measured discharge by adding Gaussian errors (bounds calculated by consideration of uncertainty sources). Optimized parameter sets using different error realizations were then compared to show induced parameter uncertainty. However, these two studies are both restricted to sequential consideration of alternative rating curves, as opposed to admission of uncertainty during model calibration.

This paper sets out to build on these previous methods in three ways. Firstly, to extend the 'envelope curve' method suitability to rivers where there is significant uncertainty in cross-section form due to scour and redeposition of sediment, and hence sequential gauging measurements may not all belong to a single rating curve. Secondly, to produce an explicit PDF of discharge for any given stage, as opposed to upper and lower limits on acceptable discharge. Lastly, to demonstrate how this empirical discharge PDF can be used to form a likelihood function, and used within a Markov Chain Monte Carlo (MCMC) method for parameter calibration with full consideration of uncertainty in the stage–discharge relationship.

CATCHMENT

The method is demonstrated for a gravel-bed river in New Zealand, the Wairau river in the northern South Island, New Zealand (Figure 1). The Wairau drains an area of 3825 km^2 and elevations in the catchment range from sea level to 2309 m. Vegetation in the Wairau includes pasture throughout the southern hills, native ever-green beech forest in the mountains to the west and southwest, a mix of native beech forest and exotic pine forest on the northern ranges, and vineyards on the Wairau plains (Figure 1). The Wairau River is a braided gravel-bed river that is approximately 100 m wide in the lower reaches. Rainfall in the Wairau is lowest over the Wairau plains and southern hills (600 mm/year) and highest over the western ranges (5000 mm/year). There is a small hydropower scheme in the middle reaches of the Wairau and some irrigation on the Wairau plains, but these have only minor effects on catchment streamflow.

The Wairau is managed for water allocation and flood mitigation purposes; these applications require long- and short-term estimates of discharge data statistics (Rae, 1987; Rae and Wadsworth, 1990; Williman, 1995). In addition to the input and model structural uncertainty which are usually assumed to dominate more stable river systems (e.g. Kavetski *et al*., 2006a,b), scour and redeposition of the bed gravels (and additionally anthropogenic gravel extraction) are known to introduce additional errors into models of the system. Current rainfall-runoff models in use in the catchment use a deterministic rating curve established from gauging data and adjusted over time to include new data points and discard older points which are no longer deemed representative (Ibbitt and Wild, 2005). This is problematic in the case of flood-stage gaugings which are rarely collected. The process relies on expert judgement to determine the frequency and extent to which the curve should be updated (Whalley *et al*., 2001), and implicitly on an assessment of the balance between gauging errors and rating-curve change.

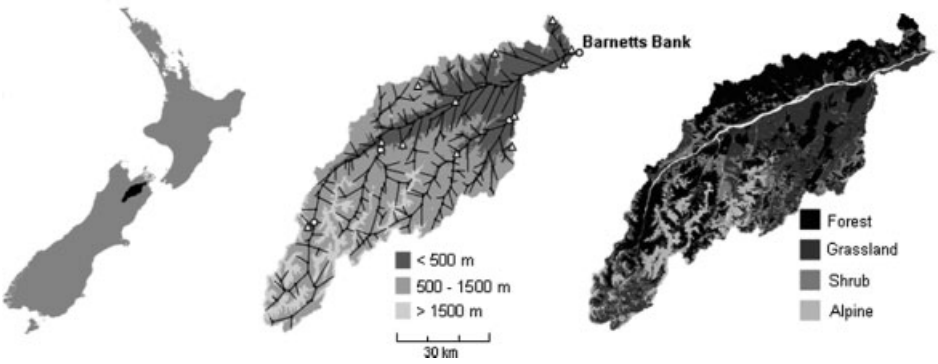

Figure 1. The Wairau River basin, showing (left) location; (middle) elevation, digital river network, location of discharge gauging sites (circles) and rainfall stations (triangles); and (right) land cover. For TopNet simulations the Wairau Basin is disaggregated into 380 sub-catchments, linked with the digital river network (middle). Figure reproduced from Clark *et al*. (2008)

The Wairau therefore presents a good example of a catchment where the assumption of zero uncertainty in the rating curve is unjustified and a rainfall-runoff model calibration technique that is able to account for rating-curve estimation errors would be a valuable tool.

DATA AND MODEL

Flow data

Flow gauging has been undertaken at various locations on the Wairau River since 1937, using stage recorders backed up by gaugings to determine the rating curve (Rae, 1987). The catchment outlet site at Barnett's Bank is used in this study, and represents the longest and most reliable record for the Wairau. Despite this, there is considerable scatter in the stage–discharge relationship: refer to Figure 2 for a photograph of the gauged cross section and Figure 4 which shows individual gaugings. This scatter in part represents the difficulties associated with flow gauging in braided, gravel-bed rivers. For the majority of its length the Wairau has a mobile gravel bed, where frequent movement of gravel changes the cross section of the river and determines the relative flow in each of the river braids. This includes the Barnett's Bank gauging site, where records show that between 2005 and 2008 the river thalweg switched from the braid nearest to the true left bank where the stage recorder is located, to the true right bank, and back again. The impact of river bed movement at this location outweighs any hysteresis effects which are minor due to the relatively steep gradient of the Wairau. The site is also used for gravel extraction which alters the channel cross section.

Frequent gaugings go some way in identifying such changes in flow regime and hence in the required rating curve; however, undertaking a new set of gaugings at a full range of river flows is an extended process which may not keep pace with river bed movement. Gaugings are taken by wading at low flows (stage height less than 3 m on the gauge) and the exact cross section used varies depending on braid locations to ensure the safety of the field team. At high flows gaugings are taken from the road bridge crossing the Wairau close to the gauge, and hence may record cross-section changes due to scour around the bridge piers. For flood flows (stage heights over 5 m), acoustic Doppler current profiler (ADCP) gauging from a jet-boat is the preferred method, although gaugings from the bridge are still used in some cases. It is particularly hard to identify ratings for high-flow events where scour and fill is continuously occurring during the event (the effects of this can be seen as 'sawtooth' patterns in stage recordings relating to waves of gravel passing the recorder, and are also recorded as a non-zero bed velocity during ADCP gaugings; not shown). Clearly multiple gaugings would be required to fully characterize the uncertainty at flood flows; however, practical constraints mean that a limited number of such gaugings can be collected.

Figure 2. Photograph of the gauged cross section at Barnett's Bank

Model

The distributed rainfall-runoff model TopNet was used in this study to provide flow predictions in the Wairau. TopNet was developed by combining TopModel (Beven and Kirkby, 1979; Beven *et al*., 1995), which is most suited to small watersheds, with a kinematic wave channel routing algorithm (Goring, 1994) so as to have a modelling system that can be applied over large watersheds, using smaller sub-basins as model elements (Ibbitt and Woods, 2002; Bandaragoda *et al*., 2004; Clark *et al*., 2008). TopNet uses TopModel concepts for the representation of sub-surface storage controlling the dynamics of the saturated contributing area and baseflow recession, with additional components for evapotranspiration, interception (based on the work of Ibbitt, 1971), infiltration (using a Green-Ampt mechanism; Mein and Larson, 1973) and soil zone. Kinematic wave routing moves the sub-basin inputs through the stream channel network. Complete model equations are provided by Clark *et al*. (2008).

The model uses input precipitation and climate data from Tait *et al*. (2006) who interpolated data from over 500 climate stations in New Zealand across a regular 0·05° latitude–longitude grid (approximately 5 km × 5 km), including data from 12 climate stations within the Wairau catchment. These data are provided at daily timesteps, and are disaggregated to hourly data before use in the model, based on an interpolation of the sub-daily distribution at the climate stations. In this study, we use data from the winter months of 2004 and 2006, in both cases including flood peaks where discharge exceeded the mean annual flood.

To apply TopNet in the Wairau, TopNet requires information on catchment topography, physical and hydrological properties. This information is available from a variety of sources. The New Zealand River Environment Classification (REC; Snelder and Biggs, 2002) includes a digital network of approximately 600 000 river reaches and related sub-basins for New Zealand. A 30-m Digital Elevation Model (DEM) provides topographic properties. Land cover and soil data are available from the

Table I. TopNet model parameters

	Name	Estimation
Sub-basin parameters		
f (m^{-1})	Saturated store sensitivity	Constant = 12·4 (multiplier calibrated)
K_0 (m/h)	Surface saturated hydraulic conductivity	Constant = 0·01 (multiplier calibrated)
$\Delta\theta_1$	Drainable porosity	From soils (multiplier calibrated)
$\Delta\theta_2$	Plant available porosity	From soils (multiplier calibrated)
D (m)	Depth of soil zone	From soils (multiplier calibrated)
C	Soil zone drainage sensitivity	1
φ (m)	Wetting front suction	From soils
V (m/s)	Overland flow velocity	Constant = 0·1 (multiplier calibrated)
CC (m)	Canopy capacity	From vegetation
Cr	Intercepted evaporation enhancement	From vegetation
A	Albedo	From vegetation
Lapse (°C/m)	Lapse rate	0·0065
Channel parameters		
N	Mannings n	Constant = 0·024 (multiplier calibrated)
A	Hydraulic geometry constant	0·00011
B	Hydraulic geometry exponent	0·518
State variables		Initialization
z' (m)	Average depth to water table	Saturated zone drainage matches initial observed flow
SR (m)	Soil zone storage	0·02
CV (m)	Canopy storage	0·0005

New Zealand Land Cover Database (LCDB) and the New Zealand Land Resource Inventory (LRI; Newsome *et al.*, 2000). The river basin was first disaggregated into individual sub-catchments, each one of which becomes a model element. We use the Strahler 3 sub-catchments from the REC, which have a typical size of 10 km^2, and split the Wairau basin into 380 elements. The REC also provides the geometrical parameters of the river network. Frequency distributions of the topographic wetness index and distance to streams are calculated from the DEM. The wetness index is formulated as $\ln(a/\tan\beta)$, where a is the contributing upstream area and β is surface slope (Beven and Kirkby, 1979). Average soil and land-cover parameters are derived from the LRI and LCDB, respectively. In total, 12 parameters are required for each sub-catchment, of which 6 may be specified using the information described above; the remaining 6 must be calibrated (refer to Table I for descriptions of all the parameters). In addition, the Manning's n value for the sub-catchment channel section must also be calibrated.

UNCERTAINTY QUANTIFICATION

As previously described, discharge is derived at Barnett's Bank using a rating curve to transform stage measurements into discharge estimates (Ibbitt and Wild, 2005). In order to quantify discharge uncertainty, we use the concept of an 'uncertain rating curve' which decomposes into a PDF of discharge for any given stage measurement. To create the uncertain rating curve, we account for the three components of uncertainty that were considered most important at this site:

1. Lack of knowledge about the current cross-section state chiefly due to bed movement, but also possibly affected by seasonal growth of vegetation.
2. Uncertainty in individual gauging measurements, via inaccuracies of stage and velocity measurement, and interpolation between point velocity measurements.
3. Uncertainty as to the correct form of the rating curve, leading to its approximation by a functional type, e.g. power law.

We now explore each of these components in more detail to define our methodology:

Cross-section state

The first component relates to the scatter in the set of {stage, discharge} data points. These data represent snapshots of river state during the continuous process of bed movement and channel cross-section change, and hence cannot be lumped into a single rating curve. Instead, we assume that the most significant changes in bed form occur during flood events, and hence divide the complete gauging series into coherent sets between major events, each of which represents a more stable phase in the bed evolution. We used a 0·5-year return period as the threshold to define a 'major event'; however, this measure is subjective, and should be set with knowledge of the individual gauging site. These gauging sets are each assumed to represent a possible state of the current cross section, and hence are used to construct possible rating curves. Low flows may be additionally affected by sedimentation between floods but this would be captured by a spread of gaugings and hence higher uncertainty within the gauging set. The number of individual points in each gauging set varies from 4 to 12, depending on the length of the stable phase and the frequency of gauging during that time (refer to Figure 4 which differentiates the gauging sets). Phases where no high stage measurement is made contribute to greater uncertainty at high flows; this

is in contrast to the previous deterministic rating curves used at Barnett's Bank which were all forced through the highest recorded gauging. These greater uncertainties are retained when the uncertain rating curve is constructed, accounting for the component of uncertainty relating to rating curve extrapolation. Similarly, the uncertainty component relating to shifts in hydraulic geometry is captured through the multiple low-flow gauging sets. When making future predictions, no distinction is made as to which gauging set is most representative, as it is recognized that rapid changes in bed form due to gravel transportation during a flood event may significantly alter the cross section in a short period of time. In rivers where bed form changes only over long timescales, it might be more appropriate to weight recent gaugings more highly than past gaugings.

Uncertainty in gauging measurement and rating curve form

These two points are considered together. To account for the uncertainty in gauging measurements and hence rating-curve shape, our method builds on the idea of a fuzzy rating curve developed by Pappenberger *et al.* (2006) and Krueger *et al.* (2009). We accept the common assumption that the gauging discharge measurements are corrupted by an error of size proportional to the discharge magnitude, here approximated as a truncated Gaussian distribution centred on the true discharge [Equation (1)]

$$Q_{\text{Measured}} \propto N(Q_{\text{True}},\ \sigma^2) \quad \text{where } |Q_{\text{Measured}} - Q_{\text{True}}| < 3\sigma$$
$$= 0 \quad \text{where } |Q_{\text{Measured}} - Q_{\text{True}}| \geq 3\sigma \tag{1}$$

where Q_{Measured} is the measured discharge, Q_{True} is the true discharge. The variance of the distribution is chosen so as to give a 95% confidence interval at 8% of the true discharge, a typical value for discharge uncertainty which is individually calculated for each Barnett's Bank gauging by the hydrometrists taking into account equipment and method accuracy. Hence, we set the standard deviation $\sigma = 0{\cdot}04Q_{\text{True}}$. It is preferable to set the variance according to site-based knowledge, as here; however, alternatively, standard values could be used such as those provided by Pelletier (1988) or Whalley *et al.* (2001) which are comparable with the value used here.

The distribution is truncated at 3σ (12%) error, which captures >99% of the distribution, while avoiding very large error values which are not considered reasonable (the 12% bound only represents possible error for a single gauging and does not include error due to rating curve interpolation/extrapolation or cross-section change). Given this error form, the probability distribution for Q_{True} can then be calculated numerically for a given gauging measurement of Q_{Measured}; this results in a skewed distribution due to the assumption that error magnitude increases with discharge. Accepting the common assumption that stage error is invariable with stage value, and again using typical uncertainty values recorded for Barnett's Bank gaugings by the hydrometrists on-site, true stage is modelled using a Gaussian error centred on the measured stage and with standard deviation of 0·02 m. As with discharge uncertainty, this value is comparable or conservative with respect to previous studies (Van der Made, 1982; Petersen-Øverleir and Reitan, 2005).

Given error PDFs for discharge and stage, random samples may be drawn from these distributions to give many possible pairs of 'true discharge' and 'true stage' for each gauging point in the rating set. Using a Monte Carlo approach, multiple sample sets are taken to approximate the true joint distribution of the gauging points. Each sample set now becomes the basis for fitting of a rating curve.

To fit the rating curve, a variation on the method proposed by Krueger *et al.* (2009) is used. The method is illustrated in Figure 3 and relies on using each combination of three sample points in the set in turn (allowing exact fitting via a three-parameter power law) to produce multiple possible rating curves:

1. Loop through all combinations of three sample points.
2. Fit the power-law equation $Q = a\ (h + b)c$ exactly to these three points (this is solved numerically).
3. Retain fitted rating if the curve intersects the error PDFs for all remaining sample points in the gauging set other than the three chosen in step 1.

Through this approach, we limit the individual rating curves to a single-segment, power-law form. This low-complexity approach is commensurate with the often limited number of sample points in a single sample set. If a denser set of gauging measurements was available, formulations such as alternative models or multi-segment curves could be considered. If prior information were available as to the expected rating curve parameter values, it could also be included here via a Bayesian approach. Although the individual curve form is thus restricted, the final uncertain rating curve may take a free and unparameterized shape as it combines thousands of individual curves (refer to following section and Figure 4 for an example), accounting for uncertainty in rating-curve form.

Constructing the uncertain rating curve

The fitted power-law curves for each sample set and for each gauging set are combined to produce the uncertain rating curve, as follows. First, the rating curves are weighted such that each gauging set has equal total weight (as previously stated, no distinction is made as to which gauging set is most representative) and all rating curves within the same gauging set have equal weight. Then, for each value of stage for which discharge estimation is required, the discharge values given by all the rating curves are ordered. The constraint $b > -h$ is enforced for the rating-curve parameters, i.e. gaugings

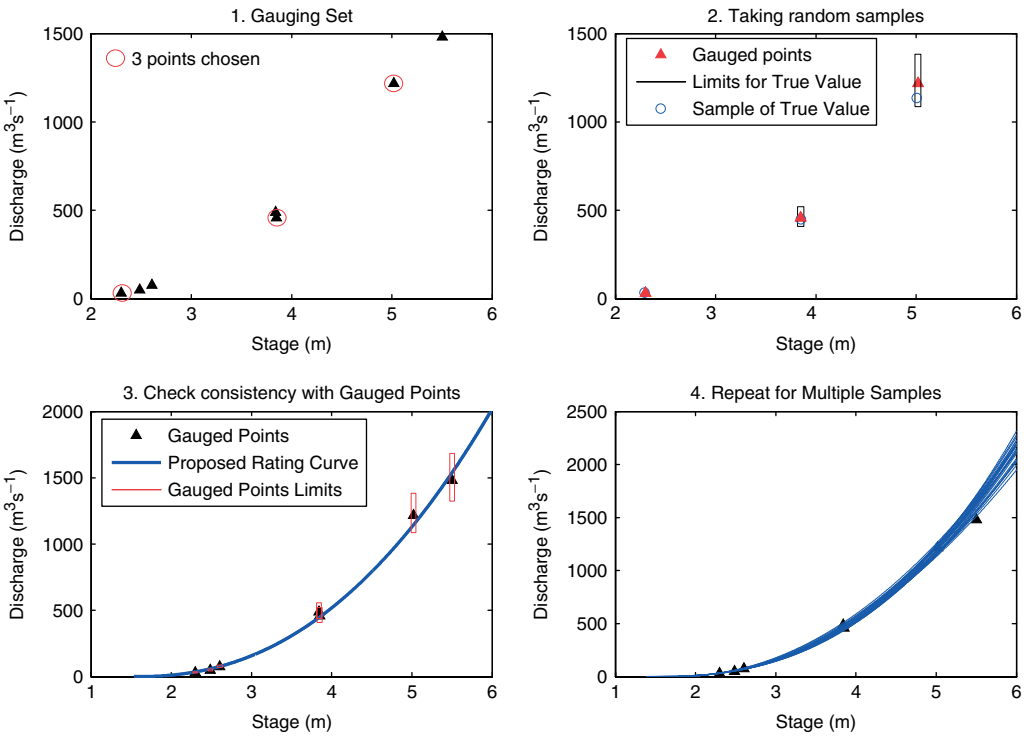

Figure 3. Illustration of the proposed Monte Carlo sampling method used to fit possible rating curves. (1) Select three points from the gauging set; (2) take a random sample of true stage/discharge; (3) fit the power-law rating curve and check consistency with remaining points; and (4) repeat for multiple samples

that were taken at a higher bed level where the gauged height h is lower than the current bed are ignored. There are approximately 105 individual rating curves. Finally, the weighted discharge values combine to provide a CDF for discharge. The form of the CDF hence represents the likelihood of each discharge value based on the distribution of the Monte Carlo samples.

RESULTS OF RATING CURVE ESTIMATION

The method described was used to estimate the form of the uncertain rating curve for the Wairau catchment outlet stage recorder at Barnett's Bank. The results are summarized in Figure 4, which shows quantiles of the estimated true discharge, plotted against the recorded stage.

The results show that the combination of gauging sets gives a unique uncertain rating curve with a form tailored to the Barnetts Bank site. For example the log space plot (Figure 4B) shows two preferred states at low flows, probably corresponding to different river braids carrying the flow thalweg. At flood flows uncertainty is high due to the limited number of gaugings, up to ±23% of the median discharge.

MODEL CALIBRATION

Evaluation measure

In order to assess model flow predictions against uncertain validation data, a performance measure must be chosen which reflects the discharge uncertainty in addition to parametric and structural uncertainty. Previous studies have used a variety of methods to do this. Pappenberger and Beven (2004) use a 'multicomponent mapping' technique, where an expected observation error structure is used to define membership values according to the distance between observed and modelled hydrographs. Krueger *et al.* (2009) define a timestep-based performance measure which scores any model prediction within the discharge envelope curve (min/max limits of the uncertainty estimation) as an exact match, and otherwise calculates the ratio of the distance between the prediction and the envelope curve to the width of the envelope curve; contrastingly, Liu *et al.* (2009) use a triangular performance measure defined via a 'limits of acceptability' approach.

The result of the rating-curve estimation for Barnett's Bank gauging station produced relatively wide uncertainty bounds due to the mobile nature of the river cross section at the gauging site (Figure 4). However, to avoid overstating the uncertainty in the observed discharge, we require a performance evaluation method which retains maximal information content from the gauging data. The method should therefore discriminate between values within the envelope of possible true discharge, using the CDF produced by the rating-curve estimation procedure. Hence, we use a timestep-based method, which stores the conditional probability of the modelled flow, given the observed flow for each timestep: this is an empirical function based on the gauging data and does not have an analytical form. The CDF for estimated discharge at Barnett's Bank is

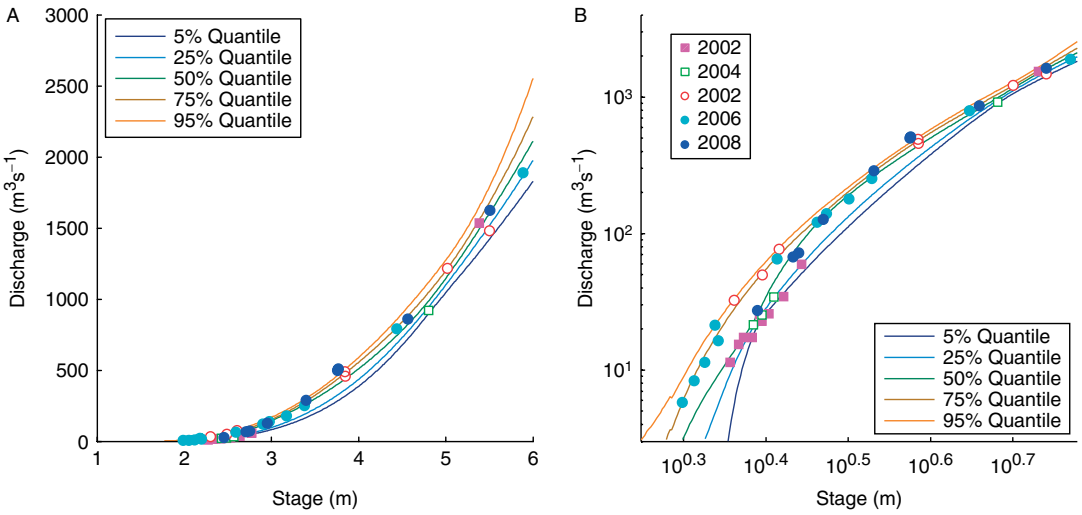

Figure 4. Quantiles of estimated discharge at Barnett's Bank gauging site at the catchment outlet of the Wairau River, shown in linear (A) and log (B) space. Markers show gauging points used, the symbols identifying discrete gauging sets

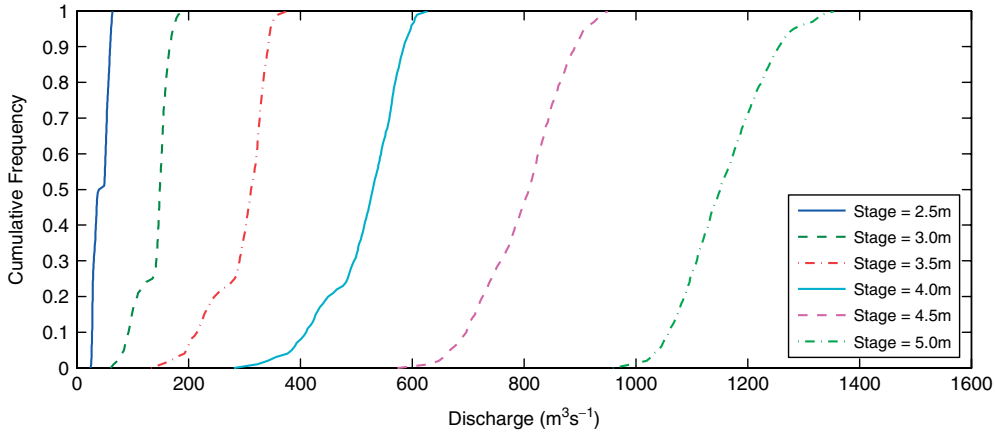

Figure 5. Empirical CDF for discharge at Barnett's Bank: examples at six stage values

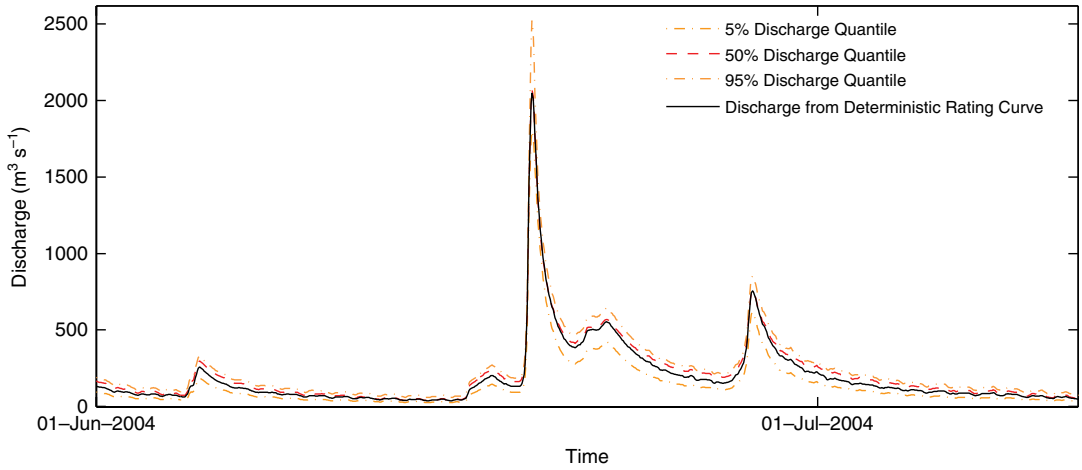

Figure 6. Illustration of discharge median and confidence bounds at Barnett's Bank, compared with discharge calculated using deterministic rating curve

shown in Figure 5 for various measured stage values, and Figure 6 shows how this information translates into discharge bound quantiles for an example section of flow record. The conditional probabilities are unique to the gauging record at the Barnett's Bank site: note, e.g. that the lower density of gauging points in the lower quantiles of the uncertain discharge curve (Figure 4) lead to skewed CDFs (Figure 5) which in turn lead to lower uncertainty limits that are wider than the upper limits in Figure 6.

A variety of methods could be used to aggregate the timestep-based likelihoods over the modelled time period: in classical Bayesian inference the product of the individual probabilities (p) would be used:

$$\mathrm{p}(\theta|\mathbf{y}) = \prod_1^N \mathrm{p}(\theta|\mathrm{y}) \tag{2}$$

were $\mathbf{y}$ is a vector of observed data of length N, and θ are the model parameters. We assume here a uniform bounded prior. However, this method assumes independence of observed data between timesteps which is unlikely for hydrological time series. Instead, we use here a modification of this product of probabilities which accounts for the reduction of information content of the data due to such autocorrelation:

$$\hat{\mathrm{p}}(\theta|\mathbf{y}) = \left[\prod_1^N \mathrm{p}(\theta|\mathrm{y})\right]^{\frac{ESS}{N}} \tag{3}$$

where *ESS* is the 'Effective Sample Size': a measure of the information content of the data series. Several explanatory notes are required:

1. To illustrate the coherence of this form of the conditional probability distribution, consider the classical assumption of independent Gaussian residuals, combined with a Jeffrey's prior on σ^2. Box and Tiao (1973) derive the likelihood function:

$$\mathrm{p}(\theta|\mathbf{y}) \propto M(\theta)^{-N/2} \tag{4}$$

where $M(\theta)$ is the sum of squared errors and N is the number of data points. Following their derivation, but substituting the revised definition $\hat{\mathrm{p}}(\theta|\mathbf{y})$ as above [Equation (3)], it can be shown that the likelihood function takes the form:

$$\hat{\mathrm{p}}(\theta|\mathbf{y}) \propto M(\theta)^{-ESS/2} \tag{5}$$

hence showing that this revision of the product of probabilities gives the expected likelihood function when used with standard assumptions.

2. We use *ESS* (Thiebaux and Zwiers, 1995; Wilks, 1997) as a measure of information content. This measure is designed to represent the equivalent number of independent data points and uses autocovariance to quantify the degree of time coherence in the data series. The ESS is calculated from the true sample size N as follows:

$$ESS = \frac{N}{\sum_{\tau=-(N-1)}^{\tau=+(N-1)} \left(1 - \frac{|\tau|}{N}\right) \rho(\tau)} \tag{6}$$

where τ is time lag and $\rho(\tau)$ is the corresponding autocorrelation function.

3. While the methods in this study account for rating-curve error, we recognize that additional uncertainty components (input uncertainty, model structural uncertainty), not explicitly characterized here, also affect model response. This has important implications for the multiplicative form of the likelihood function, as prediction for a single timestep outside the empirical discharge envelope (which is likely due to these additional error components) would give a total probability of zero. Therefore, we use the simplest method possible to incorporate those effects: a uniform (small) error constant ε is added to the response surface before the multiplicative step.

$$p(\theta|y) = \max(p(\theta|y), \varepsilon) \tag{7}$$

This addition has the result that simulations which lie outside the flow uncertainty bounds in some timesteps are disadvantaged but not rejected completely. This step is also important to improve the convergence speed of the MCMC algorithm (Section on MCMC Parameter Search Method) by allowing simulations to be properly ranked. In the extreme case where no predictions lie within the uncertainty bounds, the simulation would be ranked lower than any other realization, and quickly rejected by the MCMC algorithm.

MCMC parameter search method

As has been extensively discussed by Beven (1993; 2006; Beven and Binley, 1992) and others (Wagener and Gupta, 2005), the many sources of uncertainty in a hydrological model application, including but not limited to the measurement uncertainty discussed in this paper, lead to equifinality of parameter sets in providing acceptable model performance. Performance is judged with reference to the observed data, here using the evaluation measure described in the section on Evaluation Measure. The aim of our calibration technique is to enable an efficient search of the parameter space, identifying those regions where model performance is considered satisfactory in the light of observation error on the discharge. The task is made more difficult by the typically complex nature of the model response surface (Duan *et al.*, 1992; Sorooshian *et al.*, 1993) which may be exacerbated by artefacts of model timestep and solution techniques (Kavetski *et al.*, 2006c,d).

In response to these difficulties, MCMC methods have gained increasing popularity, providing targeted sampling of the parameter space and hence considerable efficiency savings over uniform random sampling (Blasone *et al.*, 2008). These methods enable simulation of complex multivariate distributions by casting them as the invariant distribution of a Markov Chain. We use here a popular version of the original Metropolis-Hastings MCMC algorithm: the adaptive SCEM-UA algorithm (Vrugt *et al.*, 2003) which combines the Metropolis-Hastings sampler with the SCE-UA optimization method (Duan *et al.*, 1992), using information exchange between multiple sampler chains to improve convergence rates.

MCMC algorithms have traditionally been used to sample posterior distributions derived from classical statistical likelihood functions (e.g. Thiemann *et al.*, 2001; Vrugt *et al.*, 2003). However, the careful use of 'informal' likelihood (performance) measures chosen using

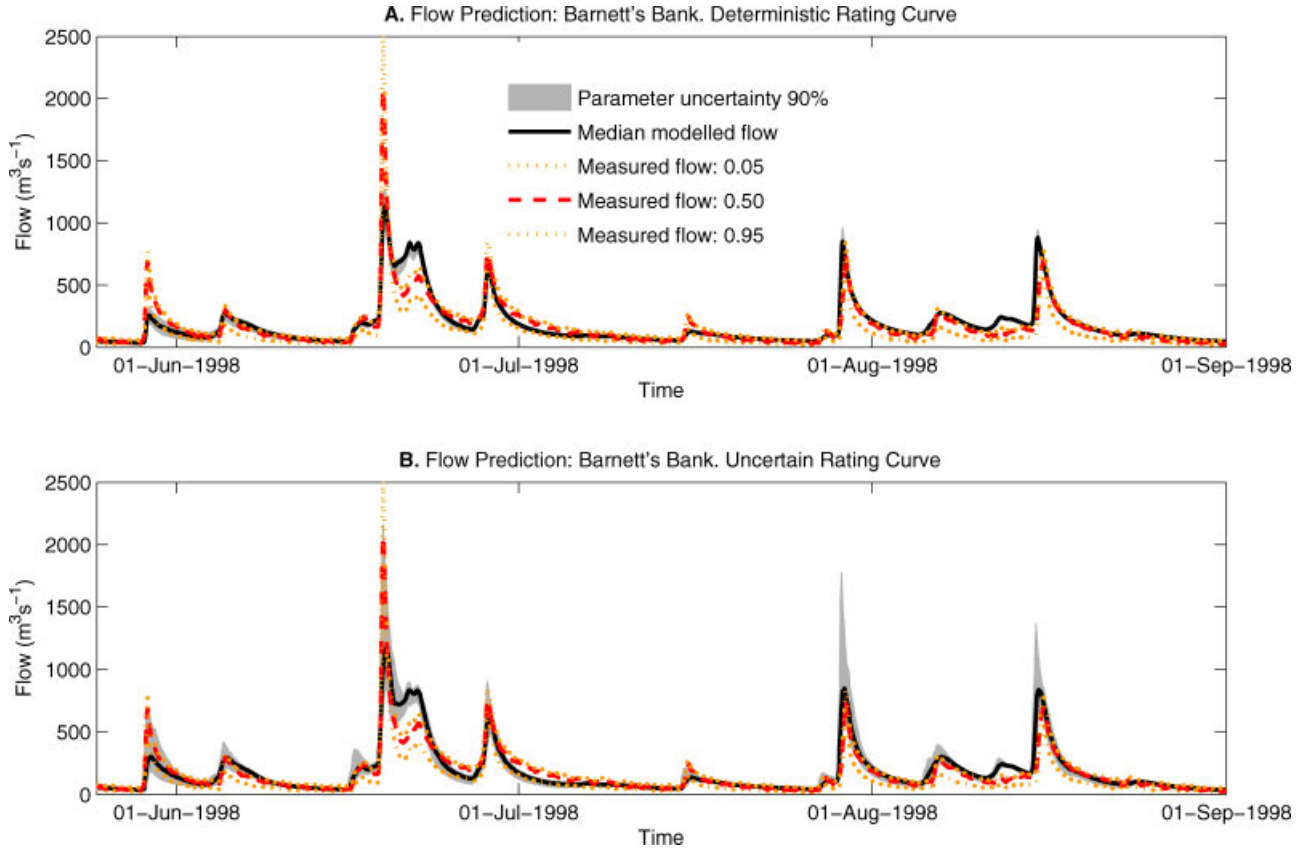

Figure 7. Parameter estimation incorporating rating-curve uncertainty: 90% confidence interval for streamflow at Barnett's Bank during example section of model calibration period. Comparison of results using deterministic rating curve (A) versus rating curve including uncertainty (B)

modeller judgement can improve the ability of the algorithm to fully explore the response surface (McMillan and Clark, 2009; Smith *et al.* (2008)). Here the evaluation measure described in the section on Evaluation Measure best reflects our knowledge of the information and uncertainty contained within the observed flow data, and hence is used to describe the response surface.

The method used within the MCMC algorithm to adjust the model parameters is via parameter multipliers. In this approach, the default TopNet model parameters (which vary spatially within the river basin) are adjusted uniformly throughout the river basin using a spatially constant set of parameter multipliers. In this method all sub-catchments receive the same multiplier, i.e. we assume that the spatial distribution of default TopNet parameters is suitable. While this approach represents a simplification, it is a valuable tool to reduce the dimensionality of the parameter estimation problem using prior knowledge of the spatial variation in catchment characteristics. The resulting estimation problem uses seven parameter multipliers, and accordingly the MCMC algorithm is run using seven parallel chains. A burn-in period of 2000 iterations is followed by a parameter estimation period of 1000 iterations. Calculation of the Gelman-Rubin convergence statistic during the burn-in period was used to confirm that the Markov Chain had converged to the stationary distribution representing the model posterior distribution.

FLOW MODELLING RESULTS

We now demonstrate how inclusion of explicit discharge uncertainty information can offer additional insights into model calibration, by applying the method described above to calibration of the TopNet model in the Wairau catchment. The calibration is done in two parts. First, TopNet is calibrated using flow data produced from the deterministic rating curve recommended for the Wairau to serve as a benchmark with which to compare the new method. Second, the calibration is repeated using the uncertain rating curve previously derived. Refer to Figure 6 for an example comparison of the resulting flow data in the two cases.

Deterministic rating curve

Model calibration was carried out for a 6-month period in winter 2004 (1 April 2004–1 October 2004). The MCMC algorithm was run using the likelihood function derived in the section on MCMC Parameter Search Method under the classical assumption of Gaussian error: $\hat{p}(\theta|\mathbf{y}) \propto M(\theta)^{-ESS/2}$ [i.e. Equation (5)] where $M(\theta)$ is the sum of squared errors and ESS is the effective sample size.

The resulting uncertainty bounds on the flow hindcast are shown in Figure 7A; note that despite the use of the effective sample size measure which reduces the peakedness of the objective function, the uncertainty bounds have a very narrow range and are barely visible except during times of high flood.

Uncertain rating curve

The model calibration was repeated with the same MCMC algorithm, but using the performance measure described in the section on Evaluation Measure which incorporates rating-curve uncertainty. The resulting flow hindcast with 90% confidence intervals is shown in Figure 7B.

Comparison of results

Both parameter estimation techniques are shown to underestimate the total uncertainty during some periods, demonstrated, e.g. during the recession of the flood peak in late June: for both techniques the estimated flow bounds lie completely outside the estimation flow uncertainty (Figure 7). However, when the uncertain rating curve is used, the flow uncertainty bounds are wider and the model is more successful in predicting the flood peaks, especially during the wetting-up period of early winter. It is especially noticeable that during the highest flood peak, the calibration using the uncertain rating curve includes the median modelled flow within the uncertainty bounds, whereas the calibration using the deterministic rating curve underestimates the flood peak by almost 50%, even at the 90% confidence level.

It is also desirable to provide a more objective measure of the ability of the two models to span the observed discharge data. For example, consider the percentage of time that the median gauged discharge lies within the bounds of the modelled discharge CDF. During the calibration period, this figure is 68% for the model using the deterministic rating curve versus 86% for the model using the uncertain rating curve, suggesting an improvement in performance in the latter case. However, this measure could be criticized as it favours models with overly wide uncertainty bounds. To overcome this, we consider a generalization of the rank histogram, which measures how well the spread of a model forecasts represents the true variability of the observations. A rank histogram, usually based on deterministic observation data, is derived by tallying, for each timestep, the quantile at which the observed data lies within the model forecast. A perfect result gives a flat histogram. A 'u' shape histogram indicates an underdispersive model, with many observations lying outside the extremes of the model prediction; conversely, a dome shape indicates that the model spread is too large. We extend this to the case of uncertain observed data, by tallying the model quantile at which each observed data quantile lies, for each observed data quantile and for each forecast timestep.

Such generalized rank histograms are shown in Figure 8, for both deterministic and uncertain rating curves, for the winter 2004 calibration period. The rank histograms show additionally the proportion of the values in the lower and upper quantiles where the observed data lie outside the model bounds. It is clear that both models are underdispersive, i.e. the uncertainty bounds are not wide enough to capture the errors between modelled and measured discharge data. However, the underdispersion

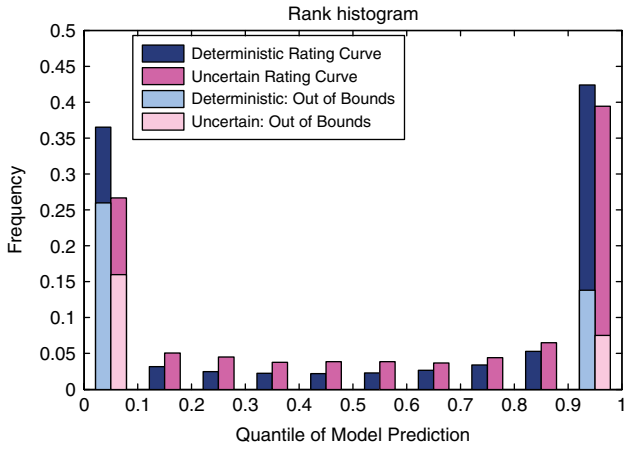

Figure 8. Generalized rank histogram showing spread of model predictions compared with variability of uncertain discharge data, during calibration period. Results shown for both deterministic and uncertain rating-curve cases

is less severe in the case of the uncertain rating curve. This is consistent with the fact that our method has taken into account one source of uncertainty in the modelling procedure, i.e. rating-curve uncertainty; but there are still many uncertainty sources not considered which contribute to the underdispersion.

Effects on behavioural model parameter sets

In order to compare the effects of the deterministic versus uncertain rating curve on model behaviour, we examine the differences in the distribution of model parameters between the two methods. Figure 9 presents histograms showing the marginal posterior probability density function for each of the TopNet model parameters, for both the deterministic and the uncertain rating-curve calibration run. We observe that, in general, parameter distributions are less constrained when the uncertain rating curve is used. This result reflects the wider range of model behaviour considered behavioural when the errors on the discharge measurements are not artificially constrained. The result suggests that the identifiability of parameters such as the TopModel f parameter may be a consequence of the artificially peaked response surface due to the use of a deterministic rating curve.

Validation

The model calibration process was tested by running the model with deterministic and uncertain rating-curve calibrations for an independent validation time period. Again a 6-month winter period was used (1 April 2006–1 October 2006), where significant flood events occurred in the Wairau and were recorded at Barnett's Bank gauge. The model predictions for these flood events are shown in Figure 10, with the model uncertainty bounds compared as before with the 5%, median and 95% quantiles of the measured flow data series. The results show that when using the deterministic rating-curve calibration, the uncertainty in the model predictions is severely underestimated in the validation phase,

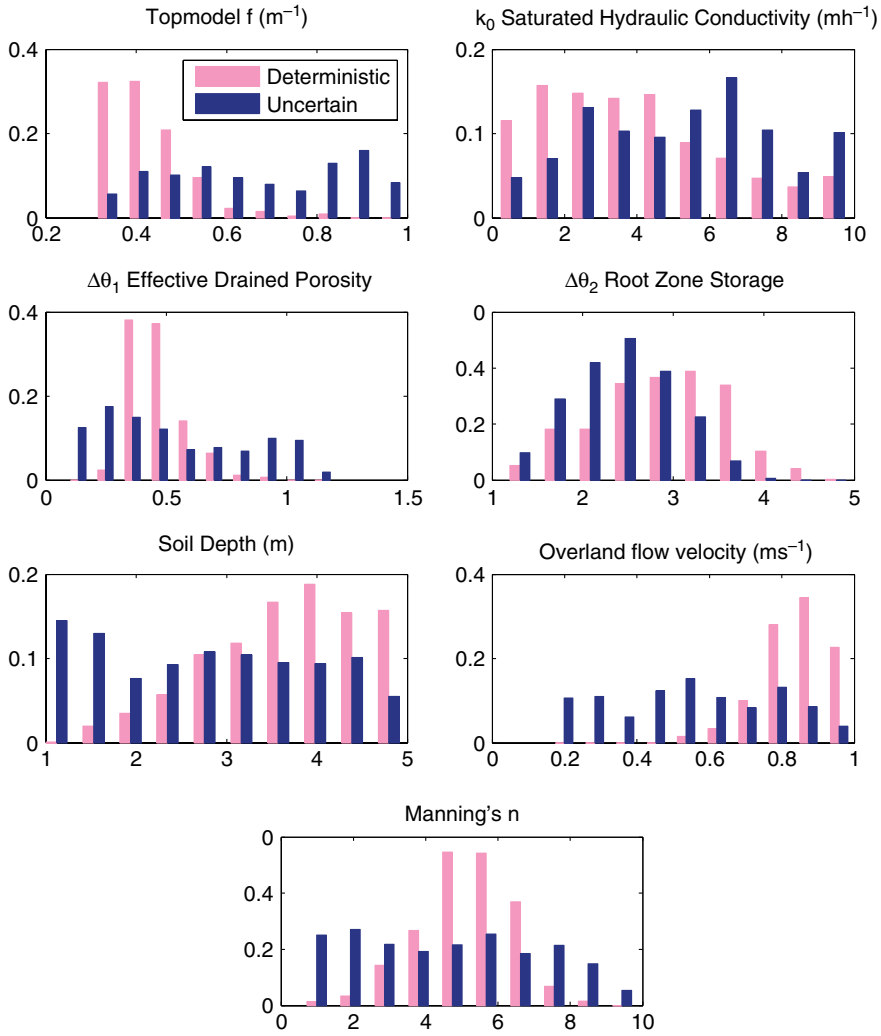

Figure 9. Parameter estimation: comparison of results using deterministic rating curve versus rating curve including uncertainty. Plots show marginal posterior probability density functions for each of the TopNet model parameters

especially during flood events (Figure 10A). When using the uncertain rating-curve calibration, there is a significant improvement in the model's ability to bracket the observed flow (Figure 10B). This improvement is at the cost of increased uncertainty in the predictions; however, the results suggest that the wider uncertainty bounds are warranted due to discrepancies between modelled and measured flow in the deterministic rating-curve case. While increased forecast uncertainty may be unwelcome for decision makers, previous research has demonstrated that model predictions including significant uncertainty can be successfully accommodated within a flood forecasting framework, using techniques such as probabilistic assessment of threshold exceedance (de Roo *et al.*, 2003; Pappenberger *et al.*, 2008).

As before, a rank histogram is calculated to show how well the model spread captures the variability in the observed data (Figure 11). Similar to the calibration period, the underdispersion of the model forecast is less severe when the uncertain rating curve is used. In part this measure demonstrates conditional bias in the model, which tends to underpredict during flood peaks and overpredict during recession periods, with the bias being more severe in the case of the deterministic rating curve. Overpredictions also arise from errors in timing of model predictions for the second flood peak in the validation phase, which may signal the influence of other unaccounted-for uncertainty sources.

DISCUSSION

This paper demonstrates the improvements in model performance, and particularly in uncertainty estimation, that can be gained by explicit recognition of uncertainty in the stage–discharge relationship embodied in the rating curve. Application of the model to a 'validation' time period showed that ignoring rating curve uncertainty could lead to significant underestimation of the uncertainty associated with the model flow predictions, particularly during flood events. While the improvement is particularly pronounced in mobile-bed rivers, such as the Wairau River considered here, all rivers gauged using a stage–discharge relationship are subject to rating-curve uncertainty. Perhaps the most important advance

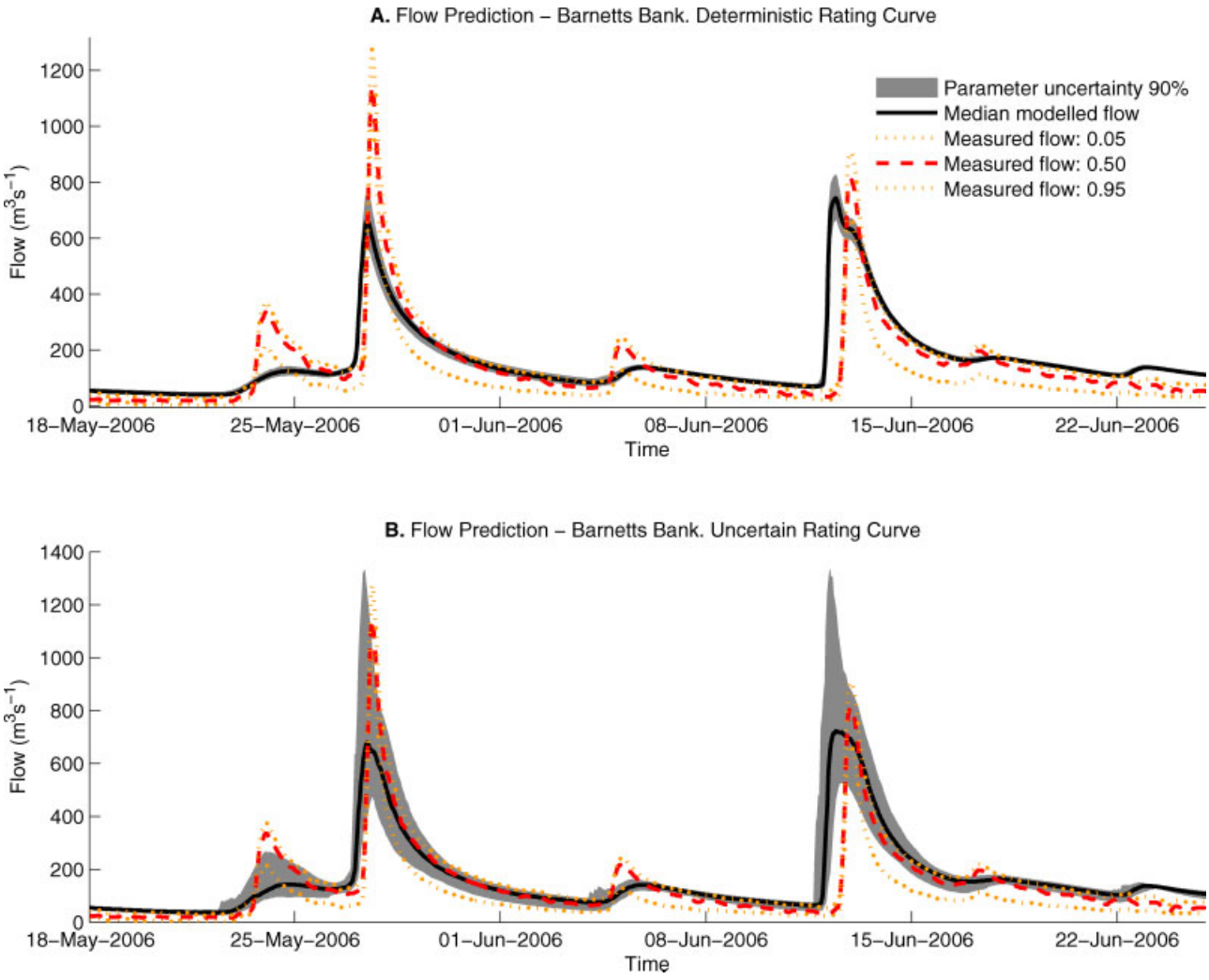

Figure 10. Model validation results: 90% confidence interval for streamflow at Barnett's Bank during model validation period. Comparison of results from model calibrated using a deterministic rating curve (A) versus rating curve including uncertainty (B)

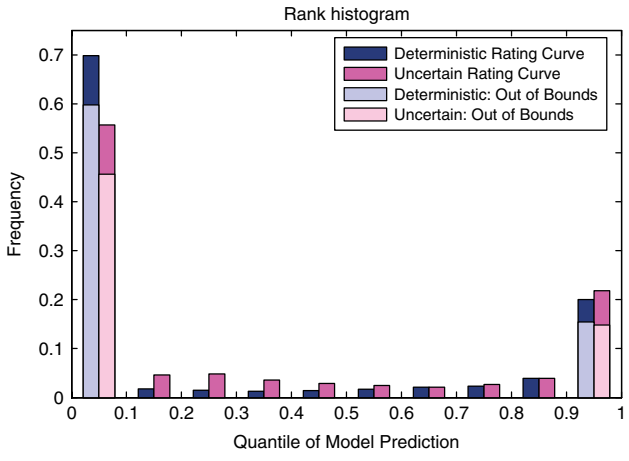

Figure 11. Generalized Rank Histogram showing spread of model predictions compared with variability of uncertain discharge data, during calibration period. Results shown for both deterministic and uncertain rating-curve cases

demonstrated by our method was the ability to produce an explicit PDF of discharge as opposed to upper and lower limits on acceptable discharge. This allowed the discharge PDF to be used to form a likelihood function which could be used within a conventional uncertainty estimation method. However, subjective choices were not completely removed from the method [for example, choice of return period for temporal segregate of the time series; choice of error distribution for individual measurements (Gaussian used); weighting of gauging sets by time] and these choices might be investigated further in future applications of the method.

It must be noted that this paper takes only one step towards the goal of total error quantification in hydrological modelling. To achieve this aim, the type of analysis suggested here must be combined with methods to quantify uncertainty due to input (precipitation) error, initial and boundary condition error, structural error and others. Until then, unaccounted-for uncertainties are implicitly mapped onto parameter uncertainty, which can lead to bias, under- or over-estimation of uncertainty in model predictions. Recognition and evaluation of rating-curve uncertainty magnitude may also help to define those situations where the uncertainty could be reduced, for example by increasing the number of verticals in manual discharge measurements. However, this study has also

demonstrated that the contribution of such 'measurement uncertainties' is often small when compared to 'natural uncertainties' such as shifts in hydraulic geometry occurring during flood events.

Note also that the Wairau, despite its challenges, is regularly gauged; in contrast, there are many parts of the world where large river systems are remote and difficult to access or monitor. In these types of environments, gaugings will be infrequent or non-existent, and discharge estimation may necessarily be undertaken using remote-sensing methods (Bjerklie *et al*., 2003, 2005). Such methods introduce new sources of discharge measurement error, and additional difficulties, such as ice cover at the gauging site, may occur in some regions (Shiklomanov *et al*., 2006). In such areas, integrated measures such as mean annual discharge may be required rather than continuous discharge measurement; however, these are also strongly affected by rating-curve error (Clarke, 1999; Clarke *et al*., 2000). Many large rivers of the world have complex and unstable morphology affected by multiple channel-changing mechanisms including floods, landslides and changes in sediment supply (e.g. Goswami *et al*., 1999; Ashworth *et al*., 2000; Sarma, 2005). Globally, significant discharge uncertainty may be the norm rather than the exception.

More general recognition of uncertainty in measured flow data will have implications for other hydrological modelling techniques which rely on flow data as input. For example when data assimilation is used to update model states based on observed flow data (as it has been in the Wairau: Clark *et al*., 2008), errors in the flow data must be explicitly specified. Good performance of the data assimilation method relies on accurate error estimates; hence, it is essential to take into account the multiple sources of rating-curve uncertainty such that the error estimates are valid even during flood events. An analysis such as that suggested here would allow those errors to be confidently specified.

Another example of the implications of uncertainty in flow data is in the recent suggestion that integrated performance measures, which evaluate a range of aspects of model behaviour via a single number, should be replaced by more meaningful 'diagnostic signatures' (Gupta *et al*., 2008). These signatures would use a specific interpretation of model output to focus the evaluation on a particular component of model structure or parameterization, and to identify deficiencies and suggest improvements to the conceptual model structure. As an example, an analysis of dQ/dt versus Q could be used to study the form of the catchment storage–discharge relationship. Where uncertainty in the stage–discharge relationship is recognized, it follows that the 'true catchment behaviour' used to define the diagnostic signature (Q or dQ/dt in this case) is not known exactly. Hence, the true 'diagnostic signatures' will become a fuzzy quantity, with consequences for the methods used to compare them with model output. In whatever form that observed data are used for model calibration, whether via diagnostic signatures, 'soft data' or expert knowledge (Seibert and McDonnell, 2002), manual or automatic calibration (Boyle *et al*., 2000), it is essential that the information, uncertainty and error within that data is evaluated, so that models are not incorrectly forced to fit uncertain data treated as though it were deterministic.

CONCLUSIONS

This paper presents a method to quantify uncertainty in river discharge measurements caused by uncertainties in the rating curve used to transform stage values into discharge values. The method was designed to assess the combined uncertainty caused by errors in stage and velocity measurements, rating-curve interpolation or extrapolation and cross-section change due to vegetation growth of bed movement.

We demonstrate how the method can be used to provide a complete PDF (and hence also confidence bounds) for measured discharge, and how this PDF can be used to form a likelihood function to enable model calibration allowing for rating-curve uncertainty. The method is tested on the Wairau River in New Zealand, and results for calibration and validation periods using both deterministic and uncertain rating curves are compared. We show that explicit consideration of the uncertainty in flow measurements leads to a flatter response surface with higher parameter uncertainty and hence wider uncertainty bounds for flow predictions. Use of the uncertain rating curve therefore provides model predictions with confidence bounds which are more successful at enclosing the measured flow during model validation, and hence we suggest provide a more realistic estimate of model uncertainty.

ACKNOWLEDGEMENTS

This work was supported by Foundation for Research, Science and Technology (FRST) (grant C01X0812) and UK Natural Environment Research Council (NERC), Flood Risk from Extreme Events (FREE) programme (grant number NE/E002242/1). We thank the NIWA field team for collecting the Wairau gauging data.

REFERENCES

Aronica GT, Candela A, Viola F, Cannarozzo M. 2006. Influence of rating curve uncertainty on daily rainfall-runoff model predictions. In *Predictions in Ungauged Basins: Promise and Progress*, IAHS Publ. 303, Sivapalan M, Wagner T, Uhlenbrook S, Liang X, Lakshmi V, Kumar P, Zehe E, Tachikawa Y (eds). 116–124.

Ashworth PJ, Best JL, Roden JE, Bristow CS, Klaassen GJ. 2000. Morphological evolution and dynamics of a large, sand braid-bar, Jamuna River, Bangladesh. *Sedimentology* **47**(3): 533–555.

Bandaragoda C, Tarboton DG, Woods R. 2004. Application of TOPNET in the distributed model intercomparison project. *Journal of Hydrology* **298**(1–4): 178–201.

Beven KJ. 1993. Prophecy, reality and uncertainty in distributed hydrologic modelling. *Advances in Water Resources* **16**: 41–51.

Beven KJ. 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* **320**(1–2): 18–36.

Beven KJ, Binley AM. 1992. The future of distributed models: model calibration and uncertainty in prediction. *Hydrological Processes* **6**: 279–298.

Beven KJ, Kirkby MJ. 1979. A physically based variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin* **24**(1): 43–69.

Beven KJ, Lamb R, Quinn P, Romanowicz R, Freer J. 1995. Topmodel. In *Computer Models of Watershed Hydrology 18*, Vol. 18 Singh VP (eds). Water Resources Publications: Highlands Ranch, CO; 627–668.

Beven KJ, Smith PJ, Freer JE. 2008. So just why would a modeller choose to be incoherent? *Journal of Hydrology* **354**: 15–32.

Bjerklie DM, Dingman SL, Vorosmarty CJ, Bolster CH, Congalton RG. 2003. Evaluating the potential for measuring river discharge from space. *Journal of Hydrology* **278**(1–4): 17–38.

Bjerklie DM, Moller D, Smith LC, Dingman SL. 2005. Estimating discharge in rivers using remotely sensed hydraulic information. *Journal of Hydrology* **309**(1–4): 191–209.

Blasone R-S, Vrugt JA, Madsen H, Rosbjerg D, Robinson MA, Zyvoloski GA. 2008. Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. *Advances in Water Resources* **31**: 630–648.

Box GEP, Tiao GC. 1973. *Bayesian Inference in Statistical Analysis*. Addison-Wesley: Reading, Mass.

Boyle DP, Gupta HV, Sorooshian S. 2000. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research* **36**: 3663–3674.

Clark MP, Rupp DE, Woods RA, Zheng X, Ibbitt RP, Slater AG, Schmidt J, Uddstrom MJ. 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Research* **31**(10): 1309–1324.

Clarke RT. 1999. Uncertainty in the estimation of mean annual flood due to rating-curve indefinition. *Journal of Hydrology* **222**(1–4): 185–190.

Clarke RT, Mendiondo EM, Brusa LC. 2000. Uncertainties in mean discharges from two large South American rivers due to rating curve variability. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques* **45**(2): 221–236.

De Roo APJ and co-authors. 2003. Development of a European flood forecasting system. *International Journal of River Basin Management* **1**(1): 49–59.

Di Baldassarre G, Montanari A. 2009. Uncertainty in river discharge observations: a quantitative analysis. *Hydrology and Earth System Sciences Discussion* **6**: 39–61.

Duan Q, Sorooshian S, Gupta V. 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research* **28**(4): 1015–1031.

Freer JE, McMillan H, McDonnell JJ, Beven KJ. 2004. Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *Journal of Hydrology* **291**: 254–277.

Goring DG. 1994. Kinematic shocks and monoclinal waves in the Waimakariri, a steep, braided, gravel-bed river. In *Proceedings of the International Symposium on Waves: Physical and Numerical Modelling,* 21–24 August, 1994. University of British Columbia: Vancouver, Canada; 336–345.

Goswami U, Sarma JN, Patgiri AD. 1999. River channel changes of the Subansiri in Assam, India. *Geomorphology* **30**(3): 227–244.

Gupta HV, Wagener T, Liu Y. 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes* **22**(18): 3802–3813.

Ibbitt RP. 1971. *Development of a conceptual model of interception*. Hydrological research progress report. No. 5. Wellington, Ministry of Works.

Ibbitt R, Wild M. 2005. *Flood forecasting for the Wairau River using a TopNet model*. NIWA client report CHC2005-020 prepared for Marlborough District Council.

Ibbitt RP, Woods R. 2002. Towards rainfall-runoff models that do not need calibration to flow data. In *Friend 2002—Regional Hydrology: Bridging the Gap between Research and Practice*. IAHS Publication no. 274, van Lanen HAJ, Demuth S (eds). 189–196.

Kavetski D, Kuczera G, Franks SW. 2006a. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research* **42**(3): W03407.

Kavetski D, Kuczera G, Franks SW. 2006b. Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research* **42**(3): W03408.

Kavetski D, Kuczera G, Franks SW. 2006c. Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts. *Journal of Hydrology* **320**(1–2): 173–186.

Kavetski D, Kuczera G, Franks SW. 2006d. Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis. *Journal of Hydrology* **320**(1–2): 187–201.

Krueger T. 2009. *Uncertainties in modelling agricultural phosphorus transfers*. PhD thesis, Lancaster University.

Krueger T, Quinton JN, Freer J, Macleod CJA, Bilotta GS, Brazier RE, Butler P, Haygarth PM. 2009. Uncertainties in data and models to describe event dynamics of agricultural sediment and phosphorus transfer. *Journal of Environmental Quality* **38**(3): 1137–1148. DOI: 10.2134/jeq2008.0179.

Liu YL, Freer J, Beven K, Matgen P. 2009. Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *Journal of Hydrology* **367**: 93–103. DOI: 10.1016/j.jhydrol.2009.01.016.

McMillan H, Clark M. 2009. Rainfall-runoff model calibration using informal likelihood measures within a Markov chain Monte Carlo sampling scheme. *Water Resources Research* **45**: W04418, DOI:10.1029/2008WR007288.

Mein RG, Larson CL. 1973. Modeling infiltration during steady rain. *Water Resources Research* **9**: 384–394.

Montanari A. 2004. *An attempt to quantify uncertainty in observed river flows: effect on parameterisation and performance evaluation of rainfall-runoff models*. AGU Fall Meeting Abstracts, 2004. http://adsabs.harvard.edu/cgi-bin/nph-bib_query?bibcode=2004AGUFM.H11G.03M.

Moyeed RA, Clark RT. 2005. The use of Bayesian methods for fitting rating curves, with case studies. *Advance in Water Resources* **28**: 807–818.

Newsome PFJ, Wilde RH, Willoughby EJ. 2000. *Land Resource Information System Spatial Data Layers*. Technical Report. Landcare Research NZ Ltd., Palmerston North, NZ.

Pappenberger F, Beven K. 2004. Functional classification and evaluation of hydrographs based on multicomponent Mapping. *International Journal of River Basin Management* **2**(2): 89–100.

Pappenberger F, Matgen P, Beven KJ, Henry JB, Pfister L, Fraipont P. 2006. Influence of uncertain boundary conditions and model structure on flood inundation predictions. *Advances in Water Resources* **29**: 1430–1449.

Pappenberger F, Bartholmes J, Thielen J, Cloke HL, Buizza R, de Roo A. 2008. New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophysical Research Letters* **35**(10): L10404.

Pelletier PM. 1988. Uncertainties in the single determination of river discharge: a literature review. *Canadian Journal of Civil Engineering* **13**: 834–850.

Petersen-Øverleir A. 2004. Accounting for heteroscedasticity in rating curve estimates. *Journal of Hydrology* **292**(1–4): 173–181.

Petersen-Øverleir A, Reitan T. 2005. Uncertainty in flood discharges from urban and small rural catchments due to inaccurate head measurement. *Nordic Hydrology* **36**: 245–257.

Rae SN. 1987. *Water and Soil Resources of the Wairau.* Marlborough catchment and regional water board: Blenheim; 301.

Rae SN, Wadsworth V. 1990. Wairau river catchment flood forecasting. *NewZealand Journal of Hydology* **29**(1): 1–17.

Reitan T, Petersen-Øverleir A. 2006. Existence of the frequentistic estimate for power-law regression with a location parameter, with applications for making discharge rating curves. *Stochastic Environmental Research and Risk Assessment* **20**(6): 445–453.

Reitan T, Peterson-Øverleir A. 2009. Bayesian methods for estimating multi-segment discharge rating curves. *Stochastic Environmental Research and Risk Assessment* **23**(5): 627–642. DOI: 10.1007/s00477-008-0248-0.

Sarma JN. 2005. Fluvial process and morphology of the Brahmaputra River in Assam, India. *Geomorphology* **70**(3–4): 226–256.

Seibert J, McDonnell JJ. 2002. On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration. *Water Resources Research* **38**: art. no.-1241.

Shiklomanov AI, Yakovleva TI, Lammers RB, Karasev IP, Vorosmarty CJ, Linder E. 2006. Cold region river discharge uncertainty—estimates from large Russian rivers. *Journal of Hydrology* **326**(1–4): 231–256.

Smith P, Beven KJ, Tawn JA. 2008. Informal likelihood measures in model assessment: Theoretic development and investigation. *Advances in Water Resources* **31**(8): 1087–1100.

Snelder TH, Biggs BJF. 2002. Multi-scale river environment classification for water resources management. *Journal of the American Water Resources Association* **38**(5): 1225–1240.

Sorooshian S, Duan Q, Gupta VK. 1993. Calibration of rainfall-runoff models: application of global optimization to the Sacramento

soil moisture accounting model. *Water Resources Research* **29**(4): 1185–1194.

Tait AB, Henderson RD, Turner RW, Zheng X. 2006. Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. *International Journal of Climatology* **26**(4): 2097–2115.

Thiebaux HJ, Zwiers FW. 1995. The interpretation and estimation of effective sample size. *Journal of Climate and Applied Meteorology* **23**: 800–811.

Thiemann M, Trosset M, Gupta H, Sorooshian S. 2001. Bayesian recursive parameter estimation for hydrologic models. *Water Resources Research* **37**(10): 2521–2535.

Thyer M, Renard B, Kavetski D, Kuczera G, Franks SW, Srikanthan S. 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Research* **45**: W00B14. DOI: 10.1029/2008WR006825.

Van der Made JE. 1982. Determination of the accuracy of water-level observations. *Proceedings of the Exeter Symposium*, July 1982 (Advances in Hydrometry). Publ. no. 134. IAHS; 172–184.

Venetis C, IASH. 1970. A note on the estimation of the parameters in logarithmic stage–discharge relationships with estimation of their error. *Hydrological Sciences Bulletin.* **15**: 105–111.

Vrugt JA, Gupta HV, Bouten W, Sorooshian S. 2003. A Shuffled Complex Evolution Metropolis algorithm for optimisation and uncertainty assessment of hydrologic model parameters. *Water Resources Research* **39**(8): 1201.

Vrugt JA, ter Braak CJF, Clark MP, Hyman JM, Robinson BA. 2008. Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resources Research* **44**: W00B09, DOI:10.1029/2007WR006720.

Wagener T, Gupta HV. 2005. Model identification for hydrological forecasting under uncertainty. *Stochastic Environ. Res. Risk Assess.*, **19**(6): 378–387, DOI:10.1007/s00477-005-0006-5.

Whalley N, Iredale RS, Clare AF. 2001. Reliability and uncertainty in flow measurement techniques—Some current thinking. *Physics and Chemistry of the Earth Part C-Solar-Terrestial and Planetary Science* **26**(10–12): 743–749.

Wilks DS. 1997. Resampling hypothesis tests for autocorrelated fields. *Journal of Climate* **10**(1): 65–82.

Williman EB. 1995. Flood frequency analysis for the Wairau river, Marlborough. *NewZealand Journal of Hydrology* **33**(2): 71–93.

