A comparison of methods for streamflow uncertainty estimation

J.E. Kiang, C. Gazoorian, Helen Mcmillan, G. Coxon, Jérôme Le Coz, I. Westerberg, A. Belleville, D. Sevrez, A.E. Sikorska, A. Petersen Overleir, et al.

AGU100 ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2018WR022708

Key Points:
- Methods for estimating the stage-discharge rating curve and its uncertainty were compared for stream gauges with varying hydraulic complexity
- Uncertainty estimates varied widely at high and low flows for the different methods
- Careful description of the assumptions behind uncertainty methods is needed

Supporting Information:
- Supporting Information S1
- Data Set S1

Correspondence to:
J. E. Kiang,
jkiang@usgs.gov

Citation:
Kiang, J. E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I. K., et al. (2018). A comparison of methods for streamflow uncertainty estimation. *Water Resources Research*, *54*, 7149–7176. https://doi.org/10.1029/2018WR022708

Received 16 FEB 2018
Accepted 9 AUG 2018
Accepted article online 21 AUG 2018
Published online 2 OCT 2018

A Comparison of Methods for Streamflow Uncertainty Estimation

Julie E. Kiang[1], Chris Gazoorian[2], Hilary McMillan[3], Gemma Coxon[4,5], Jérôme Le Coz[6], Ida K. Westerberg[7], Arnaud Belleville[8], Damien Sevrez[8], Anna E. Sikorska[9], Asgeir Petersen-Øverleir[10], Trond Reitan[11], Jim Freer[4,5], Benjamin Renard[6], Valentin Mansanarez[6,12], and Robert Mason[1]

[1]Water Mission Area, U.S. Geological Survey, Reston, VA, USA, [2]New York Water Science Center, U.S. Geological Survey, Troy, NY, USA, [3]Department of Geography, San Diego State University, San Diego, CA, USA, [4]School of Geographical Sciences, University of Bristol, Bristol, UK, [5]The Cabot Institute, University of Bristol, Bristol, UK, [6]Hydrology-Hydraulics, IRSTEA, Lyon, France, [7]IVL Swedish Environmental Research Institute, Stockholm, Sweden, [8]EDF-DTG Department of Water Monitoring, Grenoble, France, [9]Department of Geography, University of Zurich, Zurich, Switzerland, [10]Statkraft Energi AS, Oslo, Norway, [11]NVE, Oslo, Norway, [12]Department of Physical Geography, Stockholm University, Stockholm, Sweden

Abstract Streamflow time series are commonly derived from stage-discharge rating curves, but the uncertainty of the rating curve and resulting streamflow series are poorly understood. While different methods to quantify uncertainty in the stage-discharge relationship exist, there is limited understanding of how uncertainty estimates differ between methods due to different assumptions and methodological choices. We compared uncertainty estimates and stage-discharge rating curves from seven methods at three river locations of varying hydraulic complexity. Comparison of the estimated uncertainties revealed a wide range of estimates, particularly for high and low flows. At the simplest site on the Isère River (France), full width 95% uncertainties for the different methods ranged from 3 to 17% for median flows. In contrast, uncertainties were much higher and ranged from 41 to 200% for high flows in an extrapolated section of the rating curve at the Mahurangi River (New Zealand) and 28 to 101% for low flows at the Taf River (United Kingdom), where the hydraulic control is unstable at low flows. Differences between methods result from differences in the sources of uncertainty considered, differences in the handling of the time-varying nature of rating curves, differences in the extent of hydraulic knowledge assumed, and differences in assumptions when extrapolating rating curves above or below the observed gaugings. Ultimately, the selection of an uncertainty method requires a match between user requirements and the assumptions made by the uncertainty method. Given the significant differences in uncertainty estimates between methods, we suggest that a clear statement of uncertainty assumptions be presented alongside streamflow uncertainty estimates.

Plain Language Summary Knowledge of the uncertainty in streamflow discharge measured at gauging stations is important for water management applications and scientific analysis. This paper shows that uncertainty estimates vary widely (typically up to a factor of 4) when comparing seven recently introduced estimation methods. A clear understanding of the assumptions underpinning different uncertainty estimation methods and the sources of uncertainty included in their calculations is needed when selecting a method and using and presenting its uncertainty estimates.

1. Introduction

Streamflow time series data are fundamental to hydrological science and water management applications. Uncertainty in streamflow data directly translate into uncertainty in hydrologic models (e.g., Liu et al., 2009; McMillan et al., 2010), into research conclusions and management decisions (Wilby et al., 2017), and may result in significant economic costs (McMillan et al., 2017). Understanding the causes and characteristics of streamflow uncertainty, as well as how it can best be estimated, is therefore an important research task.

Streamflow time series for river gauging stations are most often computed through the use of a stage-discharge rating curve that relates measured river stage to streamflow discharge. The rating curve is usually developed using discrete, concurrent measurements of stage and discharge as calibration data. Rating curves are used because discharge is difficult to measure continuously, while methods to continuously monitor stage height are readily available. The rating curve function is typically selected to be consistent with the

physics of open channel flow for the river cross section and controlling reach in question and often combines multiple segments to represent the riffles, weirs, channel bank controls, backwater effects, and overbank flows that occur. This indirect calculation of streamflow data from stage, in addition to uncertainties in stage height and channel definition, generates discharge uncertainties that are not always apparent or reported to users of streamflow data.

Multiple sources of uncertainty impact the formulation of rating curves and hence streamflow estimation; McMillan et al. (2012) provide an in-depth review of these. There are three major components. (1) Measurement errors in the underlying stage-discharge gaugings, which are usually approximated as random errors (Coxon et al., 2015; Pelletier, 1988). (2) Imperfect approximation of the true stage-discharge relation by the rating curve model, including extrapolation of rating equations to higher and lower discharges beyond the range of the stage-discharge gaugings. (3) Ignored additional drivers that may create instability in the stage-discharge relation, for example, factors including unsteady flow, variable backwater effects, or changes to the channel cross section due to sediment transport, vegetation growth, or ice formation.

The combined effects of these uncertainty sources result in significant uncertainty in streamflow values. Typical total uncertainties (95% uncertainty intervals) have been estimated at ±50–100% for low flows, ±10–20% for medium to high flows, and ± 40% for out of bank flows (McMillan et al., 2012). Alternative methods exist that estimate streamflow and its associated uncertainty without the use of a stage-discharge rating curve. *Index velocity* methods measure local velocity directly and then apply a conversion to average cross-section velocity (Levesque & Oberg, 2012). For flood flows, noncontact techniques such as Particle Image Velocimetry can estimate velocity from observations of the flow surface (Muste et al., 2011). In many regions of the world, rivers are inaccessible or gauging technology is not easily available, and therefore, remote sensing techniques for streamflow and uncertainty estimation are desired (e.g., Bjerklie et al., 2005). However, in this paper we focus only on stage-discharge rating curve methods, as they are the most widely used.

Many different methods have been suggested to estimate uncertainty in stage-discharge rating curves. The traditional method for stage-discharge rating curve uncertainty estimation is the linear regression method (Herschy, 1999) proposed in International Organization for Standardization (ISO) international standards (ISO 1100-2:2010, ISO/PWI 18320) and World Meteorological Organization (WMO) technical regulations (World Meteorological Organization (WMO), 2006), though we are not aware of any agency having applied it routinely. This method has been extended to include additional error sources and effects of time averaging (Clarke, 1999; Dymond & Christian, 1982; Venetis, 1970); however, the research community has recently developed multiple methods for different gauging stations and types of stage-discharge relationships. These methods use a variety of different approaches for rating curve uncertainty estimation, including assessing gauging deviations (Tomkins, 2014), fuzzy methods (Shrestha et al., 2007; Westerberg et al., 2011), locally weighted regression (LOWESS) regression (Coxon et al., 2015; Mason et al., 2016), Generalised Likelihood Uncertainty Estimation (Guerrero et al., 2012), Bayesian informal (McMillan & Westerberg, 2015), Bayesian formal (Juston et al., 2014; Le Coz et al., 2014; Moyeed & Clarke, 2005; Reitan & Petersen-Overleir, 2008; Sikorska et al., 2013), dynamic rating curve analysis (Jalbert et al., 2011; Morlot et al., 2014; Reitan & Petersen-Overleir, 2011), and perturbations introduced into a hydraulic model (Di Baldassarre & Claps, 2011 Di Baldassarre & Montanari, 2009; Domeneghetti et al., 2012).

We argue that there is no single optimal method for streamflow uncertainty estimation because each method makes different assumptions about the sources and types of uncertainty and thus how the rating model is calculated. Different perceptual understandings of the dominant uncertainty sources drive the formulation of the different methods (Westerberg et al., 2017). Methods have been designed for different purposes and may range from locally specific to generalized in their application, be more suitable for smaller or larger rivers, and have lower or higher requirements for the availability of gaugings and metadata. Therefore, different methods will be preferable in different circumstances. Methods differ in how the estimated discharge uncertainty is normally presented, for example, as upper/lower bounds (Westerberg et al., 2011), distributions of discharge for each stage value (Coxon et al., 2015), or full distributions of rating curve samples (Le Coz et al., 2014; McMillan & Westerberg, 2015). These differences in output may restrict the ways in which uncertainty can be propagated to other analyses such as hydrological model calibration and hydrological signature uncertainty calculation.

Table 1
Examples of Design Questions That Must Be Answered During Development of a Discharge Uncertainty Estimation Method

Model specification	▪ Which types of rating curve models should be considered (e.g., piecewise power functions)? ▪ How much user knowledge of the model, parameters, and gauging station characteristics should be required? ▪ When utilized, how should priors be assigned, using hydraulic data or otherwise?
Changes over time	▪ How do rating curves change over time, how can changes be detected, and how can this be incorporated in the method?
Extrapolation	▪ Can the rating curve be safely extrapolated beyond current gaugings to lower or higher flows? ▪ How can hydraulic knowledge be used to constrain out of bank extrapolations?
Data	▪ How many gaugings are required for discharge uncertainty estimates and how should they be distributed across the flow range? ▪ How should outliers and questionable (or more or less certain) data be handled?

Table 1 illustrates some of the necessary decisions when designing an uncertainty estimation method. Typically, these questions do not have *correct* answers, because we lack full information on error sources and characteristics. This lack of an optimal method has been referred to as *uncertainty about uncertainty* or *uncertainty*^{2} (Juston et al., 2014).

Given the large number of estimation methods available, it is important to understand how uncertainty estimates differ between methods and how these differences depend on the assumptions and methodological choices made in each method. To date, there has been little coordination between the diverse research groups developing discharge uncertainty estimation methods. Limited previous studies have compared some pairs of methods (Mason et al. 2016; Ocio et al., 2017; Storz, 2016), but we know of no broader comparisons.

In this paper we present a first attempt to bring together and compare several streamflow uncertainty estimation methods. Intercomparison experiments have an important place in hydrology as a way to compare and benchmark competing methods. For example, the Model Parameter Estimation Experiment compared a priori methods for hydrologic/land surface parameter estimation (Duan et al., 2006); and the Hydrological Ensemble Prediction Experiment facilitated comparisons of ensemble forecasting (Schaake et al., 2007). Comparison experiments are today aided by new tools such as Virtual Science Laboratories that provide a central location for researchers to share data, metadata, models, and protocols, helping to address the issue of ensuring reproducibility of hydrologic experiments (Ceola et al., 2015; Hutton et al., 2016). We discuss our experimental design and protocols in section 2.

This paper summarizes and reviews seven methods for estimating uncertainty in rating curves (section 2), with diverse assumptions and methodological choices (Tables 2 and 3), that are actively maintained and in current use by their respective research groups. Although not exhaustive, we believe that the methods included provide a representative sample of potential differences in uncertainty estimation techniques. Because streamflow uncertainty is highly dependent on hydraulic characteristics of the gauging site, hydrologic regime, and streamflow gauging practices, we compared uncertainty estimation results from all of the methods at three gauging locations with diverse rating characteristics. We chose two sites in Europe and one in New Zealand, with stage-discharge relations that range from simple and stable to complex and time varying. The paper aims to illustrate the importance of understanding the assumptions involved in stage-discharge rating curve uncertainty estimation, to describe what characteristics these existing models have in common, and how their differences affect the resultant uncertainty estimates. It also aims to provide guidance on the suitability of each method for a variety of typical river gauging stations, and finally, we suggest critical next steps to improve our treatment of streamflow uncertainties in light of this paper's findings.

2. Experimental Design

This section describes the experimental design that we used to compare river discharge uncertainty estimates between methods. We present the rationale for the comparison experiment and the scope of our design, describe the stream gauges where we conduct the comparison, summarize the uncertainty estimation methods and their assumptions, and describe the uncertainty components that they treat.

2.1. Method Comparison Experiment

We designed the method comparison experiment during two workshops that brought together research groups who had developed their own methods. Our discussions highlighted a lack of knowledge on how

Table 2
Summary of Uncertainty Estimation Methods

Methods	Main principles	Rating equation (main model)	Type of uncertainty results	Users	Main references
BaRatin	Bayesian, MCMC sampling, matrix of controls user defined from preliminary hydraulic analysis	Piecewise/added power functions	Sampled distribution	Research + operational (SCHAPI, CNR)	Le Coz et al. (2014)
Bristol	Nonparametric LOWESS regression and multisection rating curves	Piecewise linear	Sampled distribution	Research	Coxon et al. (2015)
GesDyn	Gauging segmentation, time variographic analysis, MC sampling, and one rating estimated for each gauging	Piecewise power functions	Sampled distribution	Research + operational (EDF)	Morlot et al. (2014)
ISO/WMO	Analysis of linear regression residuals	Piecewise power functions	Standard deviations	Not known	ISO 18320, WMO (2006)
NVE/HydraSub	Bayesian, MCMC sampling, and objective segmentation	Piecewise power functions	Sampled distribution	Research + operational (NVE)	Reitan and Petersen-Overleir (2008)
BayBi	Bayesian, MCMC sampling, and bias correction	Piecewise power functions	Sampled distribution	Research	Sikorska et al. (2013)
VPM	MCMC sampling, likelihood function that accounts for random measurement errors, and epistemic errors	Piecewise power functions	Sampled distribution	Research	McMillan and Westerberg (2015)

Note. SCHAPI= France's national hydrologic service; CNR = Compagnie nationale du Rhone, French hydropower company; EDF = Electricité de France; NVE = Norwegian Water Resources and Energy Directorate; LOWESS = locally weighted regression; ISO = International Organization for Standardization; WMO = World Meteorological Organization; MCMC = Markov Chain Monte Carlo.

different assumptions and methodological decisions impact discharge uncertainty estimates at different stream gauges with different characteristics and data availability.

We therefore decided to undertake a thorough comparison of the participating discharge uncertainty estimation methods. Three stream gauges were selected, representing low, medium, and high perceived rating curve complexity (section 2.2). For each gauge, we compiled a data set including

1. a stage time series (either 15-min or hourly data),
2. stage-discharge gauging points (including gauging time, gauging method, and uncertainty of the measured discharge if this information was available),
3. official (operationally used) rating curves, and
4. gauging site characteristics (e.g., photographs, cross-section information, and catchment descriptors).

Each research group applied their own uncertainty method to each stream gauge data set. Results were returned as follows: estimated rating curve and uncertainty quantiles for specified stage values, discharge series, and associated uncertainty quantiles. These results were then compiled and compared centrally.

2.2. Stream Gauge Descriptions

Three stream gauges with sufficient minimum information to run all the models were selected to compare the discharge uncertainty estimation methods: (1) the Isère at Grenoble Campus (France), (2) the Mahurangi at College (New Zealand), and (3) the Taf at Clog-y-Fran (United Kingdom, see Figure 1). These stream gauges were chosen first to cover a range of conditions impacting uncertainties in the stage-discharge relationship ranging from simple to more complex cases and second as all three gauges had the necessary data available so that the different methods could be easily compared. These stream gauges have been well studied, for example, Bayesian approaches (Thyer et al., 2011) and the Voting Point Method (VPM) method (McMillan & Westerberg, 2015) have been used to estimate discharge uncertainty at Mahurangi, while the Bristol method (Coxon et al., 2015) and the VPM method (Westerberg et al., 2016) have been applied to the Taf using a different data period than that used here. However, discharge uncertainty estimates from multiple methods, within a consistent set of experimental protocols, have not been compared at these gauging stations.

2.2.1. The Isère River at Grenoble Campus, France

The Isère River at Grenoble Campus, in the French Alps, was chosen as the simplest case to apply the discharge uncertainty methods as it has a single, stable channel control over the entire range of stages. The Isère is the largest of the three rivers (mean discharge 179 m^3/s) with large seasonal fluctuations in flow due to rain and snow melt. It is highly regulated due to many dams and water diversions. Although there is a cableway to measure high flows at the gauge, it is very difficult to obtain accurate measurements due to fast velocities and floating debris. At the station, the Isère flows in a wide, fairly uniform alluvial channel with overbank flows strictly limited by dikes (out of bank height is approximately 6.5 m). Both operational and academic operators produce frequent gaugings over the full range of observed stages (168 gaugings between January 1998 and August 2015). The rating curve has changed at this station due to the occurrence of a major flood (16 October 2000) and in-channel gravel mining downstream of the station (2 February 2013). In this study, only gaugings after the October 2000 flood and before the February 2013 modifications were used to allow a comparison of the methods for a single stable period. For each gauging, the expected discharge gauging uncertainty (±5% or ± 7%) was provided based on the measurement method (either an acoustic Doppler current profiler or a mechanical current meter, respectively).

2.2.2. The Mahurangi River at College, New Zealand

The Mahurangi River at College allows the comparison of different discharge uncertainty estimates for a multisegment stage–discharge relationship, which has extrapolation of

Table 3
Summary of Method Differences

	Data needed			Can include gauging uncertainties explicitly		Time variation			Output				
	Stage-discharge gaugings	Official rating curves	Information on hydraulic controls	Stage	Discharge	Pooling or auto batching of gauging data	Dynamic treatment	Assumption that rating can continually change	Confidence intervals	Prediction intervals	Rating curve samples (from Monte Carlo)	Extrapolation of the rating curve	Available online
ISO	+	(+)	(+)	–	–	–	–		+	+	–	–	–
Bristol	+	(+)		+	+	+	–		–	+	–	–	–
BaRatin[a]	+	(+)	(+)	+	+	–	–		+	+	+	+	+
BayBi[a]	+	(+)	(+)	–	+	–	–		+	+	+	+	–
NVE[a]	+	(+)	(+)	+	+		–		+	+	+	+	+
VPM	+	(+)	(+)	+	+	–	–	+	–	+	+	+	–
GesDyn	+		(+)	–	+	+	+	+	+	–	+	+	–

Note. + indicates that a method either needs this particular data set or can do a particular task, (+) indicates a method can incorporate such data, blank indicates that the model is not specifically designed for the situation or does not need to use such data, – indicates that the method does not have this capability. NVE = Norwegian Water Resources and Energy Directorate; VPM = Voting Point Method.
[a]BaRatin, BayBi, and NVE have versions for dynamic treatment of time variations that were not used in this study.

the rating curve to the highest flows. It is the smallest of the three rivers (mean discharge 1.1 m^3/s) and has a fast rainfall–runoff response with few peak flows that typically last a few hours, which makes it challenging to gauge the whole flow range. It is a low-land stream with a mild-slope channel that meanders between narrow, wooded floodplains (about 20 m wide in total). Out of bank flows occurs at approximately 4–4.3 m on the right flood bank.

The station is equipped with a 90°, 600 mm-high V-notch weir nested within a wider triangular weir with a wing slope of 1:10. The official rating curve follows a three-segment model. Four successive ratings were defined due to weir break and repair (April–May 1995) and top end change due to excavator work in the channel (since 7 May 2010), however, these minor perturbations were ignored in this study and a single rating period was considered. Seventy-eight gaugings have been produced from 1985 to 2013 over a wide range of stages, but the highest flows require significant extrapolation from the highest gauging of 2.7 m to the highest recorded stage of 4.2 m. One outlying gauging was discarded after indication of adverse measuring conditions by the station manager. The gauging authority also classifies two of the high flow gaugings as poor quality because of large (12–22 cm) stage changes during gauging, which potentially reflects stage–discharge hysteresis. These two high-flow gaugings were included in the discharge uncertainty estimates as an interesting case of epistemic uncertainty. For each gauging, the expected gauging uncertainty (±5% or ± 7%) was provided based on the measurement method (acoustic Doppler current profiler or current meter, respectively).

2.2.3. The Taf River at Clog-y-Fran, Wales, United Kingdom

The Taf River at Clog-y-Fran gauging station was chosen as a challenging case for rating-curve uncertainty estimation as it has a gauged out-of-bank section (it overflows on the right flood bank at approximately 3.2–3.4 m) and multiple changes in the rating curve at low flows due to deposition of silt. The river has a mean annual discharge of 7.6 m^3/s. It is a predominantly lowland natural catchment and is underlain by low permeability bedrock resulting in a fast rainfall-runoff response. River flows from this site are used in the calibration of regional flood forecasting models and as an indicator for the onset of potential drought conditions in the UK, thus making it an important site for which to consider flow data quality and uncertainty. At the gauging station, there is a natural alluvial bed and discharge is measured using the velocity-area method with a cableway to allow high flow gauging. The official rating curve is a multisegment rating curve that has changed 13 times over 51 years of operation. In this study, we chose to focus on the period from 1990 to 2012 as the stage and discharge time series included a great deal of missing data before this point. This time period included six changes to the official rating curve and a total of 442 gaugings. There was no information available on the method and instrument used to take the gaugings or the likely gauging uncertainty. Consequently, each group specified their own gauging uncertainties for this site as described below.

2.3. Discharge Uncertainty Components and Intervals

Rating curve uncertainty arises from numerous sources (as described in section 1), and one of the challenges in this study was that each model treats the uncertainty components differently, making it difficult to directly compare discharge uncertainty results and attribute reasons for differences across methods. Consequently, for the purposes of this study, we divide the uncertainty sources into three components:

1. *Parametric uncertainty*: uncertainty arising from model parameter identification (usually, the parameters of the power law function). Estimated parameters are uncertain because of limited calibration sample size and errors affecting calibration data.

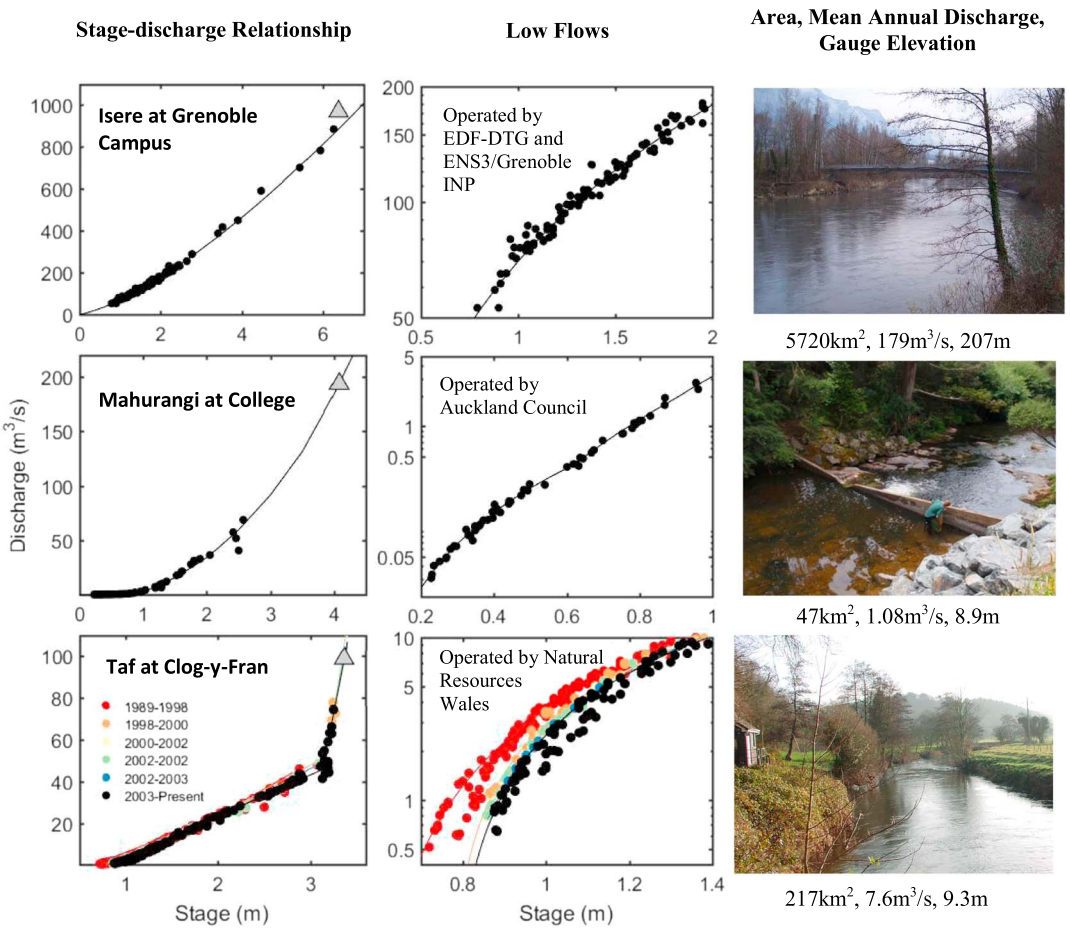

Figure 1. Stage-discharge relationships for the full flow range (left column) and for low flows (middle column). In the left column, the gray triangle denotes the maximum measured stage. Gauging station characteristics and photos of the three gauging stations are shown in the right column.

2. *Structural uncertainty*: uncertainty arising because the model is imperfect and does not include all elements required to model the channel hydraulics and the potentially time variable stage-discharge relationship.
3. *Measurement uncertainty*: uncertainty arising from imperfect measurements of stage and discharge when these measurements are used to evaluate the rating curve.

The discharge predicted with the rating curve model aims at estimating the unknown true discharge—as opposed to the discharge that can be measured only imperfectly. The preferred uncertainty intervals would therefore include parametric and structural uncertainty but not the measurement uncertainty. While we attempted to utilize outputs from each method to do so, not every method is able to provide outputs that include just these two components of uncertainty. There are two key differences among uncertainty intervals shown for each method that relate to (1) which uncertainty components are included in the estimation and (2) the assumptions about their interactions. Table 4 shows the components included by each method in the uncertainty intervals used in this paper.

As described further below, the formal Bayesian methods (Bayesian Bias correction [BayBi], Bayesian Rating curve [BaRatin], and Norwegian Water Resources and Energy Directorate [NVE]) use an explicit error model to separate the different uncertainty components and can thereby produce uncertainty intervals with any components. In producing prediction bounds for true discharge, measurement uncertainty is removed (i.e., the bounds consist only of parametric and structural uncertainty). The Bristol and VPM methods are based on the assumption that the three different components cannot be separated and include all three components jointly in the estimation and the uncertainty bounds. For the ISO

Table 4
Components of Uncertainty Included in Uncertainty Intervals for Each Method

Methods	Uncertainty components included in uncertainty intervals in this paper
ISO/WMO	Parametric, structural, and measurement
Bristol	Parametric, structural, and measurement(all components taken into account jointly)
BaRatin	Parametric and structural
BayBi	Parametric and structural
NVE	Parametric, structural, and measurement
VPM	Parametric, structural, and measurement(all components taken into account jointly)
GesDyn	Parametric and measurement

Note. ISO = International Organization for Standardization; WMO = World Meteorological Organization; VPM = Voting Point Method; NVE = Norwegian Water Resources and Energy Directorate.

method, uncertainty intervals are constructed by using the residual mean square errors. These residual errors reflect all three components of uncertainty. The GesDyn method includes parametric uncertainty and measurement uncertainty only.

We compared our methods using the uncertainty intervals computed by each method and present both 95% and 68% intervals. These percentage intervals were chosen as they correspond to one standard deviation and two standard deviations of a Gaussian distribution, and 95% is recommended by the Hydrometric Uncertainty Guidance, ISO/TS 25377:2007. It is important to note that discharge errors are typically highly autocorrelated and vary systematically with stage (i.e., not randomly within the uncertainty intervals).

2.4. Uncertainty Estimation Methods

This section summarizes the seven uncertainty estimation methods. Longer descriptions of each method from the contributing research groups are provided in Appendix A. Here we explain the main principles, similarities, and differences of the methods focusing on their basic principles, the data needed to run the methods, which key sources of uncertainty are included, and how discharge uncertainty results are provided. Table 2 introduces the name, uncertainty estimation principles, rating equation type, result format, and expected end users for each method and key reference. Table 3 summarizes the key similarities and differences between the methods.

2.4.1. Basic Principles, Similarities, and Salient Differences

All seven methods are fundamentally based on the regression of piecewise (segmented) power functions (except Bristol, see Table 2) using stage-discharge gauging data and accounting for data uncertainties (except ISO). The methods used in this paper can be very broadly classified into four categories.

Two of the methods, ISO and Bristol, could be considered to be based on a least squares regression framework. They mainly differ from each other in their stage-discharge models, where the Bristol method's local nonparametric regression method is more flexible than the ISO method. Being driven by stage-discharge gaugings, these methods can not only be applied within the gauged range but also require only minimal information to provide discharge uncertainty estimates.

Three methods, BaRatin, BayBi, and NVE, use a formal Bayesian framework and accept prior information from the user to inform the model development. These three formal Bayesian methods are very similar in their basic principles but differ in their implementation. Most importantly, BaRatin requires more user information to constrain the rating curve model and parameters than BayBi, and BayBi more than NVE. In terms of modeling philosophy, NVE uses an objective segmentation and estimation of the rating curve, whereas BaRatin builds on the subjective expertise of the field hydrologist, expressed in formal terms. As a consequence, extrapolation above the highest gauging (or below the lowest gauging) is expected to be less uncertain with BaRatin, provided that the assumed structure of the controls is correct. Another technical difference is the way structural errors are accounted for (see Appendix A). In BayBi, the standard deviation of the structural error is assumed to be constant with a rather small prior value (1% of the average recorded discharge). BaRatin offers several structural error models, the default being a linear model in which the standard deviation of the structural error is the sum of a constant term and of a term that is proportional to discharge. There is no structural error term in NVE, instead the structural error may be captured through the sampling of rating curve models with different numbers of segments.

VPM and GesDyn are in their own categories because they do not assume that the underlying stage-discharge relationship is constant over time, whereas other methods rely on the existence of stable periods where gaugings can be grouped and assumed to derive from a constant stage-discharge relationship. VPM considers unknown temporal variability in the stage-discharge relationship as a key source of uncertainty and incorporates this in the estimated uncertainty intervals, whereas GesDyn tracks temporal variability by deriving a new rating curve for each gauging and increasing its uncertainty with time up to the next gauging.

2.4.2. Data Requirements

All the methods rely on stage-discharge gaugings to produce discharge uncertainty bounds. Some of the methods (e.g., ISO and Bristol) specify minimal numbers of gaugings (20 per segment and 20 per gauge, respectively) to ensure robust discharge uncertainty bounds. The formal Bayesian methods BaRatin, BayBi, and NVE also utilize knowledge on the hydraulic controls to set informative priors on the parameters of the stage-discharge relationship (although they can be run with standard priors if no hydraulic information is available). Informative priors based on hydraulic knowledge can also be used with the VPM method (Ocio et al., 2017). Operational rating curves are used by some of the methods either to identify segments in the stage-discharge relationship (VPM) or to subset gaugings (Bristol and NVE).

2.4.3. Measurement Uncertainty in Stage and Discharge Gauging Data

Measurement uncertainty in the stage and discharge gauging data are important components of uncertainty in rating curve estimation. Most of the methods can incorporate information on the magnitude of stage and discharge gauging uncertainties for parameter estimation, except BayBi and GesDyn that only incorporate discharge uncertainties and ISO that cannot incorporate either stage or discharge gauging uncertainty. While the Bristol method can incorporate stage gauging uncertainty, in this study only discharge gauging uncertainty was included. For this study, methods did not incorporate stage uncertainty when utilizing the gaugings for parameter estimation, except for VPM for all sites and the BaRatin method for the Mahurangi.

To enable consistent comparative analyses between techniques and because of a lack of reliable data, uncertainties regarding the stage time series readings themselves were not considered by any of the models in the generation of the discharge time series and their uncertainties.

2.4.4. Uncertainty in the Rating Curve

When the rating curve consists of multiple segments, the breakpoints between the segments relate conceptually to substantial changes in the governing control at different flow depths. Such changes are typically caused by hydraulic changes (going from one section control to another section control or to channel control), geometrical alterations (one control changing shape), or significant variations in flow resistance.

Most of the methods for rating curve uncertainty estimation used in this paper represent rating curve equations as piecewise power functions and include uncertainty in the breakpoint parameters. Typically, the user must estimate the number of rating curve segments and the values of the rating curve and breakpoint parameters either using official rating curves or expert knowledge. All three formal Bayesian methods and VPM allow the user to estimate the power function parameter priors based on hydraulic principles and information on the stream gauge location. BayBi, BaRatin, and VPM expect the user to define the number of segments, whereas NVE infers the number of segments (up to a given maximum) like other parameters of the rating curve model. All three formal Bayesian methods and VPM can be run with default, noninformative parameter priors. However, BayBi typically expects prior information on segment limits (breakpoints and zero-flow stage), while BaRatin provides practical guidance to the user to compute parameter priors from common knowledge of the field sites.

The Bristol method is different in that it uses numerous piecewise linear segments that are not related to cross-section shape. Local regression methods do not have a physical basis for their rating curve model and are therefore not able to extrapolate outside the gauged stage range. For the methods using power functions, the rating curve can be extrapolated above/below the highest/lowest gauging. This assumes that there is no substantial change in channel control within the extrapolated stages, such as due to overbank flow. Uncertainty is typically higher in the extrapolated part(s) and for segments with few or no gaugings; in this case, the parameter priors play an important role in constraining the estimated uncertainty.

Gauging uncertainties are often provided as upper and lower discharge bounds, which the methods incorporate as probability intervals for an assumed distribution (e.g., uniform or Gaussian). These distributions are then incorporated in different ways. Some methods use the gauging error distributions as one of the

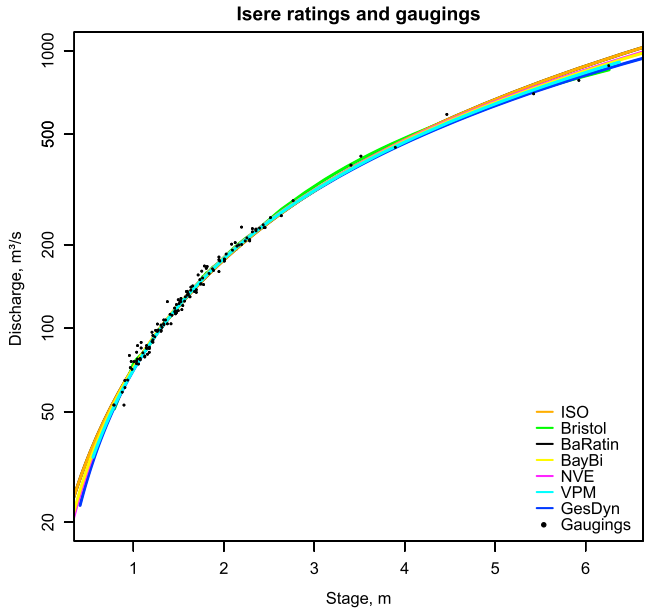

Figure 2. Stage-discharge rating curves for all methods for the Isère gauging station. ISO = International Organization for Standardization; VPM = Voting Point Method; NVE = Norwegian Water Resources and Energy Directorate.

terms in a likelihood function that defines the likelihood of the actual gauging measurements given candidate rating curve parameters (e.g., VPM, BaRatin, and BayBi). Other methods create synthetic gaugings by sampling directly from the discharge gauging errors (GesDyn) or from the stage and discharge gauging errors (Bristol) and then fit a rating curve to the synthetic gaugings.

All methods apart from the Bristol and ISO methods use Monte Carlo estimation of rating curve samples (Table 4). In this paper, to calculate discharge time series uncertainties the Monte Carlo approaches use rating curve samples from which multiple feasible discharge time series realizations are calculated (e.g., Westerberg et al., 2016), whereas the Bristol method provides prediction intervals derived from a Gaussian mixture model of the rated discharge for any given stage (although time series samples can be derived using an appropriate error model, for example, Lloyd et al., 2016). Such sets of time series samples can then be used to propagate discharge uncertainty to assess impact on any subsequent analysis by performing that analysis repeatedly for each time series sample (e.g., calculation of total annual discharge as in Juston et al., 2014).

2.4.5. Time Variable Rating Curves

As discussed in section 1, rating curves commonly vary due to changes in the river cross section and hydraulic conditions caused by, for example, sediment movement, vegetation, hysteresis, or ice. Consequently, a key characteristic of the methods is how they handle rating curve variation over time.

The most common approach, suitable for all the methods here, is to apply the uncertainty estimation separately to subsets of gauging data often corresponding to time periods defined by changes in official rating curves. The GesDyn method automatically batches the data into suitable subsets and the Bristol method pools gaugings together based on similarity of official rating curves. Apart from VPM and GesDyn, the rating curve is then assumed to be stationary within each subset (usually a contiguous time period). Residual time variation within each subset is handled differently between methods. VPM was specifically designed to capture uncertainty related to such temporal variability of the stage-discharge relation, via the design of its likelihood function. BayBi and BaRatin separate the residual error model into its component uncertainties, with a remnant error component that encompasses the effect of ignored time variability. GesDyn is the only method that explicitly models time variation. It updates the rating curve after each new gauging and uses variographic analysis to represent changes in uncertainty caused by down weighting of previous rating curves with time.

While extensions to NVE, BayBi, and BaRatin methods have been developed to provide alternative solutions for handling the time-varying nature of rating curves, these extensions have not been used for the comparisons in this paper.

3. Results

Each of the seven rating curve and uncertainty estimation methods were applied to the three sites described earlier. Note that the rating curves produced by each method are an estimate of the true stage-discharge relationship, which is unknown.

3.1. Isère River at Grenoble Campus

The stage-discharge relationship at the Isère River stream gauge is the least complex of our three sites and can be reasonably modeled as a single segment curve. Figure 2 shows the rating curves produced by each method and the gaugings used to develop the rating curves. Although GesDyn produces a new rating curve for each gauging, a single mean rating curve is shown in Figure 2, as well as for Mahurangi and Taf. Similarly, VPM and Bristol are not intended to produce a single optimal rating curve, but a median curve is shown for

the purpose of this paper. The rating curves are quite similar for midrange flows, where the bulk of individual gaugings has been made. The chief differences in the modeled ratings are at the high and low end.

Figure 3 shows the rating curve and uncertainty intervals for each method individually. The estimated uncertainty intervals differ substantially between methods, even for the midrange flows. At a stage of 2 m, near mean flow conditions where there are many gaugings, the smallest estimated uncertainties as quantified by 95% uncertainty intervals (Figure 4 and Table 5) are in the 3–4% range (NVE, BayBi, and GesDyn), while the largest are up to 17% (ISO and VPM). The differences in estimated uncertainty are more pronounced near the highest and lowest measured flows. Uncertainty intervals for GesDyn show a pronounced difference in uncertainty between the mean range of flows where there are many gaugings available and the high and low ends of the rating curve, where fewer gaugings are available.

For ISO, Bristol, and VPM, the uncertainty intervals represent the total uncertainty (parametric, structural, and measurement) and tend to be wider than for the other methods for this station. The VPM uncertainty intervals, in fact, cover all gaugings apart from one. This likely relates to its assumption of time variability, that is, that the gaugings may belong to different stage-discharge relations.

The differences in uncertainty between methods are most clearly seen in Figure 4, which plots the percentage uncertainty estimated by each method versus river stage. These percentage uncertainties can differ between methods by more than a factor of 2 throughout the range of flows and are especially pronounced at the higher and lower ends of the rating curve. Note that the percentage uncertainties are calculated against the rating curve produced by each method, and that these are different between the methods (Figure 2).

Figure 5 shows the 68% and 95% uncertainty intervals as applied to the hydrograph for 2 days in April 2012.

3.2. Mahurangi River at College

Mahurangi was chosen as our midcomplexity site and requires a rating curve with multiple segments, in comparison to the single segment at the Isère. Figure 6 shows the stage-discharge ratings computed by each method for a period in September 2012. For methods using individual segments, fitting the rating curve typically required at least three segments, corresponding to the three major changes to the control.

A particular challenge at Mahurangi is to develop the rating curve for high flows, because the highest gauging is 2.6 m (69 m^3/s on the official rating), but the highest recorded stage is nearly 4.2 m (nearly 200 m^3/s on the official rating). There is significant scatter in the high-flow gaugings potentially as a result of stage-discharge hysteresis, leading to added uncertainty in the extrapolation. Although most methods produce similar rating curves in the middle section, the difficulty at high flows is reflected in larger intermodel differences. For example, at a stage of 4 m, best estimates of discharge from the models range from approximately 110 to 190 m^3/s.

Figure 7 shows the rating curve and the 68% and 95% uncertainty intervals for each method applied to the Mahurangi, and these intervals are shown as percentage uncertainty versus stage in Figure 8. The estimated uncertainties for the Mahurangi tended to be larger than for Isère, with larger differences between methods in the ungauged range but not the gauged range (apart from BayBi). Note that the parametric uncertainty is expected to be larger for this site because rating models are more complex for the Mahurangi. For example, the BaRatin model has 5 segments and hence required 15 parameters to be estimated for Mahurangi versus 3 parameters for the single segment model used for Isère. A mitigating factor is that for methods that could incorporate it, prior information on the low-flow and mid-flow controls (weirs) was more precise.

The differences in estimated uncertainty among the methods were most pronounced at the extremes and in the extrapolated section but remained relatively large even for midrange flows. Some of the reasons for differences in estimated uncertainty can be traced back to the method assumptions. For example, where methods utilize prior information to define the rating based on hydraulic principles (BaRatin, for example), this information can be used to constrain the uncertainty in the extrapolated stage range. While the NVE method also uses hydraulic principles, different assumptions about uncertainty in the extrapolated area (see section 2.4.1) results in a large increase in uncertainty for stages above the highest gaugings. Incorporation of prior information based on hydraulic principles is not possible for methods such as ISO or Bristol. The ISO method truncates the higher end of the rating because insufficient gaugings are available to fit the rating curve.

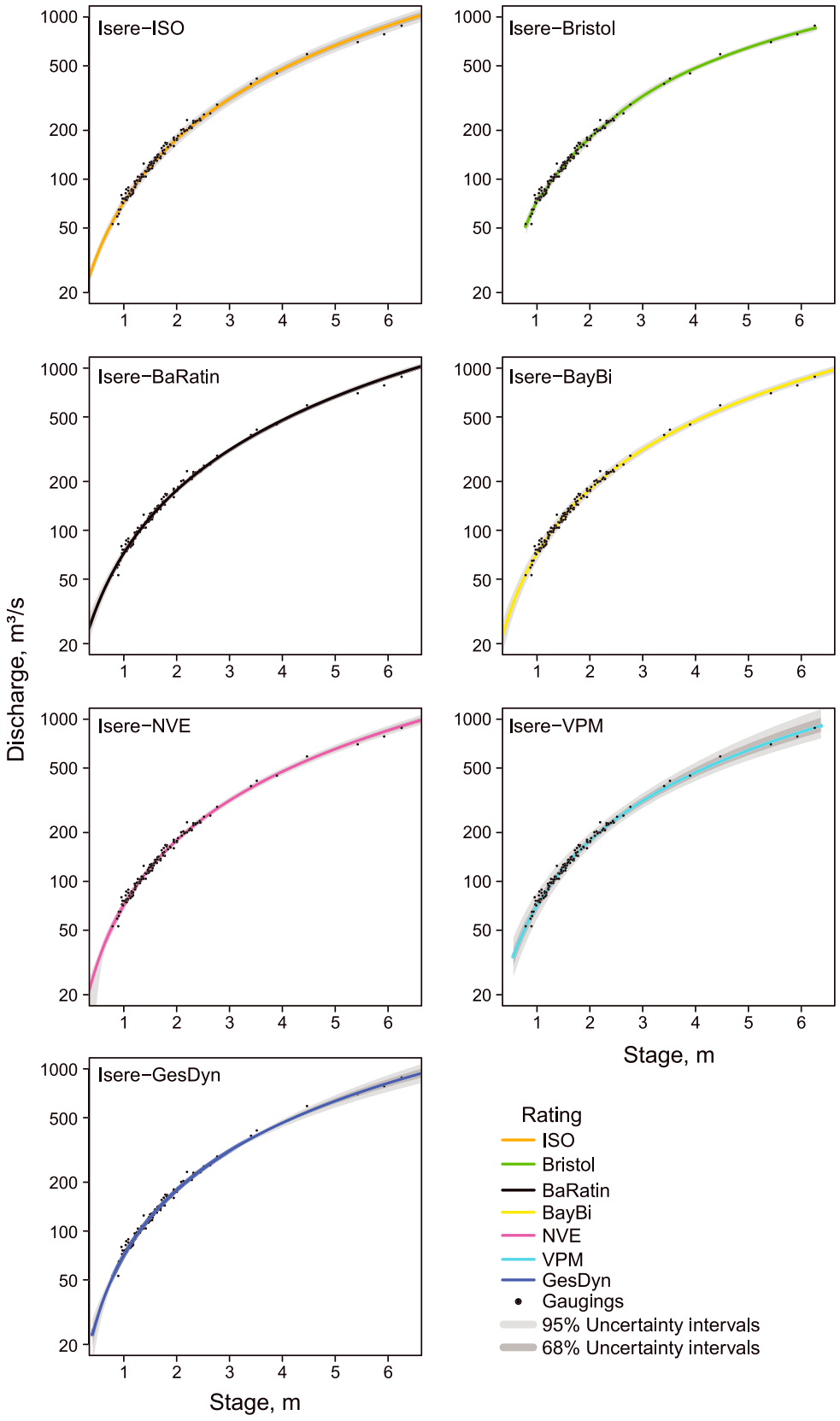

Figure 3. Rating curves and uncertainty intervals for all methods for Isère. The 68% and 95% uncertainty intervals are shown for all models. ISO = International Organization for Standardization; VPM = Voting Point Method; NVE = Norwegian Water Resources and Energy Directorate.

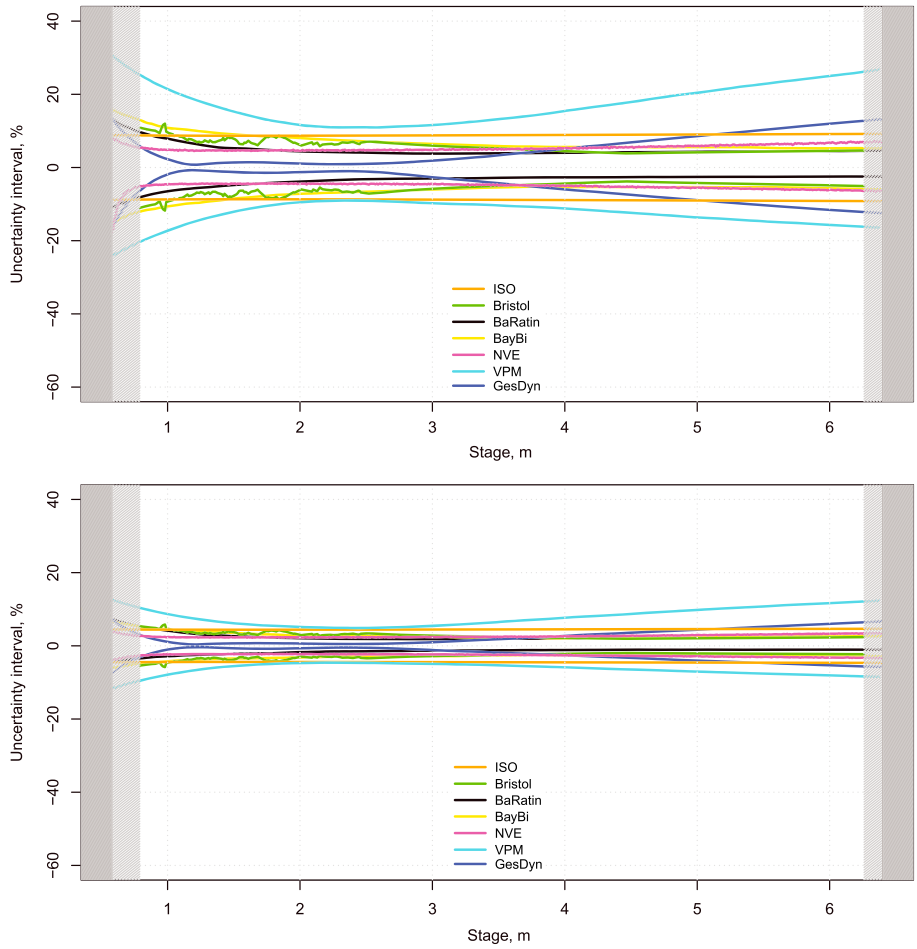

Figure 4. The 95% (top panel) and 68% (bottom panel) uncertainty intervals for the Isère River stream gauge, as computed by each model. The shaded gray areas are river stages for which gaugings are unavailable; for the period of record used, no stages were observed in the darker gray area. Note that the relative uncertainties are calculated against the estimated rating curve for each method, and that these are different. ISO = International Organization for Standardization; VPM = Voting Point Method; NVE = Norwegian Water Resources and Energy Directorate.

The VPM method has wider uncertainty bounds for high/low extremes where there is spread in the gaugings, because this model assumes that the rating may fluctuate throughout the range suggested by the gaugings (i.e., it accounts for epistemic uncertainty related to the spread in the gaugings, which at this site may be caused by hysteresis, see section 2.2.2). Other methods assume that only some of the gaugings are relevant for a specific time, typically because changes in the channel are assumed to have caused a persistent shift in the rating.

3.3. Taf River at Clog-y-Fran

The Taf River stream gauge was selected because it has an unstable channel and a gauged out-of-bank section, presenting other challenges for modeling the stage-discharge rating curve and its uncertainty.

Figure 1 shows the different gaugings used for different official rating curves over time. BaRatin, Bristol, ISO, VPM, and BayBi all utilize subperiods based on the subdivisions used in the official rating curves to develop ratings and uncertainty bounds. Had the subdivisions used by the official rating not been available, each modeler may have chosen different subperiods to divide their analyses. Whereas the official rating had six subperiods, GesDyn identified four homogeneous periods using the Hubert et al. (1989) segmentation procedure. The results for GesDyn shown in this paper are based on the mean rating curve computed from the set of gaugings in the fourth period (April 2008 to June 2012), with a few additional high flow gaugings retained from the earlier periods.

Table 5
Summary of Estimated Rating Curves and Estimated Uncertainty for All Methods

	Stage (m)	Range in discharge from rating curves of all methods (min, max; m^3/s)	CV of rated discharge	Across all methods, the minimum and maximum estimated uncertainty intervals (min, max;[a] m^3/s)	Range in full width 95% uncertainty intervals, as computed for each method[a] (min, max)
		ISERE			
Minimum stage (of gaugings)	0.79	(51.3, 55.0)	2.6%	(42.1, 66.2)	(11%, 22%)
Median stage (of time series)	1.75	(147, 150)	0.8%	(133, 168)	(2.8%, 17%)
Maximum stage (of gaugings)	6.26	(854, 940)	3.7%	(744, 1120)	(7.0%, 25%)
		MAHURANGI			
Minimum stage (of gaugings)	0.23	(0.029, 0.035)	6.8%	(0.0, 0.57)	(26%, 2000%)
Median stage (of time series)	0.65	(0.47, 0.56)	5.5%	(0.02, 1.59)	(18%, 312%)
Maximum stage (of gaugings)	2.6	(57.4, 67.1)[b]	5.7%[b]	(42.4, 75.9)[b]	(27%, 52%)[b]
Stage in extrapolated range	4.08	(126, 195)[c]	16%[c]	(49, 412)[c]	(41%, 202%)[c]
		TAF			
Minimum stage (of gaugings)	0.88	(0.74, 0.88)[d]	7.5%[d]	(0.37, 1.76)[d]	(28% 101%)[d]
Median stage (of time series)	1.15	(4.2, 4.7)	3.5%	(3.4, 6.2)	(12.9%, 55%)
Maximum stage (of gaugings)	3.27	(55.7, 69.7)[e]	7.8%[e]	(47.6, 79.4)[e]	(18%, 34%)[e]

Note. ISO = International Organization for Standardization.
[a]Note that the uncertainties estimated by each method are relative to the rating curve estimated by that method. [b]ISO results were unavailable for the stage of 2.6 m and are not included in these results. [c]Extrapolated rating curves were not available near the maximum observed stage for the ISO or Bristol methods, and these methods are not included for these entries. A stage of 4.08 m was chosen because it was the highest stage consistently reported by all other methods. [d]GesDyn results were not available for a stage of 0.88 m. [e]ISO results were not available for a stage of 3.27 m.

Figure 9 shows the resulting rating curves for the Taf. For all methods except GesDyn (see above), the rating curve shown is for the period with gaugings from 29 August 2003 to 20 June 2012. The gaugings utilized by the methods are shown as black circles in Figure 9; the remaining gaugings from earlier periods are shown as gray circles. Divergences between the rating curves are mainly seen at the extrapolated low end of the curve and at high flows. The GesDyn rating stands out from the other curves, but it is fit to a reduced set of points.

In Figure 10, the individual ratings are plotted with their 95% and 68% uncertainty intervals. Figure 11 shows the uncertainty as a percentage of the rated flow estimated by each method. At this site, the percentage uncertainties are particularly high for lower flows. At the high end of the rating, there is a distinct break in the slope of the measured gaugings where the river goes out of bank. NVE and BaRatin show a smaller change in the rating curve and uncertainty shape in the over-bank section compared to VPM, BayBi, Bristol, and GesDyn. The ISO method is truncated early, because of insufficient data in the highest segment.

At this site, gauging measurement uncertainties were not provided with the data set, so each group was required to set their own uncertainties and these choices varied. For example, on the simple end of the spectrum, the uncertainty for all gaugings was set to ±5% (GesDyn). For BayBi, a gauging distribution with a zero mean and a standard deviation equal to 1% of the mean observed discharge over the analyzed period was assumed for a discharge gauging error, but the posterior distribution was estimated from the data used in the analysis for each rating curve period independently. For VPM and Bristol, a logistic discharge gauging error that was previously estimated for the UK was used for this site (McMillan & Westerberg, 2015); it is more heavy tailed than a Gaussian distribution and varies with flow range (95% bounds on relative errors of ±13% for high flow to ±25% for low flow).

4. Discussion

Seven rating curve uncertainty-estimation methods were compared and applied to three stream gauges with increasing complexity in the stage-discharge relationship: a single segment rating (the Isère at Grenoble

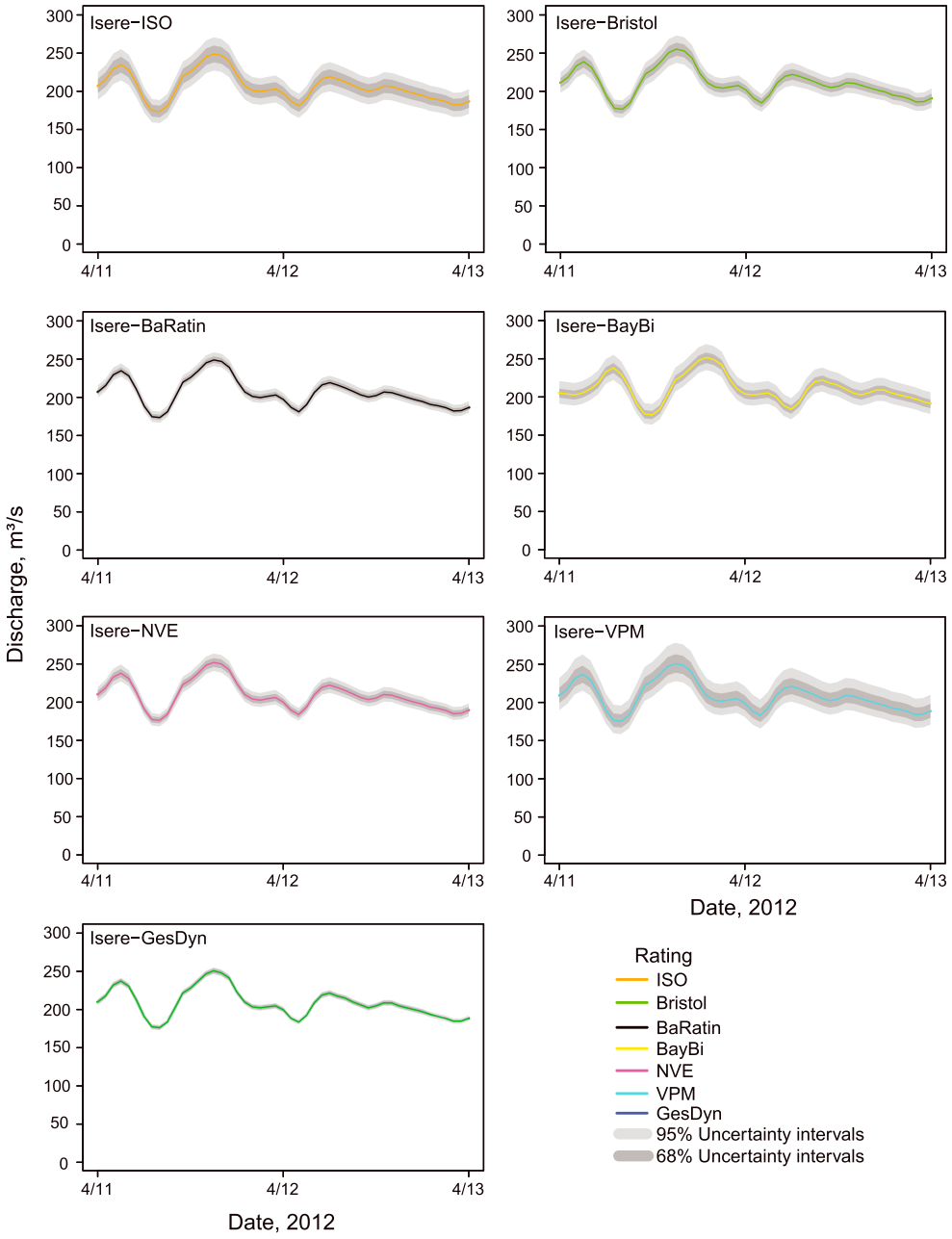

Figure 5. Uncertainty intervals for a 2-day period at the Isère gauging station. This period's discharge is within a range with many available gaugings. The 68% and 95% uncertainty intervals are shown for all methods. ISO = International Organization for Standardization; VPM = Voting point Method; NVE = Norwegian Water Resources and Energy Directorate.

Campus, France), a multiple segment rating with a substantial extrapolated range (Mahurangi at College, New Zealand), and a time-varying multiple segment rating with a distinct out-of-bank section (Taf at Clogy-Fran, United Kingdom).

As summarized in Table 5, the comparisons at the three sites showed different uncertainty intervals for the different methods, even when the estimated rating curves were similar. Because not all methods produce symmetrical uncertainty intervals, the total width of the uncertainty interval was used as the summary statistic (note that these are calculated against the discharge from the optimal curve for each method and that these are different). At the Isère, estimated rating curves were fairly similar, as seen in Figure 2. The

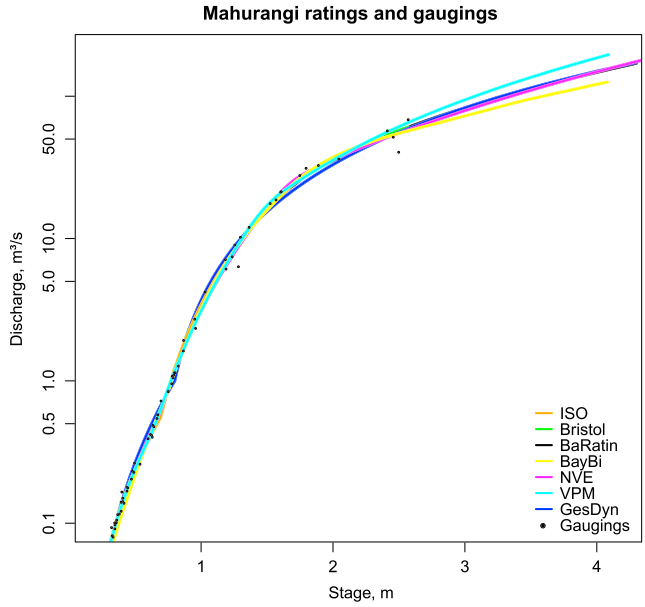

Figure 6. Mahurangi ratings curve for each method. ISO = International Organization for Standardization; VPM = Voting point Method; NVE = Norwegian Water Resources and Energy Directorate.

coefficient of variation of rated discharge between the methods was between 0.8% and 3.7% for the selected river stages in Table 5. The range of the 95% uncertainty intervals estimated by the methods was often notably higher. For example, the 95% uncertainty interval widths ranged from 7% to 25% at the maximum stage. The 95% uncertainty intervals for all flow levels were notably larger for the Mahurangi and Taf sites, where additional complexities of segmented rating curves and a mobile channel bed and out-of-bank section made estimation more challenging. This is particularly evident for very low flows.

Discharge uncertainty results from the different methods diverged most strongly for extrapolated sections of the rating curve, that is, stages below or above the set of gaugings. These sections are shown in shaded gray in Figures 4, 8, and 11, representing stages that have been recorded (and hence the rating curve has been used within these bounds), but were not gauged. Extrapolation was most apparent at Mahurangi, where the maximum gauged stage is 2.7 m while the maximum recorded stage is 4.2 m. Though less pronounced, extrapolation was also seen for low stages at Isère and high stages at Taf. Differences in the estimated rating curve from each method diverge most strongly for the out of bank conditions at Taf.

4.1. How Do Method Assumptions Influence Width of Uncertainty Intervals?

As part of this experiment, we paid careful attention to the assumptions used by each uncertainty method. Some assumptions that we expected to be influential were not so; for example, methods that included more components of the uncertainty did not consistently produce wider uncertainty intervals (Table 4). Neither was there a large degree of consistency in which method tended to produce the largest or smallest uncertainty intervals. For example, while VPM and BayBi tended to have wide uncertainty intervals, this was not true for all sites or all ranges of flows.

4.1.1. Time Variation in Rating Curves

One of the most important differences between methods is the handling of time-varying rating curves. Whereas the three Bayesian methods (BaRatin, BayBi, and NVE) assume that the underlying stage-discharge relationship is constant through time, VPM and GesDyn assume that all the gaugings do not come from the same underlying stage-discharge relationship. VPM incorporates the uncertainty related to time variability into the uncertainty estimate, leading to uncertainty intervals that were among the largest of the methods we tested. In contrast, GesDyn explicitly models the temporal variability and removes gauging points belonging to other time windows from the estimation. This subsampling of gaugings leads to GesDyn uncertainty intervals being among the smallest in the middle stage range (as fewer gauging points occur) but the largest in the tails where fewer gauging points lead to a greater proportion of the stage range being treated as extrapolation. However, GesDyn users can decide to add high flow gaugings from other time windows, as was the case for the Taf site here.

4.1.2. Expert Knowledge on Hydraulic Controls

In the extrapolated section of the rating curve, the amount of expert knowledge on hydraulic controls used within the method has a strong influence on the width of the uncertainty intervals. Within the gauged range of stages (nonshaded areas), uncertainty estimates tended to be more consistent between the methods. For example, for Mahurangi, only BayBi was very dissimilar to the other methods within this region. This is because despite differences in assumptions, results in this region are strongly constrained by the gauged data. However, outside the gauged range of stages (shaded area), uncertainty estimates are markedly different between the methods. In the absence of gauging data in this region, the width of the uncertainty bounds are controlled by differences in the underlying principles of the methods and the choice of parameter priors and in particular the extent to which prior or inferred hydraulic information is used to constrain the uncertainty. Even within the gauged range of stages, the prior affected BayBi results. A uniform prior was used for results shown in this paper.

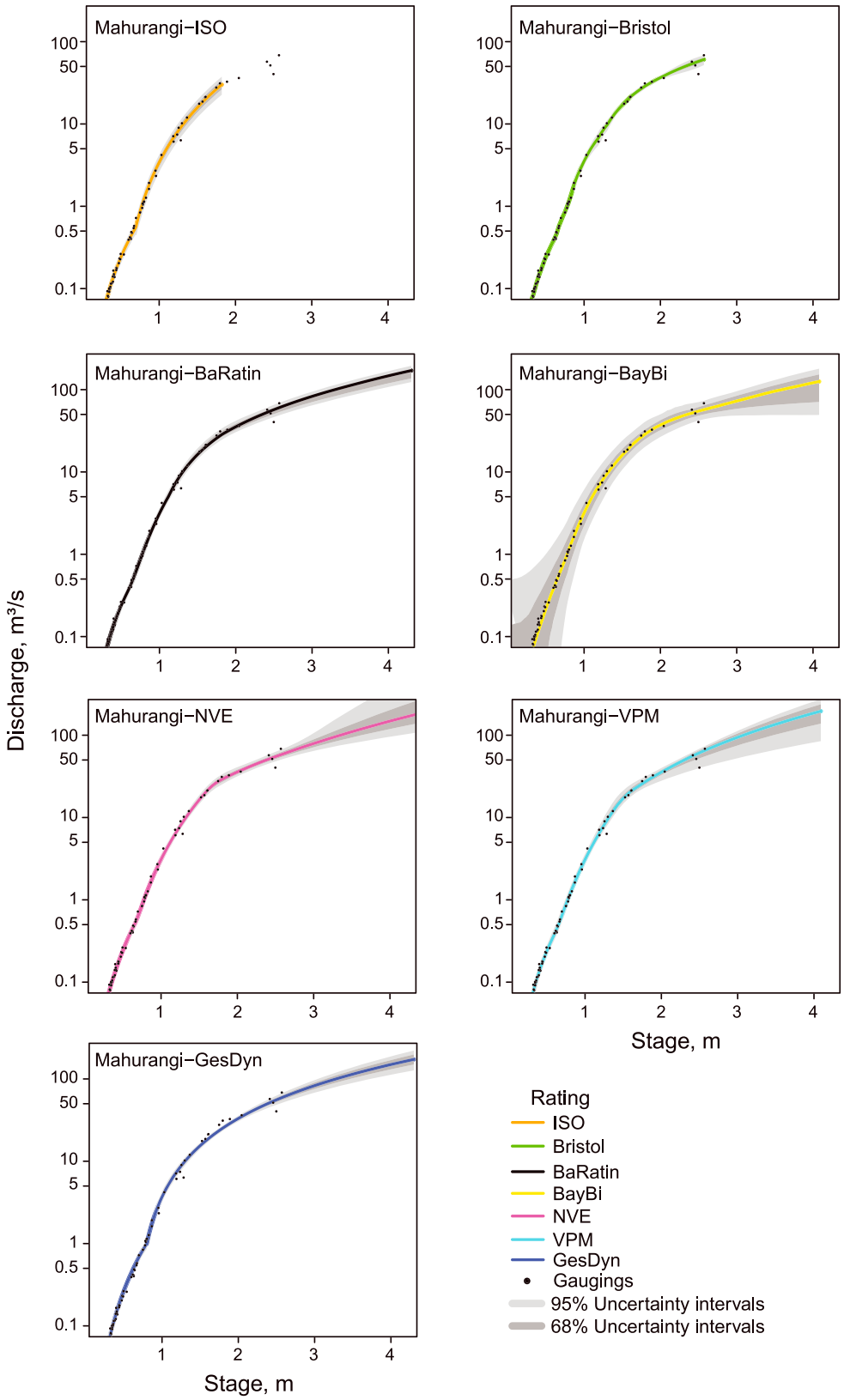

Figure 7. Mahurangi rating curve and uncertainty intervals for each method. ISO = International Organization for Standardization; VPM = Voting point Method; NVE = Norwegian Water Resources and Energy Directorate.

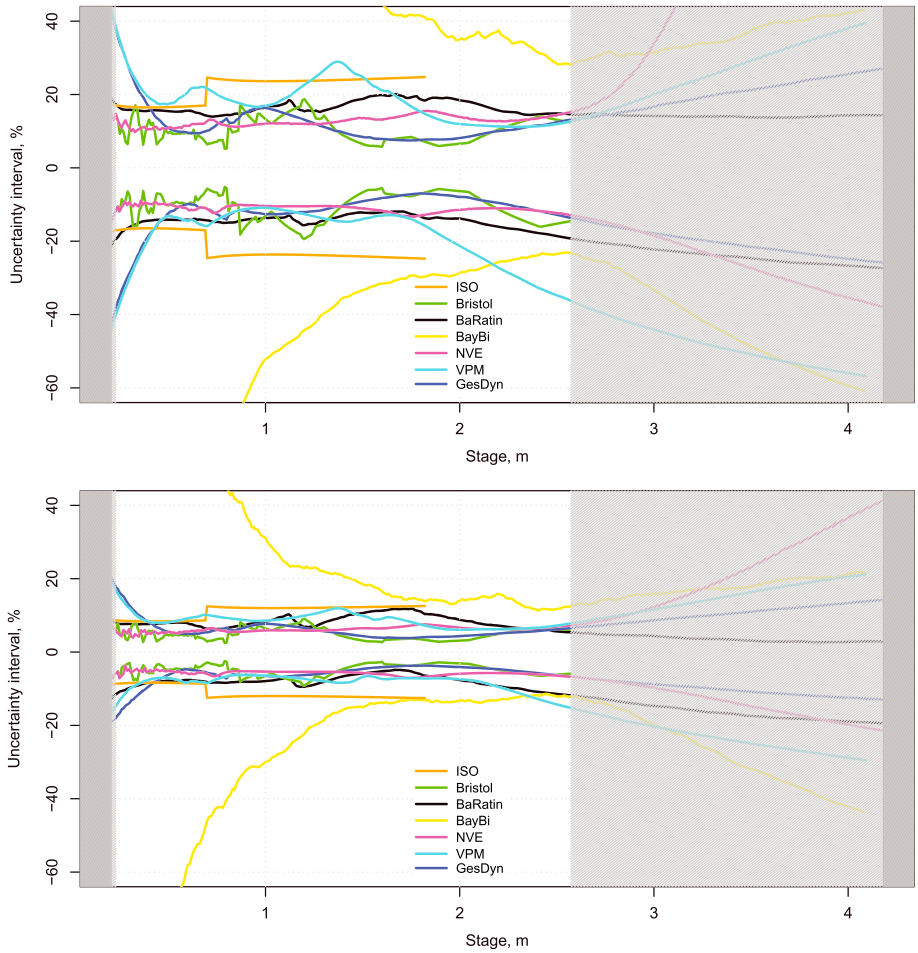

Figure 8. The 95% (top panel) and 68% (bottom panel) uncertainty intervals for the Mahurangi River stream gauge, as computed by each method. The shaded gray areas are river stages for which gaugings are unavailable; for the period of record used, no stages were observed in the darker gray area. For the Mahurangi, the highest observed stage during the period 1982 to 2015 was 4.2 m, while the highest stage during a gauging was 2.6 m, resulting in a large range where the rating model required extrapolation. Note that the relative uncertainties are calculated against the estimated rating curve for each method, and that these are different. ISO = International Organization for Standardization; VPM = Voting point Method; NVE = Norwegian Water Resources and Energy Directorate.

The Bristol and ISO method produces no rating curve for the extrapolated region because they are solely gauging-data based. The other methods capitalize on known hydraulic characteristics of the channel, allowing extrapolation beyond the observed gaugings. A clear example occurs for Taf, where the VPM uncertainty bounds increase rapidly in the extrapolated region. This occurs because VPM assumes a change in channel hydraulic control at the start of overbank flow (i.e., the top part is modeled as a separate section, following the official rating curve) and because there are few gaugings in this out-of-bank section. The uncertainty in the extrapolated part of this out-of-bank section therefore becomes dependent on the parameter priors. Because no hydraulic analysis was used to support the setting of the VPM parameter priors for this station, little can be assumed about the stage-discharge relationship in the extrapolated part, leading to rapidly increasing uncertainty bounds. The NVE had an even larger increase in uncertainty in the over-bank section than VPM (Figure 10). GesDyn shows a similar increase of uncertainty in the over-bank section but with lower uncertainty magnitude. For BaRatin, the hydraulic controls are user specified and therefore supply more information and lower uncertainty even in this ungauged region. However, a degree of caution here is needed for most stream gauging stations. The Taf has an over-bank gaugings, which is relatively rare, and the gaugings therefore show a resultant sharp change in the stage-discharge behavior. However, where no such data exists to define out-of-bank extrapolated relationships, such rating curve estimates may have significant biases at

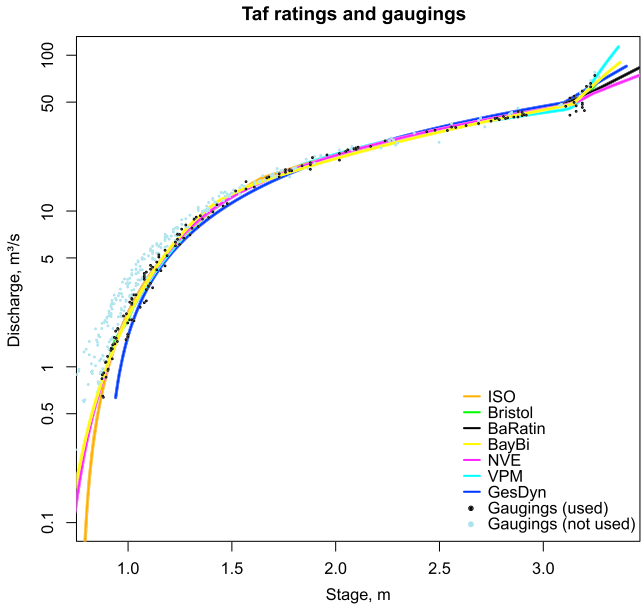

Figure 9. Rating curves for each model for the Taf River. Measurements shown in black are those made between 29 August 2003 and 20 June 2012. Earlier measurements, dating from October 1989, are shown in light blue. Note that the results from GesDyn use only gaugings from April 2008 to June 2012 and a few additional high flow gaugings, as shown in Figure 10. ISO = International Organization for Standardization; VPM = Voting Point Method; NVE = Norwegian Water Resources and Energy Directorate.

high flows. As shown in the results, this change in behavior even while observed resulted in the largest difference between methods for the final set of rating curves (see Figure 9).

In general, larger differences between methods should be expected where extrapolation is greater and there are fewer data in the higher stage range. The user should be aware of such differences and decide for themselves whether reliable and sufficiently precise parameter priors can be set without a hydraulic analysis or whether such an analysis is necessary. Very wide uncertainty bounds in extrapolated areas are implicitly showing that very little information is available, and flow estimates are uncertain. On the other hand, hydraulic analyses to support extrapolation require more information about the river cross section and more time and effort. Hydrometric stations often require a rating curve across the full range of recorded stages, and rating curves will continue to be used outside the gauged stage range. It is therefore important that the uncertainty of computed flow in the extrapolated range and the extrapolating assumptions are made explicit so they can be defended and reviewed.

4.2. Limitations in Comparing Uncertainties

The design of our study led to some limitations in the treatment of uncertainties for the three gauging stations. We provided a standard set of information about each station and asked research teams not to investigate further information sources. This decision helped us to ensure that the results were comparable between teams, but did preclude site visits, contact with gauging station staff, or other methods that could be used to elicit more information about potential error sources. Our standardization also meant that the capabilities of some of the uncertainty methods were not fully exploited. For example, we did not include propagation of uncertainties in the continuously measured stage series. Nevertheless, it is still difficult to compare uncertainty estimates across methods, as each method's uncertainty applies to the rating curve generated by that method.

In our study the rating equations and priors used by the different methods were different, which makes it hard to isolate differences that are due to other methodological aspects. Ocio et al. (2017) compared BaRatin and VPM using the same prior width for both methods (standard and hydraulic model-based priors). They found similar uncertainty ranges at one station and wider uncertainty for VPM than BaRatin at another station for the standard, but not the hydraulic model priors. This indicates that differences between methods will partly depend on data set characteristics. As mentioned earlier, substantial differences between BayBi results for the Mahurangi site were seen for different priors.

Some limitations are common across all the methods, such as the treatment of gauging errors as mutually independent between gaugings, whereas there may in practice be systematic components to these errors. Future development of the uncertainty assessment methods may enable such complexities to be included. However, inclusion of ever more detailed error descriptions implies increased information requirements, which may not be suitable for application across large numbers of sites or remote sites. While we chose three gauging sites with different characteristics of the rating curve uncertainty, future comparison experiments including larger numbers of sites would help to further tease out differences in uncertainty estimates.

4.3. Recommendations for Rating Curve Uncertainty Estimation

The findings from this study have a number of important implications for researchers and operational users who estimate discharge uncertainties for environmental data and modeling analyses, water resources planning, and environmental management. Operational users such as hydrometric agencies (who may provide discharge and associated uncertainty data to multiple clients) and commercial users such as hydropower industry or irrigation authorities are typically interested in ensuring that decisions relying on discharge data are robust in the light of potential uncertainty. Research users not only study the impact of discharge

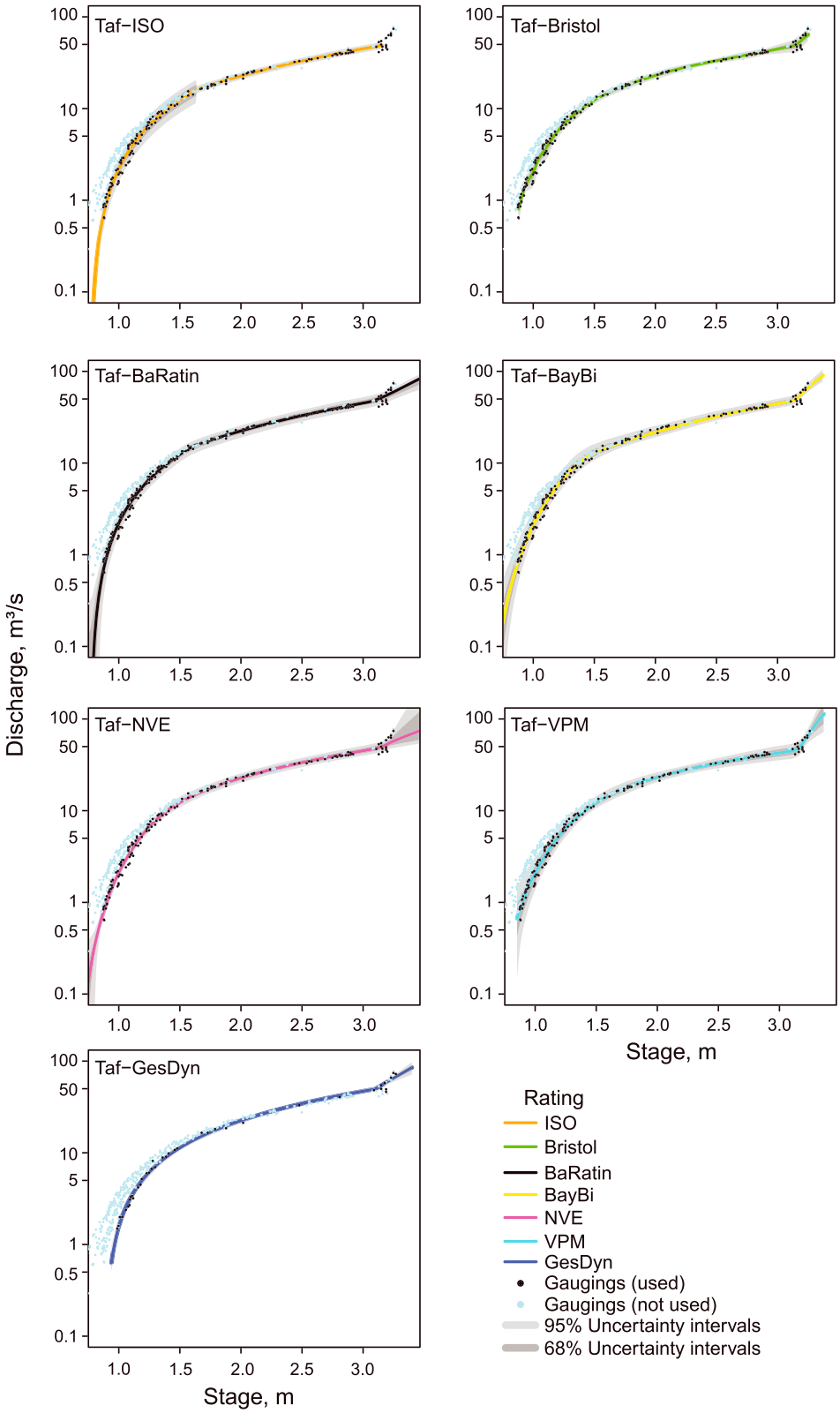

Figure 10. Rating curves and uncertainty intervals for Taf. Only the measurements shown in black were used in fitting the rating curves and estimating rating curve uncertainty. Earlier measurements, not used to fit the rating for this time period, are shown in light blue. Note that the set of measurements used for GesDyn is different than the set used for all other methods. ISO = International Organization for Standardization; VPM = Voting Point Method; NVE = Norwegian Water Resources and Energy Directorate.

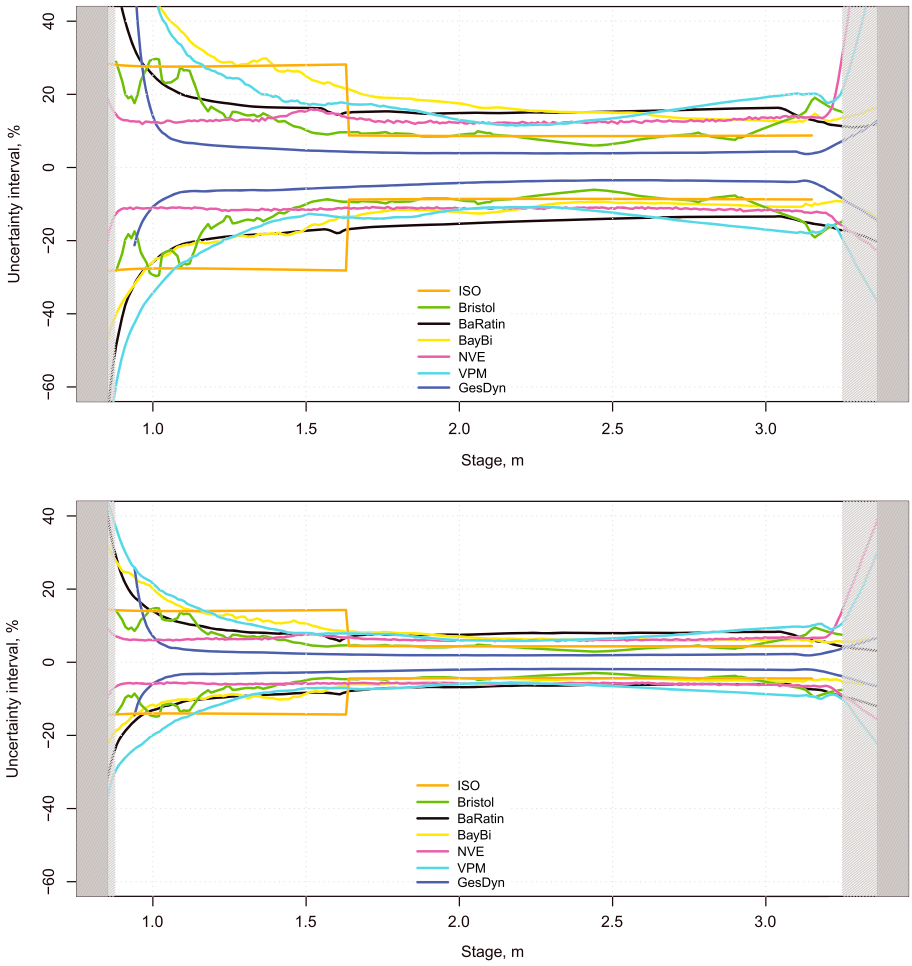

Figure 11. The 95% (top panel) and 68% (bottom panel) uncertainty intervals for the Taf at Clog y Fran stream gauge, as computed by each method. The shaded gray areas are river stages for which gaugings are unavailable; for the period of record used, no stages were observed in the darker gray area. Note that the relative uncertainties are calculated against the estimated rating curve for each method, and that these are different. ISO = International Organization for Standardization; VPM = Voting Point Method; NVE = Norwegian Water Resources and Energy Directorate.

uncertainty on hydrologic inference and model building but will also want to ensure that their modeling and data results are robust and are not unreasonably affected by discharge uncertainties.

In light of these needs, it is evident that wherever possible discharge uncertainty estimates should be derived and assessed. From the experimental design in this study, we suggest that the minimum rating curve station characteristics and data needed for discharge uncertainty estimation include (1) information on channel and controls (stable section, natural weir, etc.) and (2) gaugings with dates. Additional information such as official rating curves, photos of the gauging station, the methods used for individual gaugings, and associated uncertainty of gaugings were also useful in constraining the discharge uncertainty estimates. We suggest that when streamflow data are reported by researchers, monitoring stations, or environmental regulators, all of this additional information be included to enable improved estimation of discharge uncertainty.

Once this data are available, the end user must choose which method to use to estimate discharge uncertainties. The main deciding criteria are the need for online software, time variability in the rating, whether the user needs to specify stage and discharge measurement uncertainties, whether extrapolation of the rating curve is needed or indeed warranted under different hydraulic controls, and whether rating curve samples are needed for error propagation to subsequent analyses. Figure 12 provides a flow chart for the user to determine which methods are most suitable for their application.

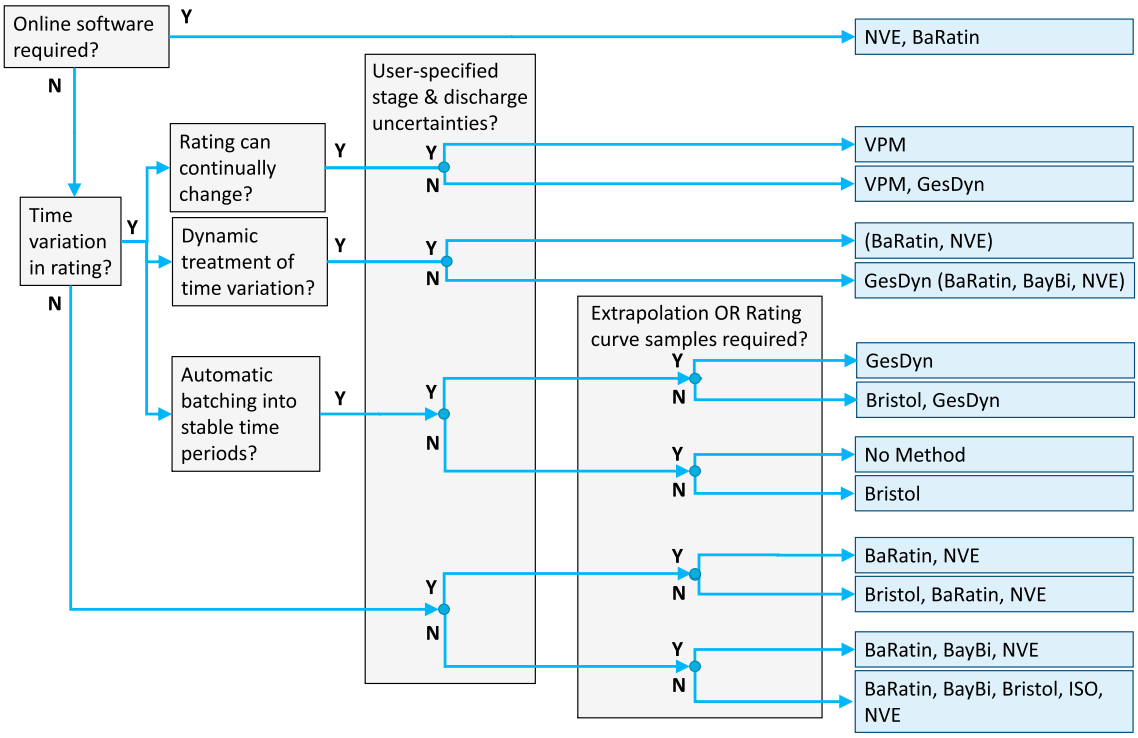

Figure 12. Flow chart to guide selection of uncertainty estimation method. Methods shown in brackets have new capabilities in time variation in development or completed but were not included in this study. ISO = International Organization for Standardization; VPM = Voting Point Method; NVE = Norwegian Water Resources and Energy Directorate.

4.4. Recommendations for Future Research

4.4.1. Ensemble Methods and Software

Currently, the choice of method is driven by user and application requirements. Some of the methods have additional capabilities in development (e.g., dynamic treatment of time variability in BaRatin, NVE, and BayBi), so that they will become more widely applicable. Others (Bristol and ISO) are very general and have limited input data requirements and so can be more easily applied to national gauging networks (e.g., Coxon et al., 2015). In these circumstances, users will have multiple methods open to them, and therefore an ensemble approach that combines the strengths of the different methods would be a valuable future development. An ensemble approach would explicitly include uncertainty due to assumptions made within each method, incorporating the large differences that we found between uncertainty estimates from different methods. This would require significant development as the methods would need to be coded in compatible software, this task in itself would provide additional benefits to users if a single software package allowed easy choice and implementation of multiple methods. The ensemble approach could use the technique of Bayesian Model Averaging (e.g., Duan et al., 2007), where methods are differentially weighted at each time step according to their calculated fit to the data (i.e., a likelihood measure) or a codified appraisal of their suitability.

4.4.2. Validation of Uncertainty Intervals

Our results demonstrated unexpectedly large differences in the uncertainty intervals estimated by different methods. A significant future research direction for our group is to investigate possible methods for *validation* of uncertainty intervals. A large collection of validation experiments would provide a basis to determine which gauging station characteristics indicate suitability of which methods. The design of validation methods is a challenging task because uncertainty intervals represent a combination of multiple unknown and time-varying uncertainty components. While a split sample analysis could determine what percentage of new gaugings lay within the derived uncertainty intervals, the question of what percentage *should* lie within the intervals is determined by the various assumptions (e.g., separation of uncertainties in gauging measurements and parametric/structural rating curve errors, and the

nature and treatment of time variation in the rating curve). These are the same differences in assumptions that cause the differences in uncertainty intervals between methods that we would be trying to test.

Methods could additionally be tested against synthetic data created using multiple different error assumptions, to determine the flexibility of each method. Another alternative is to work at locations where an alternative reference discharge is available, such as where dam releases are separately measured in the dam conduits. Dam conduit flows are typically more accurate than gauging data, with uncertainties of 2–3% (Le Coz et al., 2016). An additional benefit of this approach is that it would allow comparing discharges averaged over specific time steps (hourly, daily, and monthly), which are the basic data used in many applications. Such comparison cannot be performed using gaugings as validation data, because the latter represents sporadic measurement of instantaneous discharge. However, this type of approach would also require the development of new tools to compare two uncertain flow records, both containing a mixture of systematic and nonsystematic errors. Horner et al. (2018) include discussion of this issue.

4.4.3. Downstream Impacts on Model and Data Analysis

While many environmental modeling analyses have incorporated discharge uncertainties for rainfall-runoff model calibration (Coxon et al., 2014; Krueger et al., 2010; Liu et al., 2009; McMillan et al., 2010; Sikorska & Renard, 2017), flood forecasting (Ocio et al., 2017), flood frequency analysis (Osorio & Reis, 2016), sensitivity analyses of a flood inundation model (Savage et al., 2016), design flood estimation (Steinbakk et al., 2016), hydrological change detection (e.g., Juston et al., 2014; Lang et al., 2010), deriving hydrological signatures (Westerberg et al., 2016), and calculating nutrient flux estimates (Lloyd et al., 2016), these studies often incorporate a single methodology to derive discharge uncertainty estimates. Given the large differences in discharge uncertainty estimates between methods, an important future step to the work in this study is to assess the impact these different discharge uncertainty estimates have on hydrological model and data analyses.

Appendix A: Additional Description of Rating Curve Uncertainty Estimation Methods

A1. ISO/WMO

The current standardized method for rating curve uncertainty analysis is described in Herschy (1999) and in ISO and WMO documents (ISO 1100–2:2010; ISO/PWI 18320, 2015; WMO, 2006). This ISO/WMO method appears to be simple to apply to a set of stations and statistically sound, as it is based on traditional linear regression analysis. However, we are not aware of examples of its inclusion in operational software, use in research works, or routine application by hydrological services.

The ISO/WMO method requires a number of assumptions that actually brings important limitations and drawbacks. A fundamental assumption is stated as "The stage-discharge relationship, being a line of best fit, should be more accurate than any of the individual gaugings" (ISO/PWI 18320, 2015). This is actually questionable, and in practice both the structural errors of the rating curve model and the measurement errors of the gaugings will be mixed up in the residuals. Then, gauging uncertainties will be estimated from their deviations to the fitted rating curve. Gaugings errors are assumed to be mutually independent, and the corresponding uncertainties to be equal, in percentage of the flow. The linear regression analysis is conducted with discharge and flow depth in the log space, and approximations are required to get uncertainties back to the natural space. The uncertainty analysis is applied to each segment of the rating, previously fitted as a power function: $Q = a(h - b)^c$. A minimal number of $N = 20$ gaugings per segment is necessary, which practically makes the uncertainty computation impossible for many extrapolated or not densely gauged segments.

Uncertainty results are expressed as *standard errors* in oldest documents and as *standard uncertainties* in the most recent ISO/PWI 18320 working document, in compliance with recent uncertainty guidance and standards such as the Guide to the expression of uncertainty in measurement (JCGM, 2008) and the Hydrometric Uncertainty Guidance (ISO/TS 25377, 2007). The standard deviation of the residuals (or residual uncertainty) is computed from the differences between the gauged discharges Q_i and the discharges $Q_c(h_i) = a(h_i - b)^c$ computed at gauged stages h_i with the rating curve equation:

$$S = \sqrt{\frac{\sum_{i=1}^{N} [\ln Q_i - \ln Q_c(h_i)]2}{N - p}}$$

with p= the number of rating curve parameters estimated from the N gaugings. We considered that $p = 2$ for each segment because in manual calibration methods, the offsets b are usually preliminary estimated by the user and not included as a free parameter in the regression.

The standard uncertainty of the calculated value of $\ln Q_c(h)$ at any stage h of the rating curve segment is

$$u[\ln Q_c(h)] = S \sqrt{\frac{1}{N} + \frac{[\ln(h-b) - \mu]2}{\sum_{i=1}^{N} [\ln(h_i - b) - \mu]2}}$$

with $\mu = 1/N \sum_{i=1}^{N} \ln(h_i - b)$.

If $u[\ln Q_c(h)]$ is small enough to allow the linear approximation, it can be used to compute the relative discharge uncertainty:

$$Q_c(h) e^{\pm u[\ln Q_c(h)]} \approx Q_c(h)[1 \pm u[\ln Q_c(h)]]$$

The expanded uncertainty $U[\ln Q_c(h)] = k\, u[\ln Q_c(h)]$ is computed with a coverage factor k relative to the desired level of probability. Assuming the underlying probability distribution is Gaussian, we took k= 1.96, 1.645, and 0.9945 for 95%, 90%, and 68% uncertainty intervals. This yields the *confidence interval*, which includes parametric uncertainty only.

The standard uncertainty of the predicted value of discharge $\ln Q_p(h)$ at stage h is

$$u[\ln Q_p(h)] = \sqrt{c2\, u[\ln(h-b)]2 + S2 + u[\ln Q_c(h)]2}$$

The uncertainty of the flow depth measurement, $u[\ln(h - b)]$, is neglected in this work because propagation of stage uncertainties is not considered in the comparison of methods. The same procedure as for $u[\ln Q_c(h)]$ can be applied to $u[\ln Q_p(h)]$ to approximate the expanded uncertainty of the predicted discharge. This yields the *prediction interval*, which combines the parametric uncertainty and the residual uncertainty. Both structural and measurement uncertainties are included in the residual uncertainty.

A2. Bristol

The discharge uncertainty estimation method developed at the University of Bristol (Coxon et al., 2015) is a generalized framework designed to estimate place-specific discharge uncertainties for a wide range of different and complex stage-discharge relationships. The framework is able to account for uncertainty in the stage-discharge gaugings, multisection rating curves, changes in discharge uncertainty over time, and change in the uncertainty in the stage-discharge relationship across the flow range. The discharge uncertainty estimates aim to represent both aleatory and epistemic sources of discharge uncertainty.

Discharge uncertainties are derived using a nonparametric LOWESS and require stage-discharge gaugings and the historical rating-curve equations. Subsets of the stage-discharge data contained within a moving window are used to calculate the mean and variance at every stage point, which then define the LOWESS fitted rating curve and discharge uncertainty, respectively. Weights (w_i) are dependent upon the differences in stage and are given by the tricube weight function:

$$w_i = \left(1 - \left|\frac{(x - x_i)}{\max(x - x_i)}\right|^3\right)^3$$

where x is the central stage-discharge measurement point and x_i is the other stage-discharge measurements in the set of data points defined by the span. As the method is data based, the rating curve and its uncertainty interval cannot be computed above the highest gauging and below the lowest gauging. Stage and discharge gauging uncertainties are incorporated into the framework by randomly sampling from estimated

measurement error distributions to fit multiple LOWESS curves and then combining the multiple fitted LOWESS curves and variances in a Gaussian Mixture Model. Time-varying discharge uncertainties are not modeled explicitly, but an automatic procedure uses differences in historical rating curves to separate the stage-discharge rating data into subsets for which discharge uncertainty is estimated separately.

The process is easily applicable to any gauging station with at least 20 stage-discharge gaugings and is set up to be fully automated requiring minimal user input to operate for hundreds of gauging stations.

A3. BaRatin

BaRatin allows the construction of stage-discharge rating curves with uncertainty estimation, combining prior knowledge on the hydraulic controls and the information content of the uncertain gaugings (Le Coz et al., 2014). Uncertainties on both the discharge and stage measurements of the gaugings are considered as Gaussian distributions with mean zero. Typical discharge uncertainties were assumed depending on the gauging procedure, and stage uncertainties due to varying flow were computed (cf. Le Coz et al., 2012).

The rating curve equation is derived from the combination of power functions relating discharge Q to stage h for each of the assumed or known N_{control} controls at the site:

$$Q(h) = \sum_{r=1}^{N_{\text{segment}}} \left(1_{[\kappa_{r-1};\kappa_r]}(h) \times \sum_{j=1}^{N_{\text{control}}} M(r,j) \times a_j \left(h - b_j\right)^{c_j} \right)$$

In the above equation, $M(r,j)$ is the matrix of controls, and the notation $\mathbf{1}_I(h)$ denotes a function equal to 1 if h is included in the interval I, and 0 otherwise. The number of segments N_{segment} in the rating curve is fixed by the user while the segment limits (breakpoints) κ_r are inferred, along with the coefficients a_j and the exponents c_j of the controls. The control offsets b_j are deducted from the continuity of the stage-discharge rating curve.

The user also defines the prior distributions of the physical parameters κ_r, a_j, and c_j of that stage-discharge equation, *prior* meaning estimates made without utilizing the gaugings. The priors and the information contents of the gaugings are further combined to simulate a large set of possible parameters and rating curves. Such Bayesian simulation is based on Markov Chain Monte Carlo (MCMC) sampling of the posterior distribution of the rating curve parameters inferred from the Bayes theorem. Any physical conflicts between the results and the assumed priors should be checked and lead to question the rating curve model and the estimated uncertainties of the gaugings.

A statistical postprocessing of this bunch of rating curves, or *spaghetti*, yields the uncertainty bounds of the rating curve and of the propagated discharge time series at any level of probability. The total uncertainty combines the parametric uncertainty, derived from the spaghetti samples, and the remnant uncertainty accounting for the structural errors of the rating curve model. The remnant uncertainty may be modeled as a linear function of discharge (recommended option) or as a constant. Wide uniform distributions are used as reasonably noninformative priors for remnant uncertainty parameters that will be estimated to account for the mismatch between the observations (gaugings) and the model (rating curve) that cannot be explained by the uncertainties of the gaugings. In the comparison, uncertainty interval is computed with the total uncertainty and does not include the measurement uncertainty. The maximum a posteriori rating curve is computed using the set of parameters with the highest joint probability.

BaRatin and its graphical environment BaRatinAGE have been released in French and English with a free, individual license. While not used for this study, the propagation of rating curve uncertainties and stage series uncertainties to flow series uncertainties (Horner et al., 2018) is embedded in the latest release.

A4. BayBi

The BayBi method, developed at the University of Zurich following assumptions of Sikorska et al. (2013, 2015), is designed to estimate rating curves with associated uncertainty. Three different uncertainty sources in rating curves are considered: parametric, measurement error of discharge (random), and the structural error (bias) due to the chosen rating curve form. In addition, the measurement uncertainty in stages can also be incorporated into the method but was not considered in this work.

The BayBi is based on the stage-discharge relationships in the form of power equations, where multiple segments are allowed but the number of segments must be explicitly defined by the user. A switch between

different segments is controlled by breaking points that are included into the parameter inference. The inference relies on the Bayes' theorem (Gelman et al., 2013), for which the prior and the likelihood have to be determined. For rating curve parameters, we use a noninformative uniform distribution, while priors on breakpoints and the stage equal to zero discharge should be defined based on hydraulic data (e.g., cross-section profiles) or experts' knowledge. Also, prior information on the rating curve error and the discharge gauging error needs to be specified, for which either site-specific information can be incorporated or standard errors based on measurement practice can be assumed. As for the rating curve error, an error with a zero mean and an unknown standard deviation is assumed a priori. This standard deviation was assumed to be site specific and a priori as equal to 5% of the mean discharge observed over the analyzed time period at each station.

As a result, the joint posterior distribution on all parameter is derived during the Bayesian inference. By simply sampling from this posterior, rating curve simulations can be derived and different uncertainty intervals can be given. Due to the consideration of three different uncertainty components, the parametric, systematic structural, and random measurement error (total uncertainty) can all be calculated. For predictions based on the estimated rating curve, the measurement error is not of interest, and thus, the uncertainty intervals used in this paper do not include this measurement uncertainty.

The estimation of the posterior is performed using a MCMC approach, while uncertainty intervals are approximated by Monte Carlo sampling from the estimated posterior.

A5. NVE

The NVE method for fitting static rating curves rests on a Bayesian analysis of a multisegment power law model, see Reitan and Petersen-Øverleir (2009). Unlike other Bayesian methods like BayBi or BaRatin, the number of segments is also subject to statistical inference (Petersen-Overleir & Reitan, 2005). MCMC samples are fetched for all segmentation model, as characterized by the number of segments, up to a given upper limit. A posterior probability for the number of segments is calculated, so that the credibility intervals for the curve includes both the uncertainty in the curve parameters and the number of segments. The user can fine tune the prior distribution both for parameters and number of segments, in case hydraulic expertise exists. If not, a default prior is used.

The measurement model is $\log(Q_i) = a_{s(i)} + b_{s(i)} \log(h_i - h_{0,\,s(i)}) + \sigma\varepsilon_i \log(h_i - h_{0,\,s(i)}) + \sigma\varepsilon_i$, where Q_i and h_i are the discharge and stage of measurement i, respectively, σ is the standard deviation of the measurement noise on log scale, $s(i)$ is the segment which measurement i belongs to (which is determined by how the stage relates to yet another parameter set namely the segmentation limits, $h_{s,\,k}$ where $k \in \{1, \ldots, numseg\}$) with $numseg$ as the number of segments and ε_i is independent standard normal noise. The parameter $a_{s(i)}$ is sacrificed for all except the lowest segment, in order to ensure continuity in the stage-discharge relationship. Because this is an additive model on the log scale, it is a multiplicative model on the original scale, which means that discharge gaugings are assumed to be positive and the larger the discharge the larger the measurement noise. Both credibility intervals for the curve itself and credibility intervals for curve and measurement noise can be given, though the program has a focus on curve uncertainty. The possibility of structural errors is not examined, but segmented power laws can be viewed as a global approximator for a positive stage-discharge relationship starting at a given stage when the number of segments increase.

The segmentation intensity in the data set (and the uncertainty of that) is taken into account when extrapolating beyond the limits of the data. The intensity distribution gives a probability distribution for new segments, where the distributions of the curve parameters revert to the prior. These can give quite wide uncertainties when the stage goes far beyond the limits of the data set. It is however also possible to insert prior uncertainty bands within the extrapolated area for given stage values, if extra hydraulic knowledge exists.

A6. VPM

The *VPM* for discharge uncertainty estimation accounts for both random and epistemic errors in the rating curve, with epistemic error sources that can include weed growth, gravel scour/deposition, or unconfined high flows (McMillan & Westerberg, 2015). It allows the rating curve to fit a subset of gaugings only, to

account for the typical epistemic error consequence that more than one rating curve shape is consistent with the gauged data.

The VPM is based on MCMC sampling of piecewise power law rating curves, using a likelihood function that accounts for random measurement errors and epistemic errors. The random error for each gauging point is represented using a logistic or normal distribution (for discharge), and a uniform distribution (for stage). The VPM likelihood is based on the number of gaugings intersected by the proposed rating curve (these points are *voting* for the curve), weighted by the measurement error distributions and the proportion of the stage and discharge range that is spanned by the voting points. Multisegment ratings use the product of the likelihoods of each section.

The VPM does not explicitly account for time-varying errors, because it is assumed that the user may not know how, why, or exactly when the rating curve error might change over time. Instead, the method enables generation of multiple possible rating curve samples that represent the total uncertainty, thus implicitly accounting for epistemic uncertainty about temporal change. These samples can be used for subsequent discharge analysis that needs total uncertainty estimation, for example, regionalization studies or calculation of hydrological signatures. The method was primarily designed for estimation of discharge uncertainty and not for estimation of a best-estimate/optimal rating curve, even if such a curve can be extracted using the median or optimum likelihood realization from the MCMC. Because of the formulation of the likelihood, the optimal likelihood realization curve is not necessarily optimal for the whole flow range. For this comparison study, we used the median rating curve from the posterior distribution.

To use VPM, the user provides the gauging points and a rating curve in equation form (typically, the official rating curve, where available) that are used to set the initial parameter values and the number of rating curve sections in the MCMC algorithm. Prior behavioral ranges for the power law parameters are given a default value but should be checked and modified if necessary by the user. The prior bounds can be set based on hydraulic model analyses for high flows (Ocio et al., 2017), where such information is available.

A7. GesDyn

GesDyn uses, as many other methods, the fit of piecewise power functions to historical gaugings (Morlot et al., 2014). It is a dynamic method, with a rating curve and its uncertainty model adjusted for each gauging. The GesDyn's uncertainty model takes into account three components: the stage measurement uncertainty (which is not considered in the present study), the fitting uncertainty (taking into account gauging uncertainty), and the uncertainty due to temporal variability, based on a temporal variographic analysis (Jalbert et al., 2011) so as it increases along the time. The GesDyn method is currently used as official fitting method at Electricité de France (EDF) for operational management of 260 gauging stations.

The first step is the automatic identification of chronological homogeneous gauging samples so that each sample describes a homogenous mean state of the hydraulic control. A mean rating curve is then fitted, with piecewise power functions (using the least squares method), on each homogeneous sample. This step is also used to define and adjust some hydraulic parameters (cease to flow stage range, breakpoint stage, extrapolation assumptions). Thus, mean hydraulic controls are defined, but thin variations of hydraulic control (river bed vertical oscillations, for instance) are not described yet.

This is the reason why GesDyn method builds a rating curve for each gauging, as the adaptation of the mean rating curve described before, so as to take into account the thin variations in hydraulic control. A gauging subsample is therefore selected for each gauging, assumed that it described a precise state of hydraulic control. Selected gaugings are called hydraulic analogs. Each subsample is then used to fit a piecewise power function rating curve. Curves obtained are called dynamic rating curves to illustrate the dynamic adjustment of the first mean rating curve.

The fitting uncertainty is then computed with a MC simulation. For each set of hydraulic analogs, N subsets are randomly generated by performing N MC simulations based on the uncertainty gauging model. Then, N rating curves are fitted on the N subsets. The N curves quantiles directly give the fitting uncertainty.

To finish, a temporal variographic analysis is performed with the N dynamic rating curves to model the time variation in rating curves. It provides the idea of aging of stage-discharge relationship for each water level

and allows uncertainty to increase as a function of time. As explained, the variographic analysis is performed for each water level of a considered hydrometric station as it is expected that temporal variations may be different for low flows, mean flows, and flood flows (see Morlot et al., 2014 for more details).

Acknowledgments

This paper is based on discussions and subsequent work after two workshops on Discharge Uncertainty Analysis, held at TU Vienna, Austria, in 2015 and 2016. Special thanks to Alberto Viglione and Jose Luis Salinas who went out of their way to make the logistics for these workshops possible. The paper also draws from discussions at the USGS John Wesley Powell Center for Analysis and Synthesis. The work benefited from the framework of the Panta Rhei Research Initiative of the International Association of Hydrological Sciences (IAHS), working groups *Hydrologic Services and Hazards in Multiple Ungauged Basins* and *Epistemic Uncertainty*. I. W. and V. M. acknowledge the support of The Swedish Research Council Formas (Svenska Forskningsrådet Formas; 942-2015-321). G. C. was supported by NERC MaRIUS: Managing the Risks, Impacts and Uncertainties of Droughts and Water Scarcity, grant NE/L010399/1. The authors much appreciate the willingness of the hydrometric agencies to share their data with us for this study. Data were provided for the Isère River by EDF, ENSE3/Grenoble-INP, and IGE (Université Grenoble Alpes, CNRS, IRD, Grenoble-INP). Data were provided for the Mahurangi River by Auckland Council, New Zealand. Data were provided for Taf River by Natural Resources Wales, UK. These data are available in the accompanying data release. Finally, we appreciate the thorough reviews we received from three journal reviewers and a USGS colleague reviewer. Their comments were gratefully received and have improved this manuscript.

References

Bjerklie, D. M., Moller, D., Smith, L. C., & Dingman, L. (2005). Estimating discharge in rivers using remotely sensed hydraulic information. *Journal of Hydrology*, *309*(1-4), 191–209. https://doi.org/10.1016/j.jhydrol.2004.11.022

Ceola, S., Arheimer, B., Baratti, E., Blöschl, G., Capell, R., Castellarin, A., et al. (2015). Virtual laboratories: New opportunities for collaborative water science. *Hydrology and Earth System Sciences*, *19*(4), 2101–2117. https://doi.org/10.5194/hess-19-2101-2015

Clarke, R. T. (1999). Uncertainty in the estimation of mean annual flood due to rating-curve indefinition. *Journal of Hydrology*, *222*(1–4), 185–190. https://doi.org/10.1016/S0022-1694(99)00097-9

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. *Hydrological Processes*, *28*(25), 6135–6150. https://doi.org/10.1002/hyp.10096

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

Di Baldassarre, G., & Claps, P. (2011). A hydraulic study on the applicability of flood rating curves. *Hydrology Research*, *42*(1), 10–19. https://doi.org/10.2166/nh.2010.098

Di Baldassarre, G., & Montanari, A. (2009). Uncertainty in river discharge observations: A quantitative analysis. *Hydrology and Earth System Sciences*, *13*(6), 913–921. https://doi.org/10.5194/hess-13-913-2009

Domeneghetti, A., Castellarin, A., & Brath, A. (2012). Assessing rating-curve uncertainty and its effects on hydraulic calibration. *Hydrology and Earth System Sciences*, *16*(4), 1191–1202. https://doi.org/10.5194/hess-16-1191-2012

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., et al. (2006). Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *Journal of Hydrology*, *320*(1–2), 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031

Dymond, J. R., & Christian, R. (1982). Accuracy of discharge determined from a rating curve. *Hydrological Sciences Journal/Journal des Sciences Hydrologiques*, *27*(4), 493–504. https://doi.org/10.1080/02626668209491128

Gelman, A., Carlin, J. B., Hal, S. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (p. 675). London: Chapman & Hall.

Guerrero, J. L., Westerberg, I. K., Halldin, S., Xu, C. Y., & Lundin, L. C. (2012). Temporal variability in stage-discharge relationships. *Journal of Hydrology*, *446–447*, 90–102.

Herschy, R. (1999). *Hydrometry* (p. 376). New York: Wiley.

Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H. K., & Pierrefeu, G. (2018). Impact of stage measurement errors on streamflow uncertainty. *Water Resources Research*, *54*, 1952–1976. https://doi.org/10.1002/2017WR022039

Hubert, P., Carbonnel, J., & Chaouche, A. (1989). Segmentation des séries hydrométéorologiques. Application à des séries de précipitations et de débits de l'Afrique de l'Ouest [Segmentation of hydro-meteorological series. Application to precipitation and streamflow series in West Africa, in French]. *Journal of Hydrology*, *110*(3–4), 349–367. https://doi.org/10.1016/0022-1694(89)90197-2

Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C., & Arheimer, B. (2016). Most computational hydrology is not reproducible, so is it really science? *Water Resources Research*, *52*, 7548–7555. https://doi.org/10.1002/2016WR019285

ISO 1100–2:2010 (2010). *Hydrometry measurement of liquid flow in open channels—Part 2: Determination of the stage-discharge relationship* (p. 28). Geneva, Switzerland: International Organization for Standardization.

ISO/PWI 18320 (2015). *Hydrometry—Determination of the stage-discharge relationship* (Revision of ISO 1100–2:2010). Geneva, Switzerland: International Organization for Standardization.

ISO/TS 25377 (HUG) (2007). *Hydrometric Uncertainty Guidance (HUG)*. Geneva, Switzerland: International Organization for Standardization.

Jalbert, J., Mathevet, T., & Favre, A.-C. (2011). Temporal uncertainty estimation of discharges from rating curves using a variographic analysis. *Journal of Hydrology*, *397*(1–2), 83–92. https://doi.org/10.1016/j.jhydrol.2010.11.031

JCGM100-2008 (GUM) (2008). *Evaluation of measurement data—Guide to the expression of uncertainty in measurement*. JCGM (Joint Committee for Guides in Metrology) (p. 120). Sèvres, France: BIPM.

Juston, J., Jansson, P.-E., & Gustafsson, D. (2014). Rating curve uncertainty and change detection in discharge time series: Case study with 44-year historic data from the Nyangores River, Kenya. *Hydrological Processes*, *28*(4), 2509–2523. https://doi.org/10.1002/hyp.9786

Krueger, T., Freer, J., Quinton, J. N., Macleod, C. J. A., Bilotta, G. S., Brazier, R. E., et al. (2010). Ensemble evaluation of hydrological model hypotheses. *Water Resources Research*, *46*, W07516. 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*(6), 883–898. https://doi.org/10.1080/02626667.2010.504186

Le Coz, J., Blanquart, B., Pobanz, K., Dramais, G., Pierrefeu, G., Hauet, A., & Despax, A. (2016). Estimating the uncertainty of streamgauging techniques using field interlaboratory experiments. *Journal of Hydraulic Engineering*, *142*(7), 04016011. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001109

Le Coz, J., Camenen, B., Peyrard, X., & Dramais, G. (2012). Uncertainty in open-channel discharges measured with the velocity-area method. *Flow Measurement and Instrumentation*, *26*, 18–29. https://doi.org/10.1016/j.flowmeasinst.2012.05.001

Le Coz, J., Renard, B., Bonnifait, L., Branger, F., & Le Boursicaud, R. (2014). Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach. *Journal of Hydrology*, *509*, 573–587. https://doi.org/10.1016/j.jhydrol.2013.11.016

Levesque, V. A. & Oberg, K. A. (2012). Computing discharge using the index velocity method, U.S. Geol. Surv. Tech. Methods, 3–A23, 148 pp. Retrieved from http://pubs.usgs.gov/tm/3a23/

Liu, Y. L., 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*(1–2), 93–103. https://doi.org/10.1016/j.jhydrol.2009.01.016

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*(1), 135–152. https://doi.org/10.1002/hyp.10574

Mason, R.R. Jr., Kiang, J.E., Cohn, T.A. (2016). Rating curve uncertainty: An illustration of two estimation methods, IAHR River Flow conference, St. Louis, Missouri, USA, 12–15 July, 2016, 729–734.
McMillan, H., Freer, J., Pappenberger, F., Krueger, T., & Clark, M. (2010). Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrological Processes*, *24*(10), 1270–1284. https://doi.org/10.1002/hyp.7587
McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384
McMillan, H., Seibert, J., Petersen-Overleir, A., Lang, M., White, P., Snelder, T., et al. (2017). How uncertainty analysis of streamflow data can reduce costs and promote robust decisions in water management applications. *Water Resources Research*, *53*, 5220–5228. https://doi.org/10.1002/2016WR020328
McMillan, H. K., & Westerberg, I. K. (2015). Rating curve estimation under epistemic uncertainty. *Hydrological Processes*, *29*(7), 1873–1882. https://doi.org/10.1002/hyp.10419
Morlot, T., Perret, C., Favre, A.-C., & Jalbert, J. (2014). Dynamic rating curve assessment for hydrometric stations and computation of the associated uncertainties: Quality and station management indicators. *Journal of Hydrology*, *517*, 173–186. https://doi.org/10.1016/j.jhydrol.2014.05.007
Moyeed, R. A., & Clarke, R. T. (2005). The use of Bayesian methods for fitting rating curves, with case studies. *Advances in Water Resources*, *28*(8), 807–818. https://doi.org/10.1016/j.advwatres.2005.02.005
Muste, M., Ho, H. C., & Kim, D. (2011). Considerations on direct stream flow measurements using video imagery: Outlook and research needs. *Journal of Hydro-Environment Research*, *5*(4), 289–300. https://doi.org/10.1016/j.jher.2010.11.002
Ocio, D., Le Vine, N., Westerberg, I., Pappenberger, F., & Buytaert, W. (2017). The role of rating curve uncertainty in real-time flood forecasting. *Water Resources Research*, *53*, 4197–4213. https://doi.org/10.1002/2016WR020225
Osorio, A. L. N. A., & Reis, D. S. (2016). A Bayesian approach for the evaluation of rating curve uncertainties in flood frequency analyses, World Environmental and Water Resources Congress, West Palm Beach, Florida, USA, May 22-26, 482–491.
Pelletier, P. (1988). Uncertainties in the single determination of river discharge: A literature review. *Canadian Journal of Civil Engineering*, *15*(5), 834–850. https://doi.org/10.1139/l88-109
Petersen-Overleir, A., & Reitan, T. (2005). Objective segmentation in compound rating curves. *Journal of Hydrology*, *311*(1–4), 188–201. https://doi.org/10.1016/j.jhydrol.2005.01.016
Reitan, T., & Petersen-Overleir, A. (2008). Bayesian power-law regression with a location parameter, with applications for construction of discharge rating curves. *Stochastic Environmental Research and Risk Assessment*, *22*(3), 351–365. https://doi.org/10.1007/s00477-007-0119-0
Reitan, T., & Petersen-Øverleir, A. (2009). Bayesian methods for estimating multi-segment discharge rating curves. *Stochastic Environmental Research and Risk Assessment*, *23*(5), 627–642. https://doi.org/10.1007/s00477-008-0248-0
Reitan, T., & Petersen-Overleir, A. (2011). Dynamic rating curve assessment in unstable rivers using Ornstein–Uhlenbeck processes. *Water Resources Research*, *47*, W02524. https://doi.org/10.1029/2010WR009504
Savage, J. T. S., Pianosi, F., Bates, P., Freer, J., & Wagener, T. (2016). Quantifying the importance of spatial resolution and other factors through global sensitivity analysis of a flood inundation model. *Water Resources Research*, *52*, 9146–9163. https://doi.org/10.1002/2015WR018198
Schaake, J. C., Hamill, T. M., Buizza, R., & Clark, M. (2007). HEPEX: The hydrological ensemble prediction experiment. *Bulletin of the American Meteorological Society*, *88*, 1541–1548. https://doi.org/10.1175/BAMS-88-10-1541
Shrestha, R. R., Bárdossy, A., & Nestmann, F. (2007). Analysis and propagation of uncertainties due to the stage–discharge relationship: A fuzzy set approach. *Hydrological Sciences Journal*, *52*(4), 595–610. https://doi.org/10.1623/hysj.52.4.595
Sikorska, A. E., Del Giudice, D., Banasik, K., & Rieckermann, J. (2015). The value of streamflow data in improving TSS predictions—Bayesian multi-objective calibration. *Journal of Hydrology*, *530*, 241–254. https://doi.org/10.1016/j.jhydrol.2015.09.051
Sikorska, A. E., & Renard, R. (2017). Calibrating a hydrological model in stage space to account for rating curve uncertainties: General framework and key challenges. *Advances in Water Resources*, *105*, 51–66. https://doi.org/10.1016/j.advwatres.2017.04.011
Sikorska, A. E., Scheidegger, A., Banasik, K., & Rieckermann, J. (2013). Considering rating curve uncertainty in water level predictions. *Hydrology and Earth System Sciences*, *17*(11), 4415–4427. https://doi.org/10.5194/hess-17-4415-2013
Steinbakk, G. H., Thorarinsdottir, T. L., Reitan, T., Schlichting, L., Hølleland, S., & Engeland, K. (2016). Propagation of rating curve uncertainty in design flood estimation. *Water Resources Research*, *52*, 6897–6915. https://doi.org/10.1002/2015WR018516
Storz, S. M. (2016). Stage-discharge relationships for two nested research catchments of the high-mountain observatory in the Simen Mountains National Park in Ethiopia, (Master thesis). (p. 87). Switzerland: Bern University.
Thyer, M., Renard, B., Kavetski, D., Kuczera, G., & Clark, M. (2011). Improving hydrological model predictions by incorporating rating curve uncertainty. 34th IAHR World Congress, Brisbane, Australia.
Tomkins, K. (2014). Uncertainty in streamflow rating curves: Methods, controls and consequences. *Hydrological Processes*, *28*(3), 464–481. https://doi.org/10.1002/hyp.9567
Venetis, C. (1970). A note on the estimation of the parameters in logarithmic stage-discharge relationships with estimate of their error. *Bulletin of the International Association of Scientific Hydrology*, *XV*(2), 105–111.
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*(4), 603–613. https://doi.org/10.1002/hyp.7848
Westerberg, I. K., Di Baldassarre, G., Beven, K. J., Coxon, G., & Krueger, T. (2017). Perceptual models of uncertainty for socio-hydrological systems: A flood risk change example. *Hydrological Sciences Journal*, *62*(11), 1705–1713. https://doi.org/10.1080/02626667.2017.1356926
Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., & Freer, J. (2016). Uncertainty in hydrological signatures in gauged and ungauged catchments. *Water Resources Research*, *52*, 1847–1865. https://doi.org/10.1002/2015WR017635
Wilby, R. L., Clifford, N. J., De Luca, P., Harrigan, S., Hillier, J. K., Hodgkins, R., et al. (2017). The 'dirty dozen' of freshwater science: Detecting then reconciling hydrological data biases and errors. *WIREs Water*, *4*, e1209. https://doi.org/10.1002/wat2.1209