Journal of Hydrology 651 (2025) 132537

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Evaluating the applicability of the generalized power-law rating curve model With applications to paired discharge-stage data from Iceland, Sweden, and the United States

Rafael Daníel Vias [a,*], Birgir Hrafnkelsson [a], Timothy O. Hodson [b], Sölvi Rögnvaldsson [a], Axel Örn Jansson [c], Sigurdur M. Gardarsson [d]

[a] Department of Mathematics, Faculty of Physical Sciences, University of Iceland, Reykjavik, Iceland
[b] U.S. Geological Survey Water Resources Mission Area, Urbana, IL 61801, USA
[c] Department of Physics, Faculty of Physical Sciences, University of Iceland, Reykjavik, Iceland
[d] Faculty of Civil and Environmental Engineering, University of Iceland, Reykjavik, Iceland

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Szilágyi József, Associate Editor.

Dataset link: https://github.com/RafaelVias/discharge-stage-data

Keywords:
Bayesian hierarchical models
Chézy's formula
Generalized power-law
Manning's formula
Model selection
Rating curves
Segmented power-law

ABSTRACT

Hydrologic research and operations make extensive use of streamflow time series. In most applications, these time series are estimated from rating curves, which relate flow to some easy-to-measure surrogate, typically stage. The conventional stage-discharge rating takes the form of a segmented power law, with one segment for each hydrologic control at the stream gauge. However, these ratings are difficult to estimate with numerical methods, such that most are still developed manually. A few automated algorithms have emerged, but their use is sporadic, and their relative merits have not been rigorously assessed. One recently developed approach, the generalized power-law, avoids the segmenting problem by representing the power-law exponent as a Gaussian process. On one hand, this representation is more flexible and easier to fit, but that flexibility might also allow unrealistic solutions. This study evaluates the operational viability of the generalized power-law rating curve model under a range of conditions, using observations from 180 streams in Iceland, Sweden, and the United States. Overall, the model proved flexible and computationally robust, generating convincing rating curves across a range of geographic settings and was comparable to curves generated by a segmented rating model. Lastly, we propose a model-selection algorithm based on information theory to help identify the best rating curve model for a particular stream gauge.

1. Introduction

Organizations globally operate tens of thousands of streamgauges, providing crucial data for water management and infrastructure decisions. At these gauges, streamflow is typically not monitored directly but predicted based on stage measurements through rating curves, which describe the relationship between water surface level (stage) and flow through a particular cross section of a stream. Various approaches exist for constructing these rating curves, but selecting the most appropriate model for flow prediction remains a fundamental challenge. As Kiang et al. (2018) note in their seminal review, their is currently no overarching methodology for selecting a rating model. While various performance characteristics might be relevant, such as extrapolation capability, robustness, and model complexity, we argue that information criteria, which estimate out-of-sample performance and penalize over-fitting, should be the principal metric for model selection.

The power-law model, proposed by Venetis (1970), is well-established in hydrology. It assumes a log-linear relationship between flow and flow depth, where flow depth $(h - c)$ is the difference between the water surface level and the stage of zero flow. While often applicable, the model is unable to provide an adequate description of the discharge-stage relationship in a large proportion of streams (e.g., Hrafnkelsson et al., 2021). The prevailing practice in statistical rating curve fitting for such cases is to use a segmented power-law rating curve, which is essentially a piecewise linear regression, where two or more power-law rating curves are joined together (e.g., Reitan and Petersen-Øverleir, 2008). The total number of segments and the specific stage values where the segments join,

* Correspondence to: Department of Mathematics, Faculty of Physical Sciences, School of Engineering and Natural Sciences, University of Iceland, Sæmundargata 2, 102, Reykjavik, Iceland.
E-mail address: raffidv@gmail.com (R.D. Vias).

https://doi.org/10.1016/j.jhydrol.2024.132537
Received 1 August 2024; Received in revised form 9 November 2024; Accepted 28 November 2024
Available online 20 December 2024

referred to as the change- or breakpoints, must be predefined or estimated. Knowledge of the correct changepoint values is rare. In most cases, they must be inferred; however, they are notoriously hard to estimate because of the non-regularity and multimodality of the likelihood function (Petersen-Øverleir and Reitan, 2005).

Kiang et al. (2018) review several approaches for fitting more complex stage discharge ratings with uncertainty. Most methods incorporate ancillary hydraulic knowledge at the gauging station, like the geometry of the channel (Wickert et al., 2024; Vatanchi and Maghrebi, 2024a), the types of controls and their transitions (Le Coz et al., 2014), or even pre-fit rating curves (Coxon et al., 2015). This knowledge can be helpful for certain situations, like providing additional constraints during optimization or simulating other affects like hysteresis (Vatanchi and Maghrebi, 2024b), but these parameters can also be difficult to measure or unavailable for historical periods. For context, the U.S. Geological Survey (USGS) has more than 400,000 discharge-stage observations, with some records spanning over 100 years. Many of these gauges lack detailed surveys of the historical channel, so in practice, analyses are limited to methods that only require discharge-stage observations.

In general, these methods come in one of two types: those that are more causal, meaning their basic functional form is based on the physics of open-channel flow, like segmented-power laws (e.g., Reitan and Petersen-Øverleir, 2008); and those that are more statistical ("data-driven") in nature, like B-splines (e.g., Hrafnkelsson et al., 2012). Seeking to incorporate the best of both, Hrafnkelsson et al. (2021) introduced a novel extension of the power-law rating curve, referred to as the generalized power-law rating curve, which replaces the constant power-law exponent with a stage-dependent exponent derived from the hydraulics of open channel flow, specifically from the mean velocity formulas of Manning and Chézy (Chow, 1959). A statistical model was constructed around the generalized power-law using a Bayesian framework, in which the power-law exponent was represented as a Gaussian process. A Markov chain Monte Carlo (MCMC) sampling scheme was proposed to estimate the posterior density. The model was designed to be flexible and computationally robust, with constraints on the parameter space grounded in the physics of open channel flow.

While the theoretical framework of the generalized power-law model shows promise, its practical applicability has yet to be verified through large-scale evaluation, and a direct comparison with the widely used segmented power-law model has not been conducted. This study addresses these gaps by conducting a large-scale evaluation across 180 stream gauges and demonstrating a systematic approach to model comparison using information criteria. We also provide the first direct comparison between the generalized and segmented power-law models for selected cases where the simple power-law proves inadequate.

The primary objectives of this paper are: (i) to assess the applicability and computational robustness of the generalized power-law rating curve model of Hrafnkelsson et al. (2021) across 180 streams in Iceland, Sweden, and the United States; (ii) to conduct a comparative analysis between the generalized and segmented power-law rating curve models, primarily examining datasets where the simple power-law rating curve proves inadequate; and (iii) to demonstrate a systematic approach for ranking rating curve models using information theory, specifically the *widely applicable information criterion* (WAIC; Watanabe, 2010).

The paper focuses on evaluating static ratings: the ideal condition in which the stage-discharge relationship is stable during the period of observation. In practice, a rating curve can undergo transitory or persistent changes or "shifts" from vegetation growth, ice, debris, erosion, or deposition. In theory, the generalized power law could be extended to time-varying rating, but we leave this topic for future research.

2. Background

This section presents an overview of the generalized power-law rating curve model, beginning with a brief description of the classical power-law model and the segmented power-law approach.

2.1. Power-law rating curve model

The classical power-law rating curve, introduced by Venetis (1970), describes the relationship between discharge, Q, and stage, h, with a power-law

$$Q = a(h - c)^b, \tag{1}$$

where a, b, and c are unknown constants. The parameter c can be interpreted as the stage at which discharge is zero (or "stage of zero flow"), a is the discharge when the stage is one meter above the stage of zero flow (or the value of Q when $h = c + 1$, regardless of the unit system), and b is the constant power-law exponent. A natural logarithmic transformation of the model in (1) results in a linear relationship between $\ln(Q)$ and $\ln(h-c)$, with intercept and slope parameters $\ln(a)$ and b, respectively. The transformed model can be used as a basis for a statistical model of the form

$$\ln(Q_i) = \ln(a) + b\ln(h_i - c) + \epsilon_i, \tag{2}$$

where $i = 1, \ldots, n$, for n observations, and (Q_i, h_i) are the ith paired observations of discharge and stage, and ϵ_i is the corresponding error term, most commonly assumed to be mean-zero normally distributed, and independent of other error terms.

While this model can often adequately describe the stage-discharge relationship, complex channel geometries at the measurement site may require additional flexibility. For instance, when the water surface level rises above the top of a sharp-crested weir, causing a sudden and substantial increase in stream width, the log-linear assumption of the power-law model is unlikely to hold for stages both below and above this level.

The prevailing method in such cases is to use a segmented power-law rating curve, where multiple power-law rating curves are joined together at points (stage values) called changepoints (e.g., Reitan and Petersen-Øverleir, 2008). While flexible, estimating these changepoints presents significant challenges due to non-regular and multimodal likelihood functions (Petersen-Øverleir and Reitan, 2005). Several variations of the segmented power law have emerged, trading off between ease of fitting and reproducing the exact functional form typically used during manual rating development (ISO 18320:2020, 2020). For our analysis, we use the parameterization proposed by Hodson et al. (2024), which divides the channel cross-section into vertical segments.

2.2. Generalized power-law rating curve model

An extended version of the power-law rating curve, referred to as the generalized power-law rating curve, was proposed by Hrafnkelsson et al. (2021), in which the exponent term is a function of stage, represented by a Gaussian process. The model is given by

$$Q = a(h - c)^{b+\beta(h)}, \tag{3}$$

where Q, h, a, and c are as before, and $b + \beta(h)$ is the extended power-law exponent. The construction of $b + \beta(h)$ is based on the Chézy and Manning hydraulic formulas of open channel flow (Chow, 1959), which are of the form

$$Q = KP^{-x}A^{x+1}S^{1/2}, \tag{4}$$

where K is the friction constant, x is the Manning or Chézy coefficient, A is the cross-sectional area, P is the wetted perimeter, and S is the bed slope.

By setting $c = 0$, assuming the friction along the wetted perimeter is constant, and equating (3) and (4), Hrafnkelsson et al. (2021) derived, without loss of generality, that $b + \beta(h)$ is a function of x, A, and P, taking the form

$$b + \beta(h) = \frac{(x+1)\ln(A(h)/A(1)) - x\ln(P(h)/P(1))}{\ln(h)}, \tag{5}$$

for $h > 0$, and $h \neq 1$. Here, $A(h)$ and $P(h)$ are the area and wetted perimeter, respectively, of the water stream cross section, below a given

stage, h. Hrafnkelsson et al. (2021) used Manning's constant, $x = 2/3$, based on the theoretical derivation of Manning's empirical formula by Gioia and Bombardelli (2001) using the phenomenological theory of turbulence.

A single channel may have multiple hydraulic controls: riffles, banks, floodplains, etc. As stage rises or falls, the channel may transition from one control to another, changing the parameters of the power-law relation. The generalized power-law represents these transitions by making the exponent a function of stage ($b + \beta(h)$ in Eq. (3)). Representing the exponent with a smoothly varying function has two advantages. It is accurate in the sense that around the transition point, both controls interact to form a smooth curve, as opposed to an abrupt break point. Furthermore, representing controls as discrete segments becomes non-convex (Reitan and Petersen-Øverleir, 2006, 2008; Hodson et al., 2024), such that algorithms typically require additional hydrologic information to constrain the solution.

The natural logarithmic transformation of the model in (3) was proposed as a basis for a statistical model of the form

$$\ln(Q_i) = \ln(a) + (b + \beta(h_i)) \ln(h_i - c) + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2(h_i)), \quad (6)$$

where $i = 1, \ldots, n$, for n observations, and Q_i and h_i are as before. This formulation assumes that the friction along the wetted perimeter is constant. If friction varies with stage, then the physical interpretation of the model parameters should be made with caution.

The error terms, on the logarithmic scale, are assumed to be independent and normally distributed ($\mathcal{N}$) with mean zero, and variance $\sigma_\epsilon^2(h)$ that is allowed to vary with stage. The log-error variance is modeled as an exponential of a B-spline curve defined on the interval of stage observations in the data. The curve is a linear combination of six basis functions $\{B_k\}_{k=1}^6$, and the variance is modeled as

$$\sigma_\epsilon^2(h) = \exp\left(\sum_{k=1}^{6} \eta_k B_k(h)\right) = \prod_{k=1}^{6} \exp(\eta_k B_k(h)),$$

for $h \in [h_{\min}, h_{\max}]$, where $h_{\min}$ and $h_{\max}$ are, respectively, the minimum and maximum observed stage values in the data, and $\eta_1, \ldots, \eta_6$ are unknown parameters. At the minimum stage, the log-error variance is defined as $\sigma_\epsilon^2(h_{\min}) = \exp(\eta_1)$. At the maximum stage, it is defined as $\sigma_\epsilon^2(h_{\max}) = \exp(\eta_K)$. If all η_k are equal, then the variance is a constant function of stage. The B-spline's interior knots are equally spaced on $[h_{\min}, h_{\max}]$, with additional endpoint knots. For predictions outside the observed range, $\sigma_\epsilon^2(h) = \exp(\eta_1)$ for $h < h_{\min}$ and $\sigma_\epsilon^2(h) = \exp(\eta_K)$ for $h > h_{\max}$, allowing the model to extrapolate beyond observed water elevations.

The stage-dependent deviations in the power-law exponent, $\beta(h)$, are modeled as a two times mean-square differentiable Gaussian process governed by a Matérn covariance function with smoothness parameter 2.5, marginal standard deviation σ_β, and range parameter ϕ_β (Matérn, 1986). The (i, j)th element of the Matérn covariance matrix, Σ_β, is given by

$$\{\Sigma_\beta\}_{i,j} = \mathrm{cov}(\beta(h_i), \beta(h_j)) = \sigma_\beta^2 \left(1 + \frac{\sqrt{5} v_{i,j}}{\phi_\beta} + \frac{5 v_{i,j}^2}{3 \phi_\beta^2}\right) \exp\left(-\frac{\sqrt{5} v_{i,j}}{\phi_\beta}\right), \quad (7)$$

where $v_{i,j} = |h_i - h_j|$ denotes the absolute difference between the stage values h_i and h_j.

As presented in Hrafnkelsson et al. (2021), the statistical model is set up within a Bayesian hierarchical modeling framework. The priors for the range parameter ϕ_β and the marginal standard deviation σ_β in (7) were chosen to ensure that $b + \beta(h)$ remains within the interval $[1.0, 2.67]$ with high probability. This interval was determined by analyzing how different cross-sectional shapes in natural streams affect the power-law exponent in (5). To ensure stable inference, the parameter b is fixed at 1.835, the central value of this interval. The combination of a fixed b and carefully chosen priors for $\beta(h)$ ensures the model remains flexible enough to adapt to various stream geometries while staying within physically realistic bounds. It is important to emphasize that the Manning-Chézy formulas in (4) are used only to inform these prior distributions; the model itself makes no assumptions about specific types of hydraulic controls governing the flow.

2.2.1. Nested models - Generalized power-law class

The statistical model in (6) has three nested models that are used in the model-selection algorithm presented in Section 4. We will refer to this class of four models as the *generalized power-law class*. The models can be written out as in (8), where for each model, the ith discharge observation, Q_i, conditional on its corresponding stage measurement, h_i, is modeled as a lognormal random variable. The first and most complex model (GPLM) is the generalized power-law rating curve model in (6). The three nested models are attained by assuming one or both of the following model features. Either $\beta(h)$ is set to equal zero, thus turning the power-law exponent into a constant (PLM and PLM0), or the log-error variance is assumed to be constant (GPLM0 and PLM0).

$$\begin{aligned}
\text{GPLM}: &\quad \ln(Q_i) = \ln(a) + (b + \beta(h_i)) \cdot \ln(h_i - c) + \epsilon_i, &\quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2(h_i)) \\
\text{GPLM0}: &\quad \ln(Q_i) = \ln(a) + (b + \beta(h_i)) \cdot \ln(h_i - c) + \epsilon_i, &\quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2) \\
\text{PLM}: &\quad \ln(Q_i) = \ln(a) + b \cdot \ln(h_i - c) + \epsilon_i, &\quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2(h_i)) \\
\text{PLM0}: &\quad \ln(Q_i) = \ln(a) + b \cdot \ln(h_i - c) + \epsilon_i, &\quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)
\end{aligned} \quad (8)$$

Here, $i = 1, \ldots, n$, for n observations; ϵ_i is the error term for the ith discharge observation; $\sigma_\epsilon^2(h_i)$ and σ_ϵ^2 are the *stage-varying* and *constant* variance of the error terms on a logarithmic scale, respectively; and $b + \beta(h_i)$ and b are the *stage-varying* and *constant* power-law exponents, respectively. Note that by fixing $\beta(h) = 0$, the expected value of the discharge observations is made to follow the power-law rating curve, showing that the power-law rating curve is a special case of the generalized power-law rating curve.

Hrafnkelsson et al. (2021) take into account the nested structure of the generalized power-law class by assigning penalizing-complexity (PC) priors to the hyperparameters governing the stage-varying power-law exponent and log-error variance (Simpson et al., 2017; Fuglstad et al., 2019; Hrafnkelsson and Bakka, 2023). The priors are constructed such that the stage-varying exponent or log-error variance (or both) reduces to a constant (as a function of stage) if the data suggest these terms are not stage-dependent. Thus, the PC prior penalizes increased complexity, making the model less prone to overfitting.

Only GPLM and PLM are described in detail in Hrafnkelsson et al. (2021). However, GPLM0 and PLM0 can be made from GPLM and PLM by simply modeling the log-error variance as a constant. This is done in the R package `bdrc` by Hrafnkelsson et al. (2023a), where the constant log-error standard deviation, σ_ϵ, is assigned the same prior density that Hrafnkelsson et al. (2021) assign to $\sigma_\epsilon(h_{\min})$, where $h_{\min}$ is the smallest stage measurement in the data at hand. That is, σ_ϵ is assigned an exponential prior density with rate parameter $\lambda = 28.78$ (Hrafnkelsson et al., 2023b).

Appendix A.1 provides a brief overview of the Bayesian inference scheme used by Hrafnkelsson et al. (2021) to fit the generalized power-law model.

3. Data

This paper looks at datasets of paired observations of discharge, Q, in cubic meters per second, and stage, h, in meters. The measurements come from 180 streams in Iceland ($N = 60$), Sweden ($N = 60$), and California, USA ($N = 60$) and were gathered respectively by the Icelandic Meteorological Office (IMO), the Swedish Meteorological and Hydrological Institute (SMHI), and the USGS (U.S. Geological Survey, 2024). Fig. 1 shows the distribution of the number of observations per stream gauge. The smallest dataset contains 20 observations, and the largest one 256. We focus on modeling stable streams with little to no apparent shift in the discharge-stage relationship. For those streams that exhibited some shift, measurements from specific time periods were selected to ensure stability.

Data from two stations, Skogsliden (Sweden) and Hóll (Iceland), are depicted in Fig. 2. Whereas the power-law rating curve model

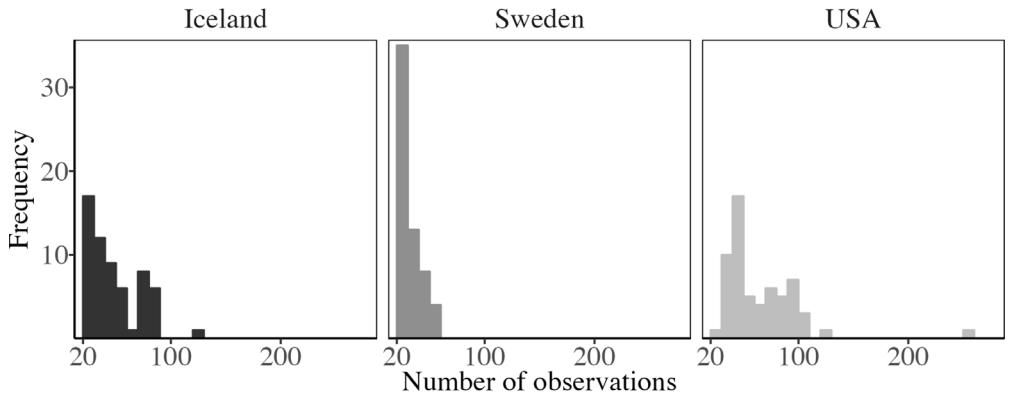

Fig. 1. A histogram of the number of observations across the 180 datasets.

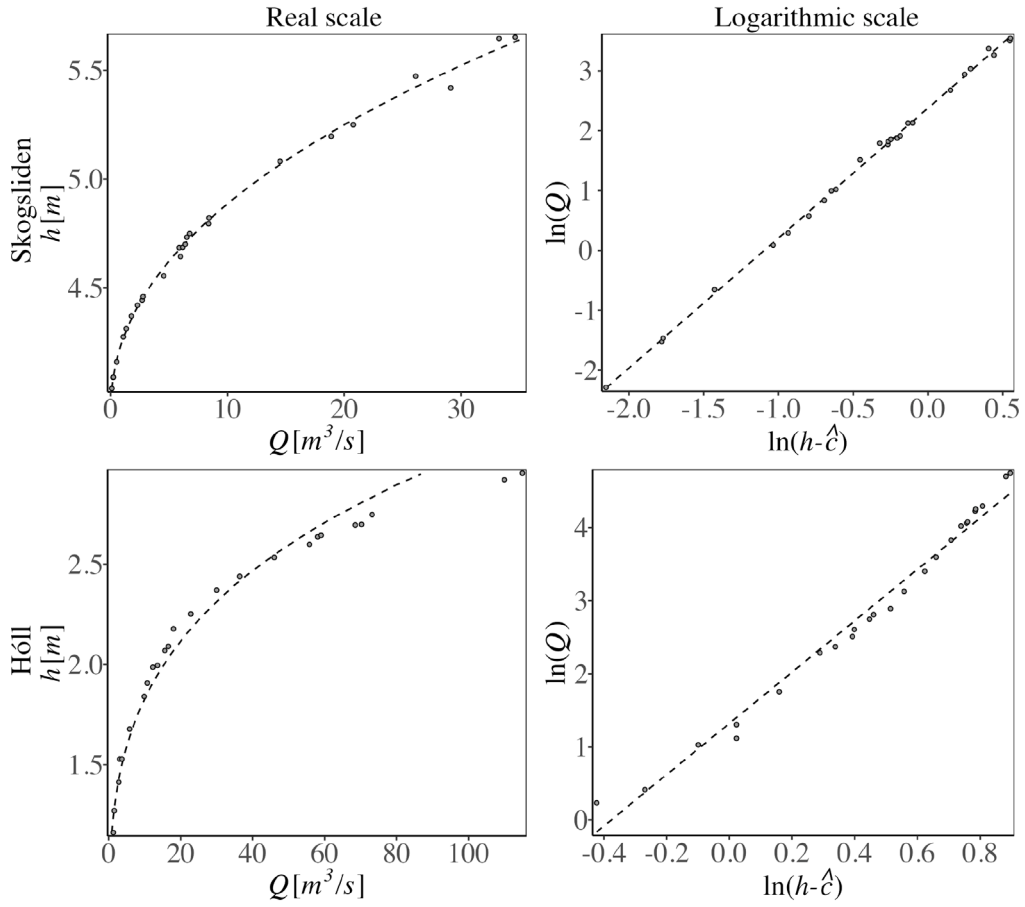

Fig. 2. Data from stations Hóll and Skogsliden, in Iceland and Sweden, respectively. Left column: stage and discharge on the real scale. Right column: discharge and flow depth ($h - \hat{c}$) on the logarithmic scale, where $\hat{c}$ is the posterior median of the c parameter estimated with the plm0 function in the bdrc R-package (Hrafnkelsson et al., 2023a). The dashed line shows the fitted power-law rating curve.

fits well to the Skogsliden data, it does not adequately capture the discharge-stage relationship at Hóll, where more flexible models like the segmented or the generalized power-law rating curve models are required to achieve a convincing fit.

In the results section, Section 5, a total of five datasets are used as examples. Data from Östra Norn and Skogsliden in Sweden were chosen to demonstrate the model-selection method presented in Section 4. Data from Stórhylur and Hóll in Iceland, and Ransta, Östra Norn, and Skogsliden in Sweden were selected to compare the generalized and segmented power-law rating curve models. All of these five datasets, except Skogsliden, are such that the power-law rating curve cannot give a convincing fit, and a more flexible model is needed. The datasets were otherwise arbitrarily chosen. The data from Ransta is unique in that documentation provided by the SMHI made it possible to approximate the water stream's cross-sectional shape at the measurement site. The documents depict a V-notch weir plate control located just downstream from where the measurements are gathered. This type of measurement site is arguably very close to meeting the assumptions of the segmented power-law rating curve model—making for an interesting comparison with the generalized power-law rating curve model in Section 5.

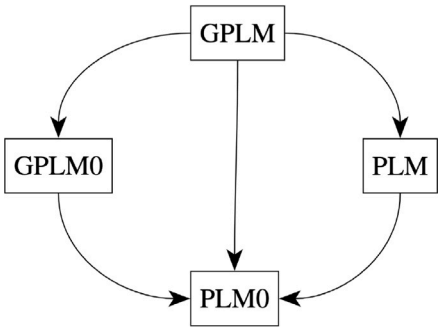

Fig. 3. A diagram showing the hierarchy of model complexity in the generalized power-law class. Arrows point from the models to their nested models.

When plotting the data on the real scale, stage is shown on the vertical axis and discharge on the horizontal axis, as is common practice in hydrology (e.g., ISO 18320:2020, 2020) However, when showing the data on a logarithmic scale, the transformed discharge is shown on the vertical axis and the transformed stage on the horizontal axis, as it is conventional in statistics to show the dependent and independent variables on the vertical and horizontal axes, respectively.

4. Model-selection method

In this section, we propose a model-selection method referred to as the *power-law tournament* that uses the widely applicable information criterion (WAIC) (Watanabe, 2010) to compare the estimated expected out-of-sample prediction error of the models in the generalized power-law class and select a parsimonious model for a given dataset.

4.1. Power-law tournament

The power-law tournament sets up the four models in the generalized power-law class in a single-elimination style tournament. The WAIC, which estimates the models' out-of-sample prediction error, is used to identify the most parsimonious model for the data at hand; see details in Appendix A.2. In total, three comparisons are made. The first two comparisons determine what models get through to the final comparison, where the most appropriate model is chosen.

The motivation for using a single-elimination style tournament setup and not simply selecting the model with the best WAIC is to give the less complex models a slight advantage. In each comparison, a model and one of its nested models go head-to-head. With the exception of one pairing, for any two models in the generalized power-law class, one is nested in the other; see Fig. 3. The only pairing for which this is not the case is GPLM0 and PLM. The tournament takes into account this hierarchy of model complexity, slightly favoring the nested model in each comparison.

Recall that there are two ways to simplify the most complex model (GPLM). Either convert the stage-varying power-law exponent to a constant (PLM) or the stage-varying log-error variance to a constant (GPLM0). If both changes are made, the model turns into the least complex model (PLM0).

The single-elimination style tournament setup is shown in Fig. 4. In each comparison, we are essentially asking if we are justified in increasing the complexity to the less complex model. In the first comparison, the models with a stage-varying power-law exponent (GPLM and GPLM0) are compared. In the second comparison, the models with a constant power-law exponent (PLM and PLM0) are compared. In these first two comparisons, we evaluate if modeling the log-error variance as a function of stage (versus constant) has a large enough effect on the predictive accuracy of the models. The models selected in the first two comparisons are then set up against each other in the third and final comparison. In the final comparison, we evaluate if modeling the power-law exponent as a function of stage (versus constant) has a large enough effect on the predictive accuracy of the models. The model that comes out on top in this final comparison is deemed the most parsimonious model for the data at hand.

As mentioned above, the motivation for using the single-elimination style tournament setup is to favor the nested model slightly in each comparison. This is done by requiring the difference in WAIC to reach a specific threshold value if the more complex model is to be selected instead of its nested rival; see details on the selection criteria in Section 4.2.

In the first two comparisons, the models with a stage-varying log-error variance are the more complex. In the third and final comparison, the model with a stage-varying power-law exponent is always said to be the more complex. This is, in fact, true for all possible model pairings in the third comparison, with one exception, namely GPLM0 and PLM, as previously mentioned. Nonetheless, we opt for calling GPLM0 the more complex model when comparing these two models. This is because making the power-law exponent stage-dependent increases the flexibility of the rating curve to a much greater extent than making the log-error variance stage-dependent.

In the following subsection, we specify the criteria for judging if a more complex model is more appropriate than its nested rival.

4.2. Selection criteria for the power-law tournament

Writing on the *Akaike information criterion* (AIC) (Akaike, 1974), another estimator of prediction error founded on information theory, Burnham and Anderson (2002) presented useful rules of thumb (particularly for nested models) to judge the degree to which there is empirical support for one model over another. As with WAIC, a smaller AIC value indicates better predictive accuracy. These rules of thumb state that if two models have an absolute difference in AIC between 0–2, both should be considered. An absolute AIC difference between 4 and 7 constitutes considerably less empirical support for the model with a greater AIC value. And a difference greater than 10 reflects essentially no support for the 'worse' model. Spiegelhalter et al. (2002) found these rules of thumb to also work reasonably well for the *deviance information criterion* (DIC), a generalization of the AIC for hierarchical modeling.

As WAIC is closely related to DIC, we base the tournament selection criteria on these rules of thumb. Here we define the WAIC difference to be $\Delta_{\text{waic}} = \text{WAIC}_B - \text{WAIC}_A$, where WAIC_A and WAIC_B are the WAIC estimates for the more and less complex models, respectively. In any given comparison, the more complex model is selected if, and only if, the following two criteria are met. We require that the difference in WAIC is greater than two in favor of the more complex model and that the WAIC difference minus one standard error is positive. That is, we require that $\Delta_{\text{waic}} > 2$, and that $\Delta_{\text{waic}} - \text{se}(\Delta_{\text{waic}}) > 0$. The latter requirement is to ensure some certainty to the claim that the more complex model is superior.

When selecting an appropriate rating curve model, we are less concerned with type I errors than with type II. This is because, to a varying degree, the discharge-stage relationship can vary through time due to changes in channel-caused erosion, deposition, vegetation growth, etc. For streams with very unstable channels, like those with sandy banks, frequent observations are needed to maintain an accurate rating curve. For this reason, we have not imposed stricter constraints on the complex models in the tournament.

In Appendix A.3, we explore an alternative selection criteria for the power-law tournament that uses the posterior model probabilities, computed with the Bayes factor, to select the appropriate model.

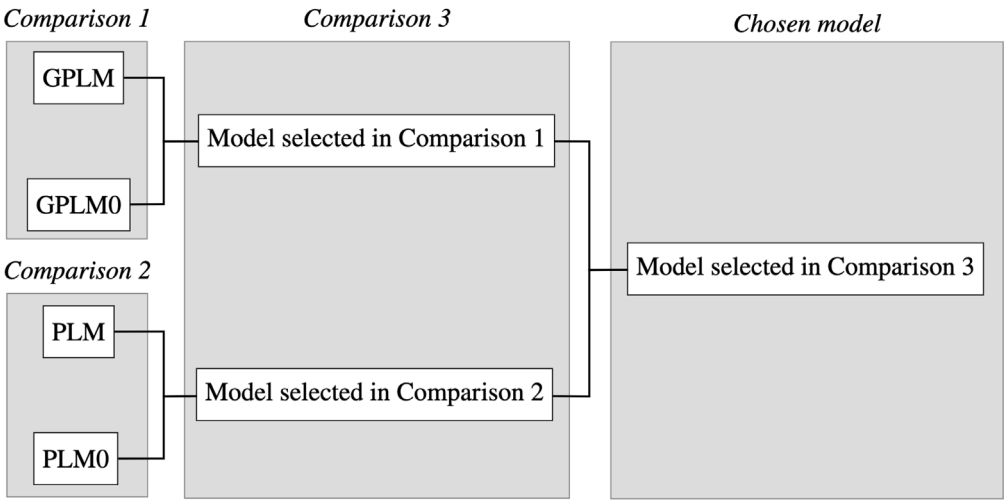

Fig. 4. A diagram showing the setup of the single-elimination style power-law tournament.

5. Results

In this section, we compare the segmented and generalized power-law rating curve models on five datasets from Iceland and Sweden, fit the four models in the generalized power-law class to two Swedish datasets, and perform a meta-analysis of the power-law tournament when employed to the 180 datasets introduced in Section 3.

5.1. Comparing generalized and segmented power-law rating curves

In this subsection, we compare the widely used segmented power-law rating curve to the generalized power-law rating curve. We use WAIC to estimate the expected out-of-sample prediction error of the models, which is valid only for predictions within the range of observed data. The models' ability to extrapolate beyond this range will be examined in Section 5.2. First, we compare the rating curves with the data from Ransta, where documents from the measuring site allow us to use highly informative priors for the changepoint parameters when fitting the segmented power-law model. Then, we compare the rating curves with four more datasets where the cross-sectional shapes at the measurement sites are unknown, forcing the use of non-informative changepoint priors.

5.1.1. Informative changepoint priors

The data from the stream gauge at Ransta in Sweden is such that the measurements are gathered just upstream from a V-notch weir plate control. Documentation provided by the SMHI was used to create an approximation of the cross section at the measurement site; see Fig. 5. A simple power-law rating curve cannot adequately describe the discharge-stage relationship at this site because of the relatively complex shape of the cross section. However, the parts of the cross section corresponding to each of the disjoint stage segments $H_1 = [8.7, 9.0]$, $H_2 = [9.0, 9.3]$, and $H_3 = [9.3, 10.0]$ are simple in form, thus suggesting the use of a three-segment power-law rating curve with changepoints at $h = 8.7$ m, $h = 9.0$ m and $h = 9.3$ m.

We use these stage values to construct informative priors for the changepoint parameters in the three-segment power-law model and compare it with the generalized power-law model. Although the cross-sectional shape is approximated and may lack complete precision, this is not critical, as it serves solely to form a reasonable prior distribution for the changepoint parameters in the segmented model. For comparison, we also include a three-segment power-law model with uninformative changepoint priors. Both versions of the generalized power-law model are analyzed: the model with constant log-error variance (GPLM0) and the model with stage-dependent log-error variance (GPLM).

Fig. 6 shows the three-segment power-law and the generalized power-law rating curve (GPLM) fit to the data from Ransta. Highly informative Gaussian priors ($\sigma \approx 1.5$ cm) were used for the changepoint parameters when fitting the segmented model, essentially telling the model the true values. Both rating curves fit well to the data.

When comparing the predictive accuracy of the models, unsurprisingly, the three-segment model (SPLM3) with highly informative Gaussian changepoint priors has the lowest estimated out-of-sample prediction error (WAIC), roughly three WAIC units lower than the next best model, GPLM; see Fig. 7. Between the models with no prior changepoint information (GPLM, GPLM0, and SPLM0 with uniform changepoint priors) the results are very similar, with the WAIC estimates falling within a two-unit interval.

These results show that when prior knowledge on changepoints exists, and the cross-sectional shape between each pair of adjacent changepoints is simple, the segmented power law with informative changepoint priors performs well. Moreover, for this cross-sectional shape, ideal for a segmented power law, the generalized power law gives a comparable fit, outperforming the three-segment power-law model when using non-informative (uniform) changepoint priors; see Fig. 7.

Unlike the segmented model, the generalized power-law model does not require several runs with differing numbers of segments and a subsequent model comparison to evaluate the appropriate segmentation. With a single run of the generalized power-law models, the stage-dependent power-law exponent appears sufficiently flexible to account for the varying hydraulic controls. For illustration, we compare two approaches to determining the exponent's variation with stage at Ransta. One computed directly from the cross-sectional geometry shown in Fig. 5 using Eq. (5), and another estimated from the actual measurements using GPLM; see Fig. 8. While the directly computed exponent relies on the approximated cross-sectional geometry and should be interpreted cautiously, it is noteworthy that it falls within the 95% credible interval of the posterior distribution of the power-law exponent estimated from the observed discharge-stage relationships.

5.1.2. Non-informative changepoint priors

In many cases, the data come from measuring sites where the cross-sectional shape is irregular, giving no clear indication of the correct number of the changepoints or their specific stages. It is common practice in such cases to fit multiple segmented power-law models with differing numbers of changepoints and compare the models. Here, we

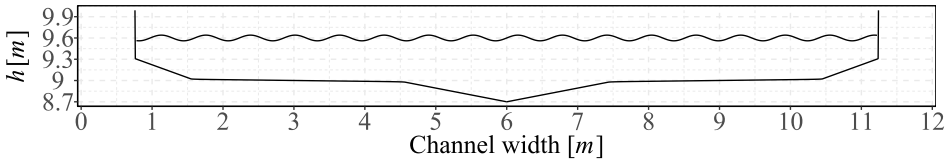

Fig. 5. An approximation of the cross section at the measuring site at Ransta, Sweden. Drawings provided by the SMHI were used to create the approximated cross section. Stage is shown in meters on the vertical axis, and the channel width is presented in meters on the horizontal axis, with the midpoint arbitrarily set at 6 m. The roughness along the wetted perimeter is assumed constant.

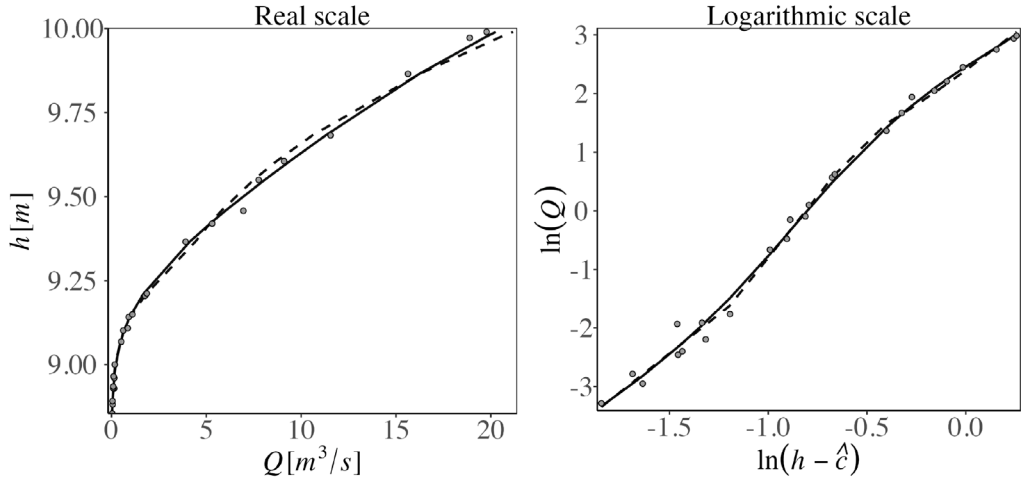

Fig. 6. A three-segment power-law (dashed line) and a generalized power-law (solid line) rating curves fitted to the data from Ransta (dots). The segmented rating curve was modeled with highly informative Gaussian changepoint priors. Left panel: stage plotted against discharge. Right panel: discharge and flow depth $(h - \hat{c})$ on the logarithmic scale, where $\hat{c}$ is the posterior median of the c parameter inferred with the generalized power-law model.

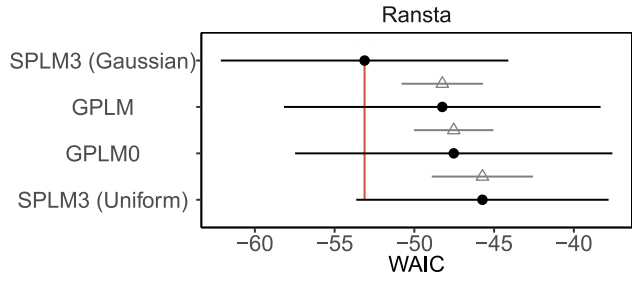

Fig. 7. A forest plot showing the WAIC estimates (black dots), plus and minus one standard error (black error bars), of the models fitted to the Ransta data. The models are arranged on the vertical axis by WAIC. The standard error of the estimated WAIC difference between each model and the model with the lowest WAIC is depicted with gray error bars centered at the WAIC of the model with the higher WAIC (gray triangles). The superiority of the top model to any other can be considered less uncertain the further the lower end of the gray error bar is above the WAIC of the top model (red line) (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.).

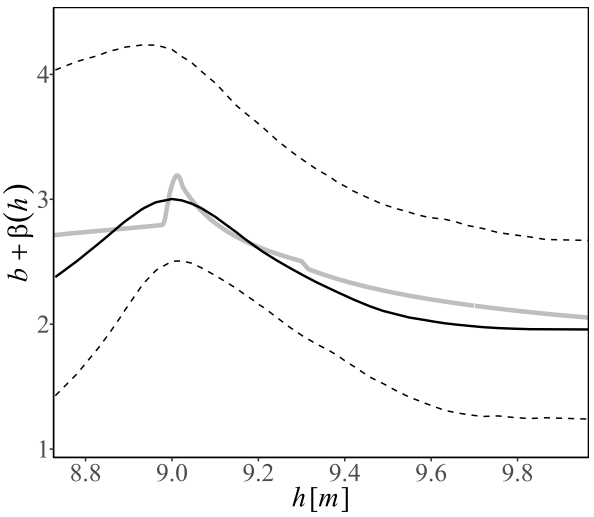

Fig. 8. The posterior distribution of the power-law exponent, $b + \beta(h)$, when applied to the data from Ransta, summarized by the median value (solid black line) and the 95% posterior intervals (dashed lines). The gray line shows the power-law exponent calculated directly with Eq. (5) using Manning's constant ($x = 2/3$).

use four datasets to compare the four models in the generalized power-law class to a two-, three- and four-segment power-law model, named SPLM2, SPLM3, and SPLM4, respectively. The datasets are Hóll and Stórhylur from Iceland and Östra Norn and Skogsliden from Sweden.

The model comparison results are shown in Fig. 9. Only the data from Skogsliden is such that the simple power-law rating curve is the most parsimonious model. A more flexible rating curve model is needed for the other three datasets. So, for Skogsliden, PLM0 has the lowest WAIC of all the models. The models in the generalized power-law class have very similar WAIC because the penalized-complexity priors reduce the more complex models to the simpler versions when the data suggest there is no need for the added complexity. On the other hand, the WAIC of the segmented models increases with every additional unnecessary segmentation.

The results for the remaining three models show that the best generalized power-law model perform as well, or, as in the case of Östra Norn, slightly better than the best segmented power-law model.

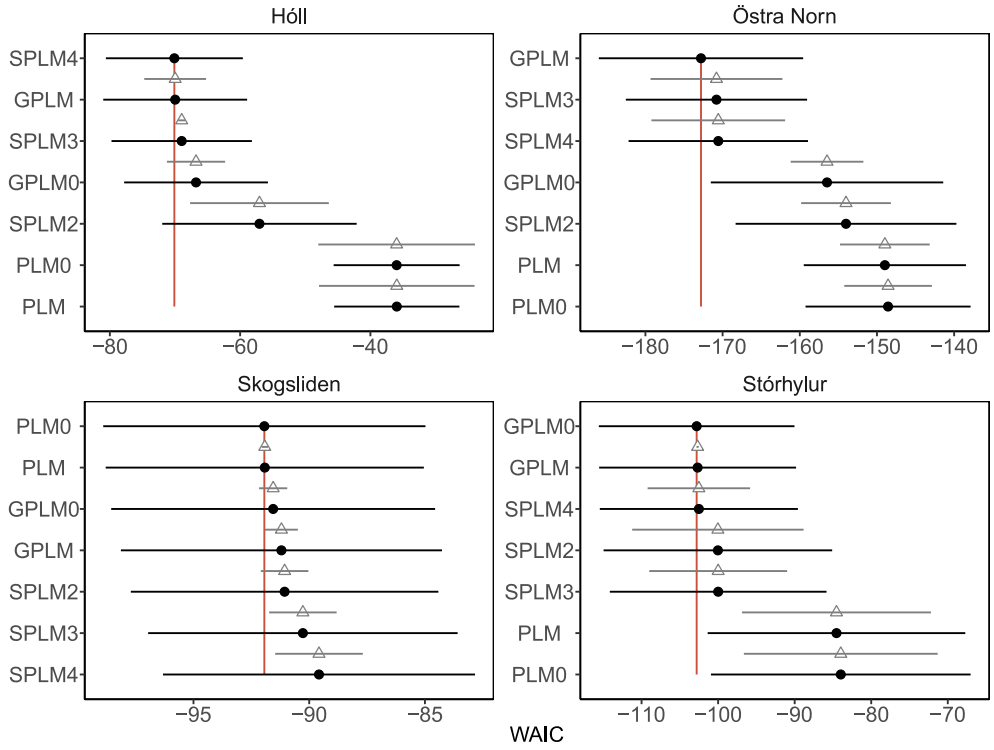

Fig. 9. A forest plot showing the WAIC estimates (black dots), plus and minus one standard error (black error bars), of the models fitted to the data from Hóll, Östra Norn, Skogsliden, and Stórhylur. The two-, three- and four-segment power-law models are named SPLM2, SPLM3 and SPLM4, respectively. The models are arranged on the vertical axis by WAIC. Thus, out of these seven models, the models on top are arguably the best for each dataset. The standard error of the estimated WAIC difference between each model and the top model is depicted with a gray error bar centered at the WAIC of the model with the higher WAIC (gray triangles). The superiority of the top model to any other can be considered less uncertain the further the lower end of the gray error bar is above the WAIC of the top model (red line) (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.).

5.2. Upward extending the generalized power-law rating curve

While the comparative analysis in the previous section evaluated predictive performance within the range of observed data, extrapolating beyond the highest observed stage is often necessary in practice, though such predictions should be approached with caution. This is particularly true when no information about channel characteristics above the observed range is incorporated into the predictions. In this section, we examine how both models perform when extrapolating to higher stages, using only the observed discharge-stage relationships.

The four datasets in the previous sections that required a more complex model than the classical power-law (Ransta, Östra Norn, Stórhylur, and Hóll) were chosen for this comparison. For each dataset, GPLM is fit to the entire set of observations, as well as the segmented model which had the lowest WAIC out of SPLM2, SPLM3, and SPLM4. For Östra Norn, SPLM3 had the lowest WAIC, whereas, SPLM4 was lowest for the other three.

Another set of rating curves is then fitted to a subset of the data. The observations that land within the interval $[h_{\min}, h^*_{\max}]$ are used as a training dataset to fit a rating curve that we then use to predict up from $h^*_{\max}$ to $h_{\max}$, where $h_{\min}$ and $h_{\max}$ are the smallest and largest stage observations in the data, respectively, and $h^*_{\max}$ is the largest stage observation below $h^* := h_{\min} + 0.8(h_{\max} - h_{\min})$. Like before, GPLM is fitted to each dataset, and the segmented models with the lowest WAIC are chosen in each case. This time around, Stórhylur was the only dataset where SPLM3 had the lowest WAIC, and for the other three, SPLM4 was the lowest.

The results are shown in Fig. 10. The extrapolations are all reasonably good, considering they extend far above the maximal discharge observations in the training set and no information about channel characteristics above these observations was incorporated in the prediction. While these results are promising, extrapolation beyond observed flow regimes should always be approached with caution, particularly for overbank flows where channel characteristics may change significantly. The generalized power-law model, being set up within a Bayesian framework, could potentially incorporate prior knowledge about channel characteristics in overbank conditions, though this extension is left for future research.

5.3. Application of the power-law tournament

In this subsection, we fit the models in the generalized power-law class to two Swedish datasets from Östra Norn and Skogsliden. We analyze the model results and employ the power-law tournament to both datasets.

5.3.1. Fitting the models

The four statistical models in the generalized power-law class are applied to the data from the Östra Norn and Skogsliden stations. The models are first applied to the data from Östra Norn. The estimated rating curves and the corresponding residual plots can be seen in Fig. 11. The generalized power-law rating curves (GPLM and GPLM0) provide a convincing fit to the data, unlike the power-law rating curves (PLM and PLM0).

The assumption of mean zero normally distributed measurement errors (on a logarithmic scale) is not met for PLM and PLM0 when applied to the Östra Norn dataset, as the residuals show a clear nonlinear pattern as a function of stage. The residual plots for GPLM and GPLM0, however, indicate that the mean of the model captures the underlying mean quite well. Furthermore, the prediction intervals of GPLM indicate that the standard deviation, modeled as a function of stage, is able to capture the variability in the error terms.

The estimated power-law exponents and standard deviations of the error terms (on a logarithmic scale) are shown as functions of stage

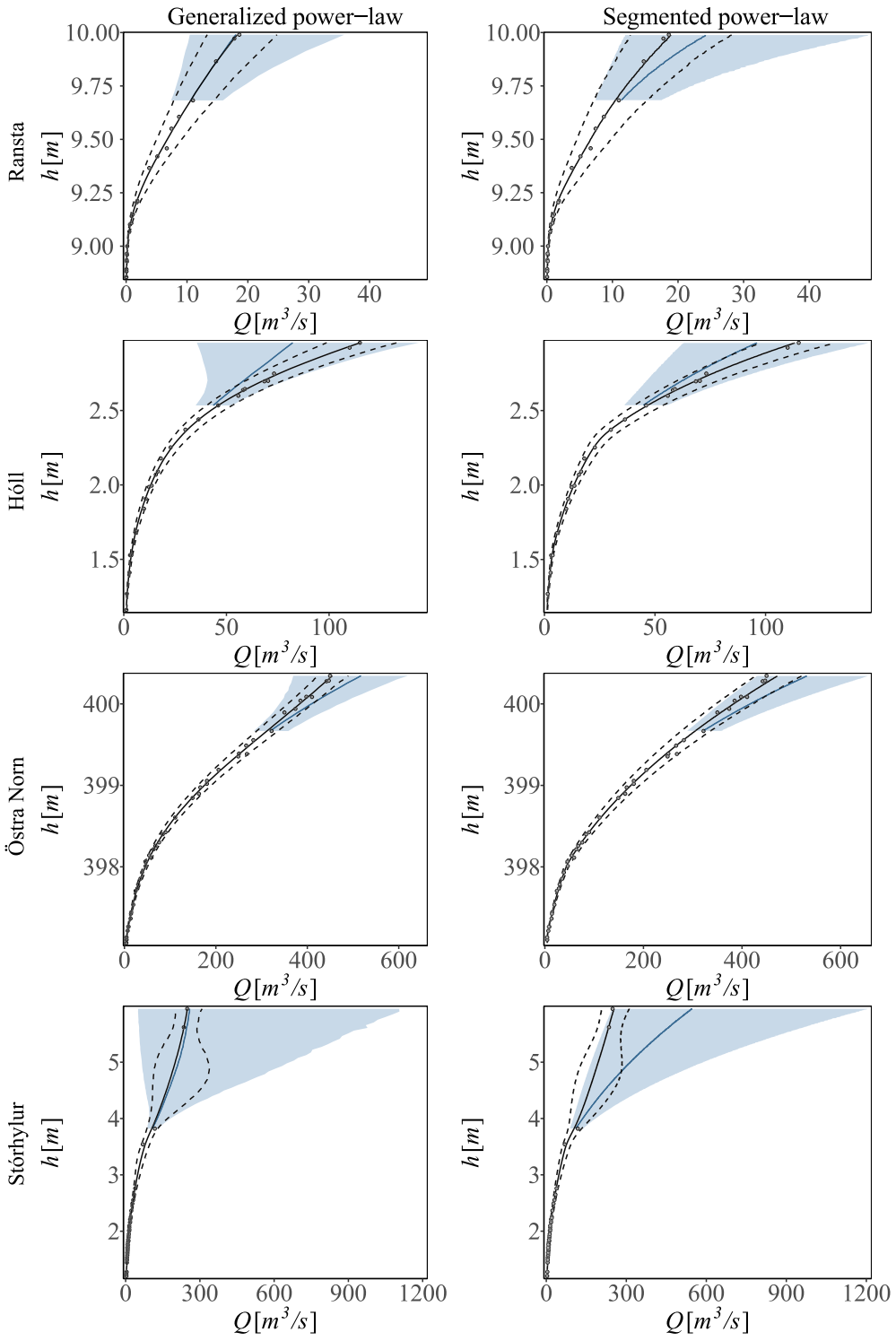

Fig. 10. Upward extrapolations of the generalized (left column) and segmented (right column) power-law rating curves for the data from Ransta, Östra Norn, Stórhylur, and Hóll. The solid black lines show the rating curves estimated with the entire dataset. The dashed lines for the generalized models are the 95% posterior predictive interval, whereas, for the segmented models, they represent $\exp(\mu \pm 1.96\tau)$, where μ is the expected value of the rating curve on a logarithmic scale, and τ is the standard deviation of the posterior predictive samples of the rating curve on a logarithmic scale. The dark blue lines and the blue ribbons show these same quantities but are upward extrapolations from a rating curve fitted only with a subset of the data (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.).

in Fig. 12. The exponents for PLM and PLM0, which are modeled as constants, are both estimated to be around 2.5. On the other hand, the exponents for GPLM and GPLM0 are non-linear, and these model components are being put to work to achieve a better fit. The two bottom-right graphs show the constant (PLM0) and stage-varying (PLM) standard deviation for the models based on the power-law rating curve, and these estimates are quite similar. The two upper-right graphs showing the standard deviation for the models based on the generalized power-law rating curve are very different. In this case, the stage-varying standard deviation captures the heteroscedasticity in the measurement

Table 1
The tournament results for the data from Östra Norn. The model selected in each comparison is pointed out with an arrow.

Comparison	Model	Selected	WAIC	Δ_{waic}	se(Δ_{waic})
1	GPLM	⇐	244.17	15.68	4.94
	GPLM0		259.85		
2	PLM		269.93	0.15	0.57
	PLM0	⇐	270.08		
3	GPLM	⇐	244.17	25.91	6.04
	PLM0		270.08		

Table 2
The tournament results for the data from Skogsliden. The model selected in each comparison is pointed out with an arrow.

Comparison	Model	Selected	WAIC	Δ_{waic}	se(Δ_{waic})
1	GPLM		5.78	−0.02	0.12
	GPLM0	⇐	5.75		
2	PLM		5.59	−0.11	0.11
	PLM0	⇐	5.48		
3	GPLM0		5.75	−0.27	0.64
	PLM0	⇐	5.48		

errors, which can be observed on the residual plots for GPLM and GPLM0, in Fig. 11. In conclusion, the full complexity of GPLM appears to be necessary to give an adequate description of the discharge-stage relationship at Östra Norn.

Next, we apply the models to the data from Skogsliden. Fig. 13 shows the estimated rating curves and the corresponding residual plots. In this case, the rating curves and corresponding prediction intervals are virtually indistinguishable, and the same holds true for the residual plots. The posterior distributions of the power-law exponents and standard deviations, presented in Fig. 14, are also practically identical. Whether the power-law exponent or standard deviation are modeled as a function of stage or as a constant does not seem to matter in this case. The use of the full complexity of GPLM does not appear to be justified in this case, and the least complex model (PLM0) gives a good enough description of the discharge-stage relationship at Skogliden. Note that the more complex models mimic the reduced models when the increased complexity is not needed, which is by design, as described in Section 2.2.1.

5.3.2. Comparing the models with the power-law tournament

The power-law tournament is applied to the data from Östra Norn and Skogsliden to select a parsimonious model. Table 1 presents the tournament results for the data from Östra Norn. In the first comparison, both requirements are satisfied for GPLM to be chosen. The WAIC estimate for GPLM is lower by around 15.68 (i.e., $\Delta_{waic} > 2$), and Δ_{waic} minus one standard error is greater than zero (i.e., Δ_{waic}-se(Δ_{waic}) > 0). This is not surprising since the residuals showed evidence of heteroscedasticity (see Fig. 11), so it is appropriate that a model that can capture such variability is chosen. The simple power-law models are compared in the second comparison. Here, the WAIC estimates are very similar, and the difference is not large enough to surpass the predefined critical value threshold of $\Delta_{waic} > 2$, and the less complex model, PLM0, is selected. Again, this seems appropriate when we consider the residuals of PLM and PLM0 in Fig. 11. Due to the poor fit of these models, the large movement of the residual mean over the stage values drowns out the actual noise in the data, making it impossible to model the underlying heteroscedasticity, resulting in similar estimates. Then, in the final model comparison, the chosen models from the first two comparisons, GPLM and PLM0, are compared. The WAIC for GPLM is lower by around 25.91 and is therefore declared the most parsimonious model. In conclusion, the power-law tournament indicates that the full generalized power-law rating curve model, GPLM, is the appropriate model for the data from Östra Norn, which is what one would expect from looking at the model results in Figs. 11 and 12.

Table 2 shows the tournament results for the data from Skogsliden. The WAIC for all the models are very similar, with a maximum difference of 0.27. Since in every comparison the difference is not large enough for the more complex model to be chosen, the least complex model, PLM0, is selected. The power-law rating curve with an assumption of a constant log-error variance gives a good enough description of the data so it is reasonable to select that model without introducing unnecessary complexity.

Table 3
The number of times each model was selected in each comparison in the power-law tournaments when applied to the 180 datasets. The selection rate is the total number of times a model was selected divided by the total number of datasets.

Comparison	Model	Selected (n)	Selection rate (%)
1	GPLM	42	23.3
	GPLM0	138	76.7
2	PLM	39	21.7
	PLM0	141	78.3
3	GPLM	26	14.4
	GPLM0	64	35.6
	PLM	15	8.3
	PLM0	75	41.7

5.4. Meta analysis

This section provides an overview of the results obtained from the power-law tournament when applied to the 180 real-world datasets introduced in Section 3. The meta-analysis gives insights into the variations in model-selection rates across the three different model comparisons and countries.

Employing the power-law tournament across the 180 datasets showed the generalized power-law class to be robust and versatile. In all cases, covering a wide range of discharge-stage relationships, the tournament was able to select a rating curve model that provided a convincing fit. The fitted rating curves and tournament summary statistics for each of the 180 datasets are provided as supplementary material.

The tournament results are categorized by comparison in Table 3 and Fig. 15. In the first and second comparisons, models incorporating a stage-dependent log-error variance (GPLM and PLM) were chosen for approximately 23% and 22% of the datasets, respectively. In the third and final comparison, the simple power-law rating curve (PLM0), proposed by Venetis (1970), was identified by the power-law tournament as the most parsimonious model for 41.7% of the datasets. For exactly half of the datasets, GPLM or GPLM0 was deemed the most parsimonious model. The tournament deemed GPLM or PLM most parsimonious for 22.8% of the datasets. That is, choosing a model that allows the log-error variance to vary with stage decreased substantially the estimated prediction error for 22.8% of the datasets.

A country-specific analysis is presented in Table 4 and Fig. 16. In Iceland and Sweden, 48.3% and 45.0% of the datasets, respectively, necessitate the use of generalized power-law models, slightly lower than the USA's percentage at 56.7%. Notably, 16.7% of the Icelandic datasets require models incorporating a stage-dependent log-error variance (GPLM or PLM), in contrast to 26.7% and 25.0% for Sweden and the USA, respectively.

Lastly, we use the WAIC to calculate Akaike weights (e.g., Burnham and Anderson, 2002) for the four models at each stream gauge. These weights represent the probability that each model is the "best" for a particular gauge, where "best" is formally defined as the model minimizing the Kullback–Leibler divergence between the modeled and observed streamflow. The distribution of the weights is shown in

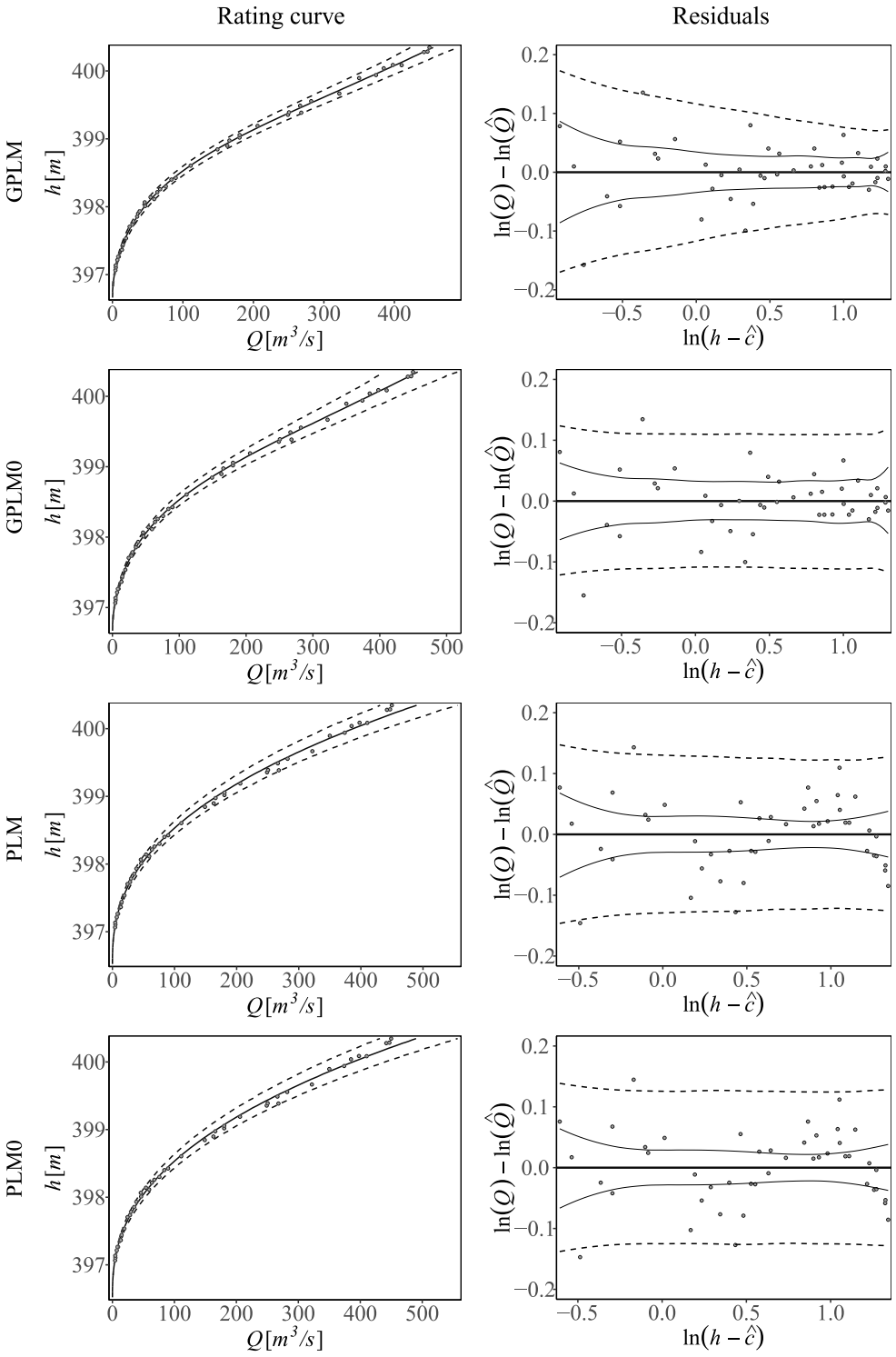

Fig. 11. The rating curves and corresponding residual plots of the four models in the generalized power-law class fit to the data from Östra Norn. Each row illustrates the results from a single model. The left column shows the estimated rating curves (solid line) and the 95% posterior predictive interval (dashed lines). The right column shows the residual plots on a logarithmic scale. The horizontal line corresponds to the posterior median of the logarithmically transformed rating curve, $\ln(\hat{Q})$. The solid curves around the horizontal line corresponds to the 95% posterior interval for the logarithmically transformed rating curve median. The dashed curves correspond to the 95% posterior predictive interval for the logarithmically transformed rating curve. The estimate for the c parameter, $\hat{c}$, is the posterior median of c.

Fig. 17. These results indicate that GPLM is never a poor model. That is, the WAIC of GPLM is never much greater than the WAIC of the best model. In fact, in the few cases where neither GPLM nor GPLM0 had the lowest WAIC, the difference of $\max(w_{\text{GPLM}}, w_{\text{GPLM0}})$ and $\max(w_{\text{PLM}}, w_{\text{PLM0}})$ was at most ~0.27, where w denotes the WAIC-based Akaike weight of a model (Fig. 18).

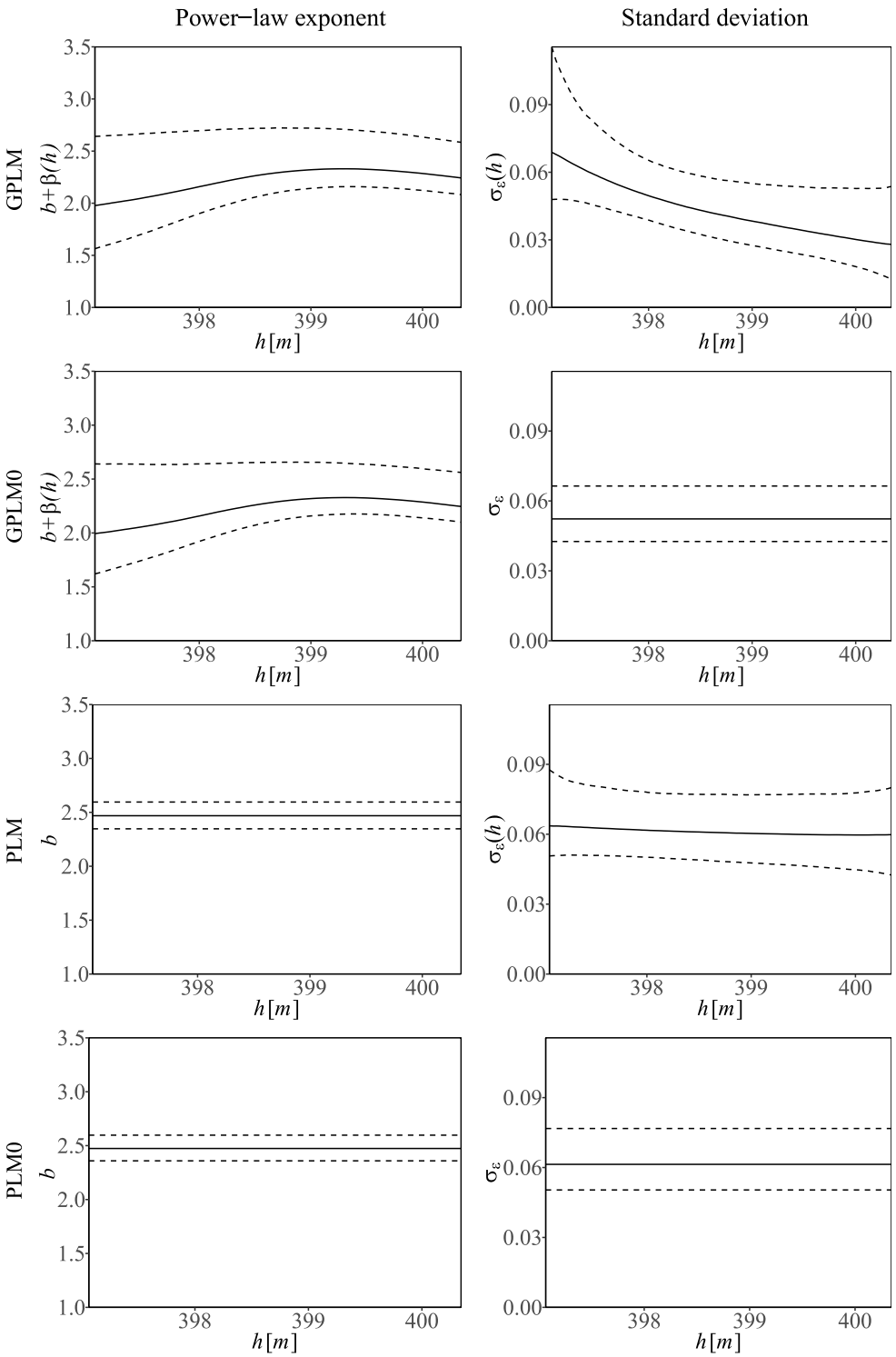

Fig. 12. The estimated power-law exponents and the log-error standard deviations of the four models in the generalized power-law class fit to the data from Östra Norn. Each row illustrates the results from a single model. The left column shows the posterior distributions of the power-law exponents, summarized by the median (solid line) and 95% credible interval (dashed lines). The right column shows the posterior distributions of the log-error standard deviations, also summarized by the median (solid line) and 95% credible interval (dashed lines).

6. Discussions

Water resources research and management make extensive use of estimated streamflow values derived from stage measurements with rating curves. Although the functional form of those ratings is grounded in open-channel hydraulics, a substantial amount of operational practice has emerged ad hoc, with most rating curves still developed "by hand" with the aid of computer software rather than by statistical modeling and optimization.

The generalized power-law rating curve model offers a new approach to statistical methods for rating curve development. By incorporating hydraulic theory through its stage-dependent exponent and modeling log-error variance as a function of stage, the model achieves flexibility in representing complex discharge-stage relationships. When

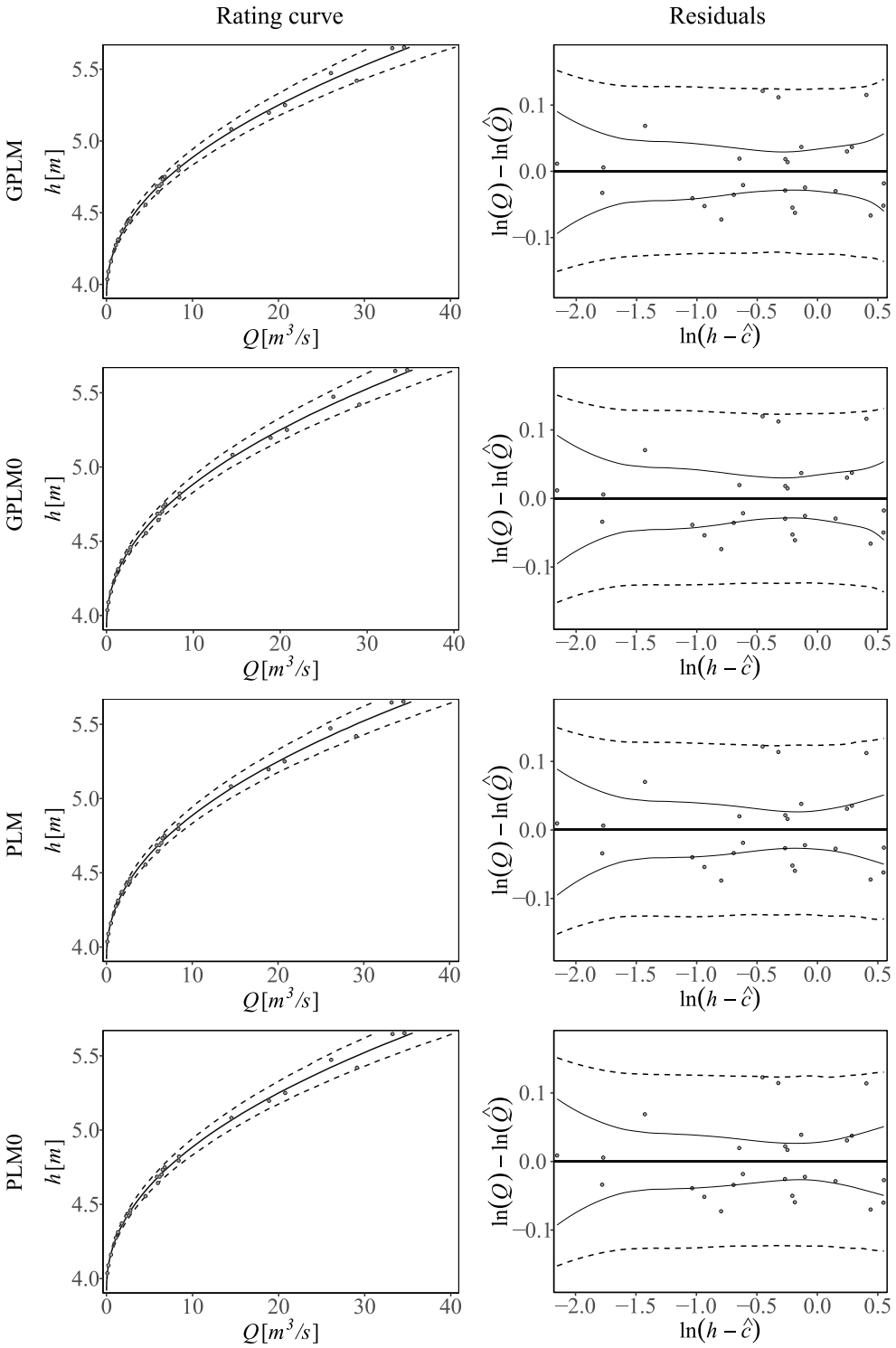

Fig. 13. The rating curves and corresponding residual plots of the four models in the generalized power-law class fit to the data from Skogsliden. Each row illustrates the results from a single model. The left column shows the estimated rating curves (solid line) and the 95% posterior predictive interval (dashed lines). The right column shows the residual plots on a logarithmic scale. The horizontal line corresponds to the posterior median of the logarithmically transformed rating curve, $\ln(\hat{Q})$. The solid curves around the horizontal line corresponds to the 95% posterior interval for the logarithmically transformed rating curve median. The dashed curves correspond to the 95% posterior predictive interval for the logarithmically transformed rating curve. The estimate for the c parameter, $\hat{c}$, is the posterior median of c.

the friction is independent of stage, the model maintains physical interpretability. However, when the friction factor varies with stage, the model parameters cannot be physically interpreted. Future research might extend the generalized power-law model to allow the friction parameter, a, to vary with stage.

The model's practical implications are particularly significant for global hydrology. Outside of a few high-income countries, like the United States, most countries lack the resources to maintain dense stream gauge networks with frequent field measurements necessary to constrain when ratings change. For these countries, the generalized

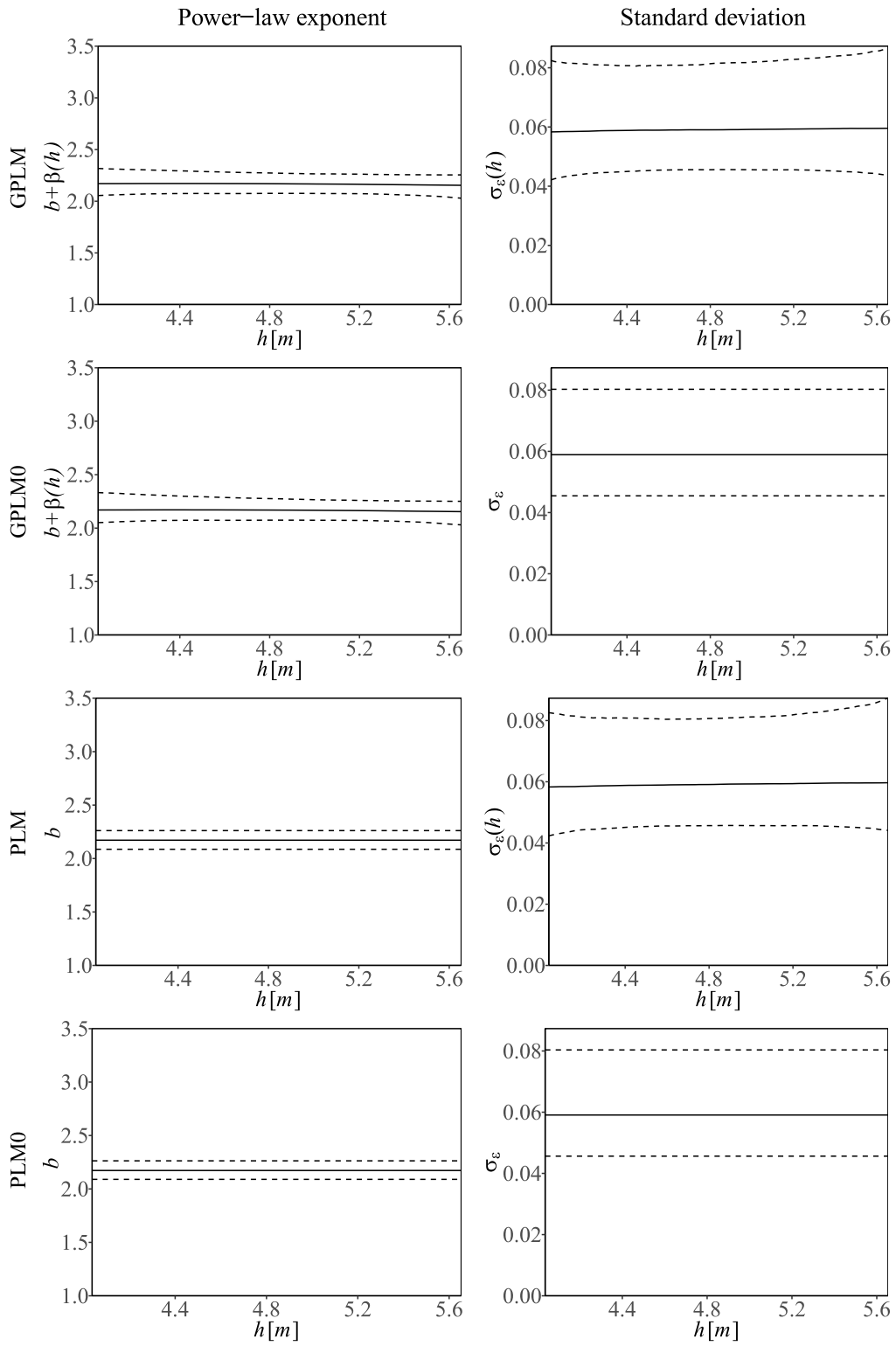

Fig. 14. The estimated power-law exponents and the log-error standard deviations of the four models in the generalized power-law class fit to the data from Skogsliden. Each row illustrates the results from a single model. The left column shows the posterior distributions of the power-law exponents, summarized by the median (solid line) and 95% credible interval (dashed lines). The right column shows the posterior distributions of the log-error standard deviations, also summarized by the median (solid line) and 95% credible interval (dashed lines).

power-law model could provide a valuable alternative to current rating development practices. In countries that make frequent measurements and apply "shifts" when channel conditions change, the model could serve as a basis for developing base ratings.

The comparison of GPLM and GPLM0 reveals important practical considerations. Theoretically, GPLM should be preferred because errors are never perfectly homoscedastic. However, when observations are limited, GPLM0 may provide a more robust fit due to its simpler

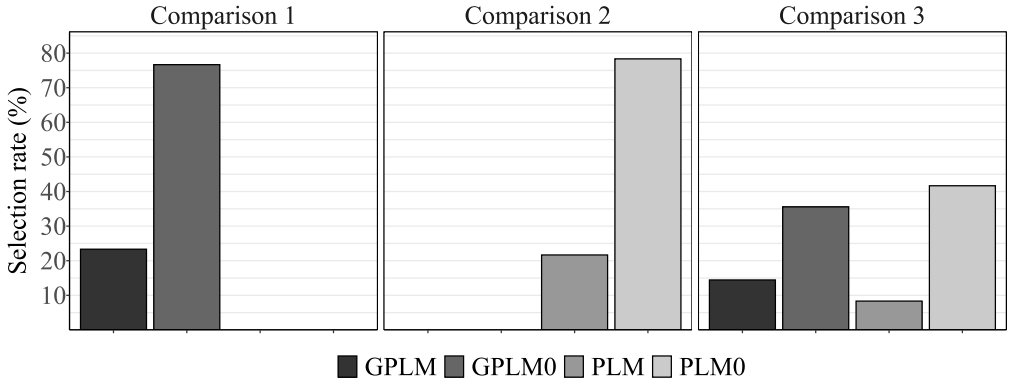

Fig. 15. Bar chart illustrating the model-selection rates for each model, broken down by comparison. The selection rate is the total number of times a model was selected divided by the total number of datasets (180).

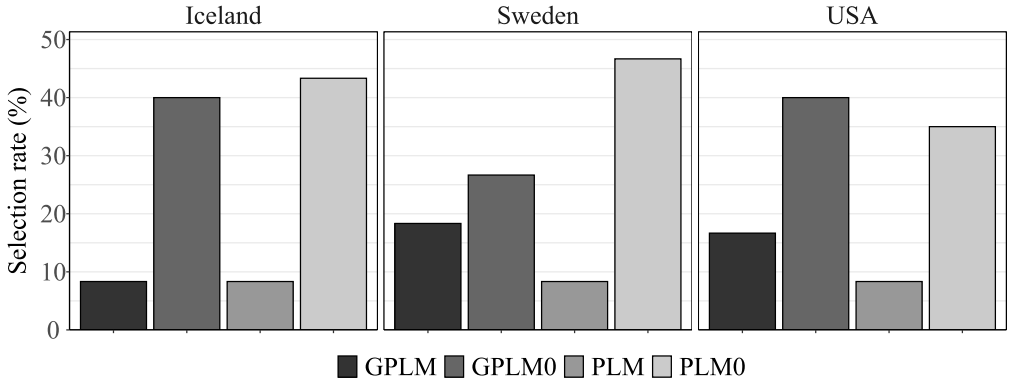

Fig. 16. Bar chart illustrating the model-selection rates for each model, broken down by country. The selection rate is the number of times a model was selected divided by the total number of datasets from that country.

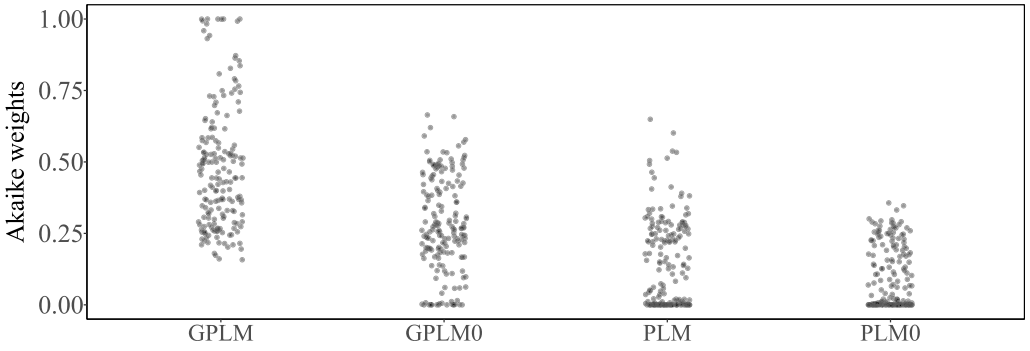

Fig. 17. A jitter plot showing the distribution of the WAIC-based Akaike weights for each model and dataset.

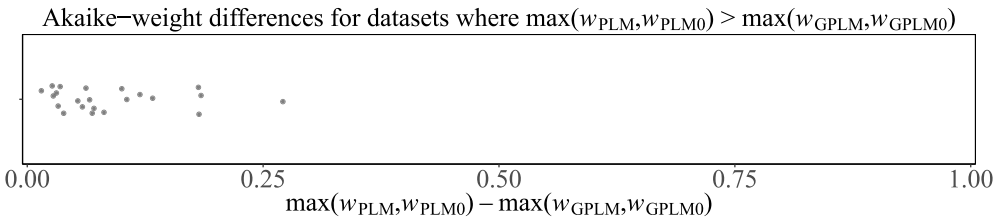

Fig. 18. A jitter plot showing the difference of the highest Akaike weight of PLM and PLM0 on the one hand (i.e., w_{PLM} and w_{PLM0}), and GPLM and GPLM0 on the other (i.e., w_{GPLM} and w_{GPLM0}), for the few datasets where the WAIC of neither GPLM nor GPLM0 where lower than the WAIC of either PLM0 or PLM.

Table 4
The number of times each model was selected as the appropriate model in the third and final comparison, broken down by country. The selection rate is the number of times a model was selected divided by the total number of datasets from that country.

Country	Model	Selected (n)	Selection rate (%)
Iceland	GPLM	5	8.3
	GPLM0	24	40.0
	PLM	5	8.3
	PLM0	26	43.3
Sweden	GPLM	11	18.3
	GPLM0	16	26.7
	PLM	5	8.3
	PLM0	28	46.7
USA	GPLM	10	16.7
	GPLM0	24	40.0
	PLM	5	8.3
	PLM0	21	35.0

parameterization, as GPLM can effectively reduce to GPLM0 when evidence for stage-varying log-error variance is weak.

While our evaluation indicates the model is ready for operational use, several areas for improvement remain. The model does not currently incorporate flow-measurement error, which can be substantial, particularly for high flows estimated by indirect techniques. The model could also be extended to account for time-varying behavior, such as hysteresis or transient shifts in channel conditions, possibly by incorporating shifts as a time-dependent Gaussian process. However, many stream gauges globally lack sufficient data to constrain such additions, and in these cases, the current formulation may already offer improvement over existing practices.

7. Conclusions

This paper evaluates the applicability of the generalized power-law rating curve model across 180 stream gauges in diverse geographic settings. The statistical model, incorporating a stage-dependent exponent and log-error variance into the classical power-law model, demonstrated computational robustness across all applications.

The comparative analysis showed that the generalized model proved competitive with the segmented power-law model in terms of predictive performance (as measured by WAIC), especially when no prior information on changepoints was available. Extrapolations with the generalized model, based solely on observed discharge-stage data, showed promising results, although incorporating channel characteristics into such predictions requires further research.

Our WAIC-based model-selection algorithm selected the generalized models (GPLM and GPLM0) as the most parsimonious models for half of the datasets. The classical power-law models (PLM and PLM0) were selected as the most parsimonious for the other half. However, the Akaike weights (based on WAIC) indicated that GPLM and GPLM0 remained competitive options across all datasets.

CRediT authorship contribution statement

Rafael Daníel Vias: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Birgir Hrafnkelsson:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Timothy O. Hodson:** Writing – review & editing, Validation, Software, Methodology, Conceptualization. **Sölvi Rögnvaldsson:** Writing – review & editing, Software. **Axel Örn Jansson:** Software. **Sigurdur M. Gardarsson:** Writing – review & editing, Resources, Project administration, Funding acquisition, Conceptualization.

Software and data availability

The `bdrc` software package, developed by Hrafnkelsson et al. (2023a) for the R programming language (R Core Team, 2023), was used to fit the models from the generalized power-law class as outlined in Section 2.2.1. The package is readily available on the *Comprehensive R Archive Network* (CRAN) at https://CRAN.R-project.org/package=bdrc. The package uses the statistical methods of Hrafnkelsson et al. (2021) and offers functionality for implementing the model-selection method presented in Section 4. For additional information, those interested can visit the `bdrc` package's homepage at https://sor16.github.io/bdrc/ and explore its GitHub repository at https://github.com/sor16/bdrc.

The segmented rating models were fit using the `ratingcurve` Python package (Hodson et al., 2024) that relies on the probabilistic programming library PyMC and uses the parameterization mentioned earlier. Recent versions of the software are available from the Python Package Index (PyPI) and `conda-forge`, and the latest source code and documentation are available at https://github.com/thodson-usgs/ratingcurve.

All 180 datasets used in this study are publicly available. The U.S. data were obtained from the USGS database, and permission has been granted to publish the Icelandic (Þórarinsson, 2024) and Swedish (Wennerberg, 2024) datasets. The complete collection of discharge-stage measurements is housed in a GitHub repository at https://github.com/RafaelVias/discharge-stage-data.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Grammarly in order to improve the language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rafael Daniel Vias reports financial support was provided by The Icelandic Student Innovation Fund of the Icelandic Centre for Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to express their gratitude to the Swedish Hydrological and Meteorological Institute, the Icelandic Met Office, and the U.S. Geological Survey for providing the datasets used in this paper. The authors would also like to thank the Icelandic Student Innovation Fund of the Icelandic Centre for Research, the Science Institute of the University of Iceland, and the University of Iceland Research Fund for their support. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Appendix A

A.1. Bayesian inference

The models described in Hrafnkelsson et al. (2021) are set up within a Bayesian hierarchical modeling framework. A joint density is defined over the data and all unknown parameters in the model. The joint density can be written as $p(\boldsymbol{y}, \boldsymbol{\psi})$, where $\boldsymbol{y} = (\ln(Q_1), \ldots, \ln(Q_n))^{\mathsf{T}}$ is

the vector of log-transformed discharge observations and $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\phi})$ is the parameter vector containing the latent and hyperparameters, respectively.

By Bayes' theorem, the posterior density can be presented as

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}|\boldsymbol{\phi}) p(\boldsymbol{\phi}) \ ,$$

where $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi})$ is the likelihood; $p(\boldsymbol{\theta}|\boldsymbol{\phi})$ is the prior density of the latent parameters, conditional on the hyperparameters; and $p(\boldsymbol{\phi})$ is the prior density of the hyperparameters. Hrafnkelsson et al. (2021) made a lognormal assumption at the response level, a normal assumption at the latent level, and specified various prior densities for the hyperparameters.

Hrafnkelsson et al. (2021) proposed an efficient Markov chain Monte Carlo scheme to sample from the posterior distribution of the rating curve models. The sampling scheme, which is motivated by the work of Knorr-Held and Rue (2002), proposes the latent and hyperparameter jointly when sampling from the posterior density.

Once the posterior samples, $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(L)}$, have been drawn from the posterior density, where L is the total number of samples, the samples can be used to compute summary statistics for the model parameters and to construct the estimated rating curve with credible intervals. In addition, the posterior predictive density can be estimated from the posterior samples and used to predict unobserved discharge values corresponding to unobserved stage values. For a new observation $\tilde{y}$ corresponding to an unobserved stage value $\tilde{x}$, given the discharge observations $\boldsymbol{y}$ and the observed stages, $\boldsymbol{x} = (h_1, \ldots, h_n)^\top$, the estimated posterior predictive density is given by

$$p(\tilde{y}|\tilde{x}, \boldsymbol{x}, \boldsymbol{y}) = \frac{1}{L} \sum_{l=1}^{L} p(\tilde{y}|\tilde{x}, \boldsymbol{\psi}^{(l)}) \ ,$$

where $p(\tilde{y}|\tilde{x}, \boldsymbol{\psi})$ is the Gaussian response density for the new observation, and $\boldsymbol{\psi}^{(l)}$ is the lth posterior sample. Summing over the posterior samples incorporates the posterior uncertainty into the predictive density. The posterior predictive density can then be used to assess the quality of the model since approximately $\alpha\%$ of the observations should fall within the $\alpha\%$ interpercentile range of the posterior predictive density.

A.2. WAIC - Measure of predictive accuracy

The power-law tournament uses the WAIC to estimate and compare the expected out-of-sample prediction error of the models (Watanabe, 2010; Gelman et al., 2013). The WAIC, which quantifies the trade-off between fit and complexity, is founded in information theory and based on two quantities: the *log pointwise predictive density*, lppd, a goodness-of-fit measure, given by

$$\text{lppd} = \sum_{i=1}^{n} \ln \int p(y_i|\boldsymbol{\psi}) p(\boldsymbol{\psi}|\boldsymbol{y}) d\boldsymbol{\psi},$$

and a penalization term, p_{waic}, referred to as the *effective number of parameters*, given by

$$p_{\text{waic}} = \sum_{i=1}^{n} \text{var}_{\boldsymbol{\psi}|\boldsymbol{y}}(\ln p(y_i|\boldsymbol{\psi})),$$

where $\text{var}_{\boldsymbol{\psi}|\boldsymbol{y}}(\cdot)$ is the variance where the expected values are taken in terms of the posterior distribution.

When WAIC is computed, samples $\boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(L)}$, from the posterior distribution, $p(\boldsymbol{\psi}|\boldsymbol{y})$, are used to evaluate lppd and p_{waic}. The *computed log pointwise predictive density*, $\widehat{\text{lppd}}$, is given by

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \ln \left(\frac{1}{L} \sum_{l=1}^{L} p(y_i|\boldsymbol{\psi}^{(l)}) \right),$$

where $\boldsymbol{\psi} = (\boldsymbol{\theta}, \boldsymbol{\phi})$ is the parameter vector containing the latent and hyperparameters, respectively; $p(y_i|\boldsymbol{\psi})$ is the likelihood of the ith observation; and $\boldsymbol{\psi}^{(l)}$ is the lth sample from the posterior distribution. The *estimated effective number of parameters*, $\hat{p}_{\text{waic}}$, is calculated as

$$\hat{p}_{\text{waic}} = \sum_{i=1}^{n} \left(\frac{1}{L-1} \sum_{l=1}^{L} \left(\ln p(y_i|\boldsymbol{\psi}^{(l)}) - \bar{p}_i \right)^2 \right),$$

where $\bar{p}_i$ is the sample average of $\ln p(y_i|\boldsymbol{\psi}^{(l)})$ over the posterior samples. Usually, the estimated effective number of parameters is smaller than the actual number of parameters in the model. However, if the prior constraints on the parameter space lend little to no information to the posterior distribution, and all the posterior information comes from the data, then the effective number of parameters will be close to the true number of parameters. Transformed onto the deviance scale, WAIC is defined as

$$\text{WAIC} = -2 \ \widehat{\text{elppd}}_{\text{waic}} \ ,$$

where $\widehat{\text{elppd}}_{\text{waic}} = \widehat{\text{lppd}} - \hat{p}_{\text{waic}}$ is the *estimated expected log pointwise predictive density*.

When comparing two models, we are only interested in their WAIC difference. Here, the WAIC difference is defined as $\Delta_{\text{waic}} = \text{WAIC}_B - \text{WAIC}_A$, where WAIC_A and WAIC_B are the WAIC estimates for the more and less complex models, respectively. Therefore, because a smaller WAIC is indicative of better predictive accuracy, a positive Δ_{waic} favors the more complex model.

Having fitted two models, M_A and M_B, to n observations, the standard error of Δ_{waic} can also be computed. We have already defined $\text{WAIC} = \sum_{i=1}^{n} \text{WAIC}_i$, where

$$\begin{aligned} \text{WAIC}_i &= -2 \ \widehat{\text{elppd}}_{\text{waic},\, i} \\ &= -2 \left(\ln \left(\frac{1}{L} \sum_{l=1}^{L} p(y_i|\boldsymbol{\psi}^{(l)}) \right) - \frac{1}{L-1} \sum_{l=1}^{L} \left(\ln p(y_i|\boldsymbol{\psi}^{(l)}) - \bar{p}_i \right)^2 \right). \end{aligned}$$

Then the standard error of Δ_{waic} is calculated as $\sqrt{n}$ times the standard deviation of the WAIC differences for each observation, or

$$\text{se}(\Delta_{\text{waic}}) = \sqrt{n V_{i=1}^{n} \left(\text{WAIC}_i^B - \text{WAIC}_i^A \right)} \ .$$

Here, $\sqrt{V_{i=1}^{n}(d_i)}$ is the standard deviation of the d_i's, where $d_i := \text{WAIC}_i^B - \text{WAIC}_i^A$. That is, $V_{i=1}^{n}(d_i) = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$.

A.3. Exploring alternative selection criteria for the power-law tournament

Rules of thumb, similar to those described by Burnham and Anderson (2002) for AIC, were presented by Kass and Raftery (1995) for a model comparison using the Bayes factor. In this section, we will explore the relationship between the posterior model probabilities (computed with the Bayes factor, estimated with the harmonic mean estimator) and Δ_{waic} in the context of the power-law tournament. More specifically, we draw a connection between Δ_{waic} and the posterior model probabilities and explore whether model-selection criteria based on the rules of thumbs presented by Kass and Raftery (1995) yield a similar threshold value for the model-selection criteria presented in Section 4.2.

A.3.1. Posterior model probability computed with Bayes factor

The posterior model probability (computed with Bayes factor) of a model M_A when compared to another model M_B, assuming a priori that the model probabilities are $\Pr(M_A)$ and $\Pr(M_A)$ (see Jeffreys, 1961; Kass and Raftery, 1995), is given by

$$\begin{aligned} \Pr(M_A|\boldsymbol{y}) &= \frac{\Pr(M_A) \int p_A(\boldsymbol{y}|\boldsymbol{\psi}_A) p(\boldsymbol{\psi}_A) d\boldsymbol{\psi}_A}{\sum_{s \in \{A,B\}} \Pr(M_s) \int p_s(\boldsymbol{y}|\boldsymbol{\psi}_s) p(\boldsymbol{\psi}_s) d\boldsymbol{\psi}_s} \\ &= \left(1 + \frac{\Pr(M_B)}{\Pr(M_A)} \times \frac{1}{B_{AB}} \right)^{-1} \ , \end{aligned}$$

where B_{AB} is the Bayes factor (see Kass and Raftery, 1995) for models M_A and M_B given by

$$B_{AB} = \frac{\int p_A(\boldsymbol{y}|\boldsymbol{\psi}_A) p(\boldsymbol{\psi}_A) d\boldsymbol{\psi}_A}{\int p_B(\boldsymbol{y}|\boldsymbol{\psi}_B) p(\boldsymbol{\psi}_B) d\boldsymbol{\psi}_B} \ .$$

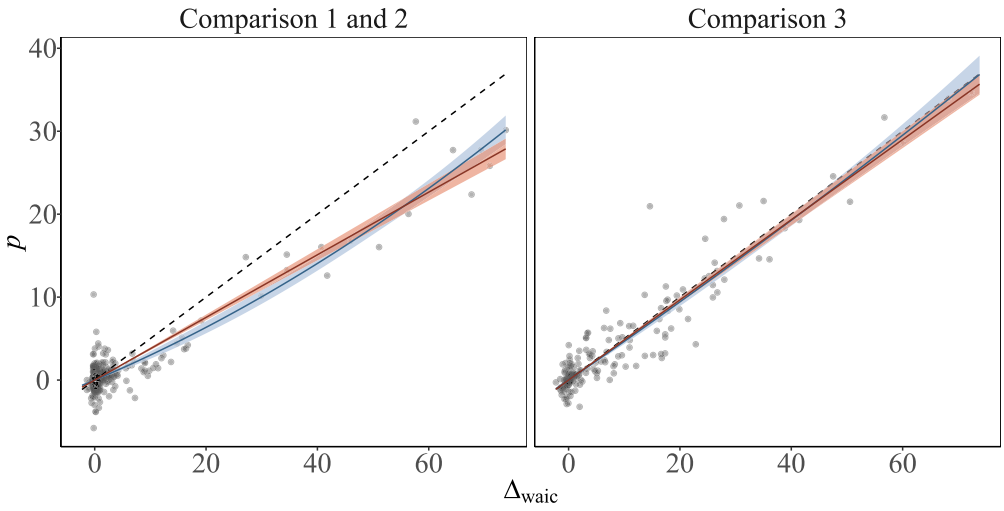

Fig. 19. A scatterplot depicting $p := \text{logit}(\Pr(M_A|\boldsymbol{y}))$ and Δ_{waic} for all 180 datasets presented in Section 3, broken down by model comparison (as described in Section 4.1). The dashed line shows the theoretical linear relationship in Eq. (9), whereas, the red lines and blue curves (and 95% confidence bands) show the results from the first and second order polynomial regression, respectively (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this article.).

Table 5
The Δ_{waic} values corresponding to a posterior model probability of 0.75 obtained from the theoretical linear relationship, and the first ([1]) and second ([2]) order polynomial regression models for comparisons one and two, and three. The values from the regression models were bootstrapped and the mean (SD) is shown rounded to one decimal place.

Comp.	Theor.	Regr.[1]	Regr.[2]
1 and 2	2.2	2.9 (0.1)	3.9 (0.4)
3	2.2	2.3 (0.1)	2.4 (0.2)

Because the Bayes factor can be hard to calculate, it is often estimated by approximating the integrals $\int p_s(\boldsymbol{y}|\boldsymbol{\psi}_s)p(\boldsymbol{\psi}_s)d\boldsymbol{\psi}_s$, $s \in \{A, B\}$, with the harmonic mean of the likelihood values, that is

$$\left\{ \sum_{l=1}^{L} \frac{1}{p_s(\boldsymbol{y}|\boldsymbol{\psi}_s^{(l)})} \right\}^{-1},$$

where $\boldsymbol{\psi}_s^{(l)}$ is the lth posterior sample of $\boldsymbol{\psi}_s$ in the model M_s; see Newton and Raftery (1994) for details.

A.3.2. Connection with Δ_{waic}

Kass and Raftery (1995) argued that, for a given dataset, it is reasonable to assume that a model is substantially more appropriate than another model if its posterior model probability is 0.75 or greater.

The quantity B_{AB} is somewhat analogous to $\exp(-\Delta_{\text{waic}}/2)$ (Burnham and Anderson, 2002). Thus, if $\Pr(M_A) = \Pr(M_B) = 1/2$, then this implies that

$$\text{logit}(\Pr(M_A|\boldsymbol{y})) \approx \Delta_{\text{waic}}/2, \tag{9}$$

and $\Pr(M_A|\boldsymbol{y}) = 0.75$ corresponds to $\Delta_{\text{waic}} \approx 2.2$. The theoretical linear relationship in (9) is depicted in Fig. 19, along the results of a first- and second-order polynomial regression performed on the $\text{logit}(\Pr(M_A|\boldsymbol{y}))$ and Δ_{waic} values for all 180 datasets presented in Section 3 and each model comparison in the power-law tournament (as described in Section 4.1). In all cases, the regression lines were made to go through the origin. The observed values in the first two comparisons have a similar distribution and were, therefore, joined together. The theoretical linear relationship in (9) holds well for the third comparison, whereas it gives a poor fit for the first two.

Table 5 shows the Δ_{waic} values corresponding to a posterior model probability of 0.75, transformed using the theoretical linear relationship in (9) and the first- and second-order polynomial regression models. Let Δ^*_{waic} be the Δ_{waic} value corresponding to a posterior model probability of 0.75. As mentioned above, the relationship in (9) returns a value of $\Delta^*_{\text{waic}} \approx 2.2$ for all three comparisons. For the third comparison, the regression models are almost identical and similar to the theoretical linear relationship, with $\Delta^*_{\text{waic}} \approx 2.4$. Lastly, the Δ^*_{waic} values attained with the polynomial regression models for the first two comparisons are somewhat larger. The better fitting model, the second-order polynomial regression curve, gives a value of $\Delta^*_{\text{waic}} \approx 4.5$, double that of the theoretical linear relationship.

The results from the regression analysis should be taken with some skepticism. The posterior model probabilities depicted in Fig. 19 are calculated using the harmonic mean estimator. In some instances, the theoretical variance of the harmonic mean estimator is infinite (Newton and Raftery, 1994). Even when this variance is finite, there is a potential for substantial enlargement since a small subset of the posterior samples can correspond to a very small likelihood, thereby exerting a substantial influence on the harmonic mean estimate. Moreover, the harmonic mean estimator can be biased despite being asymptotically unbiased (Calderhead and Girolami, 2009; Raftery et al., 2007).

Hrafnkelsson et al. (2021) compared the GPLM and PLM models with the posterior model probabilities by approximating the Bayes factor with the harmonic mean estimator, as we have done here. They found that the Bayes factor approximations would vary considerably if the models had similar predictive accuracy for the data at hand.

For this reason, the DIC was also calculated and used to compare the models. In addition, they presented the results of a simulation study that assessed the variability of these model-comparison statistics, confirming a far superior stability of the DIC estimates. These findings motivated the use of WAIC for the power-law tournament, which, as the name implies, is more widely applicable than DIC. It also produces stable estimates of the expected out-of-sample prediction error, but unlike DIC, does not assume approximate normality of the posterior distribution (Spiegelhalter et al., 2002; Watanabe, 2010).

By using the theoretical relationship in (9) to transform the posterior model probability threshold value of 0.75 to a corresponding value of Δ_{waic}, the resulting threshold value is very similar to the one chosen for the power-law tournament in Section 4.2. The regression analysis suggests that the Δ_{waic} threshold value for first two comparisons,

corresponding to a posterior model probability of 0.75, is somewhat larger.

Appendix B. Supplementary data

A catalog of the 180 datasets is provided as supplementary material. Each dataset is assigned a data page, four model pages, and a model comparison page. The data pages show the paired discharge-stage observations, a time series of the discharge observations, and selected information about the datasets. The four model pages present the rating curves fitted with GPLM, GPLM0, PLM, and PLM0, respectively, as well as posterior estimates of the model parameters and convergence diagnostics for the Markov chains, specifically $\hat{R}$ and the *effective number of samples* (for details, see, e.g., Gelman et al. (2013)). The fifth and final page for each dataset shows the results of the power-law tournament, along with the winning model. The winning model is then upward extrapolated as described in Section 5.2.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jhydrol.2024.132537.

Data availability

We have made all the data publicly accessible. The link is in the manuscript in the "Software and data availability" section. The link is: https://github.com/RafaelVias/discharge-stage-data.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723. http://dx.doi.org/10.1109/TAC.1974.1100705.

Burnham, K.P., Anderson, D.R. (Eds.), 2002. Information and likelihood theory: A basis for model selection and inference. In: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer New York, New York, NY, pp. 49–97. http://dx.doi.org/10.1007/978-0-387-22456-5_2.

Calderhead, B., Girolami, M., 2009. Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Statist. Data Anal. 53 (12), 4028–4045. http://dx.doi.org/10.1016/j.csda.2009.07.025.

Chow, V., 1959. Open-Channel Hydraulics. McGraw-Hill, New York.

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 <scp>UK</scp> gauging stations. Water Resour. Res. 51 (7), 5531–5546. http://dx.doi.org/10.1002/2014wr016532.

Fuglstad, G.A., Simpson, D., Lindgren, F., Rue, H., 2019. Constructing priors that penalize the complexity of Gaussian random fields. J. Amer. Statist. Assoc. 114 (525), 445–452. http://dx.doi.org/10.1080/01621459.2017.1415907.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2013. Bayesian Data Analysis, third ed. Chapman & Hall/CRC, http://dx.doi.org/10.1201/b16018.

Gioia, G., Bombardelli, F.A., 2001. Scaling and similarity in rough channel flows. Phys. Rev. Lett. 88 (1), http://dx.doi.org/10.1103/physrevlett.88.014501.

Hodson, T.O., Doore, K.J., Kenney, T.A., Over, T.M., Yeheyis, M.B., 2024. Ratingcurve: A python package for fitting streamflow rating curves. Hydrology 11 (2), http://dx.doi.org/10.3390/hydrology11020014.

Hrafnkelsson, B., Bakka, H., 2023. Bayesian latent Gaussian models. In: Hrafnkelsson, B. (Ed.), Statistical Modeling using Bayesian Latent Gaussian Models : With Applications in Geophysics and Environmental Sciences. Springer International Publishing, Cham, pp. 1–80. http://dx.doi.org/10.1007/978-3-031-39791-2_1.

Hrafnkelsson, B., Ingimarsson, K., Gardarsson, S., Snorrason, A., 2012. Modeling discharge rating curves with Bayesian B-splines. Stoch. Environ. Res. Risk Assess. 26 (1), 1–20. http://dx.doi.org/10.1007/s00477-011-0526-0.

Hrafnkelsson, B., Rögnvaldsson, S., Jansson, A.Ö., Vias, R.D., 2023a. Bdrc: Bayesian discharge rating curves, r package, version 1.1.0. http://dx.doi.org/10.32614/CRAN.package.bdrc.

Hrafnkelsson, B., Sigurdarson, H., Rögnvaldsson, S., Jansson, A.Ö., Vias, R.D., Gardarsson, S.M., 2021. Generalization of the power-law rating curve using hydrodynamic theory and Bayesian hierarchical modeling. Environmetrics 33 (2), http://dx.doi.org/10.1002/env.2711.

Hrafnkelsson, B., Vias, R.D., Rögnvaldsson, S., Jansson, A.Ö., Gardarsson, S.M., 2023b. Bayesian discharge rating curves based on the generalized power law. In: Hrafnkelsson, B. (Ed.), Statistical Modeling using Bayesian Latent Gaussian Models : With Applications in Geophysics and Environmental Sciences. Springer International Publishing, Cham, pp. 109–127. http://dx.doi.org/10.1007/978-3-031-39791-2_3.

ISO 18320:2020, 2020. Hydrometry–Measurement of Liquid Flow in Open Channels—Determination of the Stage-Discharge Relationship. International Organization for Standardization.

Jeffreys, H., 1961. Theory of Probability, third ed. Oxford University Press, Oxford.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90 (430), 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572.

Kiang, J.E., Gazoorian, C., McMillan, H., Coxon, G., Le Coz, J., Westerberg, I.K., Belleville, A., Sevrez, D., Sikorska, A.E., Petersen-Øverleir, A., Reitan, T., Freer, J., Renard, B., Mansanarez, V., Mason, R., 2018. A comparison of methods for streamflow uncertainty estimation. Water Resour. Res. 54 (10), 7149–7176. http://dx.doi.org/10.1029/2018wr022708.

Knorr-Held, L., Rue, H., 2002. On block updating in Markov random field models for disease mapping. Scand. J. Stat. 29 (4), 597–614. http://dx.doi.org/10.1111/1467-9469.00308.

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. J. Hydrol. 509, 573–587. http://dx.doi.org/10.1016/j.jhydrol.2013.11.016.

Matérn, B., 1986. Spatial Variation. Springer New York, http://dx.doi.org/10.1007/978-1-4615-7892-5.

Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. Ser. B Stat. Methodol. 56 (1), 3–26. http://dx.doi.org/10.1111/j.2517-6161.1994.tb01956.x.

Þórarinsson, Ó., 2024. Icelandic meteorological office (IMO). Permission to publish the Icelandic data was granted in written communication on October 24, 2024..

Petersen-Øverleir, A., Reitan, T., 2005. Objective segmentation in compound rating curves. J. Hydrol. 311 (1), 188–201. http://dx.doi.org/10.1016/j.jhydrol.2005.01.016.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Raftery, A.E., Newton, M.A., Satagopan, J.M., Krivitsky, P.N., 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: Bayesian Statistics 8: Proceedings of the Eighth Valencia International Meeting June 2–6, 2006. Oxford University Press, http://dx.doi.org/10.1093/oso/9780199214655.003.0015.

Reitan, T., Petersen-Øverleir, A., 2006. Existence of the frequentistic estimate for power-law regression with a location parameter, with applications for making discharge rating curves. Stoch. Environ. Res. Risk Assess. 20 (6), 445–453. http://dx.doi.org/10.1007/s00477-006-0037-6.

Reitan, T., Petersen-Øverleir, A., 2008. Bayesian methods for estimating multi-segment discharge rating curves. Stoch. Environ. Res. Risk Assess. 23 (5), 627–642. http://dx.doi.org/10.1007/s00477-008-0248-0.

Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H., 2017. Penalising model component complexity: A principled, practical approach to constructing priors. Statist. Sci. 32 (1), http://dx.doi.org/10.1214/16-sts576.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (4), 583–639. http://dx.doi.org/10.1111/1467-9868.00353.

U.S. Geological Survey, 2024. National water information system data available on the World Wide Web (USGS water data for the nation). http://dx.doi.org/10.5066/F7P55KJN.

Vatanchi, S.M., Maghrebi, M.F., 2024a. Estimating streamflow by an innovative rating curve model based on hydraulic parameters. Environ. Earth Sci. 83 (9), http://dx.doi.org/10.1007/s12665-024-11493-6.

Vatanchi, S.M., Maghrebi, M.F., 2024b. Hysteresis-influenced stage-discharge rating curve based on isovel contours and jones formula. Stoch. Environ. Res. Risk Assess. 38 (7), 2829–2840. http://dx.doi.org/10.1007/s00477-024-02716-0.

Venetis, C., 1970. A note on the estimation of the parameters in logarithmic stage-discharge relationships with estimates of their error. Int. Assoc. Sci. Hydrol. Bull. 15 (2), 105–111. http://dx.doi.org/10.1080/02626667009493957.

Watanabe, S., 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11, 3571–3594.

Wennerberg, D., 2024. Swedish meteorological and hydrological institute (SMHI). Permission to publish the Swedish data was granted in written communication on October 22, 2024.

Wickert, A.D., Jones, J.C., Ng, G.-H.C., 2024. A double-manning approach to compute robust rating curves and hydraulic geometries. EGUsphere 2024, 1–26. http://dx.doi.org/10.5194/egusphere-2023-3118.