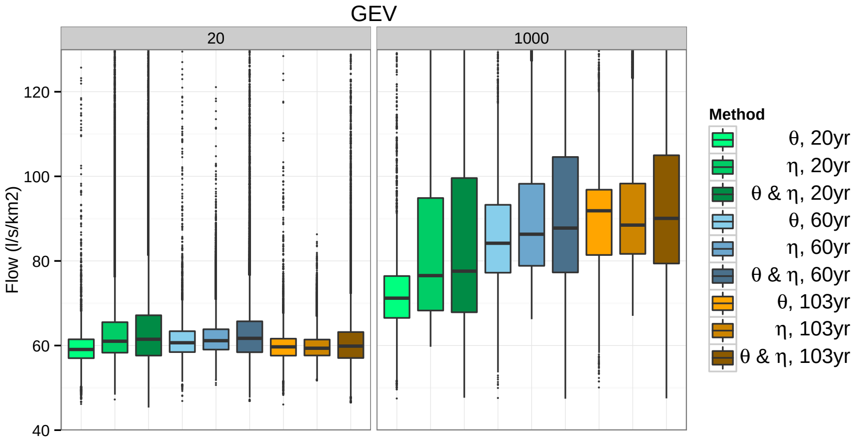
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Progagation of rating curve uncertainty in design flood estimation



Note no SAMBA/03/16

Authors
Gunnhildur Høgnadottir Steinbakk
Thordis L. Thorarinsdottir
Trond Reitan
Lena Schlichting
Sondre Hølleland
Kolbjørn Engeland

Date January 28, 2016

The authors

Gunnhildur Høgnadottir Steinbakk and Thordis L. Thorarinsdottir are working at the Norwegian Computing Center; Trond Reitan is working at the Centre for Ecological Synthesis at the Department of Biology in University of Oslo and at the Hydrological Modeling Section of the Norwegian Water Resources and Energy Directorate (NVE); Lena Schlichting and Kolbjørn Engeland are working at the Hydrological Modeling Section of NVE; Sondre Hølleland is a student at the Department of Mathematics, University of Bergen and had a summer job at Norwegian Computing Center in 2015

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Title	**Progagation of rating curve uncertainty in design flood estimation**

Authors **Gunnhildur Høgnadottir Steinbakk** <gunnhildur@nr.no>
Thordis L. Thorarinsdottir <thordis@nr.no>
Trond Reitan <trr@nve.no>
Lena Schlichting <lsc@nve.no>
Sondre Hølleland <Sondre.Holleland@student.uib.no>
Kolbjørn Engeland <koe@nve.no>

Date	January 28, 2016
Publication number	SAMBA/03/16

Abstract

Statistical flood frequency analysis is commonly performed based on a set of annual maximum discharge values which are derived from stage measurements via a stage-discharge rating curve model. However, design flood estimation techniques often ignore the uncertainty in the underlying rating curve model. Using data from seven gauging stations in Norway, we investigate the effect of curve and sample uncertainty on design flood estimation. Accounting for curve uncertainty may strongly influence the results if an extrapolation of the rating curve is necessary.

Keywords	
Target group	Anyone interested in flood frequency analysis
Availability	Open
Project	FlomQ
Project number	220662
Research field	Technology, industry and administration
Number of pages	30
© Copyright	Norwegian Computing Center

1 Introduction

The aim of flood frequency analysis is to estimate the streamflow quantiles, or design floods, for given exceedance probabilities or return periods, e.g. the 100-year flood. The design flood estimation forms the basis for hazard management related to flood risk and is a legal obligation when building infrastructure such as dams, bridges and roads close to water bodies. Flood inundation maps used for land use planning are also produced based on design flood estimates. The streamflow quantile estimates required for these applications might range from estimates of the mean annual flood to quantiles higher than the 1000-year flood.

The estimation of design floods is influenced by a range of uncertainties. Traditionally, three important sources of uncertainty are the uncertainty in the choice of a model or a distribution, the estimation uncertainty due to a small data sample, and the uncertainty in flood observations. Furthermore, the observation uncertainties contribute in a two-fold way since the observations are used both for estimating and evaluating the statistical model. In this paper, we investigate how such uncertainties influence the resulting design flood estimation with a focus on the data-related sources of uncertainty.

Applications commonly require design flood estimates for return periods longer than the sample size, resulting in estimates that are based on a large degree of extrapolation. This is one of the main contributors to a large sample uncertainty and estimates that are sensitive to the particular data used for the estimation, with large and highly skewed uncertainty bounds. To improve the robustness of design flood estimates, linear moment estimators (e.g. Hosking et al., 1985), penalized maximum likelihood (Coles and Dixon, 1999; Martins and Stedinger, 2000) and Bayesian model formulation with informative priors (e.g. Renard, 2011) have been proposed. The estimation uncertainty was originally assessed using asymptotic theory for the maximum likelihood under a normal approximation (e.g. Madsen et al., 1997; Rosbjerg and Madsen, 1995). However, resampling methods such as bootstrap or jackknife better account for the skewness of the uncertainty bounds (e.g. Engeland et al., 2005; Hall et al., 2004; Kyselý, 2008). More recently, Bayesian methods have been used to estimate uncertainty intervals as well as the predictive distribution of the flood quantiles (e.g. Lee and Kim, 2008; Renard et al., 2013, 2006).

Recently, estimation of uncertainties in streamflow data has received an increased attention in the literature, see Coxon et al. (2015), Di Baldassarre and Montanari (2009), Le Coz et al. (2014), Moyeed and Clarke (2005), Reitan and Petersen-Øverleir (2009, 2011) and Westerberg et al. (2011). Streamflow observations are derived from water level measurements by using a rating curve model to translate the measured water level to streamflow. Consequently, uncertainties in streamflow data stem from two sources. A random uncertainty relates to the accuracy of the water level measurements while uncertainty in the rating curve model translates into a systematic, or correlated, uncertainty in the output (Neppel et al., 2010). We will focus on this latter source of uncertainty.

In a natural river profile, the rating curve model is estimated using regression-like techniques based on simultaneous measurements of streamflows and water levels using a power law model with one or several profile segments or water level intervals. Important sources of uncertainty include random errors in the streamflow measurements and systematic errors related to the specification of the model, that is, the method for extrapolating beyond the observations as well as the inclusion of seasonality, temporal drift and step changes in the water level/discharge relationship (Coxon et al., 2015). In particular, the extrapolation uncertainty might be large since it is not possible to know whether the model should include a new segment beyond the observed values (Lang et al., 2010; Reitan and Petersen-Øverleir, 2009). For this reason, Le Coz et al. (2014) use hydraulic modeling to aid the extrapolation. Temporal changes can be caused by erosion or deposition of river sediments, or a seasonal effect due to vegetation growth in summer and/or ice in winter. In Norway, the effect of ice is a major modeling challenge and the streamflow data influenced by ice are systematically adjusted (ice-reduced). The changes in river profile is accounted for by using different curves for different periods or by estimation non-stationary rating curve models (Reitan and Petersen-Øverleir, 2011). Petersen-Øverleir et al. (2009) show that for 581 gauging stations in Norway, more than 93 % of the rating curves have a relative uncertainty of 10 % or larger for high flows.

In hydrological modeling there is an increasing attention to explicitly accounting for different sources of uncertainty. The effects of uncertainty in streamflow observations have been investigated in the context of precipitation-runoff modeling (e.g. McMillan et al., 2010) and in the estimation of hydrological indices and signatures (e.g. Clarke, 1999; Clarke et al., 2000; Westerberg and McMillan, 2015). For flood frequency analysis, the observational uncertainty is commonly estimated separately (Kuczera, 1996; Lang et al., 2010; Neppel et al., 2010) while one study performs an integrated analysis where the rating curve parameters and the flood frequency distribution are estimated jointly (Petersen-Øverleir and Reitan, 2009). This study concludes that the rating curve error has an important influence on the quality and the variance of the quantile estimates. Furthermore, the systematic errors caused by the rating curve model might have a larger influence on the resulting quantile estimates than random errors in water level observations (Lang et al., 2010; Neppel et al., 2010).

The results discussed above demonstrate that it is important to account for rating curve uncertainty in flood frequency estimation, and we believe there is a need to better understand both the marginal and the joint effects of sample and rating curve uncertainty. Furthermore, it is essential to assess the importance of critical data when estimating flood quantiles, i.e. to assess the added value of streamflow measurements during high floods. Such measurements may reduce the extrapolation uncertainty in the rating curve model and, subsequently, reduce the uncertainty of design flood estimates. To gain a better insight into the effect of sample and rating curve uncertainties in flood frequency estimation, we aim to answer the following research questions:

(i) How large is the contribution of the sample uncertainty and the rating curve uncertainty, individually and combined, when estimating design floods?

(ii) What is the effect of length of the annual maximum series on the design flood estimation uncertainty?

(iii) What is the added value of large streamflow measurements at the high end of the rating curve and in the annual maximum data series?

The remainder of the paper is organized as follows. A description of the rating curve model and the techniques used for the flood frequency estimation is given in Section 2. In the next Section 3, data and results for seven gauged catchments in Norway are presented. The paper closes with a discussion in Section 4 and details of the estimation algorithms are provided in the appendix.

2 Methods

We employ Bayesian inference techniques to investigate the uncertainty associated with parameter estimation in statistical models for stage-discharge rating curves and the subsequent flood frequency analysis. The general framework for such an analysis may be described as follows. Denote the available data set by $\mathcal{W} = (w_1, \ldots, w_n)$. The data is assumed to be independent and identically distributed such that the likelihood, or the sampling density, is given by

$$p(\mathcal{W} \,|\, \boldsymbol{\gamma}) = \prod_{i=1}^{n} p(w_i \,|\, \boldsymbol{\gamma}),$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$ denotes the vector of unknown parameters.

The parameter vector $\boldsymbol{\gamma}$ is associated with a prior distribution $p(\boldsymbol{\gamma})$ and, subsequently, we calculate the posterior distribution of the parameters given the data,

$$p(\boldsymbol{\gamma} \,|\, \mathcal{W}) = \frac{p(\mathcal{W} \,|\, \boldsymbol{\gamma}) p(\boldsymbol{\gamma})}{\int p(\mathcal{W} \,|\, \boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\gamma}}.$$

The posterior distribution is commonly not available in a closed form which requires the use of e.g. Markov chain Monte Carlo (MCMC) simulation techniques (Robert and Casella, 2004). The posterior distribution, either directly or as represented by a sample, may then be used to infer the parameter uncertainty given the statistical model and the data $\mathcal{W}$. Similarly, we may use the information provided by the posterior distribution to assess the uncertainty of the quantiles of the sampling distribution, or any other function of the parameters.

Below, we outline the specific uses of this general framework for the rating curve model of Reitan and Petersen-Øverleir (2009) and for flood frequency analysis using the two parameter Gumbel distribution and the three parameter generalized extreme value distribution.

2.1 Stage-discharge rating curves

Observed streamflow data are usually given by stage measurements which measure the water surface height at a gauging station. These data are then converted to discharge using a stage-discharge rating curve. It is common to assume a power law relationship between the stage h and the discharge Q, that is,

$$Q = \begin{cases} 0 & \text{if } h \leq h_0 \\ \exp(a)(h - h_0)^b & \text{if } h > h_0, \end{cases}$$

where h_0 is a cease-to-flow stage parameter and the parameters a and b determine the shape of the rating curve. A more general model assumes that the profile of the river consists of multiple segments with the shape of the rating curve changing for each segment. Reitan and Petersen-Øverleir (2009) consider a multi-segment model with log-normal measurement errors. That is,

$$\begin{aligned} \log(Q) = \sum_{j=1}^{k} & \mathbb{1}\{h_{s,j-1} \leq h < h_{s,j}\} \\ & \times (a_j + b_j \log(h - h_{0,j})) + \varepsilon, \end{aligned} \tag{1}$$

where (Q, h) is a stage-discharge measurement, $\mathbb{1}$ denotes the indicator function, $h_{s,j}$ determines the transition between segments j and $j+1$ for $j = 1, \ldots, k-1$ with $h_{s,0} = -\infty$, $h_{s,k} = \infty$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The parameters a_j for $j > 1$ are determined by the first segment a_1 and continuity from one segment to the next such that $a_1 + b_j \log(h_{s,j} - h_{0,j}) = a_{j+1} + b_{j+1} \log(h_{s,j} - h_{0,j+1})$.

Reitan and Petersen-Øverleir (2009) propose a Bayesian inference procedure for the model in (1) which has been implemented as a part of the national hydrological database system at the Norwegian Water Resources and Energy Directorate. Throughout, we refer to this framework as the Bayesian multi-segment software or the Bayesian multi-segment model. Denote by $\boldsymbol{\theta}$ the vector of regression parameters in (1), including the number of segments k,

$$\begin{aligned} \boldsymbol{\theta} = (&k, a_1, \ldots, a_k, b_1, \ldots, b_k, \\ &h_{0,1}, \ldots, h_{0,k}, h_{s,1}, \ldots, h_{s,k-1}). \end{aligned} \tag{2}$$

Given a set of training data $\mathcal{D}$ consisting of d pairs of stage and discharge measurements, $\mathcal{D} = \{(Q_1, h_1), \ldots, (Q_d, h_d)\}$, the Bayesian multi-segment software returns a sample $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ from the posterior distribution of the regression parameters, $p(\boldsymbol{\theta}|\mathcal{D})$. The multi-segment model described here is used for inference within the stage-discharge measurement set, while the process model in Reitan and Petersen-Øverleir (2009) is used for handling the extra uncertainty regarding new segments outside the measurement set. Note that the number of segments, k, is sampled together with the rest of the regression parameters. This sample, together with a set of annual maximum stage measurements $\mathcal{H} = \{h_1, \ldots, h_T\}$, may then be used to infer a sample of annual maximum discharge values by setting

$$Q_t^{(i)} = \exp\left(\sum_{j=1}^{k} \mathbb{1}\{h_{s,j-1}^{(i)} \leq h_t < h_{s,j}^{(i)}\} \right. \\ \left. \left(a_j^{(i)} + b_j^{(i)} \log(h_t - h_{s,j}^{(i)})\right) \right) \tag{3}$$

for $t = 1, \dots, T$ and $i = 1, \dots, n$.

In addition, the Bayesian multi-segment software issues a single set of values $\hat{\boldsymbol{Q}} = (\hat{Q}_1, \dots, \hat{Q}_T)$ which are considered the best estimate for the actual unobserved discharge. These values are based on a set of parameter estimates $\hat{\boldsymbol{\theta}}$ given as the mode of the multivariate posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$.

2.2 Modeling annual maximum discharge

The generalized extreme value (GEV) distribution is commonly used to model block maxima such as the annual maxima, see e.g. Coles (2001). Denote by y the maximum yearly discharge Q normalized by the catchment area A, $y = Q/A$. We assume that y follows a GEV distribution, $y \sim \text{GEV}(\mu, \kappa, \xi)$, with density

$$p(y \,|\, \boldsymbol{\eta} = (\mu, \kappa, \xi)) = \kappa\, \alpha(y)^{(\xi+1)} \exp\left(-\alpha(y)\right) \tag{4}$$

where

$$\alpha(y) = \begin{cases} (1 + \xi\kappa(y - \mu))^{-1/\xi} & \text{if } \xi \neq 0 \\ \exp\left(-\kappa(y - \mu)\right) & \text{if } \xi = 0 \end{cases} \tag{5}$$

and $(1 + \xi\kappa(y - \mu)) > 0$ for $\xi \neq 0$. The GEV distribution has three parameters that in our parameterization are location $\mu \in \mathbb{R}$, inverse scale $\kappa \in \mathbb{R}_+$, and shape $\xi \in \mathbb{R}$. The distribution is often parameterized using the scale $\sigma = 1/\kappa$ rather than the inverse scale (e.g. Coles, 2001). However, the current parameterization is common in Bayesian contexts (Dyrrdal et al., 2015; Rue et al., 2009), and is chosen because derivations related to posterior densities are considerably easier in this representation.

The tail behavior of the GEV distribution is driven by the value of the shape parameter ξ and generally falls in three classes: the Fréchet type ($\xi > 0$) has a heavy upper tail, the Gumbel tail ($\xi \to 0$) is characterized by a light upper tail, and the Weibull type ($\xi < 0$) is bounded from above. The shape parameter thus provides vital information on the statistical properties of the discharge and is, concurrently, difficult to estimate because of the involved parametric form of the density in (4) as a function of ξ. For this reason, the guidelines for flood estimation in Norway (Midttømme et al., 2011) recommend the use of the three parameter GEV distribution for local flood frequency analysis only if 50 or more observations are available at the location. For 30-50 observations, the two parameter Gumbel distribution should be used, that is, the particular case in (4) and (5) where $\xi = 0$. For less than 30 observations, it is recommended to complement the local data with regional information.

We consider both the Gumbel distribution and the more general GEV distribution. That is, under the GEV model we have $\boldsymbol{\eta} = (\mu, \kappa, \xi)$ while for the Gumbel model, the parameter vector becomes $\boldsymbol{\eta} = (\mu, \kappa)$. To assess the uncertainty in the parameter estimation, we employ the Bayesian inference techniques described in the appendix. We denote by $p(\boldsymbol{\eta} \,|\, \hat{\mathcal{Y}})$ the posterior distribution under the best estimate for the rating curve model, $\hat{\boldsymbol{\theta}}$, resulting in the data $\hat{\mathcal{Y}} = (\hat{y}_1, \ldots, \hat{y}_T) = (\hat{Q}_1/A, \ldots, \hat{Q}_T/A)$. Similarly, for $\mathcal{Y}^{(i)} = (y_1^{(i)}, \ldots, y_T^{(i)})$ based on a rating curve model with parameters $\boldsymbol{\theta}^{(i)}$, we denote the posterior distribution by $p(\boldsymbol{\eta} \,|\, \mathcal{Y}^{(i)})$ for $i = 1, \ldots, n$. To assess the combined uncertainty of estimating both $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, we consider the mixture distribution

$$p(\boldsymbol{\eta} \,|\, \mathcal{Y}) = \frac{1}{n} \sum_{i=1}^{n} p(\boldsymbol{\eta} \,|\, \mathcal{Y}^{(i)}), \tag{6}$$

where $\mathcal{Y} = \{\mathcal{Y}^{(i)}\}_{i=1}^{n}$. Note that this is an approximation of the marginal posterior distribution of $\boldsymbol{\eta}$ given the data $\mathcal{D}$ and $\mathcal{H}$,

$$\begin{aligned} p(\boldsymbol{\eta} \,|\, \mathcal{D}, \mathcal{H}) &= \int p(\boldsymbol{\eta}, \boldsymbol{\theta} \,|\, \mathcal{D}, \mathcal{H}) \mathrm{d}\boldsymbol{\theta} \\ &= \int p(\boldsymbol{\eta} \,|\, \boldsymbol{\theta}, \mathcal{H}) p(\boldsymbol{\theta} \,|\, \mathcal{D}) \mathrm{d}\boldsymbol{\theta}. \end{aligned}$$

Alternatively, to isolate the effects of the parameter uncertainty in the rating curve model on the resulting flood frequency analysis, we employ a flood frequency estimation technique that returns a single best estimate. For ease of comparison with the Bayesian inference discussed above, we use maximum likelihood estimation for this purpose. That is, we obtain a maximum likelihood estimate $\hat{\boldsymbol{\eta}}^{(i)}$ for every data set $\mathcal{Y}^{(i)}$ in $\mathcal{Y}$ and use the variability of $\hat{\boldsymbol{\eta}}^{(1)}, \ldots, \hat{\boldsymbol{\eta}}^{(n)}$ to infer the effects of the uncertainty in the rating curve model only. The overall model framework is illustrated in Figure 1.

2.3 Return levels

The goal of flood frequency analysis is usually to construct estimates of design floods for given return periods. The return level z_τ associated with return period $1/\tau$ is the quantile of the sampling distribution that has probability τ of being exceeded in a particular year. For the GEV density in (4), it is given by

$$z_\tau = \begin{cases} \mu - (\kappa\xi)^{-1} \left[1 - (-\log(1-\tau))^{-\xi}\right] & \text{if } \xi \neq 0 \\ \mu - \kappa^{-1} \log(-\log(1-\tau)) & \text{if } \xi = 0 \end{cases} \tag{7}$$

which is the quantile function of the GEV distribution function for the quantile $1 - \tau$. For the Gumbel model, the return level is given by the special case in (7) where $\xi = 0$. Given a sample $\boldsymbol{\eta}^{(1)}, \ldots, \boldsymbol{\eta}^{(m)}$ from $p(\boldsymbol{\eta} \,|\, \hat{\mathcal{Y}})$ or $p(\boldsymbol{\eta} \,|\, \mathcal{Y})$ it is then straightforward to construct a posterior sample $z_\tau^{(1)}, \ldots, z_\tau^{(m)}$ from $p(z_\tau \,|\, \hat{\mathcal{Y}})$ or $p(z_\tau \,|\, \mathcal{Y})$, respectively, using (7) by setting

$$z_\tau^{(i)} = \mu^{(i)} - (\kappa^{(i)}\xi^{(i)})^{-1} \left[1 - (-\log(1-\tau))^{-\xi^{(i)}}\right]$$

if $\xi^{(i)} \neq 0$ and similar for $\xi^{(i)} = 0$.

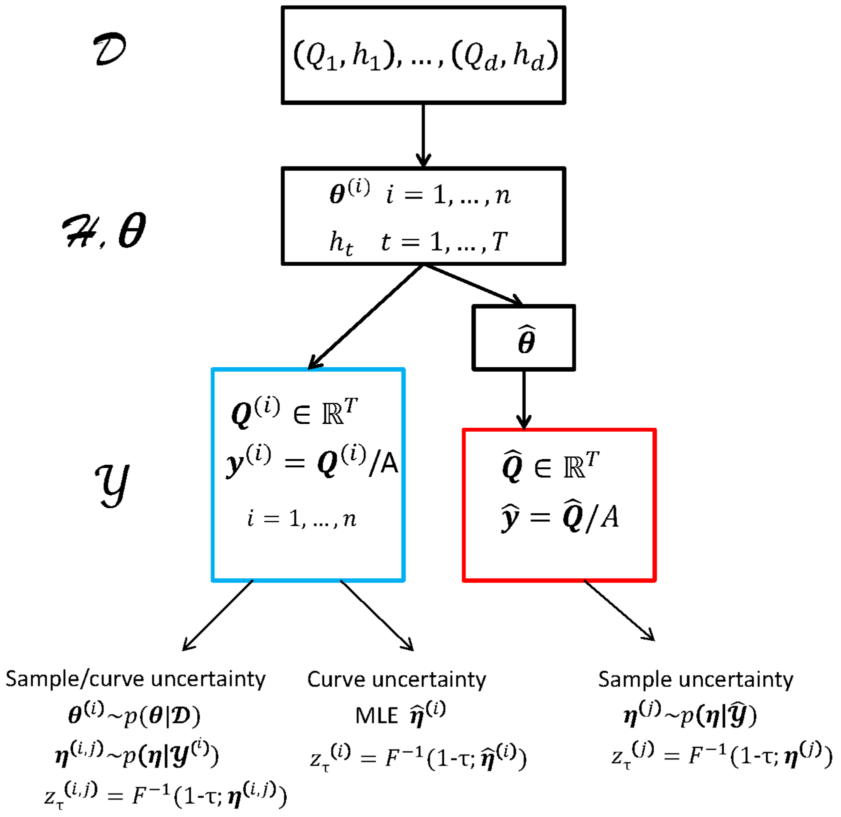

$\mathcal{D}$
$(Q_1, h_1), \dots, (Q_d, h_d)$
$\mathcal{H}, \boldsymbol{\theta}$
$\boldsymbol{\theta}^{(i)}$ $i = 1, \dots, n$
h_t $t = 1, \dots, T$
$\hat{\boldsymbol{\theta}}$
$\mathcal{Y}$
$\boldsymbol{Q}^{(i)} \in \mathbb{R}^T$
$\boldsymbol{y}^{(i)} = \boldsymbol{Q}^{(i)}/A$
$i = 1, \dots, n$
$\hat{\boldsymbol{Q}} \in \mathbb{R}^T$
$\hat{\boldsymbol{y}} = \hat{\boldsymbol{Q}}/A$
Sample/curve uncertainty
$\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$
$\boldsymbol{\eta}^{(i,j)} \sim p(\boldsymbol{\eta}|\boldsymbol{\mathcal{Y}}^{(i)})$
$z_\tau^{(i,j)} = F^{-1}(1\text{-}\tau; \boldsymbol{\eta}^{(i,j)})$
Curve uncertainty
MLE $\hat{\boldsymbol{\eta}}^{(i)}$
$z_\tau^{(i)} = F^{-1}(1\text{-}\tau; \hat{\boldsymbol{\eta}}^{(i)})$
Sample uncertainty
$\boldsymbol{\eta}^{(j)} \sim p(\boldsymbol{\eta}|\hat{\boldsymbol{\mathcal{Y}}})$
$z_\tau^{(j)} = F^{-1}(1\text{-}\tau; \boldsymbol{\eta}^{(j)})$

Figure 1. An overview of the modeling framework. $\mathcal{D}$ denotes the data used for estimating the parameters $\boldsymbol{\theta}$ in the stage-discharge rating curve relation and $\mathcal{H}$ denotes the annual maximum stage measurements. Given $\mathcal{H}$ and $\boldsymbol{\theta}$, we obtain $\mathcal{Y}$, the annual maximum discharge normalized by the catchment area A. After estimating η, the parameters of the flood frequency analysis, the return value estimates z_τ are obtained from the quantile function $F^{-1}(1-\tau;\cdot)$ accounting for both curve and sample uncertainty (left), curve uncertainty only (center), or sample uncertainty only (right).

Our main focus is on the uncertainty assessment associated with the return level estimation. As illustrated in Figure 1, the samples from $p(z_\tau \mid \mathcal{Y})$ have a two-fold uncertainty as they incorporate both the uncertainty related to the parameter estimation of the stage-discharge rating curve (curve uncertainty) and the flood frequency analysis (sample uncertainty), while $p(z_\tau \mid \hat{\mathcal{Y}})$ includes only the sample uncertainty and $\hat{z}_\tau^{(1)}, \dots, \hat{z}_\tau^{(n)}$ based on the maximum likelihood estimates $\hat{\boldsymbol{\eta}}^{(1)}, \dots, \hat{\boldsymbol{\eta}}^{(n)}$ includes only the curve uncertainty. Given a sample $z_\tau^{(1)}, \dots, z_\tau^{(m)}$ we may assess the uncertainty of the return level estimate by calculating credible intervals, e.g. the 80% credible interval is given by the interval $[\gamma_1, \gamma_2]$ where 10% of the sample are smaller than γ_1 and 10% of the sample are larger than γ_2.

3 Data

In order to investigate the three research questions stated in the introduction, we analyze data from gauging stations at seven unregulated catchments in Norway using the methodology described in the previous section. All data were extracted from the national hydrological data base at the Norwegian Water Resource and Energy Directorate.

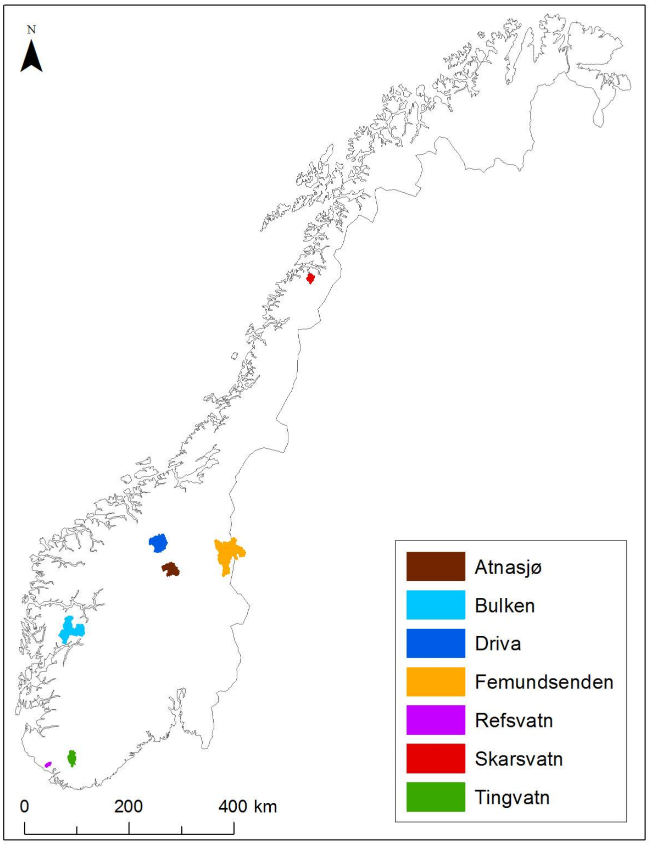


Figure 2. Map of Norway with the seven gauged catchments used in the analysis.

The locations of the catchments are shown in Figure 2, see also Table 1 for information on the size of the catchment areas. We analyze two data sets from each gauging station. A data set denoted by $\mathcal{D}$ consists of simultaneous stage and discharge measurements; this data is used to estimate the parameters of the Bayesian multi-segment rating curve model described in Section 2.1. The second data set, denoted by $\mathcal{H}$, contains annual maximum stage measurements which are transformed to normalized annual maximum discharge values via the Bayesian multi-segment rating curve model before they are used as an input in the flood frequency analysis described in Section 2.2.

The amount of available data and the time periods for which the annual maximum series are available are listed in Table 1. As can be seen from Table 1, many of the annual maximum series have a few missing values. We have not accounted for this in our analysis.

Table 1. Gauging stations, size of catchment areas (in km^2) and data availability. $|\mathcal{D}|$ indicates the number of stage-discharge measurements used for estimating the rating curve and the largest discharge measurement in $\mathcal{D}$ is $\max(\mathbf{Q})$. $|\mathcal{H}|$ denotes the number of annual maximum stage measurements used in the flood frequency analysis with $\max(\widehat{\mathbf{Q}})$ the largest estimated discharge value. Due to missing data, $|\mathcal{H}|$ might not equal the length of the data period. The extrapolation degree indicates the percentage of values in the full data series from which $\mathcal{H}$ is derived that exceed the largest value in $\mathcal{D}$ (in %) and the last column lists the number of estimated annual maximum discharge values that exceed the largest measured value in $\mathcal{D}$.

				Annual Maximum				
Station	Area	$\lvert\mathcal{D}\rvert$	$\max(\mathbf{Q})$	Period	$\lvert\mathcal{H}\rvert$	$\max(\widehat{\mathbf{Q}})$	Extrap.	$\#(\widehat{\mathbf{Q}} > \mathbf{Q})$
Atnasjø	463	38	139	1985-2014	30	159	0.018	1
Bulken (I)	1092	64	368	1892-1990	97	598	0.177	29
Bulken (II)	1092	31	705	1991-2014	24	680	0.000	0
Driva	745	67	19	1936-2013	78	467	0.134	25
Femundsenden	1792	83	87	1909-2013	103	134	0.780	17
Refsvatn	53	31	54	1984-2014	30	53	0.000	0
Skarsvatn (I)	145	19	54	1985-2002	17	101	0.128	4
Skarsvatn (II)	145	33	55	2003-2014	11	60	0.085	2
Tingvatn	272	51	128	1995-2014	20	138	0.026	2

The stage measurements in the two data sets $\mathcal{D}$ and $\mathcal{H}$ may partly overlap for a given location, indicating that discharge measurements were performed at the station during the annual maximum flood. However, it is common that the largest observed stage values exceed measurements for which corresponding discharge measurements are available, requiring an extrapolation of the rating curve. Table 1 lists the extrapolation degree for our data series which is given by the percentage of all available stage measurements at the station requiring an extrapolation of the rating curve. Additionally, we list extrapolation information for the particular data used in our analysis. For three stations (Bulken, Driva and Skarsvatn), over 20% of the data used for the flood frequency analysis are estimated by extrapolation.

For two stations, Bulken and Skarsvatn, two separate data series are listed. At Bulken, structural changes to the river profile were performed in 1990, requiring a new rating curve model. Similarly, the river profile changed in 2002 at Skarsvatn. For the subsequent flood frequency analysis at these stations, we combine the two annual maximum series and perform a single analysis.

4 Results and discussion

4.1 Contribution of sample/curve uncertainty

To estimate the parameter uncertainty in the Bayesian multi-segment rating curve model, we draw 10,000 MCMC samples from the posterior parameter distribution. Figure 3 shows the estimates for the annual maximum discharge at Bulken based on the Bayesian multi-segment model. For the first data period, 1892-1990, approximately 30% of the data points require an extrapolation resulting in high variability in the discharge estimates, particularly for the largest floods. Furthermore, the credible intervals are highly skew with a heavy upper tail. Conversely, for the second data period, 1991-2014, both stage and discharge were measured during the largest floods, resulting in significantly reduced uncertainty even though only about half as many data points are used for the parameter estimation compared to the first data period, see Table 1.

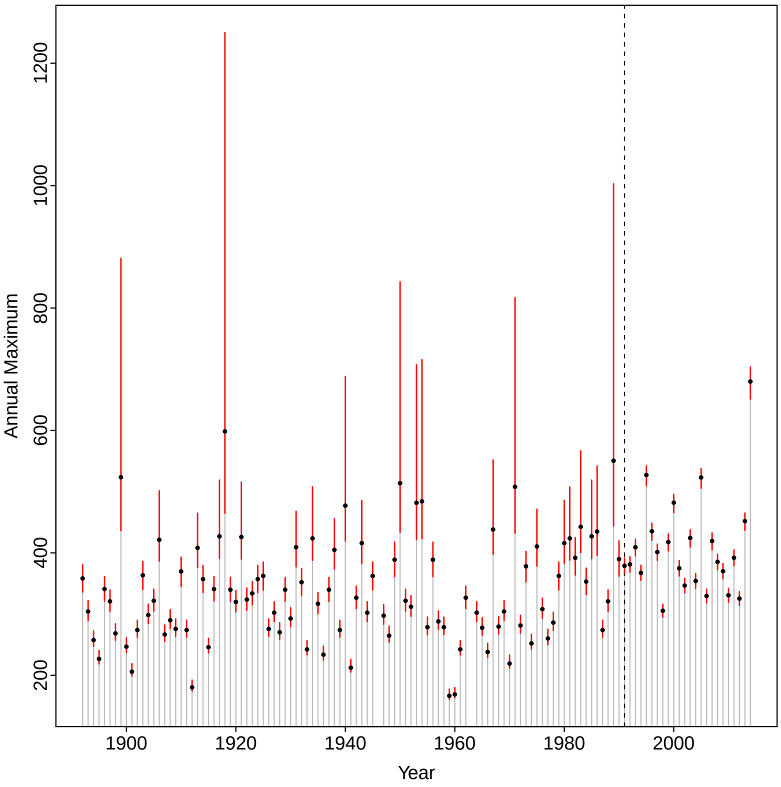


Figure 3. Estimated annual maximum discharge (in 1000 l/s) at Bulken from 1892 to 2014. The black dots show the best estimate and the red lines indicate the 99.5% credible intervals for the discharge values based on 10,000 MCMC simulations. Due to structural changes in the river profile, two separate rating curves are estimated, prior to and from 1991.

As outlined in Figure 1, we use three different approaches to obtain design flood estimates with uncertainty assessment. Firstly, for each sample $\boldsymbol{\theta}^{(i)}$, for $i = 1, \dots, 10,000$, we obtain 3,000 MCMC samples from the posterior parameter distribution of the flood frequency model (GEV or Gumbel) under the data $\mathcal{Y}^{(i)}$. This results in a sample of 3,000,000 design flood estimates where both curve and sample uncertainty are taken into account.

Secondly, we obtain a maximum likelihood estimate $\hat{\boldsymbol{\eta}}^{(i)}$ for the flood frequency parameters under each data set $\mathcal{Y}^{(i)}$. This results in a sample of 10,000 design flood estimates where only the curve uncertainty is taken into account. For two stations, Driva and Femundsenden, it was not possible to obtain maximum likelihood estimates $\hat{\boldsymbol{\eta}}^{(i)}$ for all data sets $\mathcal{Y}^{(i)}$ due to numerical issues. However, this affected less than 1% of the samples (72 samples at Femundsenden and 2 at Driva) and thus, does not affect the results presented here. Using probability weighted moments rather than maximum likelihood did not solve this issue. In a third approach, we obtain 10,000 MCMC samples from the posterior parameter distribution of the flood frequency model for a single best estimate of the normalized annual maximum series, $\hat{\mathcal{Y}}$. For this approach we computed one million design flood estimates from the posterior given $\hat{\mathcal{Y}}$ where only sample uncertainty is accounted for.

Figure 4 shows 80% credible intervals for design floods with return periods of 20, 200 and 1000 years under both the GEV and the Gumbel model for the three uncertainty settings described above. The corresponding tables are given in Appendix B. As expected, the credible intervals are generally considerably narrower under the two parameter Gumbel model than for the three parameter GEV model. Furthermore, the difference in the design flood estimates between the two models is directly linked to the estimates of the shape parameter ξ in the GEV model, cf. Figure 5. That is, ξ estimates centered around 0 yield similar design flood estimates (Femundsenden, Refsvatn and Tingvatn in our data set), when the bulk of the posterior distribution is above 0, GEV returns higher estimates than Gumbel (Atnasjø, Driva and Skarsvatn) while the opposite holds when the bulk of the distribution is below 0 (Bulken in our data set).

Overall, we obtain the largest uncertainty when both curve and sample uncertainty are accounted for, see Table 2. The length of a credible interval for a return value under sample uncertainty ranges from 55% to 94% of the length of the corresponding interval under combined curve and sample uncertainty for all the seven stations considered here. For the Gumbel model, the relative difference between curve or sample uncertainty only and combined curve and sample uncertainty is similar for six out of seven stations. At Refsvatn, the curve uncertainty is very small compared to the other two settings; this is also the only station with no rating curve extrapolation, cf. Table 1.

For the GEV model, the results are more sensitive to differences between the data sets. We observe large sample uncertainty for stations which have less than 50 measurements in $\mathcal{H}$ (Atnasjø, Refsvatn, Skarsvatn and Tingvatn), for which the GEV model is not recommended (Wilson et al., 2011). We observe that a low extrapolation degree is connected to a small shift in the posterior distribution of the ξ parameter when comparing those estimated under sample uncertianty to those estimated under both curve and sample uncertainty (Figure 5). This results in relatively narrow credible intervals for the floods (Atnasjø, Refsvatn and Tingvatn). Similarly, for the longer data series (Bulken, Driva and Femundsenden), the posterior distribution for ξ is considerably sharper under the single data set $\hat{\mathcal{Y}}$ than for the mixture in (6).

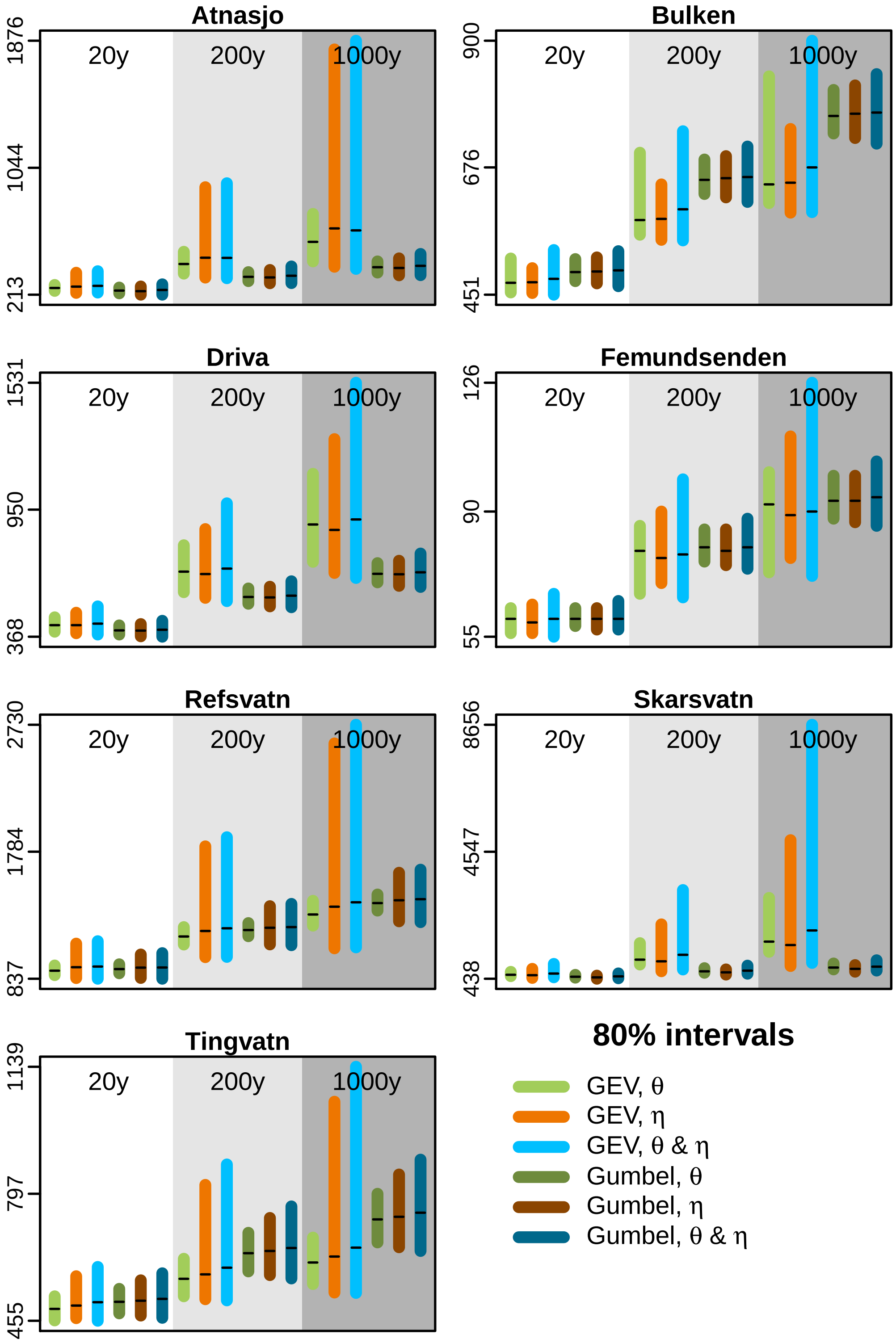


Figure 4. Estimated 80% credible intervals for return values with return periods of 20, 200 and 1000 years under curve uncertainty (light/dark green), sample uncertainty (orange/brown) and both (light/dark blue) under the GEV (lighter colors) and the Gumbel (darker colors) models for normalized annual maximum discharge. The best estimate for each instance is indicated by a black line. The return values are given in the unit of $l/s/km^2$.

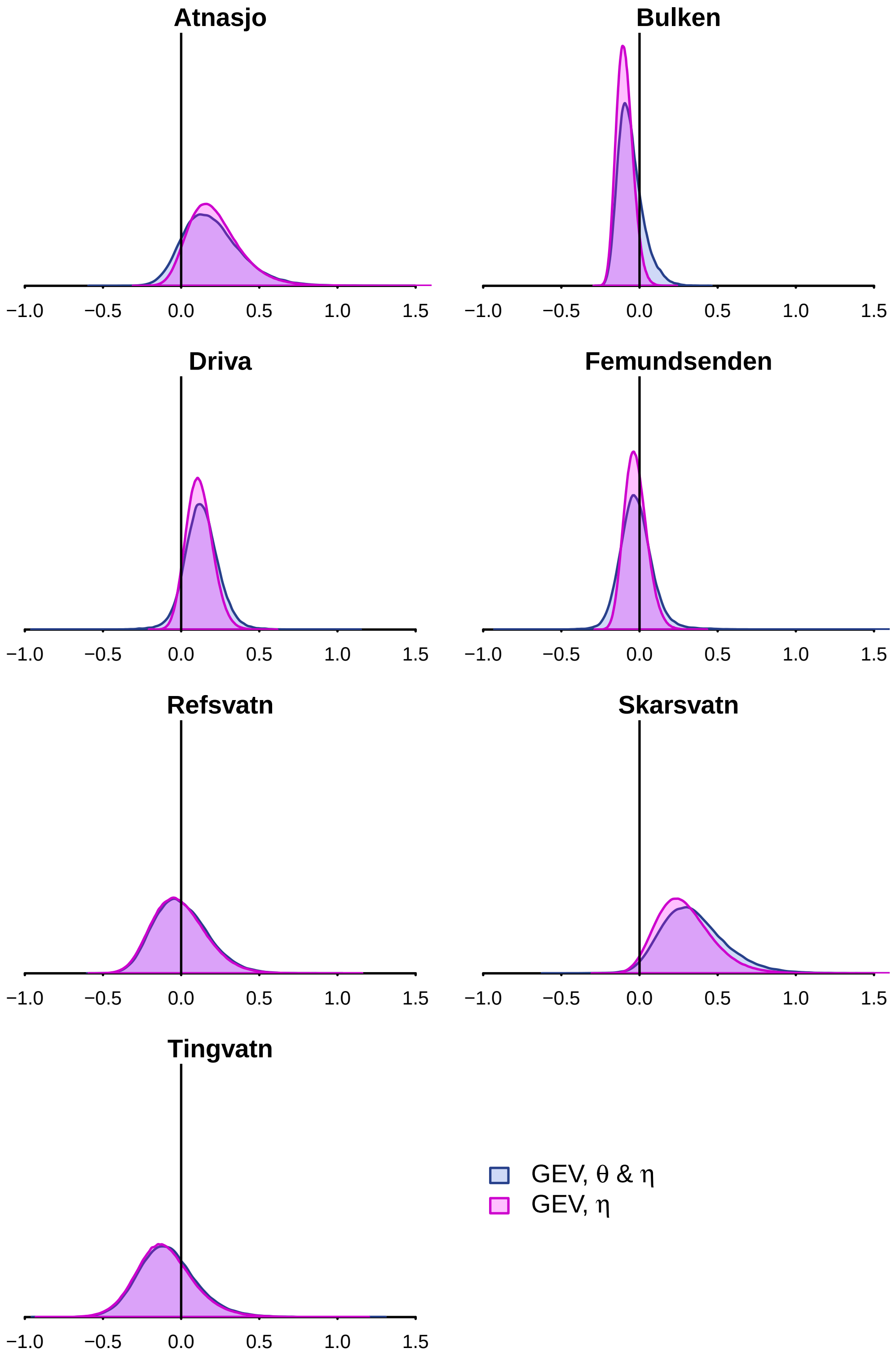


Figure 5. Posterior distributions for the shape parameter ξ in the GEV model under sample uncertainty (η) and both curve and sample uncertainty $(\theta$ & $\eta)$.

At Skarsvatn, in particular, the combination of a considerable extrapolation of the rating curve and a small sample size in $\mathcal{H}$ causes extremely large uncertainty for the combined setting. This is made clear in Table 3, where Skarsvatn has the largest relative uncertainty for all model cases. Noticeably, at Bulken, the curve uncertainty exceeds the sample uncertainty. Here, $\mathcal{H}$ contains over 100 data points while the rating curve for the first part of the series has a high extrapolation degree, see Table 1 and Figure 3.

Table 2. Ratio of 80% credible interval lengths for the 1000-year design flood. We compare, on the one hand, curve uncertainty vs. both curve and sample uncertainty ($\theta/\theta\&\eta$) and, on the other hand, sample uncertainty vs. both curve and sample uncertainty ($\eta/\theta\&\eta$).

	GEV		Gumbel	
Station	$\frac{\theta}{\theta\&\eta}$	$\frac{\eta}{\theta\&\eta}$	$\frac{\theta}{\theta\&\eta}$	$\frac{\eta}{\theta\&\eta}$
Atnasjø	0.28	0.94	0.57	0.81
Bulken	0.77	0.56	0.63	0.74
Driva	0.51	0.73	0.58	0.82
Femundsenden	0.65	0.67	0.62	0.75
Refsvatn	0.26	0.91	0.35	0.92
Skarsvatn	0.33	0.69	0.57	0.64
Tingvatn	0.45	0.78	0.55	0.80

Table 3. Ratios of 80% credible interval length and corresponding best estimate for the 1000-year design flood under curve uncertainty (η), sample uncertainty (θ), and both curve and sample uncertainty (θ & η) in the GEV and the Gumbel model.

	GEV			Gumbel		
Station	θ	η	$\theta\&\eta$	θ	η	$\theta\&\eta$
Atnasjø	0.56	2.20	2.36	0.19	0.29	0.34
Bulken	0.35	0.23	0.45	0.10	0.12	0.16
Driva	0.46	0.72	0.99	0.13	0.18	0.23
Femundsenden	0.31	0.38	0.60	0.13	0.13	0.19
Refsvatn	0.14	1.11	1.18	0.09	0.26	0.27
Skarsvatn	1.06	2.66	3.85	0.24	0.28	0.40
Tingvatn	0.20	0.82	0.93	0.18	0.27	0.33

4.2 Length of annual maximum discharge series

In order to investigate the effect of the length of the annual maximum discharge series in $\mathcal{H}$ on the design flood estimation uncertainty, we focus on the data from Femundsenden, where we have a long data series comprising 103 values derived from a single rating curve model. We separately analyze the first 20 years of the series, the first 60 years and the full series. The resulting return levels under the GEV model are given in Figure 6 with a more detailed view of the results for return periods of 20 and 1000 years under both the GEV and the Gumbel model shown in Figure 7.

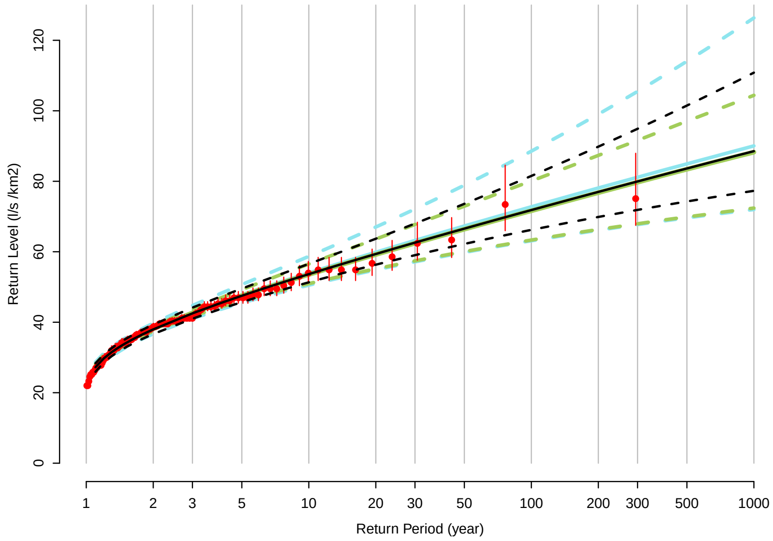


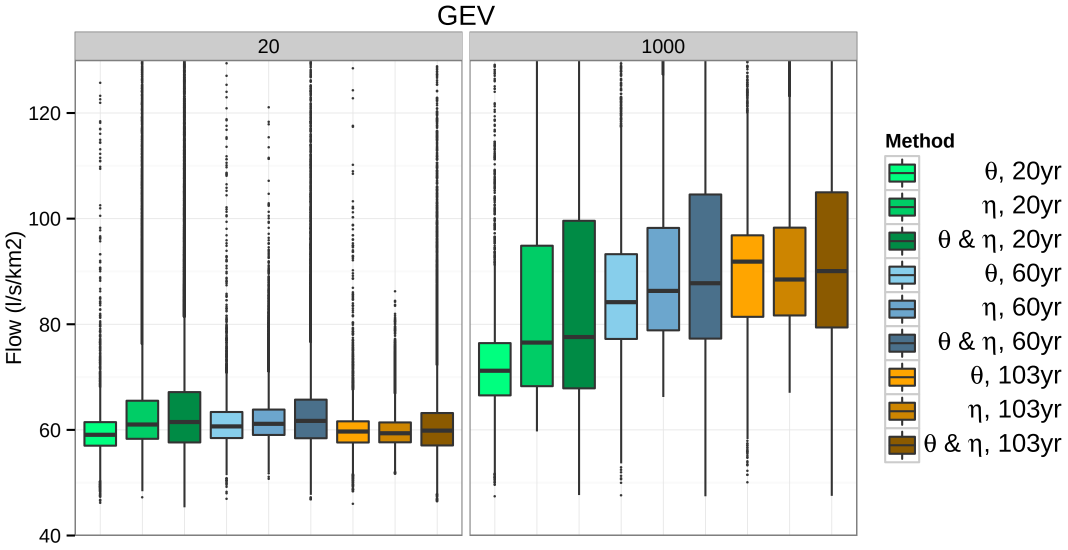
Figure 6. Estimated return level curves at Femundsenden under the GEV model with 80 % credibility intervals (dashed lines) under curve uncertainty (green), sample uncertainty (black) and both (blue). The red lines show the 80 % credible intervals for the discharge values based on 10,000 simulations from the Baysian multi segment model with corresponding best estimates (red dots).

While there is a general tendency of decreased estimation uncertainty as we include more data in the analysis, the opposite holds for the 1000 year return level in the GEV model when only curve uncertainty is accounted for, see Figure 7. Due to extrapolation, there is significantly larger uncertainty in the discharge estimation if the annual maximum flood is unusually large. At Femundsenden, the estimation of the four largest discharge values is considerably more uncertain that the estimation of the remaining values. These floods occurred in 1927, 1944, 1967 and 1995. For any additional data period, we thus also include more large values with a high degree of uncertainty. However, when sample uncertainty is accounted for, this effect somewhat cancels out by the additional benefit of having more data for the frequency analysis. The estimated level also changes as we add more data under both models. These results indicate, in particular, that the GEV distribution is not appropriate when only 20 data points are available for the parameter estimation.

4.3 Added value of large measurements

Here, we consider the added value of large measurements, both at the high end of the rating curve and in the annual maximum data series. For this, we study the latter part of the data series from Bulken, the data from 1991-2014. Originally, $\mathcal{D}$ contains 30 measurements which we compare to using a smaller data set $\mathcal{D}^-$ with the three highest stage measurements removed. Similarly, we consider the full data set $\mathcal{H}$ as well as $\mathcal{H}^-$ comprising the data from 1991-2013, leaving out the large flood in 2014 (cf. Figure 3). The resulting estimates and the associated uncertainty for 1000 year return levels are given in Figure 8.

Without extrapolation of the rating curve, the curve uncertainty is minimal. This propa-



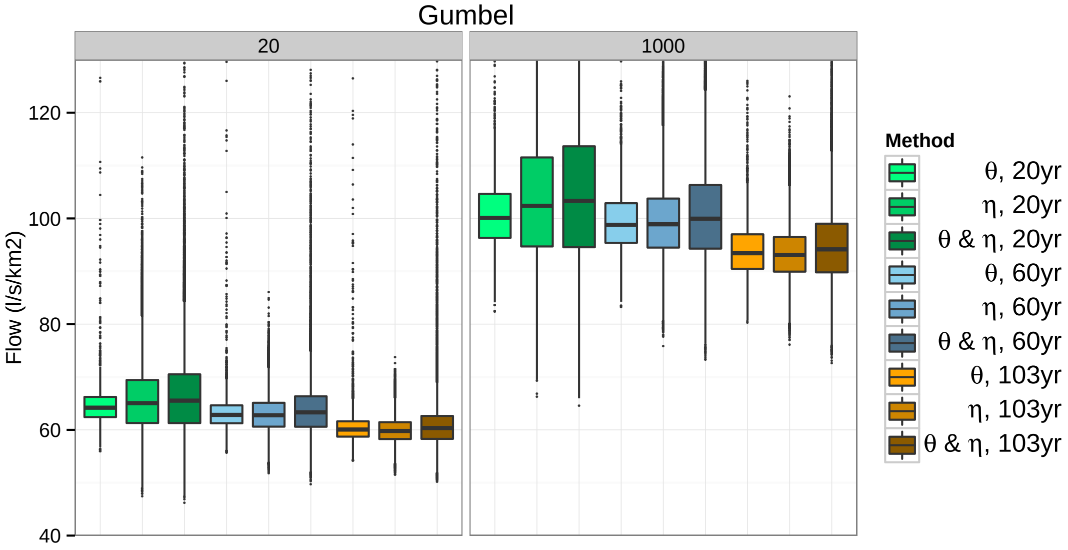


Figure 7. Box plots of estimated return levels with return periods of 20 and 1000 years at Femundsenden, based on 20, 60 and 103 years of maximum annual discharge data under the GEV (upper) and the Gumbel (lower) model. For each setting, we compare results under curve uncertainty (θ), sample uncertainty (η) and both (θ&η).

gates forward to a lower uncertainty for the combined setting, particularly for the Gumbel model which is the more appropriate model given the length of the annual maximum series. An inclusion of the 2014 measurement in $\mathcal{H}$, on the other hand, increases the resulting estimation uncertainty. The reason is that this flood is exceptionally larger than all other floods in this data set (cf. Figure 3), and the effect of increased sample size on reducing estimation uncertainty is more than compensated by the effect of larger variability in $\mathcal{H}$.

An additional effect is a significant shift: The 1000 year return level estimates are notably lower without the 2014 observation to the extend that some of the estimates are below the 2014 flood value. Similary, the estimates vary considerably for the three uncertainty settings when the inference in the rating curve model is performed with $\mathcal{D}^-$. The best estimate under curve uncertainty is higher than that under sample uncertainty. This may be explained as follows: The discharge values for the highest stage measurements follow highly skew distributions with heavy upper tails. The higher the stage, the more skewed is the distribution for discharge. Consequently, several of the data sets $\mathcal{Y}^{(1)}, \ldots, \mathcal{Y}^{(10000)}$ include more high values than $\hat{\mathcal{Y}}$, and the distribution of annual maximum streamflows becomes more skewed for several of the rating curves. This results in posterior parameter distributions shifted towards higher values with heavy upper tails, cf. Figure 9, and higher estimates of return levels. The effect of using $\mathcal{D}^-$ is less pronounced when we use $\mathcal{H}^-$ (largest flood removed), since we then exclude one observations with a large associated uncertainty.

This numerical experiment with the data from Bulken, indicates that a direct streamflow measurement at the highest flood peaks has the potential to improve both estimation bias and estimation uncertainty for the return levels and is therefore a very valuable information. On the other hand, one large observation in $\mathcal{H}$ that is based on an extrapolation of the rating curve is the most challenging case. Accounting for rating curve uncertainties then seems to result in a large over-estimation of return levels at Bulken. This effect might be reduced if we perform a joint estimation of the parameters in the rating curve and the distribution function since the estimation then would put more constraints on the likely size of the largest floods. The challenge of high outliers in extreme data has been discussed in the literature, see e.g. (Hosking et al., 1985). In our study we used estimators that are sensitive to outliers in data in order to expose the effect of data uncertainty on estimates. For more robust estimates we could have included prior information.

5 Conclusions

In this paper, we investigate the propagation of rating curve uncertainty in design flood estimation by combining results from a Bayesian multi-segment rating curve model and Bayesian flood frequency analysis under the GEV and the Gumbel distribution. This allows us to consider curve/sample uncertainty both separately and combined. Concerning our original research questions stated in Section 1, we conclude that

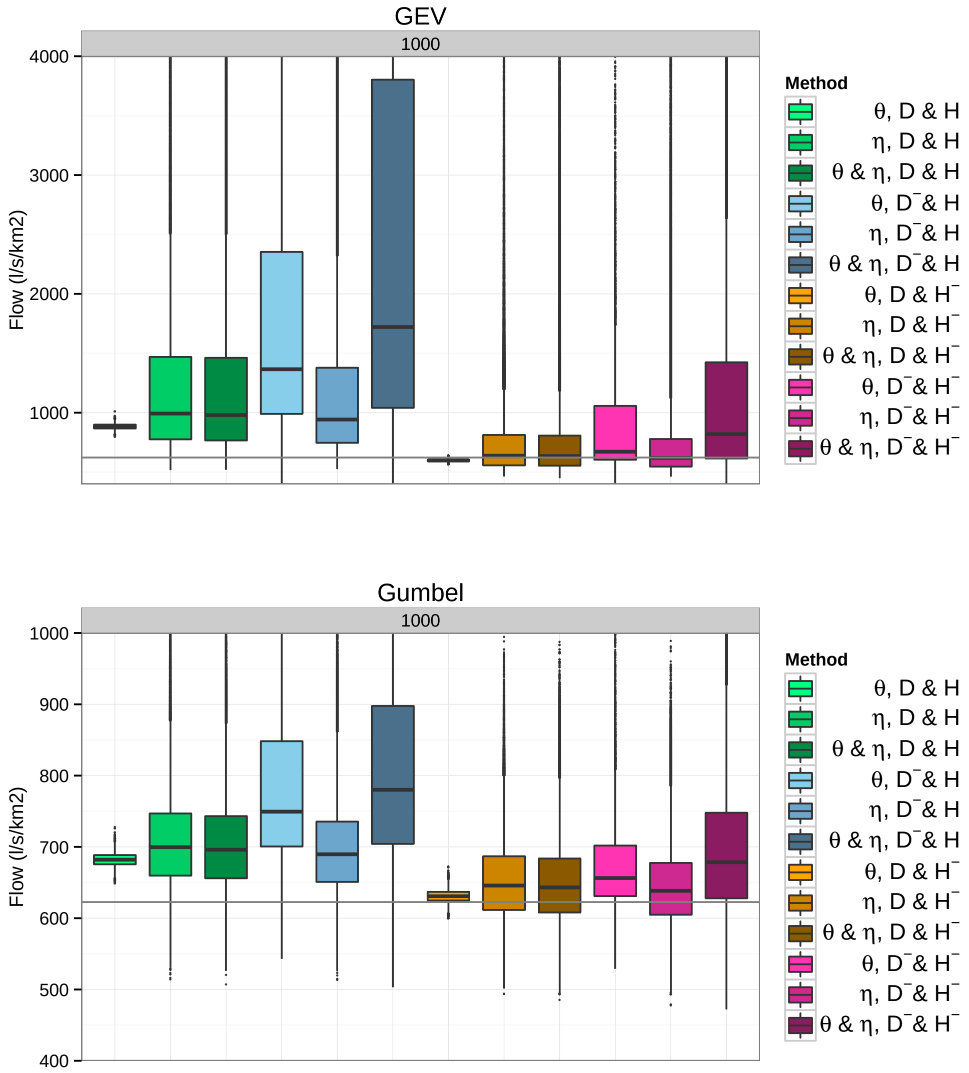


Figure 8. Box plots of 1000 year return level estimates at Bulken under the GEV (top) and the Gumbel (bottom) model, based on combinations of a full training set $\mathcal{D}$, a complete set of annual maximum discharge $\mathcal{H}$, a set $\mathcal{D}^-$ with the three highest measurements removed and a reduced set $\mathcal{H}^-$ without the 2014 flood. For each setting, we compare results under curve uncertainty (θ), sample uncertainty (η) and both ($\theta\&\eta$). The results are based on the data series from 1991-2014 only. The 2014 annual maximum flood is indicated by a dark gray line.

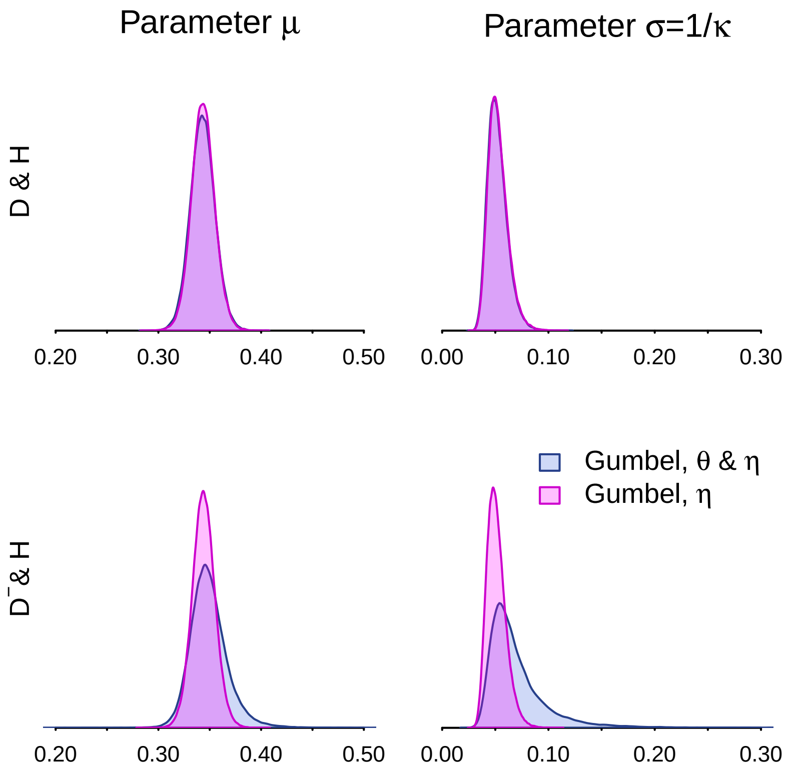


Figure 9. Posterior distributions of the parameters of the Gumbel distribution for data from 1991-2014 at Bulken under sample uncertainty ($\boldsymbol{\eta}$) and both curve and sample uncertainty ($\boldsymbol{\theta}\&\boldsymbol{\eta}$) using $\mathcal{D}$ and $\mathcal{H}$ (top panel) or $\mathcal{D}^-$ and $\mathcal{H}$ (bottom panel). The first column shows the distributions for the location parameter μ and the second column the distributions for the scale parameter $\sigma = 1/\kappa$.

(i) The sample uncertainty is generally the main contributor to uncertainty in design flood estimation. However, curve uncertainty may play an important role when extrapolation of the rating curve is necessary.

(ii) The sample uncertainty generally decreases with a longer data series. However, if a new data point has an unusually large value, the opposite may hold.

(iii) An additional high direct streamflow measurement will reduce the extrapolation degree and the rating curve uncertainty, and most likely reduce estimation biases in the return levels. A high annual maximum flood observation might, if combined with a large extrapolation degree, introduce estimation biases for return levels since the estimation is based on combining two highly skewed distributions.

Compared to the full data base of all gauging stations on Norway, our data examples are of good quality with fairly low extrapolation degree. Only the data from Femundsenden is above the median extrapolation degree (0.4%) and all the data is below the mean extrapolation degree (2%). It is thus likely that the curve uncertainty in the data used in this study is somewhat lower than Norwegian data in general. We might therefore expect that the effect of rating curve uncertainty for Norwegian data is, on average, more important than we see in this study. However, it is unclear to which extend these results generalize to other regions.

Our study can be extended in several ways. In the analysis, we have used non-informative prior distributions for the model parameters. However, the use of informative priors can improve the estimation uncertainty when additional information is available, see e.g. Parent and Bernier (2002) and Reis and Stedinger (2005). In our context is seems particularly promising to include prior information directly on the return levels, see e.g. Coles and Tawn (1996). Further, a single hierarchical model can be constructed to investigate the combined curve and sample uncertainty following the structure given in Figure 1. While this may be somewhat more appropriate from a statistical viewpoint and would potentially lower the combined estimation uncertainty, we consider such a construction outside the scope of the current paper.

References

Clarke, R. (1999). Uncertainty in the estimation of mean annual flood due to rating-curve indefinition. *J. Hydrol.*, 222:185–190.

Clarke, R., Mendiondo, E., and Bruca, L. (2000). Uncertainties in mean discharges from two large south american rivers due to rating curve variability. *Hydrol. Sci. J.*, 45:221–236.

Coles, S. G. (2001). *An Introduction to Statistical Modeling of Extreme values.* Springer Series in Statistics.

Coles, S. G. and Dixon, M. J. (1999). Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23.

Coles, S. G. and Tawn, J. A. (1996). A Bayesian analysis of extreme rainfall data. *Journal of the Royal Statistical Society, Series C*, 45(4):463–478.

Coxon, G., Freer, J., Westerberg, I., Wagener, T., Woods, R., and Smith, P. (2015). A novel framework for discharge uncertainty quantification applied to 500 uk gauging stations. *Water Resour. Res.*, 51:5531–5546.

Di Baldassarre, G. and Montanari, A. (2009). Uncertainty in river discharge observations: a quantitative analysis. *Hydrol. Earth Syst. Sci.*, 13:913–921.

Dyrrdal, A. V., Lenkoski, A., Thorarinsdottir, T. L., and Stordal, F. (2015). Bayesian hierarchical modeling of extreme hourly precipitation in norway. *Environmetrics*, 26:89–106.

Engeland, K., Hisdal, H., and Frigessi, A. (2005). Practical extreme value modelling of hydrological floods and droughts: A case study. *Extremes*, 7:5–30.

Hall, M., Boogaard, H. v. d., Fernando, R., and Mynett, A. (2004). The construction of confidence intervals for frequency analysis using resampling techniques. *Hydrol. Earth Syst. Sci.*, 8:235–246.

Hosking, J., Wallis, J., and Wood, E. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27:251–261.

Kuczera, G. (1996). Correlated rating curve error in flood frequency inference. *Water Resour. Res.*, 32:2119–2127.

Kyselý, J. (2008). A cautionary note on the use of nonparametric bootstrap for estimating uncertainties in extreme-value models. *J. Appl. Meteorol. Climatol.*, 47:3236–3251.

Lang, M., Pobanz, K., Renard, B., Renouf, E., and Sauquet, E. (2010). Extrapolation of rating curves by hydraulic modelling, with application to flood frequency analysis. *Hydrol. Sci. J.*, 55:883–898.

Le Coz, J., Renard, B., Bonnifait, L., Branger, F., and Le Boursicaud, R. (2014). Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A bayesian approach. *J. Hydrol.*, pages 573–587.

Lee, K. and Kim, S. (2008). Identification of uncertainty in low flow frequency analysis using bayesian mcmc method. *Hydrol. Process.*, 22:1949–1964.

Madsen, H., Rasmussen, P., and Rosbjerg, D. (1997). Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events: 1 at-site modeling. *Water Resour. Res.*, 33:747–757.

Martins, E. and Stedinger, J. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resour. Res.*, 36:737–744.

McMillan, H., Freer, J., Pappenberger, F., Krueger, T., and Clark, M. (2010). Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrol. Process.*, 24:1270–1284.

Midttømme, G., Pettersson, L., Holmqvist, E., Nøtsund, Ø., Hisdal, H., and Sivertsgård, R. (2011). *Retningslinjer for flomberegninger*. NVE. Retningslinjer nr. 4/2011.

Moyeed, R. and Clarke, R. (2005). The use of bayesian methods for fitting rating curves, with case studies. *Adv. Water Resour.*, 28:807–818.

Neppel, L., Renard, B., Lang, M., Ayral, P.-A., Coeur, D., Gaume, E., Jacob, N., Payrastre, O., Pobanz, K., and Vinet, F. (2010). Flood frequency analysis using historical data: accounting for random and systematic errors. *Hydrol. Sci. J.*, 55:192–208.

Parent, E. and Bernier, J. (2002). Bayesian POT modelling for historical data. *Journal of Hydrology*, 274:95–108.

Petersen-Øverleir, A. and Reitan, T. (2009). Accounting for rating curve imprecision in flood frequency analysis using likelihoode-based methods. *J. Hydrol.*, 366:89–100.

Petersen-Øverleir, A., Soot, A., and Reitan, T. (2009). Bayesian rating curve inference as a streamflow data quality assessment tool. *Water Resour. Manag.*, pages 1835–1842.

Reis, D. and Stedinger, J. (2005). Bayesian MCMC flood frequency analysis with historical information. *Journal of Hydrology*, 313:97–116.

Reitan, T. and Petersen-Øverleir, A. (2009). Bayesian methods for estimating multi-segment discharge rating curves. *Stoch. Environ Res Risk Assess*, 23:627–642.

Reitan, T. and Petersen-Øverleir, A. (2011). Dynamic rating curve assessment in unstable rivers using ornstein-uhlenbeck processes. *Water Resour. Res.*, 47:n/a–n/a.

Renard, B. (2011). A bayesian hierarchical approach to regional frequency analysis. *Water Resour. Res.*, 47:W11513.

Renard, B., Kochanek, K., Lang, M., Garavaglia, F., Paquet, E., Neppel, L., Najib, K., Carreau, J., Arnaud, P., Aubert, Y., Borchi, F., Soubeyroux, J.-M., Jourdain, S., Veysseire, J.-M., Sauquet, E., Cipriani, T., and Auffray, a. (2013). Data-based comparison of frequency analysis methods: A general framework. *Water Resour. Res.*, 49:825–843.

Renard, B., Lang, M., and Bois, P. (2006). Statistical analysis of extreme events in a non-stationary context via a bayesian framework: case study with peak-over-threshold data. *Stoch. Environ. Res. Risk Assess.*, 21:97–112.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition.

Rosbjerg, D. and Madsen, H. (1995). Uncertainty measures of regional flood frequency estimators. *J. Hydrol.*, 167:209–224.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with Discussion). *Journal of the Royal Statistical Society, Series B*, 71:319–392.

Westerberg, I., Guerrero, J., Seibert, J., Beven, K., and Halldin, S. (2011). Stage-discharge uncertainty derived with a non-stationary rating curve in the choluteca river, honduras. *Hydrol. Process.*, 25:603–613.

Westerberg, I. and McMillan, H. (2015). Uncertainty in hydrological signatures. *Hydrol. Earth Syst. Sci.*, 19:3951–3968.

Wilson, D., Fleig, A., Lawrence, D., Hisdal, H., Pettersson, L.-E., and Holmqvist, E. (2011). A review of NVE's flood frequency estimation procedures. Technical Report 9:2011, NVE, Oslo.

A Bayesian inference in flood frequency analysis

We use Bayesian inference to obtain posterior distributions for the flood frequency parameters specified in (4) and (5) via Markov chain Monte Carlo sampling (Robert and Casella, 2004). In particular, we use a Metropolis-Hastings algorithm where each parameter in $\boldsymbol{\eta}$ is updated in turn. Assume we are analyzing the data set $\mathbf{y}$. We update a parameter, say μ, by drawing a new value μ' from a proposal distribution $p(\mu \mid \mu', \cdot)$ and accept the proposal with the probability $\min\{r, 1\}$, where

$$r = \frac{p(\boldsymbol{y} \mid \mu', \cdot)p(\mu' \mid \cdot)p(\mu' \mid \mu, \cdot)}{p(\boldsymbol{y} \mid \mu, \cdot)p(\mu \mid \cdot)p(\mu \mid \mu', \cdot)}. \tag{A.1}$$

Here, $p(\boldsymbol{y} \mid \mu, \cdot)$ denotes the likelihood and $p(\mu' \mid \cdot)$ the prior distribution, both of which may depend on other parameters. The remaining parameters in $\boldsymbol{\eta}$, $\nu = \log \kappa$ and ξ are then updated in a similar manner. We use a Gaussian prior distribution with mean $\mu_0 = 0$ and standard deviation $\sigma_0 = 100$ for all three parameters.

The efficiency of the Metropolis-Hastings algorithm heavily depends on the choice of the proposal distribution. We follow e.g. Dyrrdal et al. (2015) and Rue et al. (2009) and employ a Gaussian approximation to the log-posterior density as the proposal density. If we write the marginal posterior distribution as $p(\mu' | \mathbf{y}, \cdot) \propto \exp(f(\mu'))$, a second-order Taylor expansion around the current stage μ is given by the Gaussian distribution $\mathcal{N}(c_1/c_2, 1/c_2)$, where

$$\begin{aligned} c_1 &= f'(\mu) - f''(\mu)\mu \\ c_2 &= -f''(\mu) \end{aligned}$$

(see Chapter 4.4 of Rue et al. (2009)). That is,

$$\begin{aligned} f'(\mu) &= -\frac{\mu - \mu_0}{\sigma_0} + \sum_{t=1}^{T} \frac{\partial}{\partial \mu} \log p(y_t|\mu, \cdot) \\ f''(\mu) &= -\frac{1}{\sigma_0} + \sum_{t=1}^{T} \frac{\partial^2}{\partial \mu^2} \log p(y_t|\mu, \cdot), \end{aligned}$$

under the $\mathcal{N}(\mu_0, \sigma_0^2)$ prior distribution. Similar calculations hold for $\nu = \log \kappa$ and ξ. Explicit formulas for the derivatives $f'(\cdot)$ and $f''(\cdot)$ are given below.

A.1 The case $\xi \neq 0$

Set $\alpha_t = 1 + \xi e^{\nu}(y_t - \mu)$. For $\xi \neq 0$, we obtain

$$\begin{aligned} f'(\mu) &= -\frac{\mu - \mu_0}{\sigma_0} + e^{\nu} \sum_{t=1}^{T} \alpha_t^{-1}(\xi + 1 - \alpha_t^{-1/\xi}) \\ f''(\mu) &= -\frac{1}{\sigma_0} + (\xi + 1)e^{2\nu} \sum_{t=1}^{T} \alpha_t^{-2}(\xi - \alpha_t^{-1/\xi}). \end{aligned}$$

The derivatives for the log-inverse scale parameter, $\nu = \log(\kappa)$, are

$$\begin{aligned}
f'(\nu) &= -\frac{\nu - \mu_0}{\sigma_0} + T \\
&\quad + \xi^{-1} \sum_{t=1}^{T} \alpha_t^{-1}(\alpha_t - 1)\Big\{\alpha_t^{-1/\xi} - (\xi + 1)\Big\}, \\
f''(\nu) &= -\frac{1}{\sigma_0} + \xi^{-1} \sum_{t=1}^{T} \alpha_t^{-2}(\alpha_t - 1) \\
&\qquad \times \Big\{(\xi + 1)(\alpha_t^{-1/\xi} - 1) - \xi^{-1}\alpha_t^{2/\xi}\Big\}.
\end{aligned}$$

Set $\epsilon_t = y_t - \mu$. The first derivative for ξ is given by

$$f'(\xi) = -\frac{\xi - \mu_0}{\sigma_0} + D_1 + D_2$$

where

$$\begin{aligned}
D_1 &= \sum_{t=1}^{T} \xi^{-1}\Big\{\xi^{-1}\log(\alpha_t) - (\xi + 1)e^{\nu}\epsilon_t\alpha_t^{-1}\Big\} \\
D_2 &= \sum_{t=1}^{T} \xi^{-1}\alpha_t^{-1/\xi_t}\Big\{e^{\nu}\epsilon_t\alpha_t^{-1} - \xi^{-1}\log(\alpha_t)\Big\}.
\end{aligned}$$

The second derivative for ξ is given by

$$f''(\xi) = -\frac{1}{\sigma_0} + \dot{D}_1 + \dot{D}_2 + \dot{D}_3 + \dot{D}_4$$

with

$$\begin{aligned}
\dot{D}_1 &= \sum_{t=1}^{T} \xi^{-2}\Big\{e^{\nu}\epsilon_t\alpha_t^{-1} - 2\xi^{-1}\log(\alpha_t)\Big\} \\
\dot{D}_2 &= \sum_{t=1}^{T} \xi^{-1}e^{\nu}\epsilon_t\alpha_t^{-1}\Big\{\xi^{-1} + (\xi + 1)e^{\nu}\epsilon_t\alpha_t^{-1}\Big\} \\
\dot{D}_3 &= \sum_{t=1}^{T} \alpha_t^{-1/3}\xi^{-2}\Big\{\xi^{-1}\log(\alpha_t)(2 - \xi^{-1}\log(\alpha_t)) \\
&\qquad + e^{\nu}\epsilon_t\alpha_t^{-1}(\xi^{-1}\log(\alpha_t) - 1)\Big\} \\
\dot{D}_4 &= \sum_{t=1}^{T} \xi^{-2}e^{\nu}\epsilon_t\alpha_t^{-1/\xi-1}\Big\{\xi^{-1}\log(\alpha_t) - (\xi + 1)e^{\nu}\epsilon_t\alpha_t - 1\Big\}.
\end{aligned}$$

A.2 The case $\xi = 0$

When $\xi = 0$, we set $\alpha_t = \exp(-e^{\nu}(y-\mu))$ and obtain

$$
\begin{aligned}
f'(\mu) &= -\frac{\mu-\mu_0}{\sigma_0} + e^{\nu}\sum_{t=1}^{T}(1-\alpha_t) \\
f''(\mu) &= -\frac{1}{\sigma_0} - e^{2\nu}\sum_{t=1}^{T}\alpha_t \\
f'(\nu) &= -\frac{\nu-\mu_0}{\sigma_0} + T + \sum_{t=1}^{T}\log(\alpha_t)(1-\log(\alpha_t)) \\
f''(\nu) &= -\frac{1}{\sigma_0} + \sum_{t=1}^{T}\log(\alpha_t)\left(1-\alpha_t-\alpha_t\log(\alpha_t)\right).
\end{aligned}
$$

B Return level estimates

Table B.1 lists the results also shown in Figure 4.

Table B.1. The estimated 20, 200 and 1000 year return levels (in $l/s/km^2$) for GEV and Gumbel distributed data under curve uncertainty, sample uncertainty and the combined curve and sample uncertainty. The corresponding 80% credibility intervals are given in parentheses.

	Curve Uncertainty		Sample Uncertainty		Curve & Sample Uncertainty	
20 year return level						
GEV						
Atnasjo	257	(238, 277)	266	(225, 357)	272	(227, 368)
Bulken (1991-2014)	510	(503, 518)	533	(477 ,664)	531	(475 ,661)
Bulken	472	(455, 515)	473	(454, 498)	479	(451, 529)
Driva	421	(391, 457)	421	(384, 478)	428	(378, 507)
Femundsenden	60	(56, 63)	59	(56, 64)	60	(55, 67)
Refsvatn	897	(864, 936)	923	(842, 1099)	929	(837, 1118)
Skarsvatn	567	(522, 662)	553	(465, 759)	607	(487, 916)
Tingvatn	487	(456, 521)	496	(462, 575)	505	(455, 599)
Gumbel						
Atnasjo	240	(222, 260)	237	(213, 267)	243	(214, 281)
Bulken (1991-2014)	489	(481, 497)	497	(462, 543)	495	(460, 541)
Bulken	491	(475, 514)	492	(471, 517)	494	(466, 528)
Driva	397	(378, 420)	396	(371, 430)	400	(368, 440)
Femundsenden	60	(58, 63)	60	(57,63)	60	(57,65)
Refsvatn	909	(878, 945)	920	(843, 1018)	921	(837, 1027)
Skarsvatn	503	(473, 564)	482	(438, 540)	515	(454, 613)
Tingvatn	506	(475, 541)	509	(469, 564)	514	(463, 582)
200 year return level						
GEV						
Atnasjo	414	(351, 494)	455	(325, 917)	456	(321, 961)
Bulken (1991-2014)	705	(689, 721)	764	(597 ,1387)	760	(594 ,1378)
Bulken	583	(557, 702)	585	(548, 646)	602	(547, 739)
Driva	666	(572, 787)	655	(546, 861)	680	(531, 977)
Femundsenden	79	(67, 86)	77	(70, 90)	78	(66, 99)
Refsvatn	1152	(1092, 1222)	1193	(998, 1824)	1215	(1000, 1897)
Skarsvatn	1056	(899, 1591)	1002	(680, 2198)	1219	(742, 3319)
Tingvatn	568	(521, 622)	580	(513, 821)	599	(510, 873)
Gumbel						
Atnasjo	330	(302, 361)	326	(288, 375)	337	(289, 398)
Bulken (1991-2014)	603	(593, 613)	617	(560, 692)	614	(557, 689)
Bulken	654	629, 690)	657	(623, 696)	658	(615, 713)
Driva	550	(519, 589)	549	(507, 603)	556	(503, 622)
Femundsenden	80	(76, 85)	79	(75, 85)	80	(74, 88)
Refsvatn	1200	(1154, 1252)	1217	(1092, 1378)	1222	(1087, 1393)
Skarsvatn	678	(631, 782)	645	(573,741)	700	(601, 864)
Tingvatn	637	(588, 692)	643	(578, 732)	651	(569, 762)
1000 year return level						
GEV						
Atnasjo	559	(431, 742)	647	(396, 1819)	635	(382, 1900)
Bulken (1991-2014)	883	(853,915)	988	(676, 2516)	983	(672, 2499)
Bulken	646	(613, 837)	649	(596, 744)	676	(596, 899)
Driva	882	(711, 1114)	857	(660, 1273)	904	(637, 1527)
Femundsenden	92	(73, 101)	89	(77, 111)	90	(872, 1269)
Refsvatn	1316	(1233, 1418)	1374	(1064, 2591)	1406	(1071, 2732)
Skarsvatn	1639	(1309, 3053)	1529	(850, 4922)	2014	(954, 8691)
Tingvatn	612	(554, 679)	628	(531, 1045)	650	(530, 1139)
Gumbel						
Atnasjo	393	(358, 431)	388	(340, 452)	402	(341, 480)
Bulken (1991-2014)	682	(670, 694)	700	(629, 797)	696	(624, 793)
Bulken	767	(736, 813)	771	(728, 821)	772	(717, 842)
Driva	656	(617, 705)	655	(602, 724)	663	(595, 749)
Femundsenden	93	(88, 100)	93	(87, 100)	94	(86, 104)
Refsvatn	1401	(1345,1465)	1422	(1265, 1627)	1429	(1259, 1650)
Skarsvatn	800	(741, 932)	758	(666, 881)	829	(703, 1043)
Tingvatn	728	(666, 797)	735	(653, 849)	746	(643, 888)