Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Mixture models for statistical flood frequency analysis

Note no SAMBA/31/16

Authors Silje Hindenes
Thordis L. Thorarinsdottir
Gunnhildur H. Steinbakk

Date July 8, 2016

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

The authors

Silje Hindenes is a master's student in Statistics at the Norwegian University of Science and Technology, Thordis L. Thorarinsdottir and Gunnhildur H. Steinbakk are Senior Research Scientists at the Norwegian Computing Center.

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.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Mixture models for statistical flood frequency analysis**
Authors	**Silje Hindenes** `<siljeh@nr.no>` **Thordis L. Thorarinsdottir** `<thordis@nr.no>` **Gunnhildur H. Steinbakk** `<gunnhildur@nr.no>`
Date	July 8, 2016
Publication number	SAMBA/31/16

Abstract

Statistical flood frequency analysis is commonly used to estimate design floods based on a series of annual maximum discharge data. The current guidelines for flood estimation in Norway recommend the use of a regional model when only a short series of at-site data is available, for 30-50 years of at-site data the two-parameter Gumbel distribution is the recommended statistical model while the three-parameter generalized extreme value (GEV) distribution should be used for more than 50 years of at-site data. This note investigates a mixture model approach that combines all three models with the mixture weights depending on the amount of available at-site data. The mixture weights are derived by assessing the predictive performance of the three models on out-of-sample data using proper scoring rules. In a case study of 60 discharge series from Norway it is found that the resulting weighting scheme depends heavily on the scoring rule that is used to obtain the weights. A trend similar to that of the current guidelines is most apparent when using scoring rules that focus on the upper tail, i.e. the quantile score and the quantile-weighted continuous ranked probability score. When the models are estimated based on 10-50 years of data, the weight corresponding to the regional model increases as the number of observations in the training set decreases and opposite for the local GEV and Gumbel models. The Gumbel model receives larger weights than the GEV model, indicating and overall better performance.

The drawing on the front page by John William Edy shows Drammenselva in 1800.

Keywords	Dam safety, flood frequency analysis, mixture model, proper scoring rules
Target group	Researchers in hydrology and statistics
Availability	Open
Project	FlomQ
Project number	220662
Research field	Teknologi, industri og forvaltning
Number of pages	20
© Copyright	Norwegian Computing Center

Contents

1 Introduction

In Norway, building regulations for infrastructure such as dams, bridges and roads close to water bodies include a legal obligation to account for hazard management related to flood risk in the design of the structure. For instance, bridges must be designed to withstand large floods that happen on average once every 200 years, the so-called 200 year flood, while dams should be able to withstand the 1000 year flood, see Stenius et al. (2015) and references therein.

The estimation of the design flood is commonly performed via statistical flood frequency analysis (FFA) by fitting a distribution function to a series of annual maximum discharge data. The Norwegian Water Resources and Energy Directorate (NVE) issues guidelines on how the design flood estimation should be performed, see Midttømme et al. (2011). The recommended choice of a statistical model depends on the available discharge data at the catchment, see Table 1.

Table 1. The Norwegian recommendations for design flood estimation with statistical flood frequency analysis (FFA), see also Midttømme et al. (2011).

Data	Method
$>$ 50 years	GEV estimation of at-site series
30 - 50 years	Gumbel estimation of at-site series
10 - 30 years	Use other long series in the area
$<$ 10 years	Use other long series in the area/ regional FFA
No data	Regional FFA

In this note, we will investigate a mixture model approach that combines a regional model, the Gumbel model and the generalized extreme value (GEV) model where the mixture weights change depending on the amount of available data at the catchment. The mixture weights are determined by the relative predictive performance of each model assessed on out-of-sample data using proper scoring rules (Gneiting and Raftery, 2007). Different scoring rules assess different aspects of the predictive performance and the strength of the penalization may also vary. For these reasons, we compare weighting schemes under various types of proper scoring rules using both scores that assess the full distribution as well as those that focus on the upper tail or a single quantile only. Similarly, we compare different estimation approaches for the local Gumbel and GEV models.

The remainder of the note is organized as follows. The next Section 2 describes the three statistical FFA models, the proper scoring rules we consider and the parameter estimation approaches. The data set is introduced in the following Section 3 and the results are presented in Section 4 along with discussions. Finally, conclusions are provided in the last Section 5.

2 Methods

2.1 Statistical flood frequency analysis

2.1.1 Local analysis using Gumbel and GEV distributions

Statistical flood frequency analysis (FFA) models the distribution of annual maximum discharge by fitting a statistical distribution function F to a series $y_1, \dots y_n$ of observed annual maximum discharge. Block maxima of this type are commonly modelled using the generalized extreme value (GEV) distribution or a special case thereof, the Gumbel distribution (Coles, 2001). For the GEV model, F is given by

$$F(y) = \begin{cases} \exp\big(-\big[1+\varepsilon\Big(\frac{y-\mu}{\sigma}\Big)\big]^{-1/\xi}\big) & \text{if } \xi \neq 0 \\ \exp\big(-\exp\big(-\frac{y-\mu}{\sigma}\big)\big) & \text{if } \xi = 0. \end{cases} \tag{1}$$

where $\mu \in \mathbb{R}$ is location, $\sigma > 0$ is scale and ξ is shape, and we assume that $1+\xi(y-\mu)/\sigma > 0$ for $\xi \neq 0$. The Gumbel distribution is equal to the special case in (1) where $\xi = 0$.

The aim of statistical FFA is commonly to estimate the size, or the return level, of a certain design flood. The return level for a design flood with return period of T years equals the quantile function value, $F^{-1}(p)$, at the probability $p = 1 - 1/T$. That is, the quantile that has probability $1/T$ of being exceeded in any given year. The quantile function of the GEV model is given by

$$F^{-1}(p) = \begin{cases} \mu + \frac{\sigma}{\xi}\big[1 - \big(-\log(p)\big)^{-\xi}\big] & \text{if } \xi \neq 0 \\ \mu - \sigma\log\big(-\log(p)\big) & \text{if } \xi = 0 \end{cases} \tag{2}$$

and the quantile function of the Gumbel model is the special case in (2) with $\xi = 0$.

We estimate the parameters of the GEV and the Gumbel model independently at each catchment using two frequentist estimation methods and Bayesian inference. The frequentist estimation is performed using maximum likelihood estimation and probability weighted moments (Hosking et al., 1985) as implemented in the `R` package `fExtremes`. The Bayesian inference is performed as described in Steinbakk et al. (2016).

2.1.2 Regional analysis using the GEV distribution

If design flood estimates are required at locations for which no or only a very limited amount of data is available, a regional FFA model is required. Here, we apply a regional hierarchical GEV model (Dyrrdal et al., 2015) where the parameters μ, σ and ξ of the GEV distribution in (1) are given by regression equations of the type

$$\mu_s = \theta_0^\mu + \theta_1^\mu x_{1s} + \cdots \theta_k^\mu x_{ks} + \tau_s^\mu, \quad \tau_s^\mu \sim \mathcal{N}(0, \eta_\mu^2)$$

for the catchment s, and similar for $\log(\sigma_s)$ and ξ_s. The covariates $x_{1s}, \dots, x_{ks}$ present hydrological, geographical or meteorological information on the catchment s that are available even if the discharge has not been measured. The random effect τ_s^μ is used to account for the fact that there might be some variability in the parameter values between the different catchments that has not been captured by the covariates $x_{1s}, \dots, x_{ks}$.

By pooling together data from a large number of catchments for which both discharge data and covariates are available, the parameters $\theta_0^\mu, \ldots, \theta_k^\mu$ and η_μ^2 may be estimated via Bayesian inference returning a posterior sample of size I for each parameter. These samples together with covariate values $x_{1s'}, \ldots, x_{ks'}$ may then be used to infer posterior samples of size I for $\mu_{s'}$, $\sigma_{s'}$ and $\xi_{s'}$ at a catchment s' where no discharge data is available and, subsequently, equation (2) may be used to obtain posterior distributions of return levels. The Bayesian inference procedure includes a model selection component where in each MCMC iteration, a subset of the k potential covariates is selected for each of the GEV model parameters. The resulting posterior distributions thus represent a large mixture model where the different components of the mixture include different sets of covariates for each GEV model parameter. However, due to the very large number of potential models, it is not feasible to estimate a single most likely model and we employ the full posterior mixture. For analysis purposes, the marginal inclusion probability of each covariate may be calculated for each of the GEV model parameters.

Table 2. Covariates used in the regional Bayesian GEV model and posterior inclusion probabilities (%) for the location parameter μ_s, the scale parameter σ_s and the shape parameter ξ_s.

Covariate	μ_s	σ_s	ξ_s
Constant	100	100	100
Longitude	87	100	7
Latitude	100	100	5
Effective lake percentage	99	100	1
Catchment length	14	50	3
Inflow	10	61	3
Average precipitation in May	50	97	7
Average precipitation in August	99	48	4
Average snowmelt in April	4	12	15
Average runoff in April	32	6	4
Proxy for catchment gradient	41	94	2
Average fraction of rain	2	13	68

We employ a regional model with the covariates listed in Table 2. The model is estimated at 239 locations in Norway where the catchment area is larger than 50 km^2 and more than 20 years of discharge data are available. This data set includes the data used in the current study. Instead of including the random effects τ_s in the regressions equations, we sample from $\mathcal{N}(0, \eta_\mu^2)$ to obtain out-of-sample estimates. The marginal posterior inclusion probabilities for each covariate and each GEV model parameter are given in Table 2.

2.2 Scoring rules

To assess the predictive performance of the GEV, the Gumbel and the regional model, we use scoring rules. A scoring rule is a function, $S(F, y)$, that assigns a score to a predictive distribution F based on an observation y, and is oriented such that a lower value is a better score. It is (strictly) proper if the expected score of F for the observation y is minimized if (and only if) F is the true distribution of y (Gneiting and Raftery, 2007).

A common scoring rule is the logarithmic score, or the ignorance score as it is known in meteorological applications, defined as

$$S_{\text{IGN}}(f, y) = -\log(f(y)),$$

where f is the predictive density, see e.g. Friederichs and Thorarinsdottir (2012) and references therein. Another widely used scoring rule is the continuous ranked probability score (CRPS) given by

$$S_{\text{CRPS}}(F, y) = \int_{-\infty}^{\infty} [F(z) - \mathbb{1}\{y \leq z\}]^2 \mathrm{d}z,$$

see Gneiting and Raftery (2007) and references therein. It evaluates the performance of F on its entire domain.

Since we are trying to predict extreme floods, say the 1000-year flood, we are more interested in the forecaster's ability to predict the exceedance of a certain threshold or quantile. This can be assessed using the Brier score or the quantile score respectively. The Brier score is defined as

$$S_{\text{B}}^{u}(F, y) = (p_u - \mathbb{1}\{y \geq u\})^2,$$

where u is the threshold of interest and $p_u = 1 - F(u)$ is the predicted probability of y exceeding that threshold. The quantile score for a given quantile τ is defined as

$$S_{\text{Q}}^{\tau}(F, y) = \rho_\tau(y - F^{-1}(\tau)),$$

where $\rho_\tau(u) = \tau u$ if $u \geq 0$ and $\rho_\tau(u) = (\tau - 1)u$ otherwise, see e.g. Friederichs and Thorarinsdottir (2012) and references therein.

An alternative representation of the CRPS is obtained using the quantile score

$$S_{\text{CRP}}(F, y) = 2\int_{0}^{1} \rho_\tau(y - F^{-1}(\tau))\mathrm{d}\tau, \tag{3}$$

see e.g. Friederichs and Thorarinsdottir (2012) and references therein. A weighted version of the (3) gives the quantile-weighted CRPS

$$S_{\text{qwCRP}}(F, y) = 2\int_{0}^{1} w(\tau)\rho_\tau(y - F^{-1}(\tau))\mathrm{d}\tau,$$

where $0 \leq w(\tau) \leq 1$ is a weight function, see Lerch et al. (2015) and references therein. If the interest is primarily on the upper tail of the predictive distribution, a reasonable choice for the weight function is $w(\tau) = \mathbb{1}\{\tau \geq q\}$ for some high quantile q.

2.3 Mixture model

We will investigate the possibility of replacing the guidelines in Table 1 with a seamless transition model, or a mixture model, with

$$F = \omega_1 F_{\text{GEV}} + \omega_2 F_{\text{Gumbel}} + \omega_3 F_{\text{regional}}, \tag{4}$$

where $\omega_1 + \omega_2 + \omega_3 = 1$. Here, we will assume that the weights ω_1, ω_2 and ω_3 depend on the amount of available discharge data. For instance, if a very long series is available, a large weight should be put on the GEV model F_{GEV}. Similarly, if no data are available, we should have $\omega_1 = \omega_2 = 0$ and $\omega_3 = 1$.

One way to estimate the weights is to consider the relative performance of the three models, such that the largest weight is put on the model that obtains the best average score over a test set. Let S_{GEV}, S_{Gumbel} and S_{regional} denote the respective average scores of F_{GEV}, F_{Gumbel} and F_{regional}, obtained using some scoring rule. Since scoring rules are negatively oriented, the mixture weights are then given by the ratio of the inverse scores

$$\omega_1 = \frac{S_{\text{GEV}}^{-1}}{S_{\text{GEV}}^{-1} + S_{\text{Gumbel}}^{-1} + S_{\text{regional}}^{-1}}, \tag{5}$$

$$\omega_2 = \frac{S_{\text{Gumbel}}^{-1}}{S_{\text{GEV}}^{-1} + S_{\text{Gumbel}}^{-1} + S_{\text{regional}}^{-1}}, \tag{6}$$

$$\omega_3 = \frac{S_{\text{regional}}^{-1}}{S_{\text{GEV}}^{-1} + S_{\text{Gumbel}}^{-1} + S_{\text{regional}}^{-1}}. \tag{7}$$

Another way to estimate the weights is to minimize

$$\frac{1}{n}\sum_{i=1}^{n} S(F, y_i),$$

with respect to ω_1, ω_2 and ω_3. Here y_i, $i = 1, ..., n$, denotes the test set and F is the mixture model given in equation (4). The minimization is done using `constrOptim` in the `R` package `stats`, with the constraints being $0 \leq \omega_j \leq 1$ and $\sum_j \omega_j = 1$, $j = 1, ..., 3$.

In order to investigate how the mixture weights change depending on the amount of available data, we employ the following method. We limit our study to catchments for which we have at least 75 years of data available. Given a series of annual maximum discharge data, $y_1, \ldots, y_n$, we use the last 50 observations as training data, while the preceding 25 observations are used as a test set. Based on the training set, F_{GEV} and F_{Gumbel} are estimated as described in Section 2.1.1, and the regional model is used as described in Section 2.1.2. The weights are estimated using the two approaches above with various scoring rules. We then repeat this with reduced training sets of size 40, 30, 20 and 10.

3 Data

All data used in this analysis is extracted from the national hydrological data base at the Norwegian Water Resource and Energy Directorate. To estimate the local GEV and Gumbel models, we use maximum annual discharge data from catchments greater than 50 km^2 with at least 75 years of observations and, resulting in a total of 60 catchments. The discharge values is usually given in the units of $1000\,l/s$ or $l/s/km^2$ if the data is normalized by the catchment area. For each catchment there is also hydrological, geographical or meteorological information available, which gives the covariate values used in the regional model.

4 Results and discussion

To estimate the parameters for the GEV and the Gumbel model with Bayesian inference, we draw 50,000 MCMC samples from the posterior parameter distributions. This is done independently for each catchment by using 50, 40, 30, 20 and 10 years of data. We would like to compare the results obtained with these estimates with similar estimates using both PWM and MLE. However, there seems to be an error in the implementation of `gumbelFit` in the `R` package `fExtremes`, leading to strange results. Thus, we only present the results obtained with Bayesian inference and maximum likelihood estimation.

4.1 Mixture weights estimated by relative performance

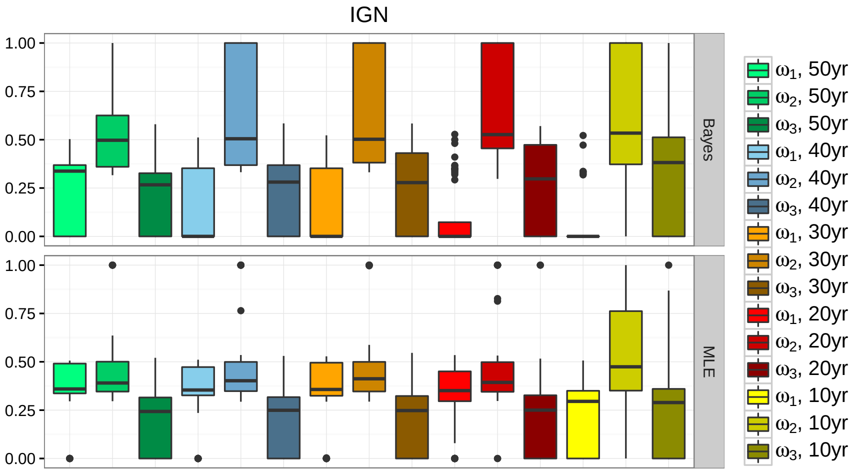

Figure 1. Box plots of the estimated weights ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model) when the ignorance score is used as the scoring rule. From left to right we see how the weights change when the years of data used in the analysis is reduced. The upper and lower results are obtained under Bayesian inference and MLE, respectively, for the parameter estimation in the local GEV and Gumbel models.

The ignorance score may yield negative scores when the range of the probability distribution of interest is small. Negative scores will make equations (5), (6) and (7) meaningless. In order to make all scores positive, we add a small constant to every score. This is not ideal, since the resulting weights will depend on the chosen constant. We choose the smallest possible constant, since a large constant will erase much of the difference in the scores. Box plots of the estimated weights using the ignorance score with this alternation is given in Figure 1. With all 50 years of data included in the training set, the Gumbel model receives the largest weight and the regional model receives the smallest weight. From left to right, ω_1 decreases while ω_2 and ω_3 increases. The weights change more when

Bayesian inference is used to estimate the parameters for the local models, compared to when MLE is used. For both estimation methods, the median of ω_3 never exceeds the median of ω_2, which does not agree with the guidelines in Table 1.

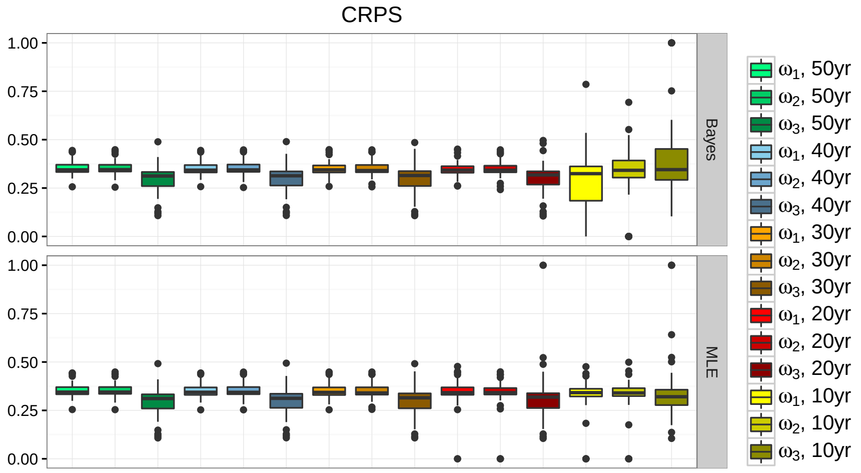

Figure 2. Box plots of the estimated weights ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model) when using CRPS as the scoring rule. From left to right we see how the weights change when the years of data used in the analysis is reduced. The upper and lower results are obtained using Bayesian inference and MLE, respectively, for the parameter estimation in the local GEV and Gumbel models.

Figure 2 presents the estimated mixture weights when CRPS is used as the scoring rule. Overall, the GEV and the Gumbel model seems to perform better than the regional model, except for the case when only 10 years of data are used to estimate the models with the Bayesian inference approach. For this particular setting, the regional model gives better estimates and there is some increase in the variation of the weights, due to numerical problems when estimating the parameters based on only 10 observations.

The difference in the three weights reduces slightly as the number of years in the training set is reduced. This somewhat agrees with the guidelines in Table 1, which suggests to use the regional model when little data is available. However, the GEV and Gumbel model seems to be approximately equally good for all settings, thus the recommended change from the GEV to the Gumbel model is not obvious in these results.

The difference in the scores assigned by CRPS does not distinguish greatly between the models. The CRPS evaluates the performance of each model on the entire predictive distribution. Thus the results in Figure 2 tells us that the GEV and the Gumbel model have similar performance in terms of predicting return levels for all return periods, and that they almost always performs better than the regional model. However, we are more interested in predicting extreme values, and it is therefore more interesting to look at scoring rules that evaluate the upper tail.

The Brier score allows us to compare the different models' ability to predict return levels above a certain threshold u. In order to make the scores from each station comparable, the threshold u must be the same for all stations. Figure 3 shows the resulting weights

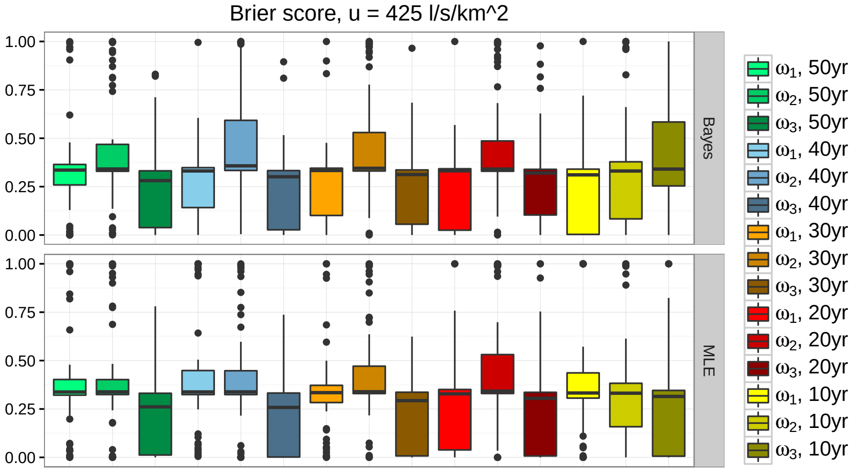

Figure 3. Box plots of the estimated weights ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model) when the Brier score with threshold $u = 425\ l/s/km^2$ is used as the scoring rule. From left to right we see how the weights change when the years of data used in the analysis is reduced. The upper an lower results are obtained under Bayesian inference and MLE, respectively, for the parameter estimation in the local GEV and Gumbel models.

under the Brier score with threshold $u = 425\ l/s/km^2$. This threshold corresponds to the return period $T = 5$ for the training data from all stations pooled together. It would be more interesting to measure the performance with a threshold corresponding to a longer return period. However, the magnitude of the annual maximum discharge varies a lot from station to station. The largest observed value from each catchment ranges from $75\ l/s/km^2$ to $1897\ l/s/km^2$, thus the threshold $425\ l/s/km^2$ is high for some stations while low for others. In total, for 21 stations the largest observed annual maximum discharge is less than $425\ l/s/km^2$. If a station has no observed values above the threshold, the Brier scores of its estimated distributions, F_{GEV}, F_{Gumbel} and F_{regional}, are approximately zero, since the predicted probabilities p_u are likely close to zero and $\mathbb{1}\{y \geq u\}$ is always zero. Each model thus receives a good score, resulting in approximately equal weights, but the performance is evaluated on a range with no observations and thus tells us little about the relative performance of the three models.

From the results in Figure 3 we see that when Bayesian inference is used the Gumbel model is given the largest median weight, except for when only 10 observations are available, when again the regional model seems to be the best choice. Overall, the three weights are approximately centered around $1/3$, with ω_1 and ω_3 skewed to the left and ω_2 to the right. This can be explained by the fact that a lot of the stations do not have observations above the threshold, resulting in approximately equal weights for the three models, as explained above. These results contribute to the median of each weight such that it is close to $1/3$. Thus the way the box plots are skewed tells us something about the relative scores for the stations which have observations above the threshold.

The quantile score measures the different models' ability to predict events above a certain quantile, where the quantile corresponds to different return values for each station. In

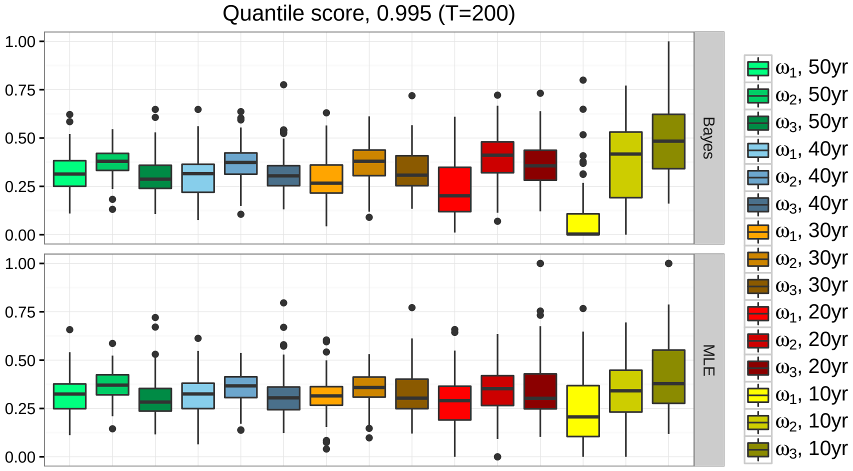

Figure 4. Box plots of the estimated weights ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model) when the quantile score at the quantile $\tau = 0.995$ is used as the scoring rule. From left to right we see how the weights change when the years of data used in the analysis is reduced. The upper and lower results are obtained under Bayesian inference and MLE, respectively, for the parameter estimation in the local GEV and Gumbel models.

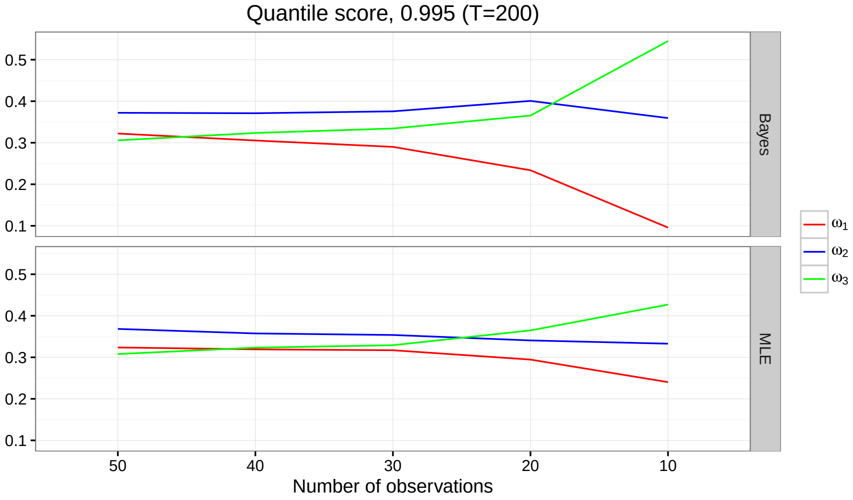

Figure 5. The average value of ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model), estimated using the quantile score with $\tau = 0.995$, plotted against the number of observations used to estimate the parameters of the GEV and the Gumbel model. The upper and lower results are obtained using Bayesian inference and MLE, respectively.

Figure 4 we see the estimated weights when this scoring rule with the quantile $\tau = 0.995$, corresponding to the return period $T = 200$, is applied. Here we see more variation in the weights as the number of years included in the training set is reduced.

Figure 5 shows how the average weights changes as the number of observations is reduced. The weights seem to follow the assumed trend, namely that the regional model receives the largest weight when less than 10 years of data is available and that the Gumbel model is preferred otherwise. When MLE is used to estimate the parameters for the local models, the largest weight is put on the regional model even when 20 years of data is available. For both MLE and Bayesian inference, the GEV starts out with a larger weight than the regional model, but this relationship shifts as the number of observations reduces from 50 to 40.

Figure 6 presents the resulting weights when the quantile-weighted CRPS is applied, with weight functions $w(\tau) = \mathbb{1}\{\tau \geq 0.8\}$, $w(\tau) = \mathbb{1}\{\tau \geq 0.9\}$ and $w(\tau) = \mathbb{1}\{\tau \geq 0.99\}$. The results for $\tau \geq 0.8$ and $\tau \geq 0.9$ are similar to the results for the CRPS, while the results for $\tau \geq 0.99$ resemble the results obtained with the quantile score. For each of these results we see some of the expected change in the weights, and this change becomes more apparent as we move our focus further out in the upper tail.

If we look at how the mean weights plotted against the amount of data used in the analysis, as displayed in Figure 7 for the quantile-weighted CRPS with$w(\tau) = \mathbb{1}\{\tau \geq 0.9\}$ and $w(\tau) = \mathbb{1}\{\tau \geq 0.99\}$, we see that ω_3 exceeds ω_1 earlier for the latter weight function. The shift in the relationship between ω_2 and ω_3 occurs at the same point in both cases.

4.2 Mixture weights estimated by minimizing the average score

The weights obtained by minimizing the ignorance score of the mixture model are given in Figure 8. Here, we see a high variability in the estimated weights for all settings. This can be explained as follows. For each catchment, the weight tends to be put on only one model. However, the model that receives the large weight changes from catchment to catchment.

It is easier to compare the results by looking at how the average weights change when the number of available observations reduces, as shown in Figure 9. The weight corresponding to the regional model, ω_3, increases as the number of observations is reduced. The two other weights alternate between increasing and decreasing, but in total they decrease from 50 to 10 observations.

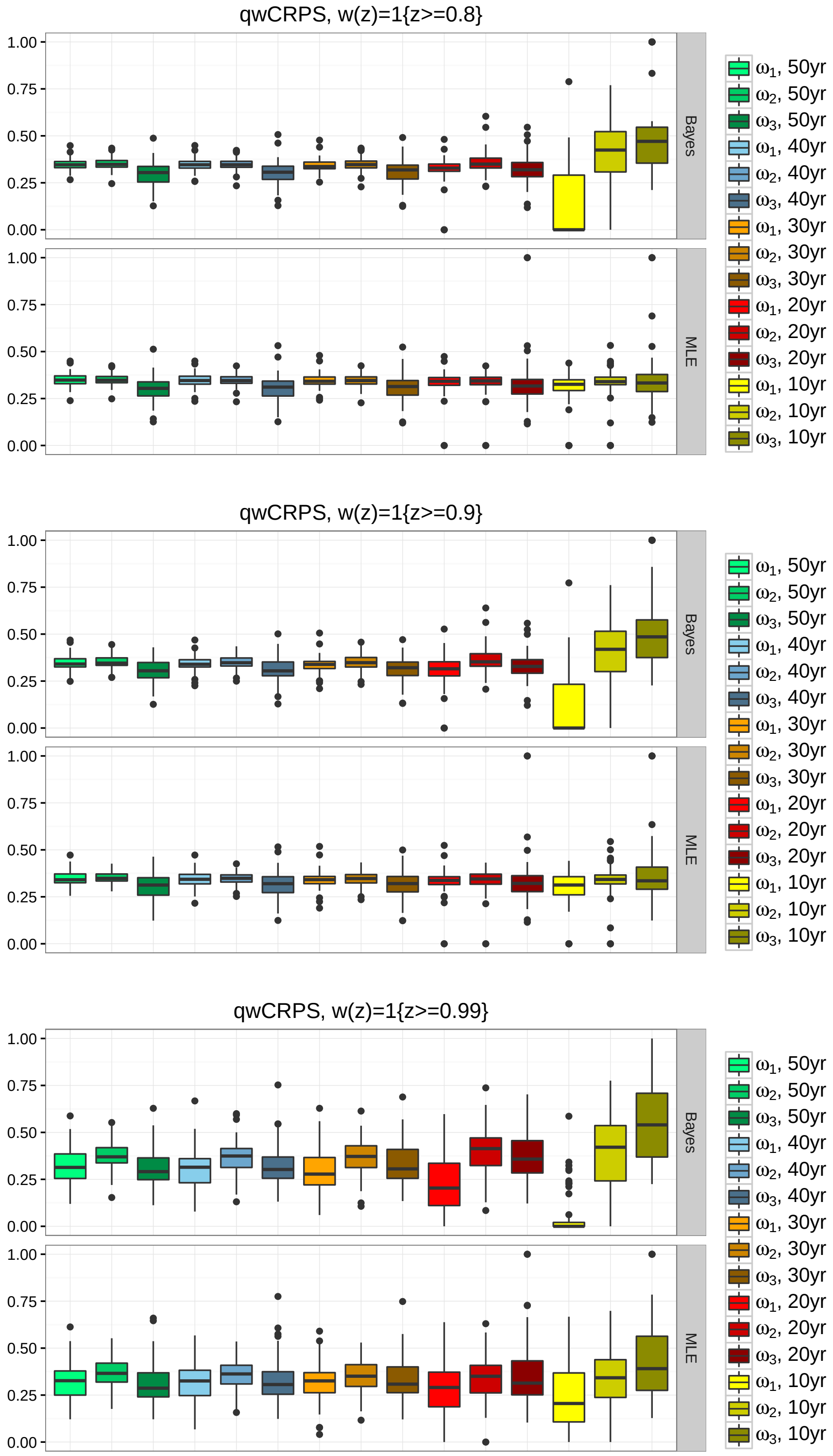

Figure 6. Box plots of the estimated weights ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model) under the quantile-weighted CRPS with weight functions $w(\tau) = \mathbb{1}\{\tau \geq 0.8\}$ (top), $w(\tau) = \mathbb{1}\{\tau \geq 0.9\}$ (middle) and $w(\tau) = \mathbb{1}\{\tau \geq 0.99\}$ (bottom). From left to right we see how the weights change when the years of data used in the analysis is reduced. In each plot, the upper and lower panels show results obtained under Bayesian inference and MLE, respectively, for the parameter estimation in the local GEV and Gumbel models.

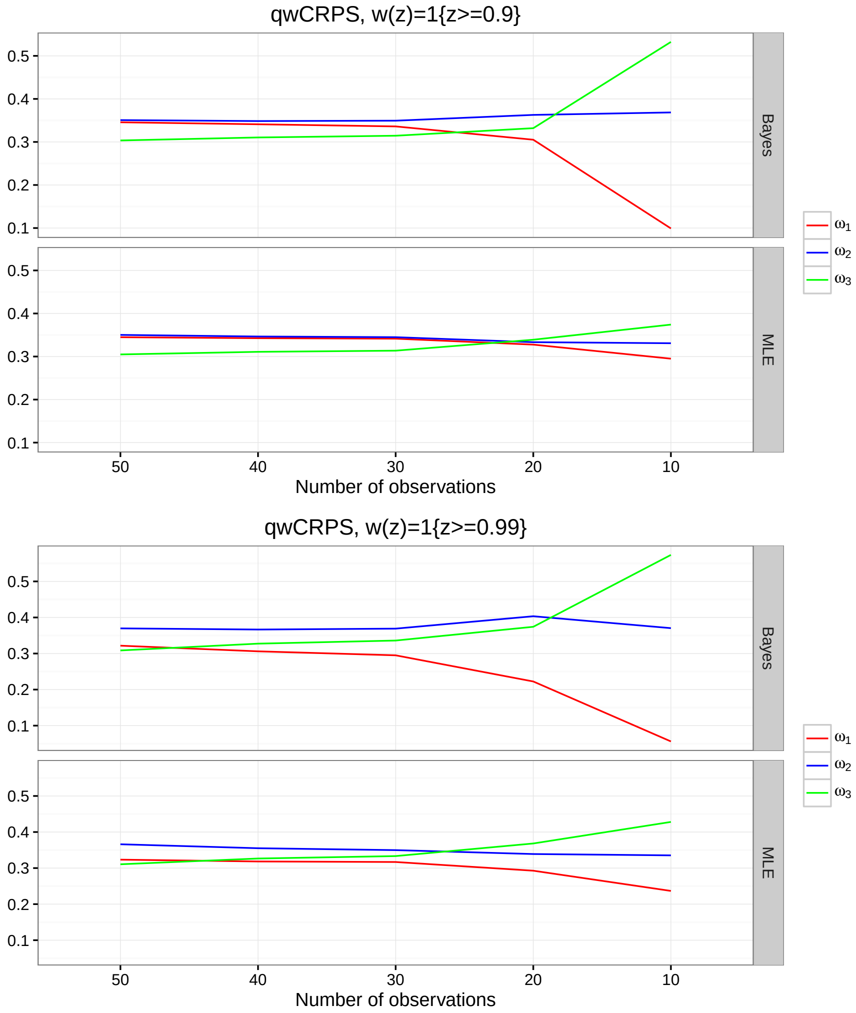

Figure 7. The average value of ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model), estimated using the quantile-weighted CRPS with $w(\tau) = \mathbb{1}\{\tau \geq 0.9\}$ (top) and $w(\tau) = \mathbb{1}\{\tau \geq 0.99\}$ (bottom), plotted against the number of observations used to estimate the parameters for the GEV and the Gumbel model. The upper and lower panel in each plot are obtained with Bayesian inference and MLE, respectively.

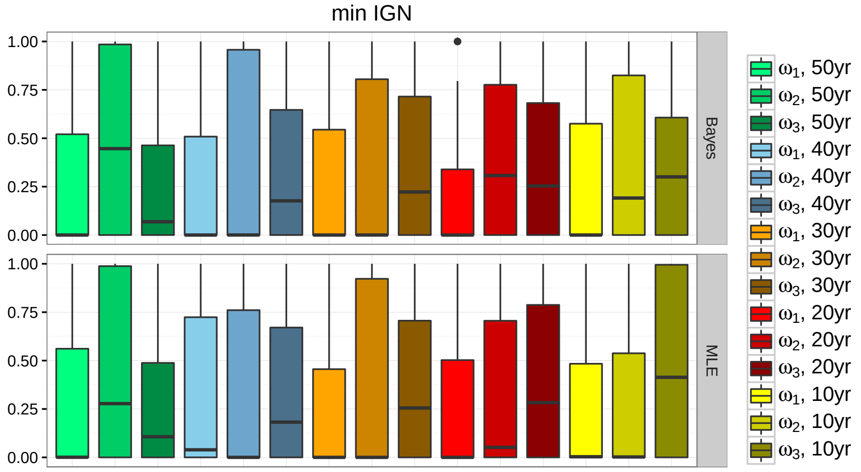

Figure 8. Box plots of the estimated weights ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model) when minimizing the average ignorance score of the mixture model over the test set. From left to right we see how the weights change when the years of data used in the analysis is reduced. The upper and lower results are obtained under Bayesian inference and MLE, respectively, for the parameter estimation in the local GEV and Gumbel models.

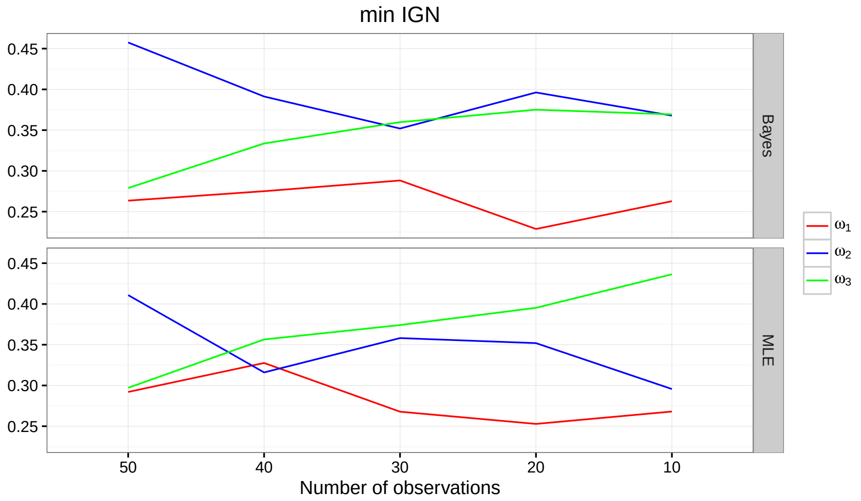

Figure 9. The average value of ω_1 (for the GEV model), ω_2 (for the Gumbel model) and ω_3 (for the regional model), estimated by minimizing the average ignorance score, plotted against the number of observations used to estimate the parameters for the GEV and the Gumbel model. The upper and lower results are obtained with Bayesian inference and MLE, respectively.

5 Conclusion

The different estimation methods for the mixture weights, with different scoring rules, suggest quite different weighting schemes. The trend we expect to see based on the current guidelines for flood frequency analysis is most apparent when using scoring rules that focus on the upper tail, i.e. the quantile score and the quantile-weighted CRPS for high quantiles. In general, the weight corresponding to the regional model increases when the number of observations in the training set is reduced, while the two other weights decrease. The Gumbel model tends to receive larger weight than the GEV model, indicating that it overall performs better than the GEV model. The weight put on the Gumbel model is relatively large also when we look at only 10 years of data.

For future work it can be interesting to investigate how the weights change if we include more than 50 years of data in the training set. This can for example be done by using all 75 observations from the catchments used in this study as training data and apply cross validation methods to asses the performance of the different models. The best estimates for the GEV, the Gumbel and the regional model might depend on the mixture weights. Thus we could try estimation approaches where we estimate all the parameters simultaneously, instead of assuming that the parameters for three models are given when estimating the mixture weights.

References

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer-Verlag, London.

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.

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with application to probabilistic peak wind prediction. *Environmetrics*, 23:579–594.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.*, 102:359–378.

Hosking, J., Wallis, J. R., and Wood, E. F. (1985). Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*, 27(3):251–261.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2015). Forecaster's dilemma: extreme events and forecast evaluation. arXiv:1512.09244.

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

Steinbakk, G. H., Thorarinsdottir, T. L., Reitan, T., Schlichting, L., and Engeland, K. (2016). Propagation of rating curve uncertainty in design flood estimation. Technical report, NR. Note nr. SAMBA/03/16.

Stenius, S., Glad, P. A., Wang, T. K., and Væringstad, T. (2015). *Veileder for flomberegninger i små uregulerte felt*. NVE. Veileder nr. 7-2015.