NTNU

Norwegian University of
Science and Technology

Mixture models for flood frequency analysis

A case study for Norway

Silje Hindenes

Master of Science in Physics and Mathematics
Submission date: June 2017
Supervisor: Ingelin Steinsland, IMF
Co-supervisor: Thordis Thorarinsdottir, Norsk Regnesentral

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This study is the result of the course *TMA4905 - Statistics, Master's Thesis* at the Norwegian University of Science and Technology (NTNU). I would like to thank my supervisor Thordis L. Thorarinsdottir at Norwegian Computing Center for her excellent guidance throughout this study, and my supervisor at NTNU, Ingelin Steinsland, for informative comments and discussions. I also want to thank Kolbjørn Engeland at the Norwegian Water Resources and Energy Directorate (NVE) for providing the data used in this analysis and for helpful comments on the hydrological aspect of this thesis.

Silje Hindenes, June 28, 2017.

Abstract

Flood frequency analysis (FFA) concerns prediction of the magnitude and corresponding frequency of extreme flood events. Extreme floods can be the result of various hydrological processes. In Norway, rainfall and snowmelt are considered to be the two main flood generating processes. The use of mixture models, to account for these different flood generating processes, are investigated for catchments in Norway. For the case of annual maximum series (AMS), a two-component mixture of Gumbel distributions is fitted, by assuming that the mixture weights are both known and unknown. Subsequently, for peaks over threshold (POT) series a two-component mixture of exponential distributions is considered. Again, the two cases of known and unknown mixture weights are studied. When assuming that the mixture weights are known, these are given by the precalculated proportion of rainfall and snowmelt contributing to each flood value. The mixture models are compared to the generalized extreme value (GEV) distribution and the Gumbel distribution for AMS, and to the generalized Pareto (GP) distribution and the exponential distribution for POT. Maximum likehood is used for parameter estimation, and for the mixture models with unknown weights the maximum likehood estimates are obtained by the expectation maximization (EM) algorithm. The predictive performance of the models are compared using various scoring rules. In addition, the stability of the models are compared. We found that although the scoring rules are not always able to differentiate between the models, the Gumbel distribution and the exponential distribution, for the case of AMS and POT respectively, often give the most reliable and stable estimates. The mixture models estimated by the EM algorithm occasionally give unexpected results and seem unfit for practical use in FFA.

Sammendrag

Flomfrekvensanalyse omhandler prediksjon av størrelse og korresponderende frekvens for ekstreme flommer. Ekstreme flommer kan være resultatet av ulike hydrologiske prosesser. I Norge er regnnedbør og snøsmelting regnet som de to dominerende flomgenererende prosesser. Bruk av blandingsmodeller, for å ta hensyn til disse ulike flomgenererende prosessene, er undersøkt for vassdrag i Norge. For modellering av årlig maksimumsserier (AMS), ser vi på en to-komponent Gumbel-fordeling hvor vi antar at komponentenes vekter både er kjent og ukjent. Videre, for modellering av flommer over en bestemt terskel ("peaks over threshold", POT), undersøker vi en to-komponent eksponential-fordeling. Også her vurderer vi tilfellet med kjente vekter og ukjente vekter. Når vi antar at vektene er kjent, er de gitt ved kalkulert andel regnnedbør og snøsmelting som bidro til hver enkelt flomobservasjon. Blandingsmodellene sammenlignes med en "generalized extreme value" (GEV) fordeling og en Gumbel-fordeling for AMS, og med en "generalized Pareto" (GP) fordeling og en eksponentialfordeling for POT. Parameterene til de ulike modellene er estimert ved "maximum likelihood". For blandingsmodellene med ukjente vekter maksimeres likelihood ved bruk av "expectation maximization" (EM) algoritmen. Prediktiv ytelse for modellene sammenlignes ved bruk av ulike "scoring rules". I tillegg er stabiliteten til de ulike modellene studert. Resultatene viser at selv om "scoring rules" ikke alltid kunne skille mellom modellene, så gir ofte Gumbel-fordelingen og eksponentialfordelingen, for henholdsvis AMS og POT, pålitelige og stabile estimat. Blandingsmodellene estimert ved EM algoritmen gir noen uventede resultater, og kan derfor ikke anbefales til bruk i flomfrekvensanalyse.

Contents

1 Introduction **1**

2 Data **5**
2.1 Flood observations 5
2.2 Flood generating process 5
2.3 Explorative analysis for three gauging stations 7

3 Modelling of annual maximum series **11**
3.1 The generalized extreme value distribution 11
3.2 Mixture model 12
3.2.1 Maximum likelihood estimation with known weights 13
3.2.2 Expectation Maximization (EM) 14
3.2.3 EM algorithm for a mixture of Gumbel distributions 16
3.3 Return level estimation 18

4 Modelling of peaks over threshold series **20**
4.1 The generalized Pareto distribution 20
4.2 Mixture model 21
4.2.1 Maximum likelihood estimation with known weights 22
4.2.2 EM algorithm for a mixture of exponential distributions . . . 22
4.3 Return level estimation 24

5 Validation **26**
5.1 Scoring rules 26
5.2 Cross validation 29
5.3 Stability of return level estimates 31

6 Results **32**
6.1 Annual Maximum Series 32
6.1.1 Detailed look at three stations 37
6.2 Peaks Over Threshold 44
6.2.1 Detailed look at three stations 49
6.3 Return level estimates 56

7 Discussion **61**

8 Conclusion **63**

1 Introduction

Predictions of magnitude and corresponding frequency of extreme flood events are important for safety and risk assessments when designing structures close to rivers, such as bridges, dams, roads and railways. If a structure cannot withstand what is expected, lives are at stake as well as economic losses. The design flood for a structure is defined as the most extreme flood which that structure is required to withstand, in terms of the frequency of occurrence. A precise estimate of the magnitude of the design flood is desired, since both underestimation and overestimation leads to excessive costs. In the case of overestimation, the initial building costs could be much higher than necessary. On the other hand, if the design flood is underestimated, rebuilding costs, in addition to costs when not operating, will occur more frequently than expected.

Design floods are often specified by the return period p. By the p-year flood we mean a flood that on average occurs every p years. Dam safety regulations in Norway require the estimation of the 500- and 1000-year flood, depending on the individual dam safety class (Lovdata, 2009). Buildings and infrastructure should resist or be protected from floods with 20, 200 or 1000 year return period, depending on the consequence of a flooding, according to building regulations in Norway (TEK 10) (Lovdata, 2010). Gauging stations in Norway commonly have about 100 years or less of recorded data. The estimation of extreme floods thus requires extrapolating far outside the range of recorded flow data.

Flood frequency analysis is a statistical approach to estimate the magnitude of such extreme floods. When a sufficient amount of historical flood data is available at the site of interest, a local analysis, which involves fitting a probability distribution to the given discharge series, can be applied. Otherwise a regional analysis must be performed, which uses discharge series from the same region in addition to hydrological, meteorological and geographical covariates to estimate the underlying distribution. Here we only consider local analysis, and therefore only study gauging stations in Norway where sufficient historical data are available.

Annual maximum series (AMS) and peaks over threshold (POT) are the most commonly used methods for flood frequency analysis. They differ in the way the flood series are constructed, and thus they apply different distributions to model the data. In the AMS approach, a distribution is fitted to a series of annual maximum flood values. Fisher and Tippett (1928) formed the theoretical basis for AMS, by showing that the limiting distribution of block maxima of identically and independently distributed (iid) random variables belongs to the generalized extreme value (GEV)

family of distributions. Later, (Gumbel, 1945) applied this theory to floods. Alternatively, POT selects all mutually independent flood peaks above a chosen threshold. A series of such flood peaks is modelled by the generalized Pareto (GP) distribution (Pickands, 1975). POT series are also referred to as partial duration series (PDS).

The GEV distribution have a location, scale and shape parameter, while the GP distribution have a scale and a shape parameter. When setting the shape parameter of these distributions to zero, the Gumbel and exponential distribution, respectively, are obtained. These distributions are also frequently used in FFA. For small sample sizes, distributions with less parameters are often preferred (see e.g. Midttømme et al., 2011). In general, a distribution with more parameters is more flexible, which implies that it is also more likely that the estimation procedure overfits the data. For AMS, Cunnane (1989) showed that the Gumbel distribution is effective for small samples, while the GEV distribution is preferred for sample sizes greater than 50. For the case of POT series, Rosbjerg et al. (1992) concluded that the exponential distribution is preferable to the theoretical correct generalized Pareto distribution, for small sample sizes and moderately long-tailed exceedance distributions.

One of the challenges with the AMS approach to extreme value modelling is that every yearly maximum might not be an extreme flood value. In addition, some years could have more than one extreme value, such that one leaves out valuable information when only considering one peak each year (see e.g. Lang et al., 1999). Peaks over threshold modelling addresses these problems by considering only peaks that exceeds some threshold. Thus low yearly maximum values might not be included in the analysis, and more than one peak from years with high flood values can be included. However, the simplicity of the AMS method makes it popular compared to POT. POT has the advantage of including more data in the analysis, but it requires the selection of a threshold and some criteria to define consecutive peaks as independent.

The data set used in our analysis is provided by The Norwegian Water Resources and Energy Directorate (NVE), and a chosen threshold is given for each POT series in the data set. NVE's method for threshold selection is based on a high quantile for each station. The quantile is adjusted such that 2-6 flood peaks are included each year (Engeland et al., 2016), which resulted in the use of the 98 percent quantile as the threshold. This agrees with current recommendations in the literature. For instance, Cunnane (1973) recommends to include at least 1.65 floods each year, while FEH (1999) suggests to include 2-6 flood peaks each year. To assure that the flood peaks are independent, NVE used a criteria for independence based on Lang et al. (1999). Two consecutive flood peaks must be separated with at least three times the time-to-rise and the discharge value must have decreased to 2/3 of the last flood peak.

AMS and POT have been compared in a variety of studies, see e.g. Bobée and Rasmussen (1995), Ferreira and de Haan (2015) and Bezak et al. (2014) for extensive reviews and comparisons of the two approaches to FFA. Cunnane (1973) compared return level estimates for AMS and POT, and found that if the POT series contains at least 1.65 times the records as the AMS, then the sampling variance of the p year flood is smaller for POT. Similar results was stated by Madsen et al. (1997). They found that for the case of maximum likelihood estimation, the POT approach yields more efficient estimators for the p-year flood. A study by Caires (2016) agrees with these results, and in addition concludes that the performance of the two methods are similar for large sample sizes (over 200 years). Overall, it seems that the POT method is preferable for small sample sizes, as long as the average number of threshold exceedances each year is greater than 1.65.

When applying the GEV and GP distributions, an assumption is that the flood values arise from the same distribution. In reality, this assumption may not be justified as a flood event can be the result of various hydrological processes (see e.g. Alila and Mtiraoui, 2002). In Norway, rainfall and snowmelt are considered to be the two main flood generating processes (FGP), which can cause an extreme flood either alone or in combination (Engeland et al., 2016). Various attempts to take FGP or seasonal variations into considerations when modelling flood data have been made by e.g. Rossi et al. (1984), Waylen and Woo (1982) and Evin et al. (2011).

Mixture models are commonly applied in cases where the data is considered as arising from multiple sub populations instead of one homogeneous population. They provide an efficient and flexible modelling tool, able to estimate e.g. multiple modes. Since it is assumed to be two dominating FGP in Norway, it is natural to consider mixture models with two components. For this reason, we investigate the use of mixture models with two components, mimicking rainfall and snowmelt, for the case of both AMS and POT data. Mixture models require the estimation of more parameters, and to limit the amount of parameters, we consider only mixtures of the Gumbel and the exponential distribution.

We estimate the parameters of the proposed mixture models both by assuming that the mixture weights are know and not known. For the case where the weights are known, they are given by the proportion of accumulated rainfall and snowmelt, respectively, in a time frame before each discharge value. When assuming that we do not know the mixing proportions, the weights will be estimated simultaneously with the other parameters by the Expectation Maximization (EM) algorithm (Dempster et al., 1977). The resulting models are compared to each other as well as to the

traditional models for FFA.

To estimate the parameters of the GEV, Gumbel, GP and exponential distribution, a number of different methods are available. NVE's guidelines for flood estimation in Norway suggest to use the method of L-moments, maximum likelihood estimation or a Bayesian analysis (Steinius et al., 2015). See e.g. Landwehr et al. (1979), Engeland et al. (2004), Hosking et al. (1985) and Gubareva and Gartsman (2010) for comparison and discussions of the commonly used parameter estimation methods. We apply the method of maximum likelihood estimation, such that the same estimation approach is used for the traditional models and for the mixture models.

In total, we compare 4 different models for AMS, namely GEV, Gumbel and a two-component mixture of Gumbels with both known and unknown mixture weights, and 4 different models for POT, GP, exponential and a two-component mixture of exponentials with both known and unknown mixture weights. The comparison is performed based on estimates from AMS and POT series at 228 gauging stations in Norway. The models for AMS and POT are compared using a 10-fold cross validation. We apply different proper scoring rules (Gneiting and Raftery, 2007) as loss functions in the cross validation procedure. For a few chosen catchments in Norway, we compare return level estimates obtained by AMS and POT.

This thesis is structured as follows. Section 2 describes the data used in our analysis, while Section 3 provides the theoretical background for the AMS approach to flood frequency analysis along with the proposed mixture models for AMS modelling. This section also presents the general theory of the EM algorithm. Similarly, Section 4 provides the theory behind POT modelling and presents the two mixture models we investigate for the POT data. The methods used for validation are introduced in Section 5, with various scoring rules given in Section 5.1 and the cross validation procedure explained in Section 5.2. Section 6 provides the results, where the results for AMS and POT modelling are given in Section 6.1 and 6.2, respectively. In Section 6.3 we compare return level estimates at different locations for the AMS and POT models. At last, a discussion and conclusion is given in Sections 7 and 8, respectively.

2 Data

2.1 Flood observations

All data used in this analysis are provided by NVE's database Hydra II. This database provides flood data in the form of annual maximum series (AMS) and peaks over threshold (POT). A total of 530 gauging stations are included in the data set, where 266 of these stations are still in use. We only consider stations where both the AMS and POT series have at least 30 records, as it is recommended to instead apply a regional analysis when limited data is available (see e.g. Midttømme et al., 2011). This results in a total of 228 gauging stations used in our analysis. Figure 1 presents histograms of the length of the AMS (left) and POT (right) series from these stations.

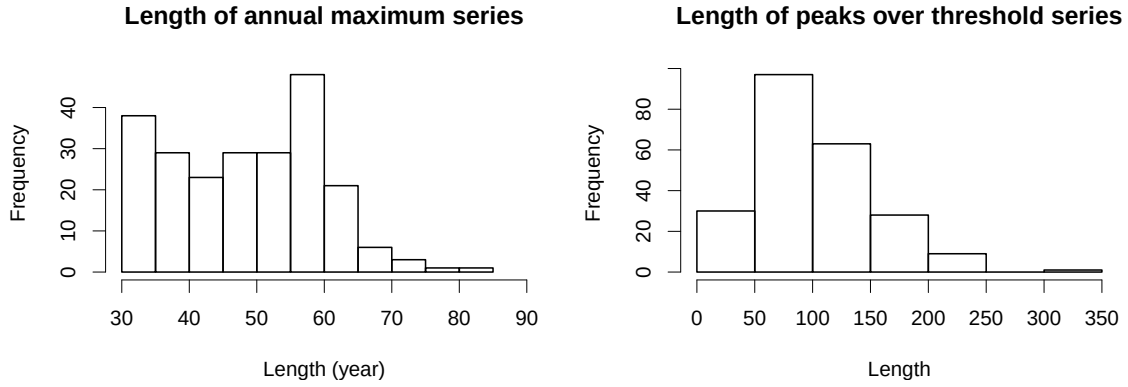

Figure 1: Histograms showing the length of the annual maximum series, to the left, and the peaks over threshold series, to the right, used in our analysis.

Both the AMS and POT data set are constructed based on daily average discharge values, given in the units m^3/s from various catchments in Norway. The AMS approach to flood frequency analysis (FFA) considers only each yearly maximum of these daily average values. Alternatively, the POT method considers all discharge values above a chosen threshold that in addition is considered to be mutually independent. A threshold is included for each POT series in the data set, as described in Section 1

2.2 Flood generating process

An extreme flood can be the result of a number of different events (either alone or in combination), such as extreme rainfall, snowmelt and landslide. In Norway,

snowmelt and rainfall are considered to be the two main flood generating processes (Engeland et al., 2016). By assuming that only rain and snowmelt contributes to flooding, NVE have estimated the proportion of rainfall and snowmelt, respectively, that contributed to each flood. This is represented by a number between 0 and 1, which indicates whether rainfall (1) or snowmelt (0), or a combination of the two, is the flood generating process for a specific flood peak. This variable, "Flood generating process (FGP)", is included in the data set for both the AMS and POT data. The FGP is estimated based on a time frame before a flood peak, by calculating the accumulated rainfall and the accumulated snowmelt in the period (Engeland et al., 2016).

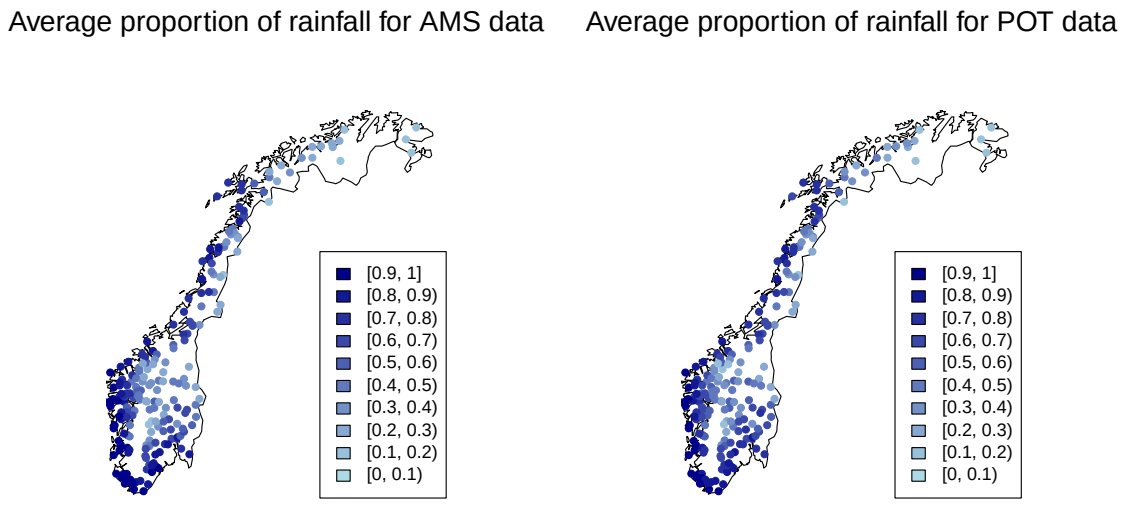

Figure 2: The average of the FGP value at each catchment for the AMS (left) and POT (right) data. 1 corresponds to a rainfall being the flood generating process, while 0 corresponds to snowmelt.

The average FGP value, where 1 and 0 corresponds to rainfall and snowmelt respectively, for each catchment used in our analysis are plotted on a map of Norway in Figure 2. From this figure we see that the coastal catchments are dominated by rainfall, while in the inland and northern parts of Norway we find catchments that are more influenced by snowmelt. The two plots in Figure 2 have the same pattern, as many of the flood peaks are included in both the AMS and POT series.

In our analysis, we only estimate the proposed models based on data where we have values for FGP, in order to make the models using the FGP variable comparable to those that do not use this variable.

2.3 Explorative analysis for three gauging stations

To illustrate the data in the data set, we take a detailed look at three specific catchments, Bulken, Atnasjø and Krinsvatn, which demonstrate various trends in the flood values. The location of these catchments are shown in Figure 3 below. Bulken is located in western Norway, an area dominated by rainfall. Atnasjø is located 701 meters above sea level in the inland and central part of Norway, where rainfall is less dominating. North of Atnasjø we find Krinsvatn, 87 meters above sea level. Krinsvatn is, similarly to Bulken, located on the coast and dominated by rainfall.

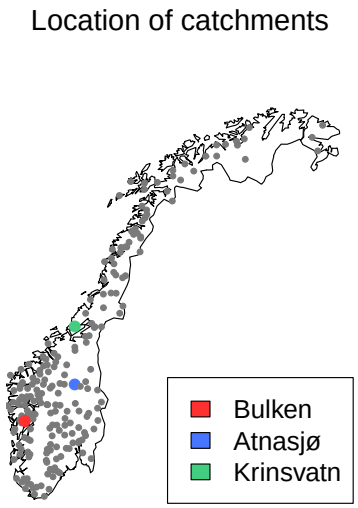

Figure 3: A map of Norway giving the locations of the catchments used in our analysis. Bulken, Atnasjø and Krinsvatn are highlighted in red, blue and green, respectively.

Table 1 below presents information about the size of the three catchments, and in what period the flood data have been recorded. In addition, the length of both the AMS and POT series, and the number of observations that also have a corresponding FGP value, for each catchment are given. Summary statistics of the flood data from the same catchments are given in Table 2. More specifically, it provides the mean, median and standard deviation (SD) of both the AMS and POT series, where the observations with no corresponding FGP value are removed from each series.

Table 1: Detailed information about three catchments, Bulken, Atnasjø and Krinsvatn. n_{AMS} and n_{POT} denotes the length of the AMS and POT series respectively, while the number in parentheses represents the corresponding number of observations with a FGP value in the data set.

Catchment	Area (km^2)	Period	n_{AMS} (FGP)	n_{POT} (FGP)	Threshold
Bulken	1092.04	1892-2015	124 (58)	298 (172)	253.91
Atnasjø	463.2	1917-2015	99 (61)	153 (87)	47.15
Krinsvatn	206.61	1916-2015	100 (65)	367 (204)	69.66

Table 2: Summary statistics for the AMS and POT series from the three catchments, Bulken, Atnasjø and Krinsvatn.

	AMS			POT		
Catchment	Mean (m^3/s)	Median (m^3/s)	SD (m^3/s)	Mean (m^3/s)	Median (m^3/s)	SD (m^3/s)
Bulken	369.19	372.47	99.00	329.71	305.45	71.50
Atnasjø	70.49	65.34	26.53	68.47	62.08	22.15
Krinsvatn	136.15	121.99	52.72	107.26	98.88	39.50

For the three catchments, Figure 4 and 5 gives the flood values and FGP values plotted against the day of year for the AMS and POT series, respectively. Bulken is dominated by rainfall floods throughout the year, except for in the summer months when snowmelt also contributes to flood peaks. Flood peaks from Atnasjø are mainly obtained in the summer. Both rainfall and snowmelt floods, in addition to a mixture of the two, are found in this period. Krinsvatn is, similarly to Bulken, dominated by rainfall floods with a few snowmelt and mixture floods in the summer. The trend in the FGP values for each catchment does not depend on whether the data is given in the form of AMS or POT series.

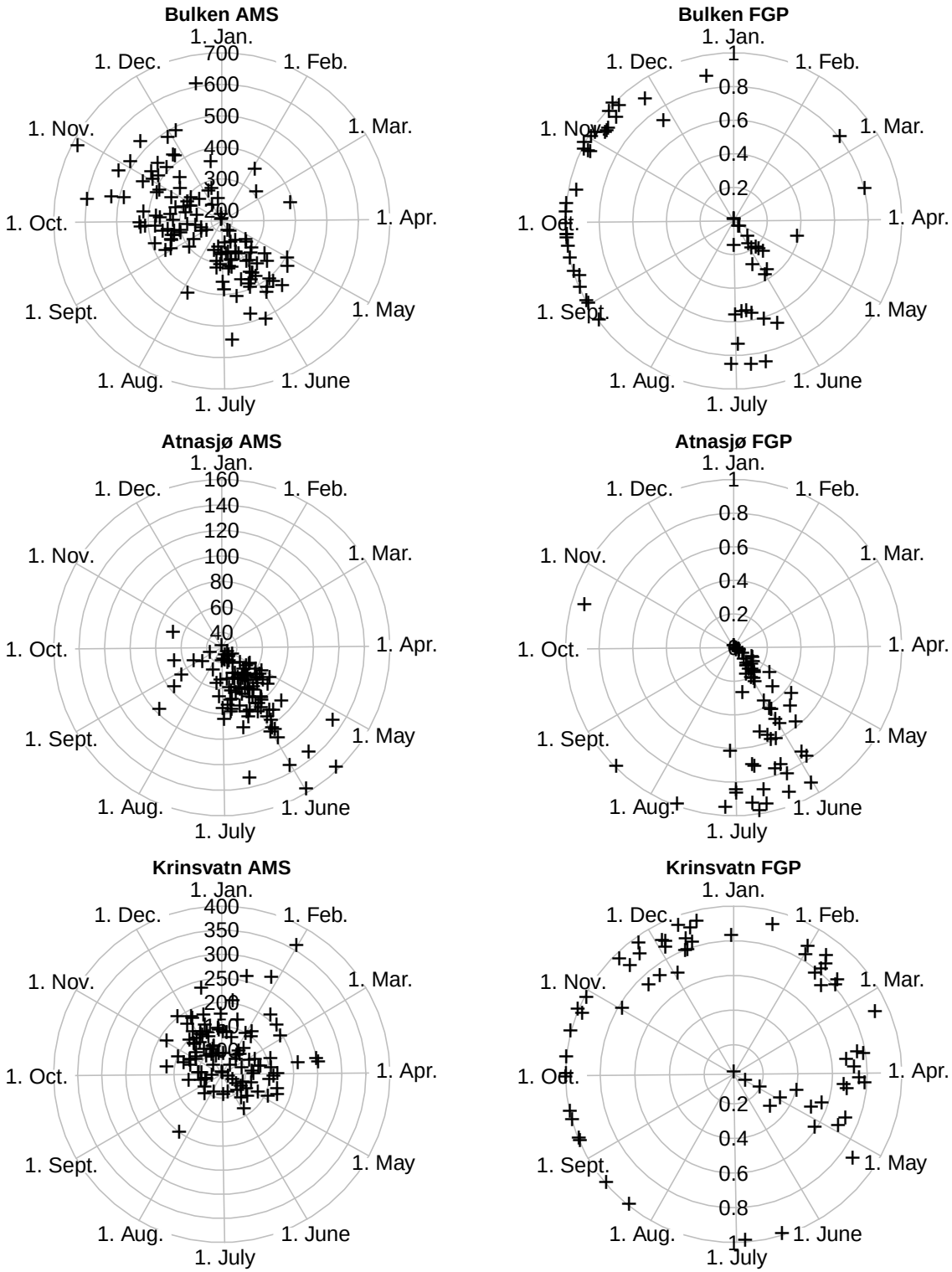

Figure 4: In the left plots, the annual maximum series from the catchments Bulken, Atnasjø and Krinsvatn respectively, are plotted against the date each flood event occurred. They are plotted in polar plots, where 360 degrees represents one year. The magnitude of the flood values are given in the units m^3/s. Similarly, to the right, the corresponding FGP values for each catchment are plotted against the date.

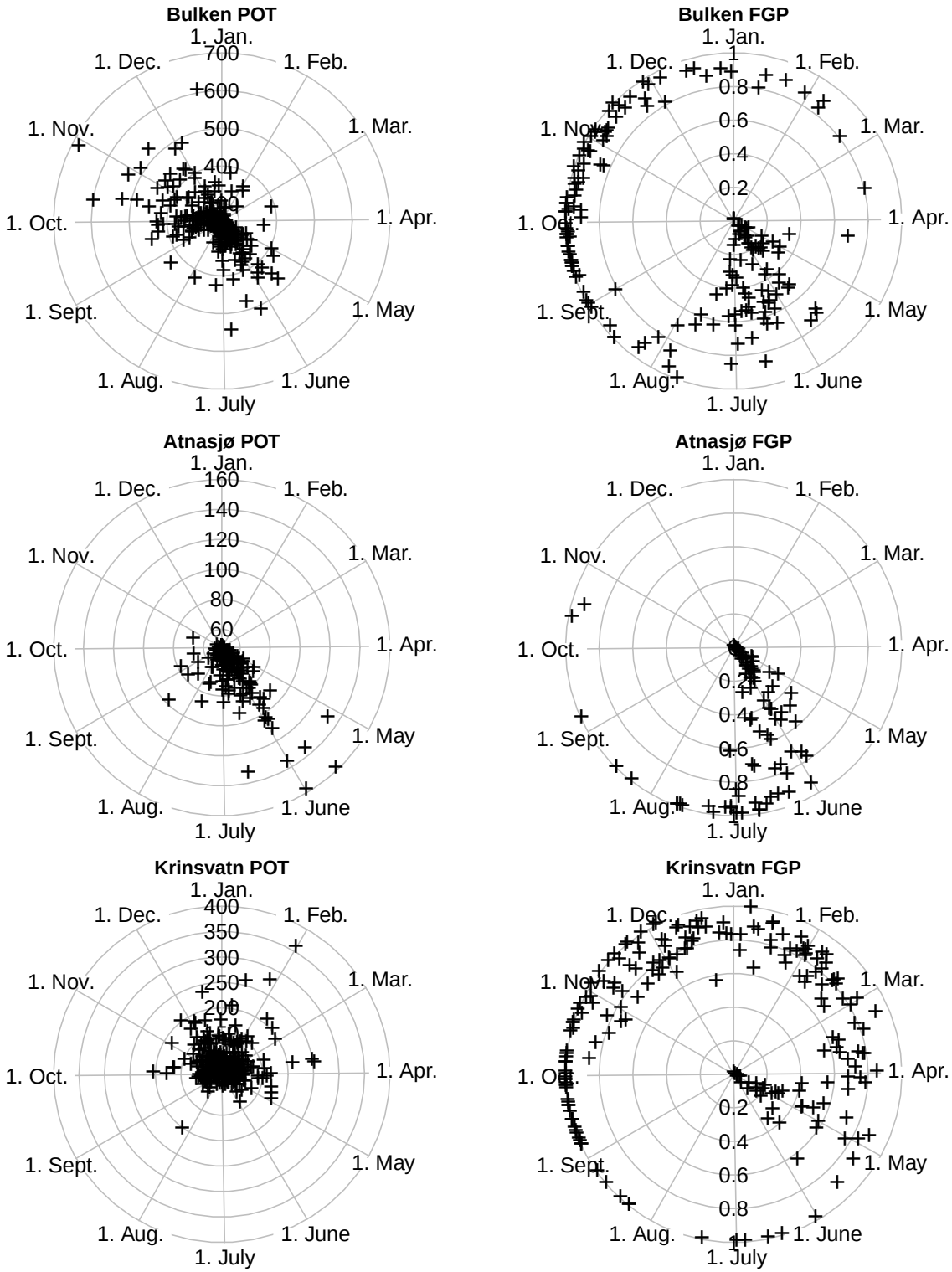

Figure 5: In the left plots, the peaks over threshold series from the catchments Bulken, Atnasjø and Krinsvatn respectively, are plotted against the date each flood event occurred. They are plotted in polar plots, where 360 degrees represents one year. The magnitude of the flood values are given in the units m^3/s. Similarly, to the right, the corresponding FGP values for each catchment are plotted against the date.

3 Modelling of annual maximum series

3.1 The generalized extreme value distribution

Consider the block maximum

$$Y_m = \max(X_1, \ldots, X_m),$$

where $X_1, \ldots, X_m$ is a series of identically and independently distributed random variables. In our context, the X_i's denotes the daily average flood values, and Y_m the yearly maximum of these. That is, we apply a block maximum period m of one year. This is commonly used in FFA, since a year is considered long enough to assume that the maximum values are independent of each other.

We are interested in the distribution of the annual maximum Y_m. Below, we follow the derivation described in Coles (2001). For the original derivation, see Fisher and Tippett (1928). If the distribution of X_i is known, e.g. $P(X_i \leq z) = G(z)$, then the distribution of Y_m is given by

$$P(Y_m \leq z) = P(X_1 \leq z) \cdots P(X_m \leq z) = G^m(z).$$

However, we do not know the distribution of the daily average flood values. Coles (2001) instead looks for a limiting distribution of $G^m(z)$ when $n \to \infty$. For a normalization of Y_m, $\frac{Y_m - b_m}{a_m}$, where $\{a_m > 0\}$ and $\{b_m\}$ are sequences of some constants, Coles (2001) states that if

$$P\left(\frac{Y_m - b_m}{a_m} \leq y\right) \to F(y), \quad \text{as} \quad m \to \infty$$

where F is a non-degenerate distribution function, then F is a member of the generalized extreme value (GEV) distribution family. Note that this does not imply that the limit must exists. However, if the limit exists it belongs to this distribution family, which is given by

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

with support $1 + \xi(y - \mu)/\sigma > 0$,where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are the location, the scale and the shape parameter, respectively. The special case $\xi = 0$ in equation (1) is called the Gumbel distribution.

When modelling AMS by the GEV distribution, one assumes that the flood values are independent and identically distributed. By only considering the block maximum , with a sufficient block period, the independence criteria is satisfied. In addition, it is assumed that all annual maximum discharge values at a specific location have the same stationary distribution, as it has not been possible to detect a clear climate change signal in the observed magnitude of annual flood events (Wilson et al., 2010).

The shape parameter ξ determines the tail behaviour of the GEV distribution. When $\xi > 0$, the distribution is bounded from below by $y = \mu - \sigma/\xi$, and when $\xi < 0$ it is bounded from above by $y = \mu - \sigma/\xi$. In the case where $\xi = 0$, there are no restrictions on y. Thus in practice, if ξ is negative, there is a finite maximal value for the annual maximum flood, while if $\xi \geq 0$ the maximum flood can be infinitely large.

Both the 3-parameter GEV distribution and the simpler 2-parameter Gumbel distribution are commonly used to to model AMS. The Gumbel distribution have the advantage of ease of fit, but is not as flexible as the GEV distribution. We consider the use of both these distributions, in addition to a mixture of Gumbels given in Section 3.2 below.

The parameters of the GEV and Gumbel distributions at each catchment are estimated using maximum likelihood. The complexity of the support of the GEV distribution makes the likelihood maximization not straightforward. Therefore we make use of the `ismev` (Heffernan and Stephenson, 2016) package in `R` (R Core Team, 2016), which have implemented procedures for ML estimation of the extreme value distributions. We apply the function `gev.fit` for the GEV distribution and `gum.fit` for the Gumbel distribution.

3.2 Mixture model

In this thesis we want to investigate the use of mixture models in FFA. For the case of AMS, we consider a mixture of Gumbel distributions. The density of a Gumbel random variable is given by

$$f(y; \mu, \sigma) = \frac{1}{\sigma} \exp\left[-\left(\frac{y-\mu}{\sigma} + \exp\big(-\frac{y-\mu}{\sigma} \big) \right) \right]. \tag{2}$$

where μ and $\sigma > 0$ denotes the location and scale parameter, respectively (see e.g. Coles, 2001). A finite mixture of k Gumbel distributions can be written as

$$f(y;\boldsymbol{\theta}) = \sum_{j=1}^{k} \omega_j f_j(y;\mu_j,\sigma_j), \tag{3}$$

$$\sum_{j=1}^{k} \omega_j = 1,$$

with each $\omega_j > 0$. $f_j(y;\mu_j,\sigma_j)$ denotes the Gumbel distribution given in (2) with location μ_j and scale σ_j, and ω_j denotes the mixture weight of the jth component.

In cases where mixture models are considered, the number of components is often unknown and the problem also involves estimating k. In Norway, as described in Section 2.2, rainfall and snowmelt are considered to be the main flood generating processes. This yields the natural selection of two components for the mixture model, and in the following we therefore only consider the case of $k = 2$.

We consider both the case where the mixture weights are known and the case of unknown weights. The FGP variable included in the data set provides a natural estimate for the mixture weights. For the case of given mixture weights, the parameters $\boldsymbol{\theta} = (\mu_1, \sigma_1, \mu_2, \sigma_2)$ are estimated using maximum likelihood, described in Section 3.2.1 below. When the mixture weights are assumed to be unknown, the parameters $\boldsymbol{\theta} = (\omega_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$ are estimated by maximizing the likelihood in an iterative procedure, using the EM algorithm given in Section 3.2.2 below. Only one of the mixture weights needs to be estimated, as $\omega_2 = 1 - \omega_1$.

3.2.1 Maximum likelihood estimation with known weights

Here we assume that the weights ω_1 and ω_2 of the mixture distribution in Equation (3), with $k = 2$, are known and given by the FGP. Given a sample of annual maximum flood values, $\mathbf{y} = y_1, \ldots, y_n$, with corresponding weights $\omega_{1,1}, \ldots, \omega_{1,n}$ and $\omega_{2,1}, \ldots, \omega_{2,n}$, the likelihood is

$$L(\mathbf{y};\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{j=1}^{2} \omega_{j,i} f_j(y_i;\mu_j,\sigma_j), \tag{4}$$

where $\boldsymbol{\theta} = ((\mu_1, \sigma_1, \mu_2, \sigma_2))$ denotes the parameters to be estimated. By taking the logarithm of (4), we obtain the log likelihood function, $l(\mathbf{y};\boldsymbol{\theta})$, given by

$$l(\mathbf{y};\boldsymbol{\theta}) = \log(L(\mathbf{y};\boldsymbol{\theta})) = \sum_{i=1}^{n} \log \sum_{j=1}^{2} \omega_{j,i} f_j(y_i;\mu_j,\sigma_j) \tag{5}$$

The parameters $\boldsymbol{\theta}$ are estimated by optimizing the log likelihood in Equation (5) using `constrOptim` in `R`, with the constraints $\sigma_1 > 0$ and $\sigma_2 > 0$.

3.2.2 Expectation Maximization (EM)

The Expectation Maximization (EM) algorithm, introduced by Dempster et al. (1977), is a procedure for maximum likelihood estimation of parameters in problems with incomplete or missing data. That is, when some part of the data is not observable, such that the observed data represents an incomplete data set. The idea of the EM algorithm is to associate the incomplete data problem with a complete data problem for which the maximizing the likelihood is more straightforward. The algorithm estimates the parameters of the proposed model by maximizing the likelihood in an iterative procedure. It is applicable in a wide range of problems, (see e.g. Meng and Pedlow, 1992), and is commonly used for estimating parameters of mixture models.

Here we give a general formulation of the algorithm, similar to the one by McLachlan and Krishnan (1996). Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote a random sample from the observable random variable Y, with distribution $f_Y(y;\boldsymbol{\theta})$. Here $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \Omega$ denotes the parameters and Ω denotes the parameter space. Further, assume that there is some unobservable data, $\boldsymbol{z}$, with random variable Z, such that $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ denotes the complete data. Let $f_{Y,Z}(y, z;\boldsymbol{\theta})$ denote the distribution of the complete data.

The estimation task is to maximize the likelihood of the complete data, $L_c(\mathbf{x};\boldsymbol{\theta})$, or equivalently, maximizing the log-likelihood $l_c = \log L_c(\mathbf{x};\boldsymbol{\theta})$. Since the log likelihood of the complete data is unobservable, its expectation given the observed data $\mathbf{y}$ is instead considered. Given the current parameter values $\boldsymbol{\theta^{(k)}}$, let

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta^{(k)}}} \left[l_c(\mathbf{x};\boldsymbol{\theta}) | \mathbf{y} \right],$$

denote this expectation. Here $E_{\boldsymbol{\theta^{(k)}}}$ denotes the expectation with parameters $\boldsymbol{\theta^{(k)}}$.

In each iteration of the algorithm there are two steps, the *Expectation step* (E-step) and the *Maximization step* (M-step), thereby the name Expectation Maximization. In its general form, the algorithm is as follows. Firstly, initial values for the parameters must be chosen. Then, for each iteration k, do the following two steps

- **E-step** Compute the expected likelihood of the complete data, given the observed data,
$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = E_{\boldsymbol{\theta}^{(k)}} \left[l_c(\mathbf{x}; \boldsymbol{\theta}) | \mathbf{y} \right].$$
- **M-step** Maximize this expected likelihood with respect to the parameters $\boldsymbol{\theta}$. That is, choose $\boldsymbol{\theta}^{(k+1)} \in \Omega$ such that
$$Q(\boldsymbol{\theta}^{(k+1)}; \boldsymbol{\theta}^{(k)}) \geq Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}) \quad \forall \quad \boldsymbol{\theta} \in \Omega.$$

The two steps above, the E-step and the M-step, are repeated until some convergence criteria is reached.

Note that k in this section, and throughout the rest of this thesis, represents the iteration. In Section 3.2 k was used to denote the number of components in the mixture model, but as we only consider two components in this study, we use the number 2 instead of k where it is applicable.

To detect convergence of the algorithm, a stopping criteria must be chosen. This stopping criteria can be based on the change in the likelihood or the change in the parameters after an iteration. The parameter estimates depend on the choice of stopping criteria as well as the choice of starting parameters (see e.g. Seidel et al., 2000). Although it is not clear which stopping criteria is best to apply, one based on the change in the likelihood is most frequently used. The relative change in the log-likelihood provides a dimensionless measure of the change, and is therefore a suitable stopping criteria when comparing different estimation methods. The relative difference is given by
$$\frac{|\log L(\mathbf{y}; \boldsymbol{\theta}^{(k+1)}) - \log L(\mathbf{y}; \boldsymbol{\theta}^{(k)})|}{|\log L(\mathbf{y}; \boldsymbol{\theta}^{(k)})|}.$$

Dempster et al. (1977) have shown that the log-likelihood of the incomplete data is non-decreasing in each iteration. That is,
$$\log L(\mathbf{y}; \boldsymbol{\theta}^{(k+1)}) \geq \log L(\mathbf{y}; \boldsymbol{\theta}^{(k)}).$$

Thus, the log-likelihood sequence nearly always converges (see e.g Wu, 1983). The likelihood may have several local or global maxima and stationary points, such that the EM algorithm can converge to a local maximum or a stationary point instead of the desired global maximum, depending on the starting values for the parameters. Its sensitivity to starting values is a drawback of the EM algorithm. To overcome this problem, it is recommended to try several runs of the algorithm with different

starting points, as small perturbations from a saddle point will cause the algorithm to diverge away from this stationary point (see e.g. McLachlan and Krishnan, 1996).

3.2.3 EM algorithm for a mixture of Gumbel distributions

Consider the mixture of two Gumbel distributions,

$$f_Y(y;\boldsymbol{\theta}) = \sum_{j=1}^{2} \omega_j f_j(y;\mu_j,\sigma_j), \qquad (6)$$
$$\omega_1 + \omega_2 = 1.$$

where $\boldsymbol{\theta} = (\omega_1, \omega_2, \mu_1, \mu_2, \sigma_1, \sigma_2)$ and $\omega_1, \omega_2 \geq 0$. This can be formulated as an incomplete data problem, where in our context $\mathbf{y} = (y_1, \ldots, y_n)$ denotes a vector of observed annual maximum flood values. We introduce Z as a hidden state variable, such that $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ denotes the missing data vector. Each $\mathbf{z}_i$ is a two-dimensional indicator vector with first and second element equal to one/zero if the observation y_i did/did not arise from the first and second mixture component, respectively. That is,

$$\begin{aligned} z_{ij} &= 1, \quad \text{if } y_i \text{ belongs to the } j\text{th component,} \\ z_{ij} &= 0, \quad \text{if } y_i \text{ does not belong to the } j\text{th component.} \end{aligned}$$

The log-likelihood of the complete data, $\boldsymbol{x} = (\boldsymbol{y}, \boldsymbol{z})$, can now be written as

$$\begin{aligned} l_c(\boldsymbol{\theta}, \boldsymbol{z}) &= \sum_{i=1}^{n} \log f_{Y,Z}(y_i, z_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log \left(f_Z(z_i;\boldsymbol{\theta}) f_{Y|Z}(y_i|z_i;\boldsymbol{\theta}) \right) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{2} \log \left(\omega_j f_j(y_i;\mu_j,\sigma_j) \right)^{z_{ij}} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{2} z_{ij} \log \left(\omega_j f_j(y_i;\mu_j,\sigma_j) \right). \end{aligned}$$

As z is unobservable, the log-likelihood of the complete data can not be computed. Instead the EM algorithm considers the conditional expectation of $l_c(\boldsymbol{\theta}, \boldsymbol{Z})$ given the

complete data and current parameter values $\boldsymbol{\theta}^{(k)}$, where $\boldsymbol{Z}$ now is considered to be a random variable. The conditional expectation is given by

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= E_{\boldsymbol{\theta}^{(k)}}[l_c(\boldsymbol{\theta}, \boldsymbol{Z})|\boldsymbol{y}, \boldsymbol{\theta}^{(k)}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{2} P(z_{ij} = 1|y_i, \boldsymbol{\theta}^{(k)}) \log \omega_j f_j(y_i; \mu_j, \sigma_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \log \omega_j f_j(y_i; \mu_j, \sigma_j), \end{aligned}$$

where we define $h_{ij}^{(k)}$ to be the probability that y_i belongs to component j, given the current parameter estimates. That is, $h_{ij}^{(k)} = P(z_{ij} = 1|y_i, \boldsymbol{\theta}^{(k)})$. In the E-step we need to compute $h_{ij}^{(k)}$, in order to obtain the expected complete log-likelihood, $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$. In the M-step we maximize this with respect to $\boldsymbol{\theta}$.

With $f_j(y_i; \mu_j, \sigma_j)$ being the Gumbel distribution given in (2), we have

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \log \left[\frac{\omega_j}{\sigma_j} \exp\left[-\left(\frac{y_i - \mu_j}{\sigma_j} + \exp\Big(-\frac{y_i - \mu_j}{\sigma_j} \Big) \right) \right] \right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \left[\log \frac{\omega_j}{\sigma_j} - \left(\frac{y_i - \mu_j}{\sigma_j} + \exp\Big(-\frac{y_i - \mu_j}{\sigma_j} \Big) \right) \right]. \end{aligned}$$

To maximize this with respect to $\boldsymbol{\theta}$, we write

$$\begin{aligned} \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \left[\log \frac{\omega_j}{\sigma_j} - \left(\frac{y_i - \mu_j}{\sigma_j} + \exp\Big(-\frac{y_i - \mu_j}{\sigma_j} \Big) \right) \right] \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \left[\frac{y_i - \mu_j}{\sigma_j} + \exp\Big(-\frac{y_i - \mu_j}{\sigma_j} \Big) + \log \frac{\omega_j}{\sigma_j} \right], \end{aligned}$$

subject to $\sum_{j=1}^{2} \omega_j = 1$. The Lagrangian for this problem becomes

$$L = \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \left[\frac{y_i - \mu_j}{\sigma_j} + \exp\Big(-\frac{y_i - \mu_j}{\sigma_j} \Big) + \log \omega_j - \log \sigma_j \right] - \lambda(\sum_{j=1}^{2} \omega_j - 1), \quad (7)$$

where λ is the Lagrange multiplier constant. By evaluating the partial derivatives of L with respect to μ_j, σ_j, ω_j and λ, we obtain the following expressions for the parameters which optimize the expected log-likelihood of the complete data,

$$\omega_j = \frac{\sum_{i=1}^{n} h_{ij}}{\sum_{i=1}^{n}\sum_{l=1}^{2} h_{il}} = \frac{\sum_{i=1}^{n} h_{ij}}{n}, \tag{8}$$

$$\mu_j = \sigma_j \log\left[1 - \sum_{i=1}^{n} h_{ij} \exp\frac{-y_i}{\sigma_j}\right].$$

A closed form expression for σ_j is not obtained. The optimal estimate of σ_j in each iteration could be estimated numerically, and then used to obtain μ_j. Instead, we choose to optimize the expression in Equation (7) with respect to μ_j and σ_j, using the estimate for ω_j in Equation (8), for $j = 1$ and $j = 2$ separately. This optimization is performed using `optim` in the `R` package `stats` (R Core Team, 2016).

The E-step consists of updating the expression for Q with the new parameter estimates, which requires the calculation of $h_{ij}^{(k+1)}$, $i = 1, \ldots, n$, $j = 1, 2$. The updated h_{ij} is given by

$$h_{ij}^{(k+1)} = p(z_j = i | y_i; \boldsymbol{\theta}^{(k)}) = \frac{\omega_j f(y; \mu_j, \sigma_j)}{\omega_1 f(y; \mu_1, \sigma_1) + \omega_2 f(y; \mu_2, \sigma_2)}, \quad j = 1, 2.$$

In an effort to overcome the problem of the algorithm converging to stationary points or local maxima, we run the algorithm with 100 different randomly generated starting values. The parameter estimates are given by the run that resulted in the largest likelihood. Random starting values for μ_1 and μ_2 are obtained by sampling from a normal distribution with mean and standard deviation equal to the sample mean and standard deviation of $\mathbf{y}$, respectively. Similarly, different starting values for σ_1 and σ_2 are generated by sampling from a normal distribution with mean and standard deviation equal to the standard deviation and 1/10 of the standard deviation of $\mathbf{y}$, respectively. For σ_1 and σ_2, the absolute value of the random starting values are used, to assure that the conditions $\sigma_1 > 0$ and $\sigma_2 > 0$ are satisfied. The initial value of ω_1 is sampled from the uniform distribution on $[0, 1]$ in each run of the algorithm.

3.3 Return level estimation

The *return level* z_p, corresponding to the *return period* p, is the magnitude of the flood with exceedance probability $1/p$ each year. Thus, for the case of AMS, it is given by

$F(z_p) = 1 - 1/p$. In terms of the quantile function F^{-1}, we obtain $z_p = F^{-1}(1 - 1/p)$. For example, the magnitude of the 1000-year flood is given by $z_{1000} = F^{-1}(0.999)$. The quantile function of the GEV distribution is given by

$$F^{-1}(q) = \begin{cases} \mu + \frac{\sigma}{\xi}[1 - (-\log(q))^{-\xi}] & \text{if } \xi \neq 0 \\ \mu - \sigma \log(-\log(q)) & \text{if } \xi = 0, \end{cases} \tag{9}$$

where again, the case $\xi = 0$ corresponds to the Gumbel distribution.

For the two-component Gumbel mixture model, an analytical expression of the quantile function is not obtainable. Therefore, in order to estimate the return level for a specific return period, we sample from the mixture model in Equation (3) and estimate the quantile function by the empirical quantile function.

Return level estimates are often evaluated by plotting the return level as a function of the return period, on a logarithmic scale. Plotting positions of the observed data are then obtained by assigning the probability $\frac{i-0.35}{n}$ to an observation of rank i (Hosking et al., 1985), where n is the number of observations.

4 Modelling of peaks over threshold series

4.1 The generalized Pareto distribution

In the threshold modelling approach to FFA we are interested in the distribution of events above some threshold u. That is, the distribution of $Y = X - u$ given $X > u$, with X being the daily average discharge values as defined in Section 3.1 and Y now denoting the threshold exceedance. In this section, we follow the derivation of this distribution given in Coles (2001). The conditional probability of $X - u$ given $X > u$ can be written as

$$P(X > u + y | X > u) = \frac{1 - G(y + u)}{1 - G(u)}, \quad y > 0, \tag{10}$$

where G denotes the cumulative distribution function of X.

If the distribution G of X is known, the threshold exceedance distribution in (10) is also known. However, since G is not known, it is in extreme value theory approximated. For large values of m, $G^m(u)$ can be approximated by the GEV distribution given in Equation (1), as described in Section 3.1. From this we have

$$m \log G(u) \approx -\big[1 + \xi\big(\frac{u - \mu}{\sigma}\big)\big]^{-1/\xi}.$$

If u is large, $n \log(G(u))$ can be approximated by $-(1 - G(u))$, which gives

$$1 - G(u) \approx \frac{1}{m}\left(1 + \xi\frac{u - \mu}{\tilde{\sigma}}\right)^{-1/\xi},$$

where μ, $\tilde{\sigma}$ and ξ denotes respectively the location, scale and shape parameters of the GEV distribution. Further,

$$\begin{aligned} P(X > u + y | X > u) &= \frac{1 - G(y + u)}{1 - G(u)} \\ &\approx \frac{\frac{1}{m}\left(1 + \xi\frac{u+y-\mu}{\tilde{\sigma}}\right)^{-1/\xi}}{\frac{1}{m}\left(1 + \xi\frac{u-\mu}{\tilde{\sigma}}\right)^{-1/\xi}} \\ &= \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \end{aligned}$$

where $\sigma = \tilde{\sigma} + \xi(u - \mu)$ (Coles, 2001). Thus the distribution of independent threshold excesses can be approximated by

$$F(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \quad y > 0,$$

with support $(1 + \xi y/\sigma) > 0$. This is known as the generalized Pareto (GP) distribution (Pickands, 1975), with parameters σ and ξ. Note that these parameters correspond to the scale and shape parameters of the GEV distribution, in the sense that the shape parameter ξ is the exact same, while the scale parameter σ is given by $\sigma = \tilde{\sigma} + \xi(u - \mu)$. Here $\tilde{\sigma}$ and μ denote the scale and location parameter of the GEV distribution, respectively.

Similarly to how the Gumbel distribution is a special case of the GEV distribution for $\xi = 0$, the exponential distribution with parameter $\lambda = \frac{1}{\sigma}$ is a special case of the GP distribution when $\xi = 0$. The exponential distribution is given by

$$F(y) = 1 - e^{-\lambda y},$$

with probability density

$$f(y; \lambda) = \lambda e^{-\lambda y}. \tag{11}$$

In this study, we consider both the case where a peaks over threshold series, $\mathbf{y} = (y_1, \ldots, y_n)$, is assumed to follow the two-parameter GP distribution and when it is assumed to follow the one-parameter counterpart, the exponential distribution.

We estimate the parameters of the GP and exponential distribution, respectively, by the method of Maximum Likelihood (ML). The Maximum Likelihood estimators for the GP distribution are not given in closed form, and the likelihood must thus be maximized numerically. The optimization is performed using `optim` in `R`. The ML estimator for the parameter λ of the exponential distribution is given by the reciprocal of the sample mean,

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} y_i}.$$

4.2 Mixture model

Similar to how we considered a mixture of two Gumbel distributions in Section 3.2 for modelling of AMS, we now consider a mixture of two exponential distributions

for modelling of POT series. The probability density of a mixture of two exponential distributions is given by

$$\begin{aligned} f(y;\lambda_1,\lambda_2) &= \sum_{j=1}^{2} f_j(y;\lambda_j) \\ &= \omega_1\lambda_1 e^{-\lambda_1 y} + \omega_2\lambda_2 e^{-\lambda_2 y}, \end{aligned} \tag{12}$$

where $f_j(y;\lambda_j)$ denotes the exponential distribution in Equation (11) with parameter λ_j and $\omega_1 + \omega_2 = 1$, $\omega_1, \omega_2 \geq 0$.

Again we consider two cases, one where the mixture weights are observed, and given by the FGP, and one where they are unobservable. The former case is described in Section 4.2.1, while the latter is described in Section 4.2.2.

4.2.1 Maximum likelihood estimation with known weights

Here we assume that the weights ω_1 and ω_2 in Equation (12) are known. Given a sample $y_1, \ldots, y_n$ of flood peaks over threshold, with corresponding weights $\omega_{1,1}, \ldots, \omega_{1,n}$ and $\omega_{2,1}, \ldots, \omega_{2,n}$, the likelihood is

$$L(y;\lambda_1,\lambda_2) = \prod_{i=1}^{n} \left(\omega_{1,i}\lambda_1 e^{-\lambda_1 y_i} + \omega_{2,i}\lambda_2 e^{-\lambda_2 y_i}\right). \tag{13}$$

By taking the logarithm of Equation (13), we obtain the log likelihood function, $l(y;\lambda_1,\lambda_2)$, given by

$$l(y;\lambda_1,\lambda_2) = \log(L(y;\lambda_1,\lambda_2)) = \sum_{i=1}^{n} \log\left(\omega_{1,i}\lambda_1 e^{-\lambda_1 y_i} + \omega_{2,i}\lambda_2 e^{-\lambda_2 y_i}\right) \tag{14}$$

The parameters λ_1 and λ_2 of the mixture distribution in Equation (12) are estimated by optimizing the log likelihood in Equation (14), subject to $\omega_1 + \omega_2 = 1$ and $\omega_1, \omega_2 \geq 0$, using `constrOptim` in `R`.

4.2.2 EM algorithm for a mixture of exponential distributions

Again, we consider the mixture of two exponential distributions in Equation (12). Here, we assume that the mixture weights ω_1 and ω_2 are not known and formulate the problem as an incomplete data problem, similar to what we did in Section 3.2.3 for Gumbel mixture.

Let $f_Y(y;\boldsymbol{\theta})$ denote the distribution of Y, assumed to be the mixture in Equation (12), with parameters $\boldsymbol{\theta} = (\omega_1, \omega_2, \lambda_1, \lambda_2)$. This can be formulated as an incomplete data problem, where in our context $\mathbf{y} = (y_1, \ldots, y_n)$ denotes a vector of observed threshold excesses. The missing data vector $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ is defined as in Section 3.2.3. That is,

$$\begin{aligned} z_{ij} &= 1, \quad \text{if } y_i \text{ belongs to the } j\text{th component,} \\ z_{ij} &= 0, \quad \text{if } y_i \text{ does not belong to the } j\text{th component,} \end{aligned}$$

for $j = 1, 2$.

The conditional expectation of $l_c(\boldsymbol{\theta}, \boldsymbol{Z})$ given the complete data and current parameter values $\boldsymbol{\theta}^{(k)}$ can now be written as

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= E_{\boldsymbol{\theta}^{(k)}}[l_c(\boldsymbol{\theta}, \boldsymbol{Z})|\boldsymbol{y}, \boldsymbol{\theta}^{(k)}] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \log \omega_j f_j(y_i; \lambda_j), \end{aligned}$$

where $h_{ij}^{(k)} = P(z_{ij} = 1|y_i, \boldsymbol{\theta}^{(k)})$. With $f_j(y_i; \lambda_j)$ being the exponential distribution with parameter λ_j, we have

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) &= \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \log\left(\omega_j \lambda_j e^{-\lambda_j y_i}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \left[\log\left(\omega_j \lambda_j\right) - \lambda_j y_i\right]. \end{aligned}$$

To maximize this with respect to $\boldsymbol{\theta}$, we write

$$\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j=1}^{2} h_{ij}^{(k)} \left[\lambda_j y_i - \log\left(\omega_j \lambda_j\right)\right],$$

subject to $\sum_{j=1}^{2} \omega_j = 1$ and $\omega_1, \omega_2 \geq 0$. This gives the following updated parameter estimates

$$\frac{1}{\lambda_j} = \frac{\sum_{i=1}^{n} h_{ij} y_i}{\sum_{i=1}^{n} h_{ij}},$$

$$\omega_j = \frac{\sum_{i=1}^{n} h_{ij}}{n},$$

$j = 1, \ldots, 2$, (see e.g. Hasselblad, 1969).

The E-step consists of updating the expression for Q, $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k+1)})$ with the new parameter estimates, which requires the calculation of $h_{ij}^{(k+1)}$, $i = 1, \ldots, n$, $j = 1, 2$. The updated h_{ij} is given by

$$h_{ij}^{(k+1)} = p(z_j = i | y_i; \boldsymbol{\theta}^{(k)}) = \frac{\omega_j^{(k)} f_j(y; \lambda_j^{(k)})}{\omega_1^{(k)} f_1(y; \lambda_1^{(k)}) + \omega_2^{(k)} f_2(y; \lambda_2^{(k)})}, \quad j = 1, 2,$$

(see e.g. Hasselblad, 1969).

Again, we run the algorithm with 100 different randomly generated starting values and choose the parameter estimates obtained in the that resulted in the overall maximum likelihood. Random starting values for λ_1 and λ_2 are obtained by sampling from a normal distribution with mean and standard deviation equal to the reciprocal of the sample mean and standard deviation of $\mathbf{y}$, respectively. The initial value of ω_1 is sampled from the uniform distribution on $[0, 1]$ in each run of the algorithm.

4.3 Return level estimation

For threshold modelling, the m-observation return level x_m is given by the solution of

$$P(X > x_m) = P(X > u)P(X > x_m | X > u) = \frac{1}{m},$$

(Coles, 2001). Here $P(X > x_m | X > u)$ is the estimated threshold exceedance distribution, which in our case is either the GP distribution, the exponential distribution or a two component mixture of exponential distributions. For the GP distribution, we obtain

$$x_m = u + \frac{\sigma}{\xi} \left((mP(X > u)^{\xi} - 1 \right),$$

and for the exponential distribution the m-observation return level is given by

$$x_m = u + \sigma \log \left(mP(X > u) \right),$$

(Coles, 2001). $P(X > u)$, the probability that a discharge value is above the chosen threshold u, can be estimated by the sample proportion of discharge values above the

threshold.

Plotting x_m against m on a logarithmic scale provides the same qualitative information as return level plots for AMS modelling. However, in order to compare them to return levels obtained using AMS modelling, it is of interest to transform these m-observation return levels to an annual scale. This is obtained by substituting m for $m = pn_y$, where p denotes the return period (in years) of interest and n_y is the number of observations per year. In our case of daily average discharge values as the raw data, $n_y = 365.25$ when accounting for leap years.

The p-year return level is given by

$$\hat{z}_p = u + \frac{\hat{\sigma}}{\hat{\xi}} \left((pn_y \cdot P(X > u))^{\hat{\xi}} - 1 \right),$$

for the GP distribution, and by

$$\hat{z}_p = u + \hat{\sigma} \log \left(pn_y \cdot P(X > u) \right),$$

for the exponential distribution. A natural estimate for $n_y P(X > u)$ is the average number of flood peaks included in the POT series each year.

For the two component exponential mixture models, a closed form expression of x_m is not obtainable. Instead we sample from

$$u + \left(\omega_1 \lambda_1 e^{-\lambda_1 y} + \omega_2 \lambda_2 e^{-\lambda_2 y} \right),$$

and estimate the return level z_p by the empirical quantile function at $1 - \frac{1}{pn_y P(X>u)}$.

Plotting positions for the observed POT series, $\mathbf{y}$, are obtained as in Section 3.3, but transformed from the m-observation scale to an annual scale by $m = pn_y$.

5 Validation

To compare the various models for AMS and POT, respectively, we consider a framework for data-based comparison of frequency analysis methods by Renard et al. (2013). The framework considers the predictive performance of the models, in terms of the reliability and the stability of the forecasts. The reliability refers to the models ability to obtain a distribution close to the unknown true distribution. This has to be measured using observed data, since the true distribution is unknown. With stability, we mean the models ability to give similar estimates when different data is used to fit the model. Stable estimates of return levels are desired, since, in practice, a structure cannot be modified whenever the estimate changes as new data is obtained. Even though stability is an important property of a forecast, it cannot alone be used to validate the models. A model can yield stable, but completely unreliable estimates. Therefore, the reliability of the models are first considered.

5.1 Scoring rules

Scoring rules provide a measure of the reliability, or calibration, of a predictive distribution. They can assess both the calibration and the sharpness of a forecast. The calibration is a joint property of the predictive distribution and the realized value, it refers to the statistical compatibility between the two. The sharpness is a property of the predictive distribution only and concerns the concentration of the forecast, or the forecasts ability to separate different situations. (Gneiting and Katzfuss, 2014).

Consider a predictive distribution $F \in \mathcal{F}$, where $\mathcal{F}$ is a class of probability distributions on $\mathbb{R}$, and denote the realized value by $y \in \mathbb{R}$. A scoring rule is a function

$$S(F, y) : \mathcal{F} \times \mathbb{R} \to \mathbb{R},$$

that assigns a numerical score to the pair (F, y). In the literature, both positively and negatively oriented scoring rules are used. Here we take scoring rules to be negatively oriented, such that a lower value means a better score.

A desired property of scoring rules is propriety. A scoring rule is said to be proper relative to the class $\mathcal{F}$ if

$$S(G, G) \leq S(F, G), \tag{15}$$

for all $F, G \in \mathcal{F}$, and strictly proper if the equality in (15) holds only for $F = G$. Here, $S(F, G)$ denotes the expected value $S(F, \cdot)$ under G, $\mathrm{E}[S(F, G)]$. Thus a proper scoring rule assures that one will report its true beliefs about the predictive distribution in order to obtain the best score. All the scoring rules that we consider in the

following are proper.

There exist a variety of proper scoring rules that penalizes and rewards different aspects of the forecast. A simple and widely used scoring rule, proposed by Good (1952), is the logarithmic score,

$$S^{\mathrm{LOG}}(F, y) = -\log f(y). \tag{16}$$

The logarithmic score gives a strong penalty to unlikely events. If $f(y)$ is close to zero, $-\log(f(y))$ goes to infinity. This is a good quality if one wants to make sure that the forecast does not assign zero probability to events that can occur. The logarithmic score is local in the sense that when evaluating how good the forecast is in terms of an observed value y, it only uses the distribution at y, $f(y)$, to calculate the score.

Since we are estimating extreme floods, e.g. the 1000-year flood, it is of interest to consider scoring rules that assess the predictive distribution's ability to predict the exceedance of a certain threshold or quantile. The Brier score and the quantile score are examples of such scoring rules. The Brier score is defined as

$$S_u^{\mathrm{B}}(F, y) = (p_u - \mathbb{1}\{y \geq u\})^2, \tag{17}$$

where u is the threshold of interest and $p_u = 1 - F(u)$ is the predicted probability of y exceeding that threshold(see e.g. Gneiting and Raftery, 2007). In the original formulation by Brier (1950) the range of the score is zero to two, while in the definition in (17) the maximum difference is 1 and the range is thus zero to one. This scoring rule requires the selection of a threshold, u, which often is given in terms of a quantile of the sample. Moreover, the quantile score evaluates how well the predictive distribution predicts the quantile τ. This scoring rule is given by

$$S_\tau^{\mathrm{QS}}(F, y) = (\mathbb{1}\{y < F^{-1}(\tau)\} - \tau)(F^{-1}(\tau) - y), \tag{18}$$

for a given quantile τ (Gneiting and Raftery, 2007).

Skill scores are often used instead of scores when comparing various models. For the scoring rules we consider, the skill score can be written as

$$SS(F, y) = \frac{S(F_{ref}, y) - S(F, y)}{S(F_{ref}, y)},$$

where F_{ref} is a reference model (see e.g. Friederichs and Thorarinsdottir, 2012). The skill score of a distribution F measures the relative gain of the this distribution with

respect to the reference distribution. Positive skill scores represent a gain in the predictive skill of the model, and a skill score of zero represents no gain in predictive skill. Also, negative skill scores indicate that the reference distribution performs better. As the GEV distribution and GP distribution are commonly applied for modelling AMS and POT series respectively, they are natural to consider as reference models.

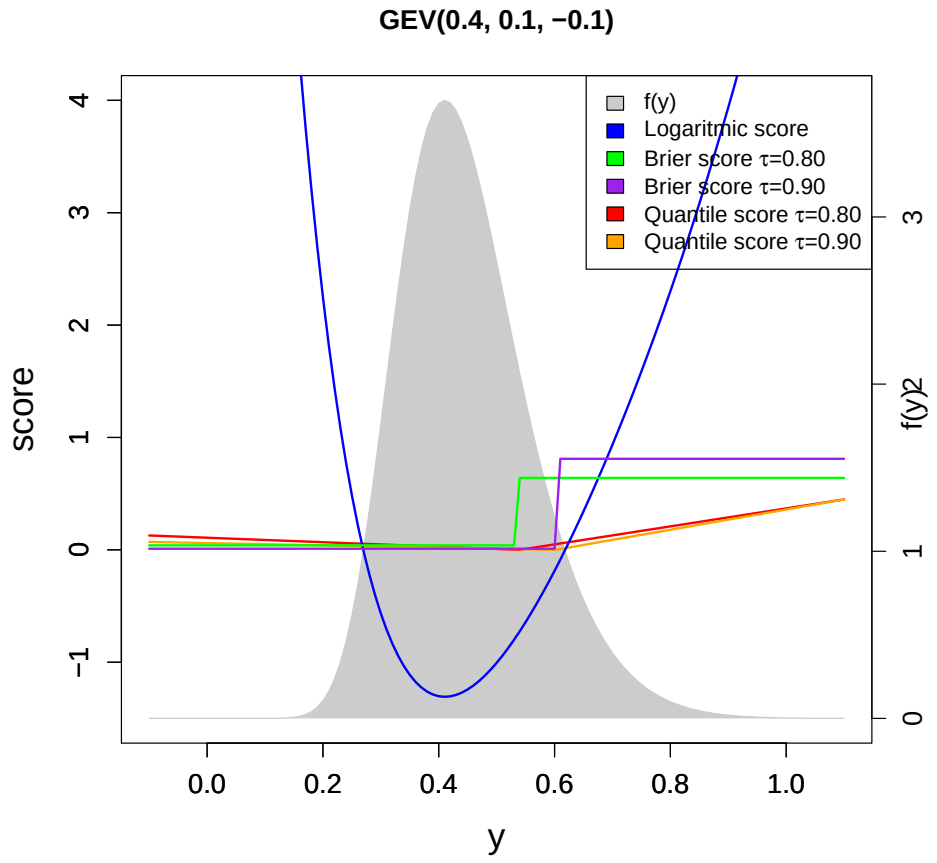

Figure 6: Various scoring rules, $S(F, y)$, as a function of y, when F is the GEV distribution with parameters $\mu = 0.4$, $\sigma = 0.1$ and $\xi = -0.1$. For comparison, the density function $f(y)$ is given in the same plot, on the right y-axis. The illustrated scoring rules are the quantile score with the quantiles 0.80 (orange) and 0.90 (red), the logarithmic score (blue) and the Brier score with thresholds corresponding to the quantiles 0.80 (green) and 0.90 (purple).

Figure 6 illustrates how the different scoring rules assess the performance of a predic-

tive distribution. Each scoring rule $S(F, y)$ is plotted as a function of the observed value y, where F in this example is the GEV distribution with parameters $\mu = 0.4$, $\sigma = 0.1$ and $\xi = -0.1$. The probability density $f(y)$ is given in the same plot. From this plot it is clear that the logarithmic score is most sensitive to outliers. The logarithmic score is optimized in the mode of the distribution F, while the quantile score takes its lowest value at the quantile of interest and penalizes deviations in the upper tail. Similar to the quantile score, the Brier score penalizes deviations in the upper tail. When integrating the Brier score over all thresholds, or the quantile score over all quantiles, the continuous ranked probability score (CRPS) is obtained (see e.g. Friederichs and Thorarinsdottir, 2012). Thus evaluating the Brier and quantile score over all threshold and quantiles, respectively, gives the same score, but as Figure 6 shows, these scoring rules are not identical for one specific quantile and the threshold corresponding to that quantile.

To obtain standard error estimates of the average scores, we apply bootstrapping. For a vector of scores $\mathbf{s} = (s_1, \ldots, s_n)$ obtained by a model at a specific catchment, we repeatedly sample n values from this vector. For each sample, the average score is calculated. In total, we resample from the score vector 1000 times, and estimate the standard error by the standard deviation of the 1000 average score estimates.

5.2 Cross validation

When applying scoring rules to assess the predictive performance of the various models, we need out-of-sample observed data, or test data, to obtain an average score for each model. To obtain such out-of-sample validations, it is common to divide the data into a training set, used to obtain the model, and a test set, used to validate the models. Ideally, one would like to test the models on large amounts of data, as well as having sufficient amounts of data to train the models. The selection of the test set size involves a bias-variance trade-off. We want to minimize the testing bias by reserving a sufficient proportion of the sample to training, such that the estimated models are as close as possible to the models obtained using the entire sample. In addition, we want a sufficient amount of data for testing, in order to minimize the testing variance. For small sample sizes, a k-fold cross validation is commonly applied to achieve this trade-off.

Cross validation is a technique for assessing the performance of a predictive model on an independent data set. It dates back to the 1930s (Larson, 1931), and an early description of the method can be found in Mosteller and Tukey (1968). The idea of cross validation is to leave out a small part of the data set, to be used for testing, while the remaining data is used to fit the model. This is performed repeatedly, such

that the model is tested on a large amount of data while not having to reserve all this data for testing at once. There exist several versions of this method. Examples are the leave-one-out or leave-p-out cross validation, where one or p observations is allocated to testing in each round.

k-fold cross validation divides the data into k equal sized subsets and, in turn, uses each subset as the test data while the remaining subsets are used to train the proposed models. The predictive performance of each model can thus be validated using the respective test set for each round of the cross validation. We apply this method, with $k = 10$, and use various scoring rules, presented in Section 5.1, to assess the performance of each model, on each of the 10 test sets. Below is a description of how we proceed to obtain scores of the various models.

- Randomly divide the sample into 10 equal sized subsets. Let $\mathbf{y}_i$, $i = 1, \ldots, 10$, denote the subsets and n_i denote the length of each subset.
- For each subset i, do the following.
  - Estimate the parameters of the models, F_j, using the other nine subsets. Here, $j = 1, \ldots, 4$ denotes the four models we consider for either the AMS or POT approach.
  - Use the current subset $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})$ as a test set, to obtain scores for each of the models obtained in the step above. That is, calculate the following average scores
    - $\frac{1}{n_i} \sum_{l=1}^{n_i} S^{\text{LOG}}(F_j, y_{il})$, for $j = 1, \ldots, 4$.
    - $\frac{1}{n_i} \sum_{l=1}^{n_i} S_u^{\text{B}}(F_j, y_{il})$, for $j = 1, \ldots, 4$ and for each of the thresholds u corresponding to the quantiles $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$ of the sample.
    - $\frac{1}{n_i} \sum_{l=1}^{n_i} S_\tau^{\text{QS}}(F_j, y_{il})$, for $j = 1, \ldots, 4$ and for each of the quantiles $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$.
- Average the scores obtained for the test set $\mathbf{y}_i$, $i = 1, \ldots, 10$, for each scoring rule.

To test for a difference between the average score of two models, we apply a paired t-test to vectors of scores from each model. The t-test is performed using `t.test` in `R`.

5.3 Stability of return level estimates

To evaluate whether return level estimates of the various models depend heavily on the sample used to estimate the models, we compare return level estimates obtained when repeatedly leaving out one year of data from the AMS and POT series. This is done for a catchment in the following manner.

Let $\mathbf{y}_{\text{AMS}}$ denote the annual maximum series, with length n_{AMS}, and $\mathbf{y}_{\text{POT}}$ the peaks over threshold series with length n_{POT}. For each observation in $\mathbf{y}_{\text{AMS}}$, we estimate the parameters of the four AMS models by leaving out this observation. Similarly, for $\mathbf{y}_{\text{POT}}$ we leave out the number of observations corresponding to one year each time, to obtain n_{AMS} different parameter estimates for each of the POT models. That is, on average we leave out $n_{\text{POT}}/n_{\text{AMS}}$ observations for each estimate. The different parameter estimates for each model are used to obtain different return level estimates for a specific return period.

In our study, we consider the return levels corresponding to the return periods $p = 100$ and $p = 1000$. The variance and magnitude of the estimates are assessed visually by analysing boxplots of the return level estimates based on different samples.

6 Results

6.1 Annual Maximum Series

In order to evaluate which of the four proposed models for AMS that overall performs the best, we look at histograms of the model that is ranked as the best model for each of the 228 catchments in the dataset. We consider the three different scoring rules presented in Section 5.1. That is, the logarithmic score, the Brier score and the quantile score. For the Brier score we consider the thresholds corresponding to the quantiles $\tau = 0.80$ and $\tau = 0.90$. The same quantiles are considered for the quantile score. The resulting histograms are given in Figure 7.

From Figure 7 we see that the logarithmic score assigns the best score most often to the Gumbel distribution. For the Brier and quantile score, there is a less obvious winner among the models, especially for the Brier score with $\tau = 0.80$. With $\tau = 0.90$, the Brier score gives the best score to the two mixture models most times. According to the quantile score, the EM mixture model and the Gumbel distribution perform well compared to the other models. The results presented in the histograms can indicate either that which model is considered to be the best among the four models depend on the scoring rule and the catchment, or that there is not much difference in the performance of the models. To get a better understanding of how the scores varies among the catchments, we study a portrait diagram of the model which obtained the best score at each catchment, in Figure 8.

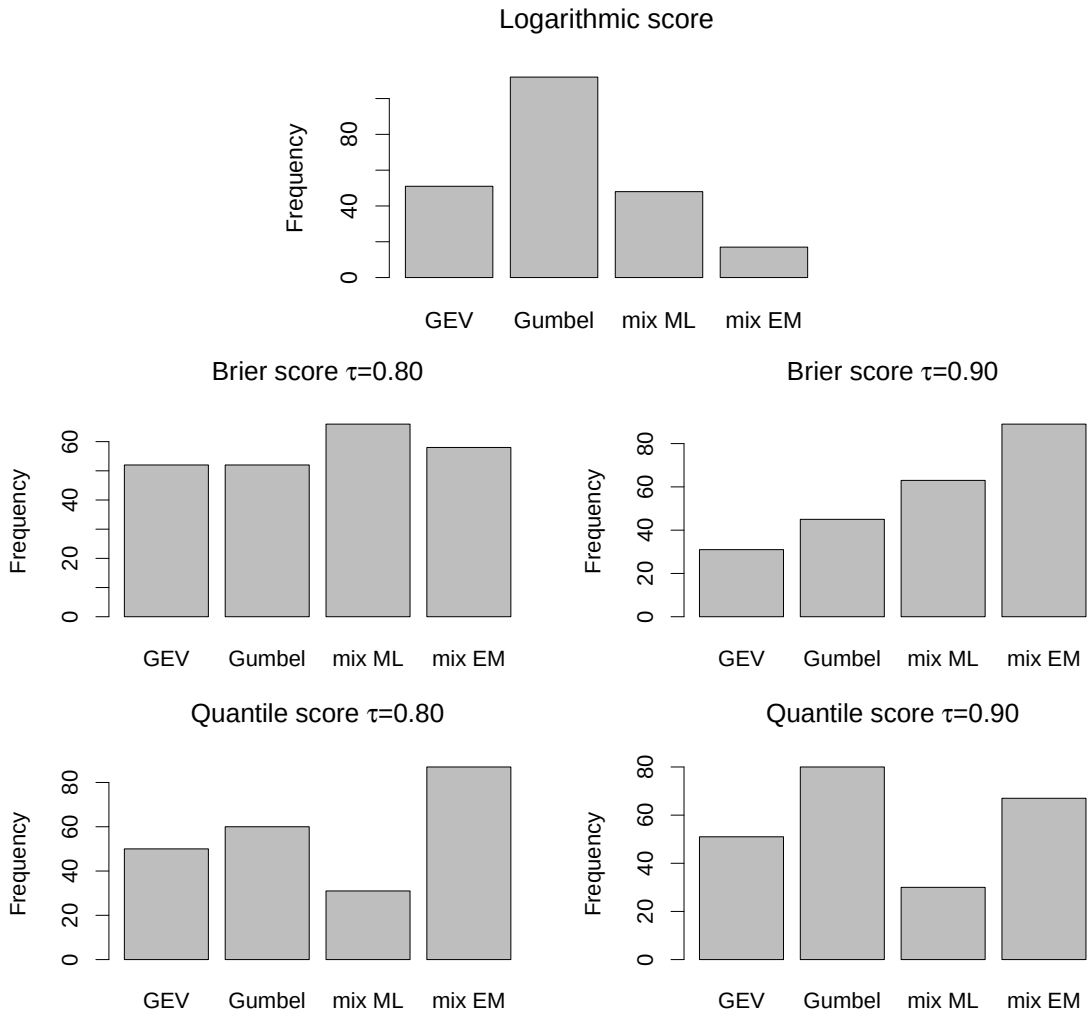

Figure 7: Histograms of the number of catchments at which each model performs the best out of the four models, when the logarithmic score, the Brier score and the quantile score are used as the scoring rules in the cross validation. For the Brier score and the quantile score the quantiles $\tau = 0.80$ (left) and $\tau = 0.90$ (right) are used.

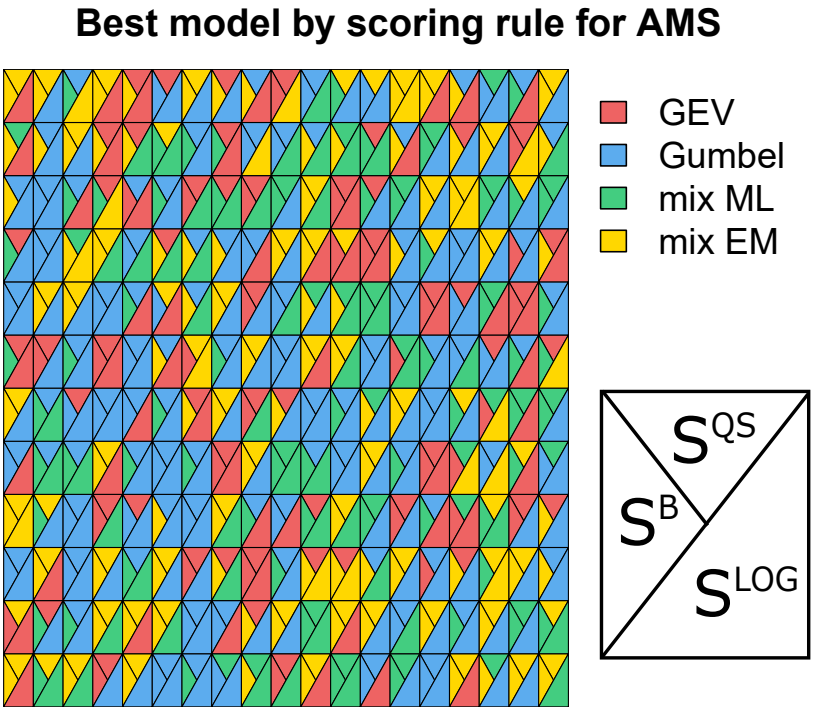

Figure 8: Portrait diagram of the best POT model at each catchment. A tile in the grid corresponds to one catchment. The lower half of the tile gives the best model according to the logarithmic score, S^{LOG}, while the upper left and upper right part gives the best model when the Brier score, S^{B}, and the quantile score, S^{QS}, with $\tau = 0.90$ are applied, respectively. The catchments are sorted, from the upper left to the lower right corner, by the increasing value of the average FGP at each catchment.

Figure 8 gives the model which obtained the lowest score at each catchment, according to the three different scoring rules, the logarithmic score, the Brier score with a threshold corresponding to the quantile $\tau = 90$ and the quantile score with the quantile $\tau = 0.90$. The catchments are sorted by the increasing value of the average FGP, from the upper left to the lower right corner. From the figure we see that various scoring rules often do not agree on which model that performs the best. There does not seem to be a particular pattern in the diagram, indicating that the model which performs the best does not depend on the FPG variable.

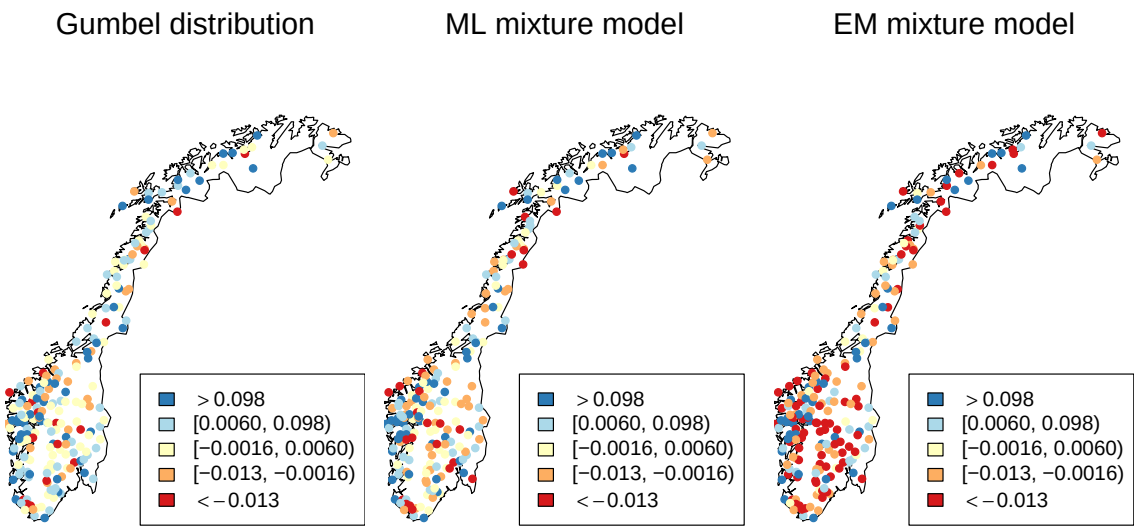

Figure 9: Maps of Norway giving the skill score of the Gumbel distribution (left), the mixture model with known weights (middle) and the mixture model with unknown weights (right), at each catchment, when using the logarithmic score and the score of the GEV distribution as the reference score.

To study the difference in the average score obtained by each model at the various catchments, we consider the skill score with respect to the GEV distribution. Figure 9 presents the logarithmic skill score of the Gumbel distribution and the two Gumbel mixtures at each catchment. We see that the EM mixture model receives more negative gains compared to the other models. For the Gumbel distribution and ML mixture, most of the positive skill scores are obtained along the coast or in northern Norway.

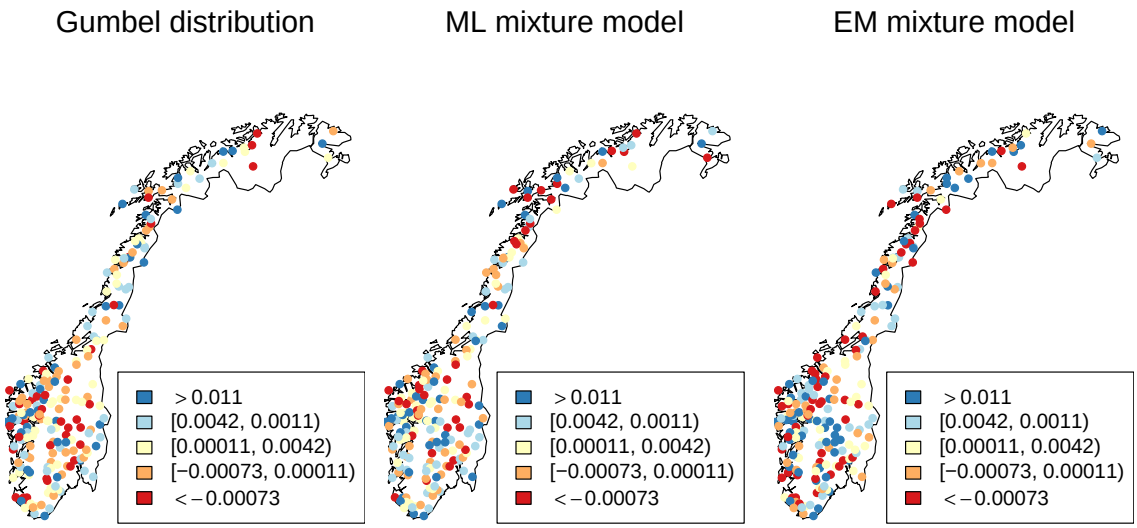

Figure 10: Maps of Norway giving the skill score of the Gumbel distribution (left), the mixture model with known weights (middle) and the mixture model with unknown weights (right), at each catchment, when using the Brier score with a threshold corresponding to the quantile $\tau = 0.90$ at each catchment and the score of the GEV distribution as the reference score.

The Brier skill score, with a threshold corresponding to the quantile $\tau = 0.90$, for each catchment are given in Figure 10. Here there is less variation between the models, compared to the results for the logarithmic score in Figure 9. In addition, the Brier skill score does not seem to depend on the location of the catchment.

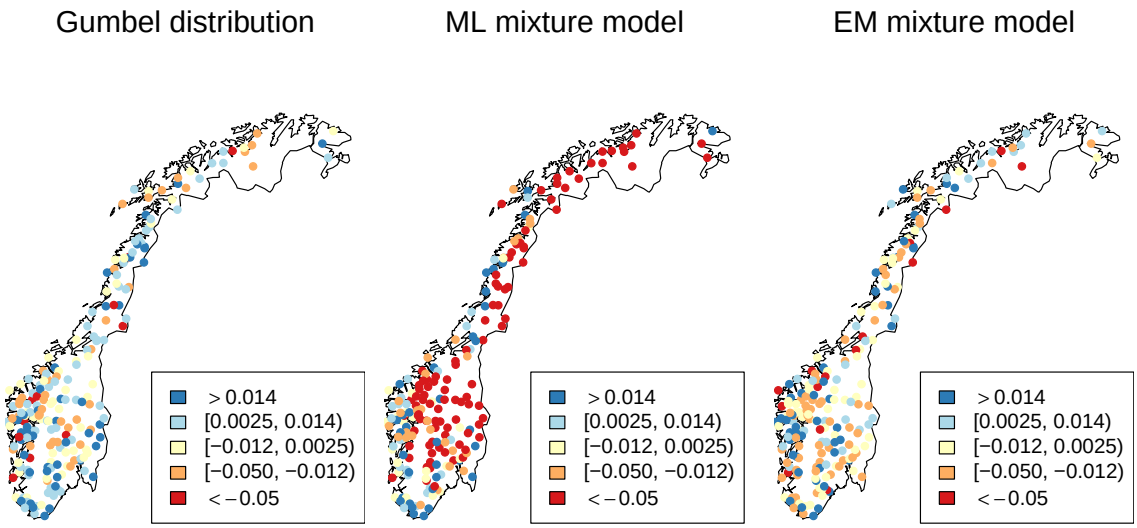

Figure 11: Maps of Norway giving the skill score of the Gumbel distribution (left), the mixture model with known weights (middle) and the mixture model with unknown weights (right), at each catchment, when using the quantile score with $\tau = 0.90$ and the score of the GP distribution as the reference score.

In Figure 11 the quantile skill score at each catchment are given for the Gumbel distribution and the two mixture models for AMS, with the GEV distribution as the reference score. The figure shows that the ML mixture model receives a negative skill score at more catchments than the two other models. Positive skill scores for this model are mainly obtained along the coast. For the two other models, the largest gains relative to the GEV distribution are also mostly found in the coastal area.

6.1.1 Detailed look at three stations

To get a better understanding of how the various models perform, we consider the estimated models for the AMS approach at the three catchment introduced in Section 2.3, Bulken, Krinsvatn and Atnasjø. First, we evaluate return level plots from each catchment. Secondly, we present the average scores obtained by the logarithmic score, the Brier score and the quantile score.

Figure 12 presents the estimated return level as a function of return period, for the four estimated models at Bulken, and the corresponding estimated parameters are given in Table 3. From the return level plot, it looks like the GP distribution and the mixture models are able to fit the observed data well, while the Gumbel distribution

overestimates the return levels. The shape parameter ξ of the GEV distribution is estimated to be negative, meaning that the flood values are bounded from above.

The estimated return level as a function of the return period, for Krinsvatn, is given in Figure 13. Here, the GP distribution fits the observed data well, while the Gumbel distribution and the mixture model with known weights underestimate the return levels for return periods greater that 10. The yellow line, corresponding to the mixture model with unknown weights, looks quite strange. Around the return period $p = 50$, the line changes shape. This is due to the large difference between the two components of the EM mixture model, see Table 4. The mixture weight of the first component is estimated to be $\omega_1 = 0.98$, thus the model assigns almost every observation to this component. From $p = 50$, the return level function is approximately a straight line through the largest observed flood value, 356.68 m^3/s. This is also the value of the location parameter μ of the EM mixture model. It turns out that for this catchment, the EM mixture model uses the second component to fit the largest observation alone, and the first component to fit the rest of the data. This model seems to be overfitting the data.

The same overfitting happens for the EM mixture model at Atnasjø. The estimated return level as a function of the return period for this catchment is given in Figure 14, and the estimated parameters of the four models are given in Table 5. Here, similar to for Krinsvatn, the EM model consists of one component that models the two largest observed values and one that models the rest of the data. Also similar to Krinsvatn, the Gumbel distribution and the mixture distribution with known weights give lower estimates for the return levels than the GEV distribution.

From the return level plots, we see that the EM mixture model tends to overfit the data when the shape parameter of the GEV distribution is positive. At Bulken, the EM mixture model also assign most observations to one distribution, but here this distribution has a smaller location parameter than the dominating component, which explains why wee do not see the same change in the return level function. For the other mixture model, this overfitting problem does not occur. It uses the FGP values as fixed weights for each observation and does not have the flexibility to assign the largest values to one component and the rest to the other component.

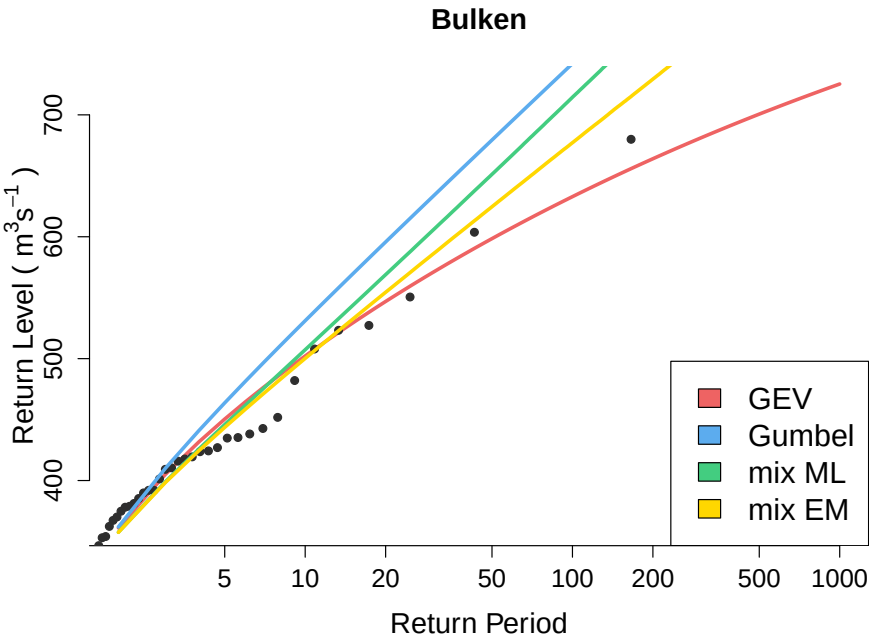

Figure 12: The estimated return level as a function of return period when the GEV distribution (red), the Gumbel distribution (blue), the mixture model with known weights (green) and the mixture model with unknown weights (yellow) are applied. The black dots denotes the observed flood values.

Table 3: The estimated parameters of the four different models for the AMS from Bulken.

Method		Parameters	
GEV	$\mu = 329.05$	$\sigma = 89.78$	$\xi = -0.14$
Gumbel	$\mu = 322.29$	$\sigma = 88.26$	
mix ML	$\omega_1 = 0.71$	$\mu_1 = 312.92$	$\sigma_1 = 93.49$
		$\mu_2 = 351.79$	$\sigma_2 = 54.89$
mix EM	$\omega_1 = 0.97$	$\mu_1 = 333.73$	$\sigma_1 = 75.19$
		$\mu_2 = 167.13$	$\sigma_2 = 0.94$

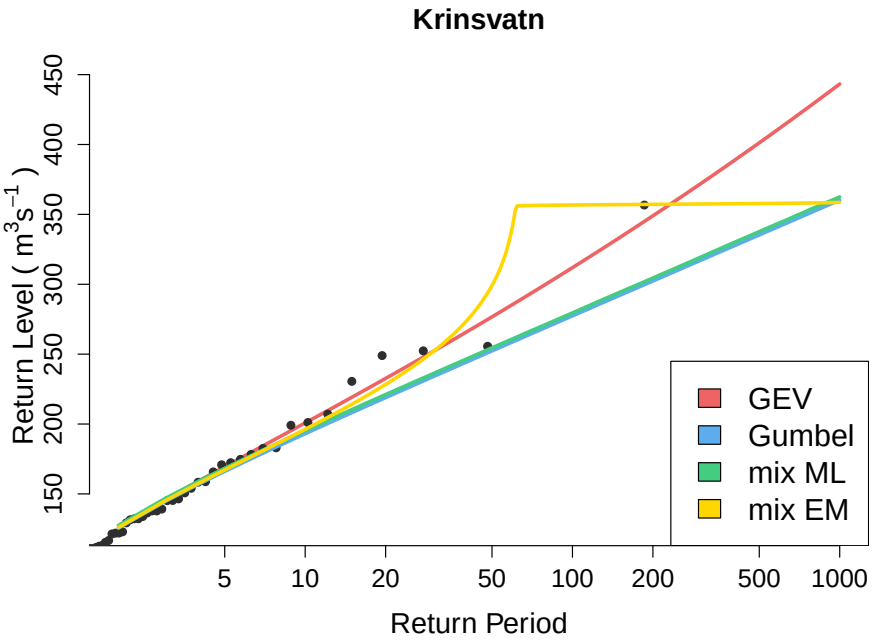

Figure 13: The estimated return level as a function of return period when the GEV distribution (red), the Gumbel distribution (blue), the mixture model with known weights (green) and the mixture model with unknown weights (yellow) are applied. The black dots denotes the observed flood values.

Table 4: The estimated parameters of the four different models for the AMS from Krinsvatn.

Method		Parameters	
GEV	$\mu = 112.31$	$\sigma = 35.89$	$\xi = 0.080$
Gumbel	$\mu = 113.88$	$\sigma = 36.91$	
mix ML	$\omega_1 = 0.79$	$\mu_1 = 118.44$	$\sigma_1 = 36.01$
		$\mu_2 = 97.85$	$\sigma_2 = 34.13$
mix EM	$\omega_1 = 0.98$	$\mu_1 = 112.47$	$\sigma_1 = 34.78$
		$\mu_2 = 356.68$	$\sigma_2 = 0.34$

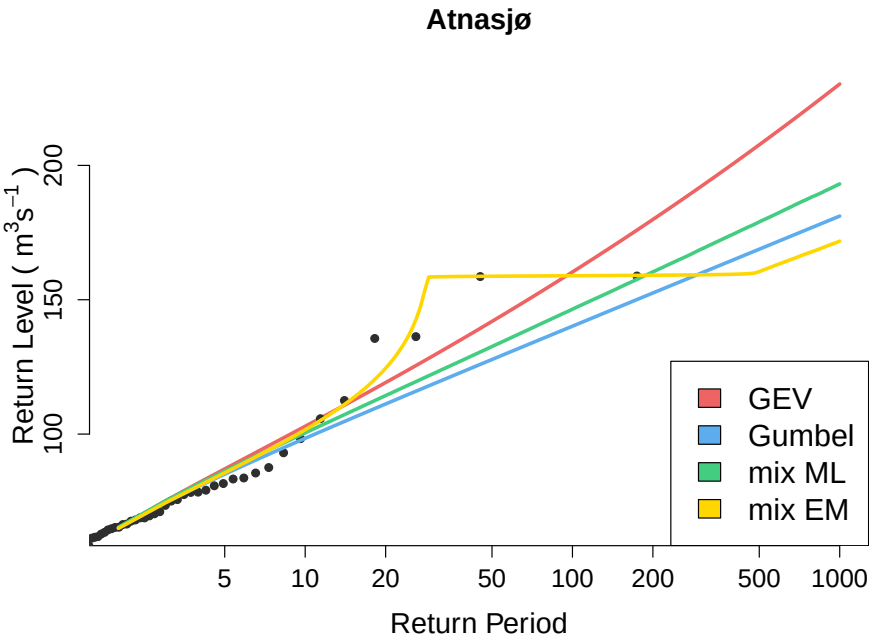

Figure 14: The estimated return level as a function of return period when the GEV distribution (red), the Gumbel distribution (blue), the mixture model with known weights (green) and the mixture model with unknown weights (yellow) are applied. The black dots denotes the observed flood values.

Table 5: The estimated parameters of the four different models for the AMS from Atnasjø.

Method		Parameters	
GEV	$\mu = 58.45$	$\sigma = 17.76$	$\xi = 0.093$
Gumbel	$\mu = 59.37$	$\sigma = 18.39$	
mix ML	$\omega_1 = 0.45$	$\mu_1 = 56.50$	$\sigma_1 = 15.19$
		$\mu_2 = 61.92$	$\sigma_2 = 20.88$
mix EM	$\omega_1 = 0.033$	$\mu_1 = 158.71$	$\sigma_1 = 0.18$
		$\mu_2 = 58.05$	$\sigma_2 = 16.57$

To further investigate the performance of the various AMS models at the three catchments, we consider the average scores obtained in the cross validation procedure with the logarithmic score, the Brier score and the quantile score. Table 6 presents the average score, and corresponding standard error estimates, of each model at each catchment, when the logarithmic score is used as the scoring rule. The mixture model with known weights receives the best score at Bulken, while the exponential distribution is assigned the best score at the other two catchments. However, the scores of each model are quite similar and the t-test concludes that there is no significant difference in the score of the various models at each catchment, at a 95% significance level.

The average scores for each model at the three catchments, obtained using the Brier score with thresholds corresponding to the quantiles $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$, are given in Table 7. The scores of the various models at each catchment does not seem to vary much. At Bulken, the mixture model with unknown weights receives the lowest score for all three thresholds, while the GEV distribution obtains the best score for all three thresholds at Krinsvatn. At Atnasjø, the EM mixture is given the lowest score for $\tau = 0.80$ and $\tau = 0.95$, while the mixture with given weights receives the best score for $\tau = 0.90$. Again, according to the paired t-test, there is no significant difference between the scores of any of the models, at all three catchments.

Table 8 presents the average score of each model when the when the quantile score with $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$ is applied. The model which receives the best score depends on the quantile and the catchment. Here, there seems to be a greater difference between the scores of the four models compared to for the Brier score, but using the paired t-test at significance level 95%, no significant difference is found.

Table 6: Cross validation with the logarithmic score at Bulken, Krinsvatn and Atnasjø, respectively. The best score for each scoring rule is highlighted in boldface. Standard error estimates corresponding to each average score, obtained by bootstrap, are given in the parentheses.

	Logarithmic score		
Method	Bulken	Krinsvatn	Atnasjø
GEV	6.07 (0.13)	5.28 (0.013)	4.55 (0.12)
Gumbel	6.06 (0.12)	**5.24** (0.013)	**4.55** (0.13)
mix ML	**6.04** (0.11)	5.26 (0.013)	4.55 (0.13)
mix EM	6.14 (0.16)	5.27 (0.014)	4.62 (0.15)

Table 7: Cross validation with the Brier score, with thresholds corresponding to $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$, at Bulken, Krinsvatn and Atnasjø, respectively. The best score for each scoring rule is highlighted in boldface. Standard error estimates corresponding to each average score, obtained by bootstrap, are given in the parentheses.

Bulken			
Method	Brier score $\tau = 0.80$	Brier score $\tau = 0.90$	Brier score $\tau = 0.95$
GEV	0.17 (0.027)	0.095 (0.032)	0.050(0.024)
Gumbel	0.17 (0.028)	0.095(0.029)	0.051 (0.025)
mix ML	0.17 (0.029)	0.094 (0.031)	0.050 (0.025)
mix EM	**0.17** (0.029)	**0.094** (0.031)	**0.050** (0.026)
Krinsvatn			
Method	Brier score $\tau = 0.80$	Brier score $\tau = 0.90$	Brier score $\tau = 0.95$
GEV	**0.16** (0.031)	**0.098** (0.032)	**0.059** (0.028)
Gumbel	0.16 (0.031)	0.098 (0.032)	0.059 (0.028)
mix ML	0.16 (0.031)	0.098 (0.031)	0.060 (0.028)
mix EM	0.16 (0.031)	0.099 (0.032)	0.059 (0.028)
Atnasjø			
Method	Brier score $\tau = 0.80$	Brier score $\tau = 0.90$	Brier score $\tau = 0.95$
GEV	0.17 (0.027)	0.10 (0.032)	0.064 (0.030)
Gumbel	0.17 (0.025)	0.10 (0.033)	0.064 (0.030)
mix ML	0.17 (0.027)	**0.10** (0.033)	0.064 (0.031)
mix EM	**0.17** (0.027)	0.10 (0.033)	**0.064** (0.029)

Table 8: Cross validation with quantile score, for $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$, at Bulken, Krinsvatn and Atnasjø, respectively. The best score for each scoring rule is highlighted in boldface. Standard error estimates corresponding to each average score, obtained by bootstrap, are given in the parentheses.

Bulken			
Method	Quantile score $\tau = 0.80$	Quantile score $\tau = 0.90$	Quantile score $\tau = 0.95$
GEV	29.06 (4.02)	20.34 (3.07)	12.77 (2.58)
Gumbel	29.31 (3.82)	20.37 (2.65)	12.94 (1.81)
mix ML	29.96 (4.88)	20.52 (3.56)	**12.63**(2.44)
mix EM	**28.73** (2.97)	**20.25** (1.76)	12.39 (2.13)
Krinsvatn			
Method	Quantile score $\tau = 0.80$	Quantile score $\tau = 0.90$	Quantile score $\tau = 0.95$
GEV	16.89(2.71)	**12.15** (2.37)	**8.09** (2.01)
Gumbel	**16.89** (2.71)	12.16 (2.50)	8.12 (1.96)
mix ML	16.93 (2.94)	12.46 (2.78)	8.82 (2.13)
mix EM	16.90 (2.753)	12.26 (2.54)	8.40 (2.04)
Atnasjø			
Method	Quantile score $\tau = 0.80$	Quantile score $\tau = 0.90$	Quantile score $\tau = 0.95$
GEV	8.42 (1.41)	6.64 (1.25)	**4.53** (0.98)
Gumbel	8.42 (1.53)	**6.56** (1.30)	4.55 (1.08)
mix ML	10.15 (2.11)	7.50 (1.98)	5.44 (1.71)
mix EM	**8.36**(1.55)	6.69 (1.35)	4.67 (1.33)

6.2 Peaks Over Threshold

Following the same procedure for presenting the results as for the AMS approach in Section 6.1, we first look at histograms of the model that is ranked best at each catchment. Again, we consider the logarithmic score, the Brier score and the quantile score, with quantiles$\tau = 0.80$ and $\tau = 0.90$. The resulting histograms are given in Figure 15.

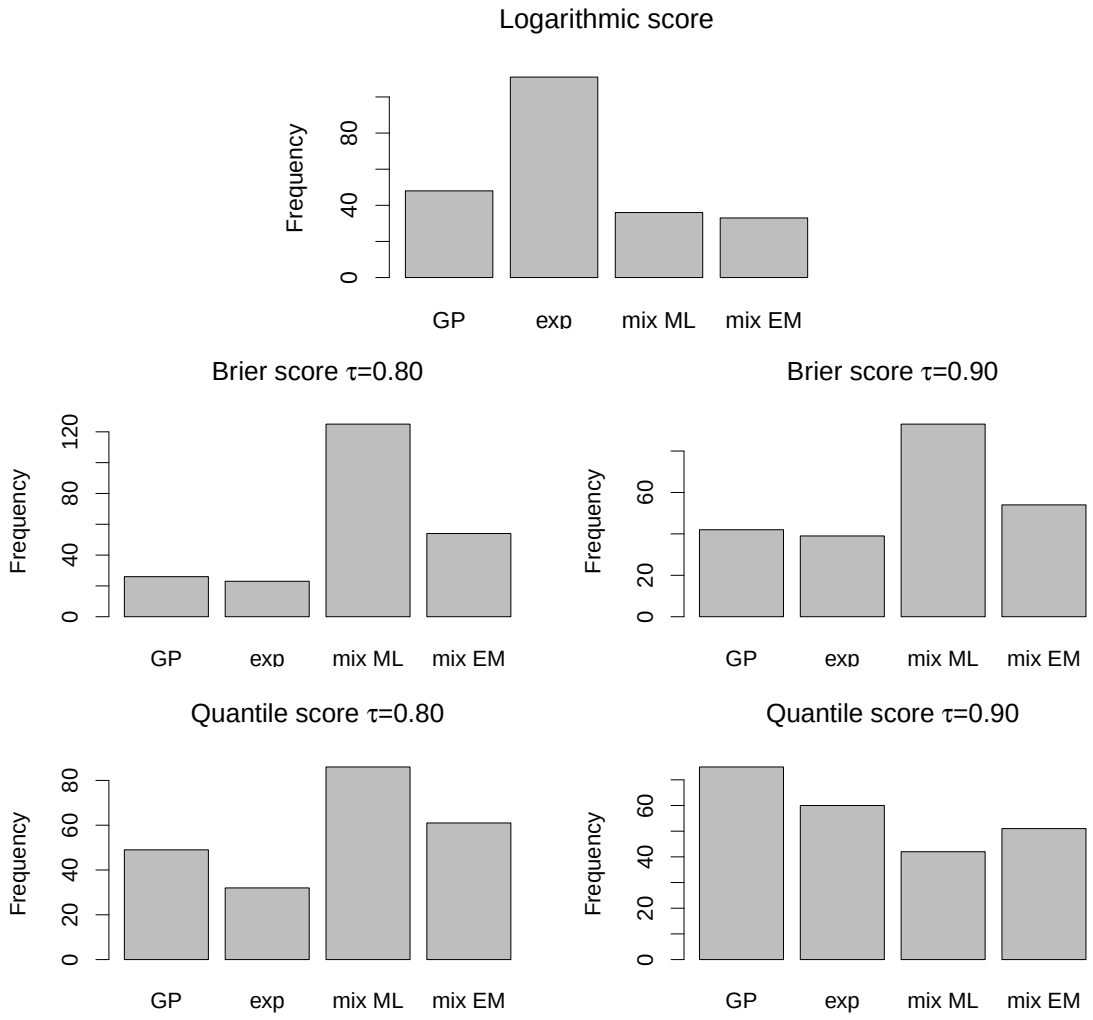

Figure 15: Histograms of the number of catchments at which each model performs the best out of the four models, when the logarithmic score, the Brier score and the quantile score are used as the scoring rules in the cross validation. For the Brier score and the quantile score the quantiles $\tau = 0.80$ (left) and $\tau = 0.90$ (right) are used.

From Figure 15 we see that, according to the logarithmic score, the exponential distribution performs the best at most locations. None of the other three models is a clear loser. For the Brier score, the mixture of exponentials with known weights performs the best at most location for both values of the threshold. When the quantile score is used, the mixture with known weights is again ranked as the best model at

most catchments for $\tau = 0.80$, while for $\tau = 0.90$ the GP distribution is considered to be the best model at most catchments. For the quantile score, the best model at each location seems to be somewhat evenly distributed among the models, especially for $\tau = 0.90$. This indicates that the quantile score might not be able to differentiate between the four models, or that the performance of each model depends on the catchment. Overall, which model that is considered to be the best depends heavily on the scoring rule used for validation.

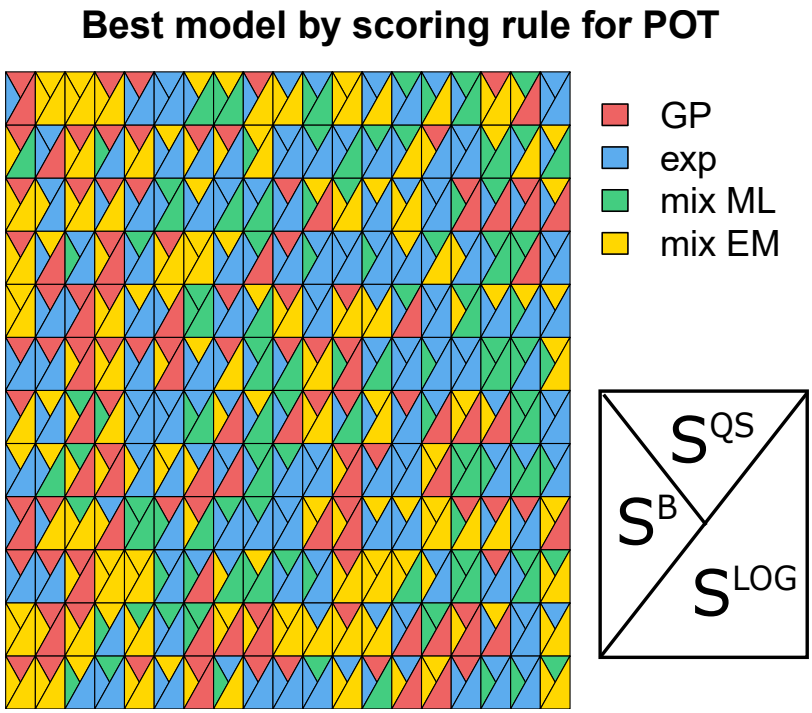

Figure 16: Portrait diagram of the best POT model at each catchment. A tile in the grid corresponds to one catchment. The lower half of the tile gives the best model according to the logarithmic score, S^{LOG}, while the upper left and upper right part gives the best model when the Brier score, S^{B}, and the quantile score, S^{QS}, with $\tau = 0.90$ are applied, respectively. The catchments are sorted, from the upper left to the lower right corner, by the increasing value of the average FGP at each catchment.

Figure 16 presents a portrait diagram of the model receiving the lowest average score at each catchment, for the logarithmic score, the Brier score with a threshold corresponding to the quantile $\tau = 0.90$ and the quantile score with the same quantile. Each tile in the diagram represents one catchment, and they are sorted by the increasing value of the FGP, from the upper left to the lower right. Again, there is no obvious

pattern in the diagram.

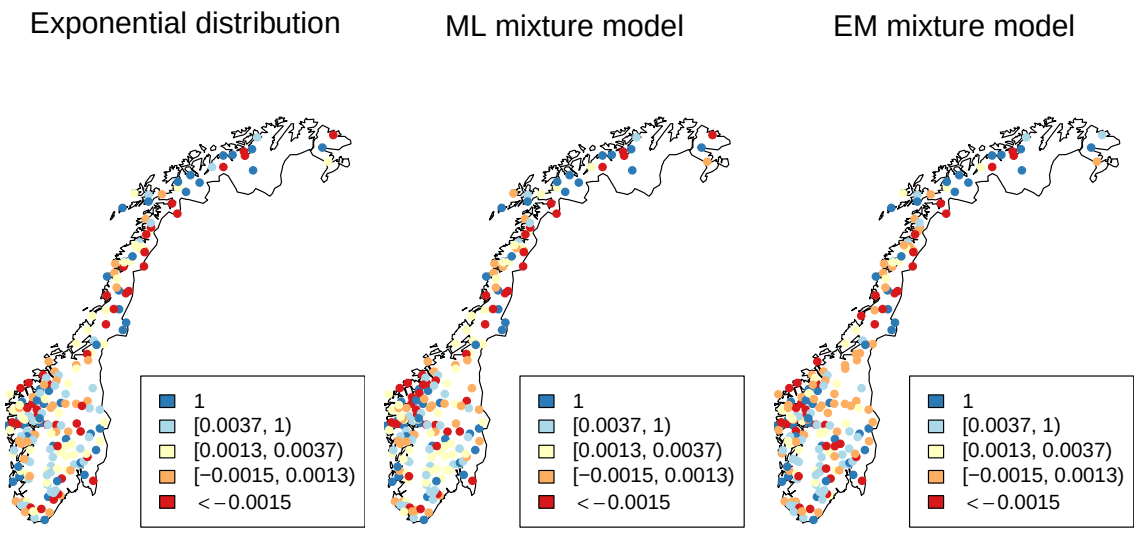

Figure 17: Maps of Norway giving the skill score of the Gumbel distribution (left), the mixture model with known weights (middle) and the mixture model with unknown weights (right), at each catchment, when using the logarithmic score and the score of the GP distribution as the reference score.

We consider the skill score with respect to the GP distribution, to further study the difference in the average score obtained by each model. The logarithmic skill score of the exponential distribution and the two exponential mixture models are given for each catchment in Figure 17. No obvious pattern in the skill score is seen from the maps in the figure. Here, the dark blue dots all represent a skill score of 1. This is obtained when the GP distribution is given an infinite score by the logarithmic scoring rule.

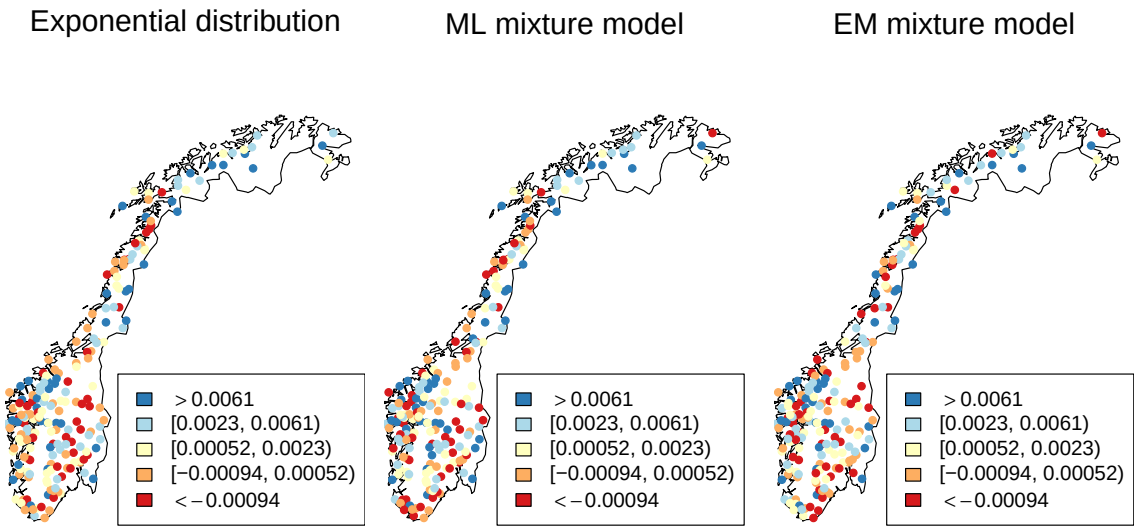

Figure 18: Maps of Norway giving the skill score of the exponential distribution (left), the mixture model with known weights (middle) and the mixture model with unknown weights (right), at each catchment, when using the Brier score with a threshold corresponding to the quantile $\tau = 0.90$ at each catchment and the score of the GP distribution as the reference score.

Figure 18 presents the Brier skill score with $\tau = 0.90$ of the exponential distribution and the exponential mixture models, at each catchment. Again, it is difficult to detect a pattern in the skill score for each model.

The quantile skill score, with $\tau = 0.90$, of each model relative to the GP distribution is given in Figure 19 below. For the exponential distribution, most of the skill scores are positive or near zero. The negative skill scores for this model are mainly obtained in the inland and northern Norway. For the two mixture models, there are more negative skill scores and no obvious pattern in the maps.

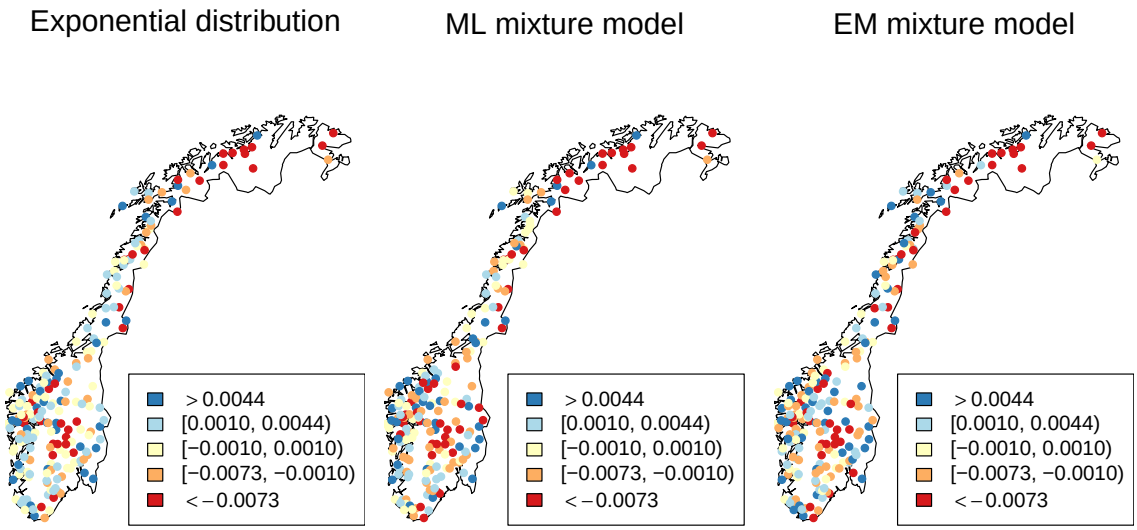

Figure 19: Maps of Norway giving the skill score of the exponential distribution (left), the mixture model with known weights (middle) and the mixture model with unknown weights (right), at each catchment, when using the quantile score with $\tau = 0.90$ and the score of the GP distribution as the reference score.

6.2.1 Detailed look at three stations

As we did for the AMS models, we now consider the estimated POT models at the three catchments, Bulken, Krinsvatn and Atnasjø. The return level plots for these catchments are presented in Figure 20, 21 and 22, respectively. Estimated parameters for the GP distribution, the exponential distribution and the two-component mixture of exponentials with both known and unknown weights, at these catchments, are given in Table 9, 11 and 10, respectively.

The results for Bulken are similar to those obtained for AMS. The GP distribution follows the observed data well, and the other three models estimate slightly higher return levels for the largest observations. However, here the estimated exponential and mixture models are approximately equal, unlike the Gumbel and mixture models for AMS at Bulken in Figure 12.

The return level plot for Krinsvatn in Figure 21 shows more variation between the models than for Bulken. Here we do not see the overfitting problem for the EM mixture model, as for Krinsvatn in the AMS approach (see Figure 13). A reason for this can be that the two-component mixture of exponentials have two less parameters

than the two-component mixture of Gumbels, which leaves it less prone to overfitting. Results from Atnasjø in Figure 22 show the same, namely that EM mixture model does not overfit the data at Atnasjø for POT modelling. The estimated models are for Atnasjø quite similar, but the difference between the return level estimates of the models increases with the return period.

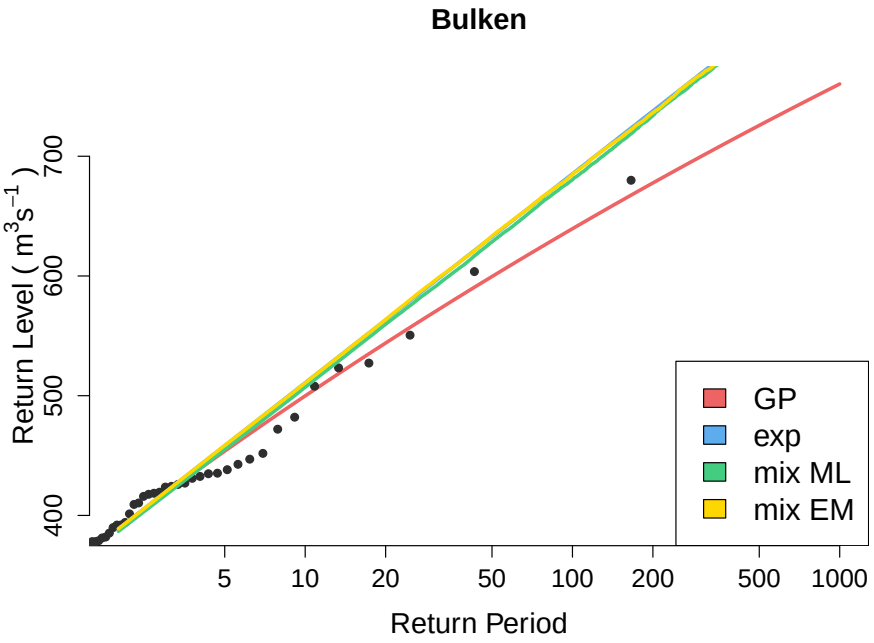

Figure 20: The estimated return level as a function of return period when the GP distribution (red), the exponential distribution (blue), the mixture model with known weights (green) and the mixture model with unknown weights (yellow) are applied. The black dots denotes the observed flood values.

Table 9: The estimated parameters of the four different models for the POT from Bulken.

Method	Parameters	
GP	$\sigma = 80.52$	$\xi = -0.063$
exp	$\lambda = 0.013$	
mix ML	$\omega_1 = 0.70$	$\lambda_1 = 0.013$
		$\lambda_2 = 0.014$
mix EM	$\omega_1 = 1$	$\lambda_1 = 0.013$
		$\lambda_2 = 0.021$

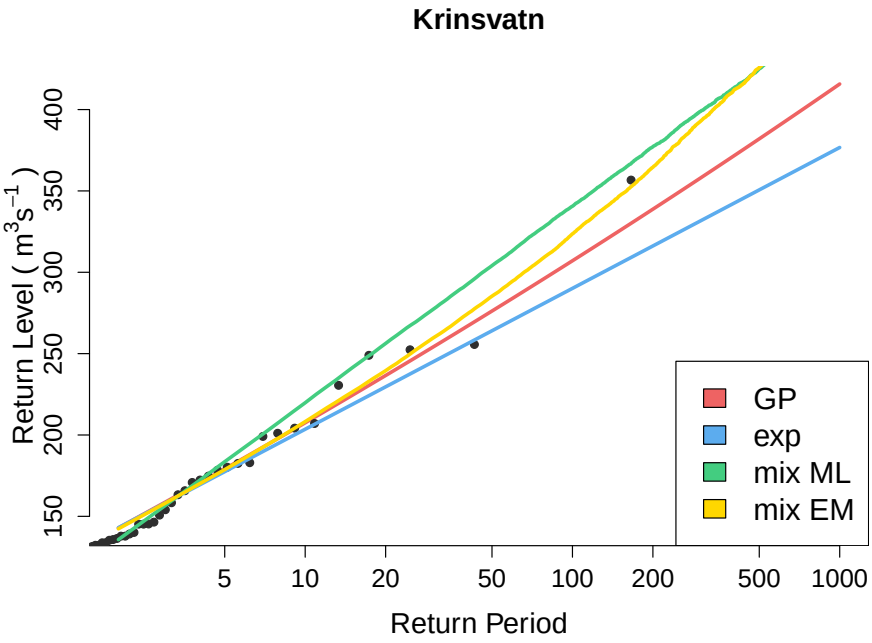

Figure 21: The estimated return level as a function of return period when the GP distribution (red), the exponential distribution (blue), the mixture model with known weights (green) and the mixture model with unknown weights (yellow) are applied. The black dots denotes the observed flood values.

Table 10: The estimated parameters of the four different models for the POT from Krinsvatn

Method	Parameters	
GP	$\sigma = 36.17$	$\xi = 0.039$
exp	$\lambda = 0.026$	
mix ML	$\omega_1 = 0.78$	$\lambda_1 = 0.023$
		$\lambda_2 = 0.056$
mix EM	$\omega_1 = 0.92$	$\lambda_1 = 0.029$
		$\lambda_2 = 0.014$

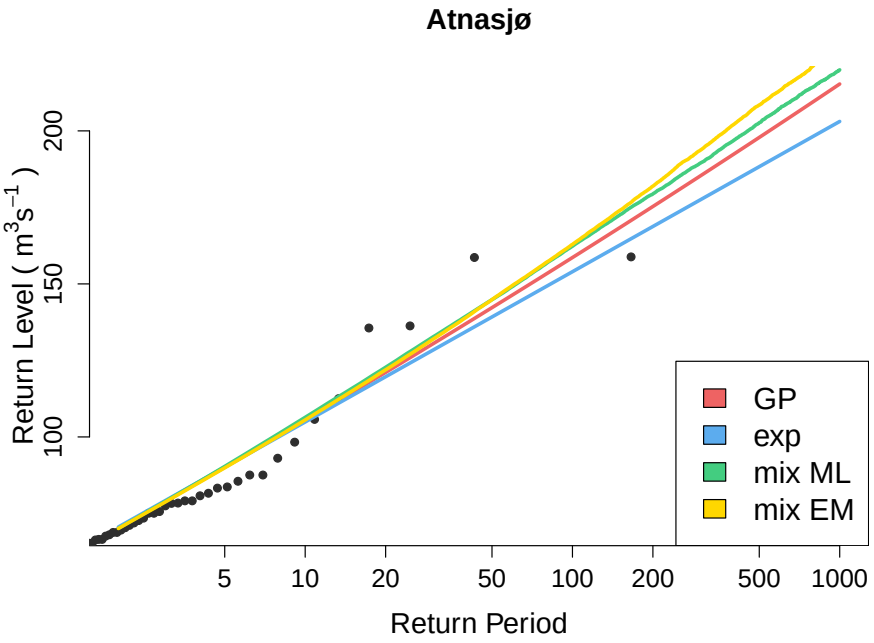

Figure 22: The estimated return level as a function of return period when the GP distribution (red), the exponential distribution (blue), the mixture model with known weights (green) and the mixture model with unknown weights (yellow) are applied. The black dots denotes the observed flood values.

Table 11: The estimated parameters of the four different models for the AMS from Atnasjø.

Method	Parameters	
GP	$\sigma = 20.72$	$\xi = 0.028$
exp	$\lambda = 0.047$	
mix ML	$\omega_1 = 0.52$	$\lambda_1 = 0.061$
		$\lambda_2 = 0.039$
mix EM	$\omega_1 = 0.82$	$\lambda_1 = 0.053$
		$\lambda_2 = 0.031$

Table 12 presents the results of the cross validation when the logarithmic score was applied as the scoring rule, for the three catchments Bulken, Krinsvatn and Atnasjø. The average scores for each model are given, and corresponding standard error estimates obtained with bootstrap are shown in the parentheses. For all three catchments, the logarithmic score judges the exponential distribution to be the best model for the POT series. However, the scores for each model, at each catchment, are quite similar. In fact, by applying a paired t-test to the scores of two different models, at a 95% significance level, a difference is only detected between the exponential distribution and the mixture model with known weights for Bulken. For Atnasjø the only significant difference is between the scores of the two mixture models. For Krinsvatn, there is no significant difference in the scores of various models, according to the paired t-test.

Table 12: Cross validation with the logarithmic score at Bulken, Krinsvatn and Atnasjø, respectively. The best score for each scoring rule is highlighted in boldface. Standard error estimates corresponding to each average score, obtained by bootstrap, are given in the parentheses.

	Logarithmic score		
Method	Bulken	Krinsvatn	Atnasjø
GP	5.34 (0.076)	4.65 (0.080)	4.08 (0.11)
exp	**5.33** (0.071)	**4.63** (0.073)	**4.07** (0.11)
mix ML	5.33 (0.073)	4.64 (0.074)	4.07 (0.12)
mix EM	5.34 (0.073)	4.64 (0.076)	4.08 (0.12)

The average scores for each model at the three catchments, obtained using the Brier score with thresholds corresponding to the quantiles $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$, are presented in Table 13. Here, there is not much variation in the scores for the various models at each catchment. For Bulken, the paired t-test at significance level 95% concludes that there is only a significant difference between the score of the EM mixture model and the other three models for $\tau = 0.95$. The same test detects no significant difference between the Brier scores of the various models at Krinsvatn and Atnasjø.

Table 13: Cross validation with the Brier score, with thresholds corresponding to $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.98$, at Bulken, Krinsvatn and Atnasjø, respectively. The best score for each scoring rule is highlighted in boldface. Standard error estimates corresponding to each average score, obtained by bootstrap, are given in the parentheses.

Bulken			
Method	Brier score $\tau = 0.80$	Brier score $\tau = 0.90$	Brier score $\tau = 0.95$
GP	0.16 (0.018)	0.094 (0.018)	**0.050** (0.015)
exp	**0.16** (0.018)	0.094 (0.019)	0.051 (0.015)
mix ML	0.16 (0.018)	**0.094**(0.018)	0.051 (0.015)
mix EM	0.17 (0.018)	0.096 (0.018)	0.054 (0.015)
Krinsvatn			
Method	Brier score $\tau = 0.80$	Brier score $\tau = 0.90$	Brier score $\tau = 0.95$
GP	0.16 (0.015)	0.093 (0.016)	0.052 (0.014)
exp	0.16 (0.015)	0.093 (0.016)	0.051 (0.014)
mix ML	**0.16** (0.015)	0.093 (0.017)	**0.051** (0.014)
mix EM	0.16 (0.015)	**0.093** (0.017)	0.051 (0.014)
Atnasjø			
Method	Brier score $\tau = 0.80$	Brier score $\tau = 0.90$	Brier score $\tau = 0.95$
GP	0.17 (0.024)	0.11 (0.024)	0.055 (0.024)
exp	0.17 (0.024)	0.11 (0.024)	**0.055** (0.022)
mix ML	**0.17** (0.024)	**0.10** (0.024)	0.055 (0.023)
mix EM	0.17 (0.024)	0.11 (0.024)	0.055 (0.023)

The average quantile scores obtained in the cross validation procedure for Bulken, Krinsvatn and Atnasjø are given in Table 14. At Bulken, the average score of the EM mixture model is somewhat higher than the three other scores, for each quantile $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$. The paired t-test at significance level 95% reports a significant difference in the score of the EM mixture model compared to the other three models for $\tau = 0.90$ and $\tau = 0.95$. For Krinsvatn, the exponential model receives the best score for $\tau = 0.90$ and $\tau = 0.95$, but there is only a significant difference in the score of the exponential model and the EM mixture model for $\tau = 0.95$. For the case $\tau = 0.80$, the EM mixture model achieves the lowest score. However, no significant difference between the scores of the four models are detected. The exponential distribution also receives the best average score at Atnasjø for $\tau = 0.90$ and $\tau = 0.95$, but there is no significant difference between the scores. For $\tau = 0.80$,

the mixture model with FGP as weights is given the lowest score, and there is a significant difference between its average score and the average scores of the GP distribution and the EM mixture model.

Table 14: Cross validation with quantile score, for $\tau = 0.80$, $\tau = 0.90$ and $\tau = 0.95$, at Bulken, Krinsvatn and Atnasjø, respectively. The best score for each scoring rule is highlighted in boldface. Standard error estimates corresponding to each average score, obtained by bootstrap, are given in the parentheses.

Bulken			
Method	Quantile score $\tau = 0.80$	Quantile score $\tau = 0.90$	Quantile score $\tau = 0.95$
GP	23.38 (2.14)	15.92 (1.82)	10.67 (1.51)
exp	**23.38** (2.14)	15.91 (1.80)	**10.60** (1.35)
mix ML	23.38 (2.23)	**15.90** (1.83)	10.60(1.34)
mix EM	23.98 (2.17)	16.98 (1.76)	11.34 (1.34)
Krinsvatn			
Method	Quantile score $\tau = 0.80$	Quantile score $\tau = 0.90$	Quantile score $\tau = 0.95$
GP	12.44 (1.38)	9.52 (1.19)	6.49 (0.99)
exp	12.45 (1.39)	**9.51** (1.20)	**6.41**(1.01)
mix ML	12.45 (1.34)	9.53 (1.13)	6.45 (0.97)
mix EM	**12.44** (1.35)	9.52 (1.22)	6.46 (0.98)
Atnasjø			
Method	Quantile score $\tau = 0.80$	Quantile score $\tau = 0.90$	Quantile score $\tau = 0.95$
GP	6.98 (1.20)	5.51 (1.04)	3.92 (0.82)
exp	6.98 (1.17)	**5.50** (1.00)	**3.91** (0.88)
mix ML	**6.95** (1.20)	5.51 (1.05)	3.93 (0.78)
mix EM	6.99 (1.17)	5.51 (1.07)	3.93 (0.83)

6.3 Return level estimates

To compare return level estimates by the various models in this study, and the stability of these estimates, we consider boxplots of return level estimates. The different estimates for each catchment are obtained as explained in Section 5.3. For this, we again consider the three catchments Bulken, Krinsvatn and Atnasjø. We choose to look at the return periods $p = 100$ and $p = 1000$ for all eight models. That is, the four models for AMS and the four models for POT.

Boxplots of the estimated return levels at Bulken are given in Figure 23. We see that, for both return periods, the EM mixture of Gumbels occasionally gives some very high estimates. Such high variation in the estimates indicate that this model is unstable. With the exception of these high estimates, this model gives return level estimates for $p = 100$ and $p = 1000$ close to the estimates by the Gumbel distribution and the mixture of Gumbels with known weights, for AMS, and the two mixture models with exponentially distributed components, for POT. For these models, the median estimate of the 100-year flood is close to the currently largest observed flood. The GEV and GP distributions give lower estimates compared to the other models, while the exponential distribution gives higher estimates.

For the return level estimates at Krinsvatn, in Figure 24, we see a large variance in the estimates by the two exponential mixture models for POT. As for Bulken, the estimates by the Gumbel distribution and the exponential distribution are fairly stable. In contrast to Bulken, the GEV and GP models here does not give the lowest estimates for the 100- and 1000-year floods. The median 100- and 1000-year flood estimate by the EM exponential mixture are almost the same, which is not realistic.

Figure 25 presents boxplots of the estimated return levels at Atnasjø, for the return periods $p = 100$ and $p = 1000$. Again, the Gumbel and exponential distributions give the most stable estimates for both return periods. Here, the exponential distribution and the two exponential mixtures tend to give lower return level estimates than the other models.

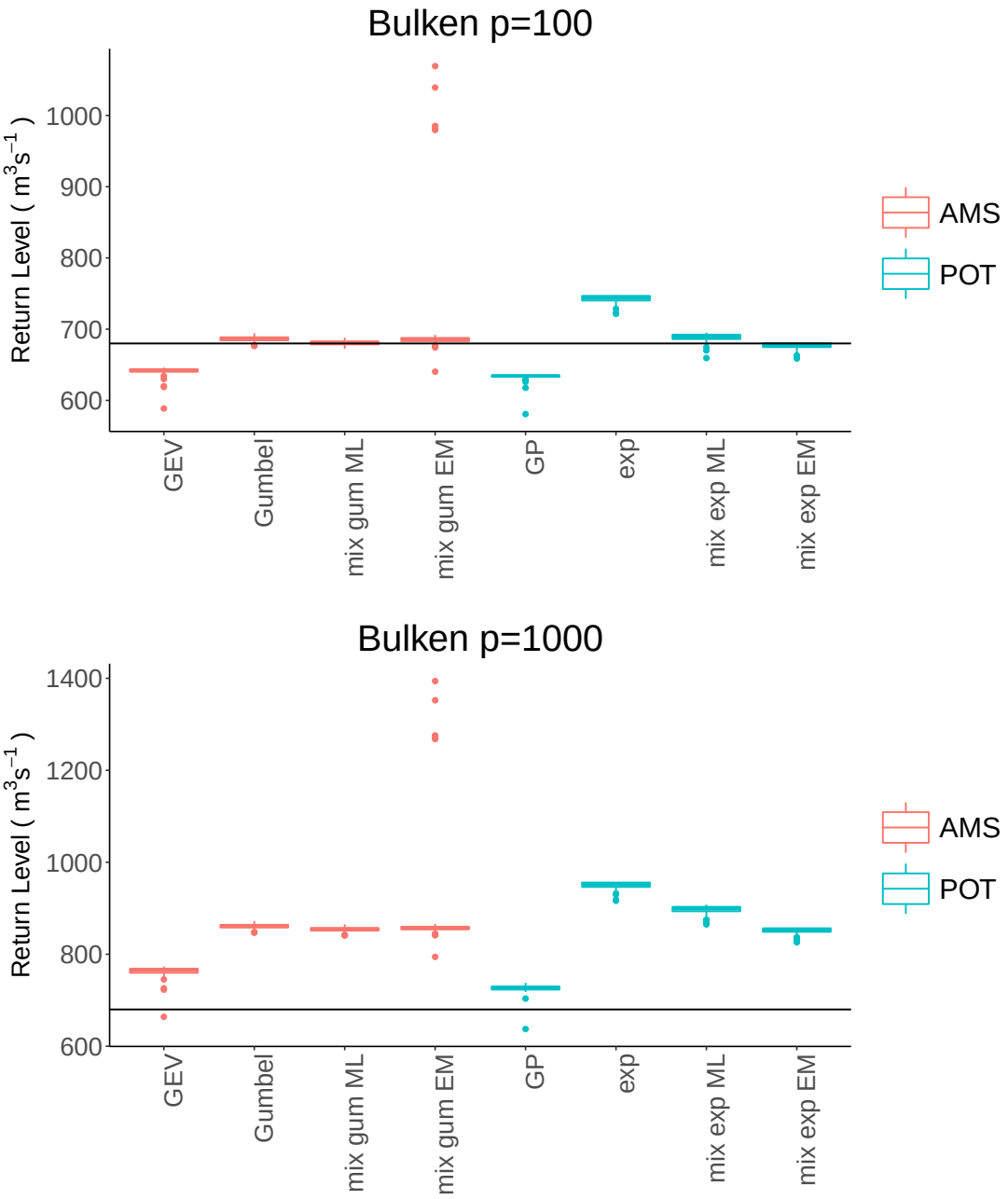

Figure 23: Boxplots of return level estimates of the 100-year (upper) and 1000-year flood (lower) at Bulken, using the four models for AMS (red) and the four models for POT (blue). The different estimates are obtained by repeatedly removing one year of data from the AMS and POT series. The horizontal line gives the largest observed flood value at Bulken.

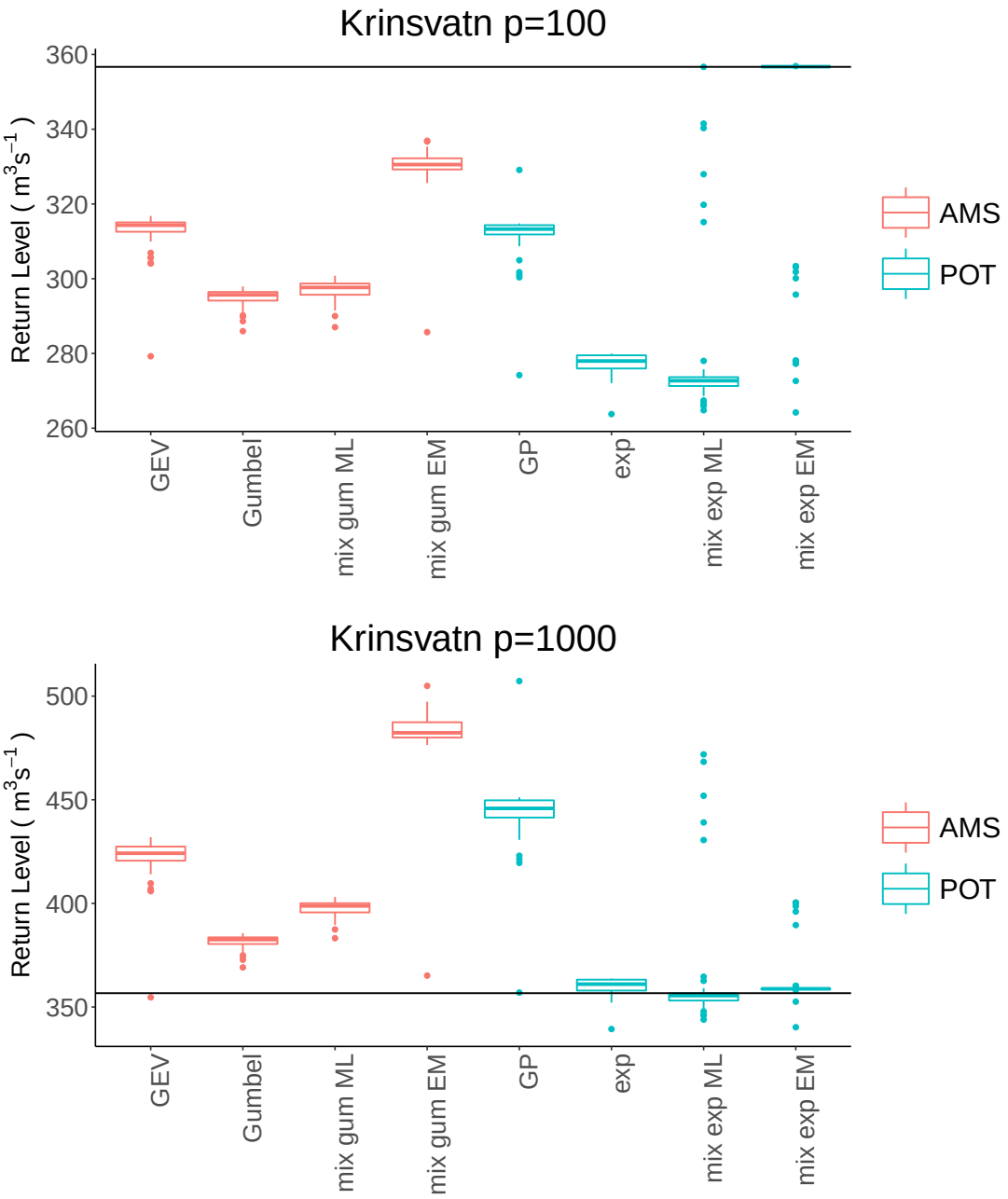

Figure 24: Boxplots of return level estimates of the 100-year (upper) and 1000-year flood (lower) at Krinsvatn, using the four models for AMS (red) and the four models for POT (blue). The different estimates are obtained by repeatedly removing one year of data from the AMS and POT series. The horizontal line gives the largest observed flood value at Krinsvatn.

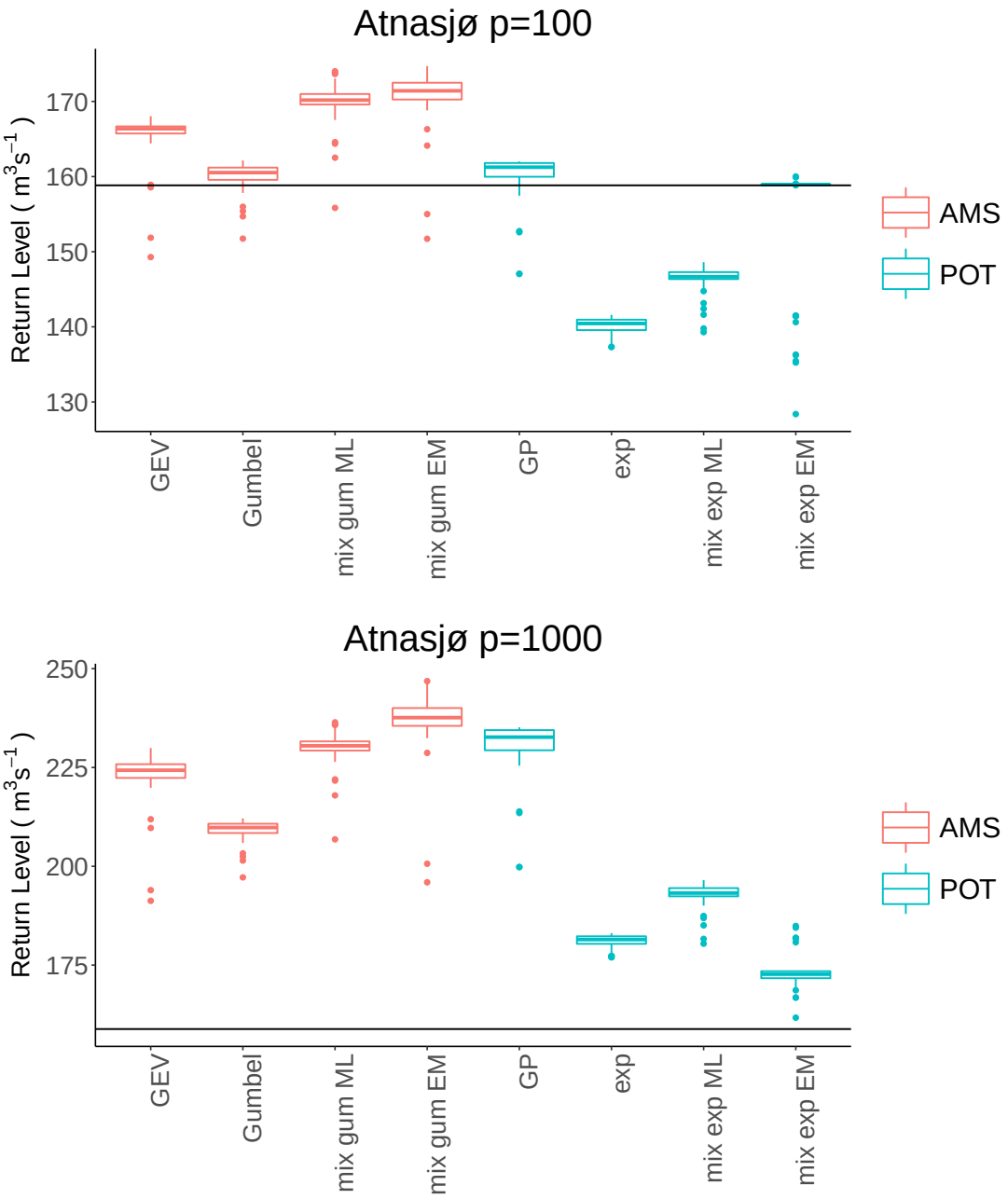

Figure 25: Boxplots of return level estimates of the 100-year (upper) and 1000-year flood (lower) at Atnasjø, using the four models for AMS (red) and the four models for POT (blue). The different estimates are obtained by repeatedly removing one year of data from the AMS and POT series. The horizontal line gives the largest observed flood value at Atnasjø.

7 Discussion

When comparing the models by scoring rules, it is difficult to differentiate between the models, both for the case of AMS and POT modelling. The model which obtained the lowest score depends heavily on the scoring rule and the catchment considered. Judging by the logarithmic score, the Gumbel distribution and the exponential distribution give most reliable estimates, for the case of AMS and POT series respectively. When using the Brier score and the quantile score, the best model for AMS depends heavily on the catchment and quantile considered. This is also the case for the quantile score of the POT models, while the Brier score most often gives the best score to the mixture with known weights. However, for the three catchments we studied in detail, there is often no significant difference between the score of each model. How the different scoring rules lead to such different results can be understood by looking at Figure 6. From this plot we see that the various scoring rules penalize and reward different aspects of the predicted model.

When applying the Brier score with high thresholds or the quantile score with high quantiles, the scores are based on small amount of data. A difference in the performance of the models is easier detected by considering skill scores. For the AMS models, the EM mixture model and the ML mixture model perform bad relative to the GEV distribution, judging by the logarithmic skill score and the Brier skill score, respectively. There is less difference in the skill score of the POT models.

From Figure 14 and 13 we see that the Gumbel mixture model estimated by the EM algorithm tends to overfit the data. This model requires the estimation of five parameters, which might be too many parameters for the amount of data available. We only use the observations which have a value for FGP, which from Table 1 we see reduces the length of the AMS and POT series by about one half. It could be of interest to investigate the performance of the EM mixture models when using all available data. Also, to prevent unrealistic parameter estimates, as e.g. the very small shape parameter of the second component in the EM mixture given in Table 4, we could impose constraints on the parameters of each mixture component.

From the return level estimates at Bulken, Krinsvatn and Atnsjø we see that the various mixture models are unstable compared to the other models. In particular, the EM Gumbel mixture gives some very high estimates at Bulken (see Figure 23) and the EM exponential mixture gives unexpected results at Krinsvatn (see Figure 24). This makes the EM mixture models unfit for practical applications. The Gumbel distribution and the exponential distribution obtain the most stable estimates at these catchments.

Of the mixture models we have studied, the ones using the FGP as mixture weights seem to be most relevant for future use. However, these require an FGP value corresponding to each flood observation. As mentioned before, this requirement reduces the amount of available flood data. In FFA the sample sizes are already considered to be small compared to return periods of interest, and leaving out valuable data from the analysis is not attractive.

As described in Section 3.2.2, the EM algorithm is sensitive to starting values. We implemented a simple random initialization procedure to start the algorithm from different points and chose the parameters that obtained the overall maximum likelihood. There exist several other methods to generate starting values for this algorithm, see e.g. Biernacki et al. (2003) for comparison of various approaches. Our implementation could have benefited from choosing a more sophisticated procedure for starting values. However, complicating the estimation procedure would make it a less attractive model to apply.

Although the EM algorithm is widely used to estimate parameters of mixture models, it clearly has some drawbacks. Other methods could be applied to estimate the parameters. For example, Shin et al. (2014) applied a metaheuristic maximum likelihood (MHML) method to fit Gaussian mixture distributions for flood frequency analysis. They concluded that this method performed better than the EM algorithm for small sample sizes.

In the form of mixture models, we have tried to incorporate knowledge about flood generating processes into the modelling of flood values. When assuming that there are two different flood generating processes, it would be natural to also consider modelling the problem as e.g. a sum of two Gumbel random variables or two exponential random variables, for AMS and POT data, respectively. Loaiciga and Leipnik (1999) derived the distribution of a sum of two independent Gumbel random variables, and Nadarajah (2008) later generalized this result to a linear combination of Gumbel random variables. However, the derived distribution is quite complex and involves an infinite sum of hypergeometric functions, limiting the practical usefulness of the distribution. Efforts have been made to approximate the distribution, see e.g. Marques et al. (2015).

Uncertainty in the flood data have not been accounted for in this analysis. The observed values from each catchment are in fact stage levels and not discharge values. A stage-discharge rating curve is used to transform the measured stage levels to discharge values. This of course implies an uncertainty in the discharge data. Steinbakk

et al. (2016) investigated the effect of ignoring this uncertainty in design flood estimates and found that ignoring these features may underestimate the potential risk of flooding. However, we followed common practice and ignored the uncertainty in the stage-discharge estimates.

8 Conclusion

In this study we investigate the use of mixture models for both the AMS and POT approach to FFA, and compare the performance of these models to commonly used distributions in FFA. The GEV distribution, the Gumbel distribution and a two-component mixture of Gumbel distributions with both known and unknown weights are considered for modelling AMS. For the case of modelling POT series, the GP distribution, the exponential distribution and a two-component mixture of exponential distributions with both known and unknown weights are studied. The performance of the different models are compared in terms of the reliability of the models and the stability of their return level estimates.

We found that the model which is considered to perform best depends on the scoring rule and the catchment. Overall, the Gumbel distribution and the exponential distribution, for the case of AMS and POT respectively, often give the most reliable and stable estimates.

References

Y. Alila and A. Mtiraoui. Implications of heterogeneous flood-frequency distributions on traditional stream-discharge prediction techniques. *Hydrological Processes*, 16 (5):1065–1084, 2002.

N. Bezak, M. Brilly, and M. Sraj. Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis. *Hydrological Sciences Journal*, 59(5):959–977, 2014.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

B. Bobée and P. F. Rasmussen. Recent advances in flood frequency analysis. *Reviews of Geophysics*, 33(S2):1111–1116, 1995.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, 78:1–3, 1950.

S. Caires. A comparative simulation study of the annual maxima and the peaks-over-threshold methods. *Journal of Offshore Mechanics and Arctic Engineering*, 138:14, 2016.

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

C. Cunnane. A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. *Journal of Hydrology*, 18(3):257 – 271, 1973.

C. Cunnane. Statistical distributions for flood frequency analysis. *Operational Hydrology Report*, (33), 1989.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

K. Engeland, H. Hisdal, and A. Frigessi. Practical extreme value modelling of hydrological floods and droughts: A case study. *Extremes*, 7:5–30, 2004.

K. Engeland, L. Schlichting, L.-E. P. Thea Wang, T. Reitan, F. Randen, E. Holmqvist, K. S. Nordtun, A. Voksø, and V. Eid. *Flomdata - Utvalg og kvalitetssikring av flomdata for flomfrekvensanalyser*. NVE, 2016. Rapport nr 85.

G. Evin, J. Merleau, and L. Perreault. Two-component mixtures of normal, gamma, and gumbel distributions for hydrological applications. *Water Resources Research*, 47(8), 2011.

FEH. *Flood Estimation Handbook,*. Institute of Hydrology, Wallingford, UK, 1999.

A. Ferreira and L. de Haan. On the block maxima method in extreme value theory: Pwm estimators. *Ann. Statist.*, 43(1):276–298, 02 2015.

R. A. Fisher and L. H. C. Tippett. Limiting Forms of the Frequency Distribution of the Largest and Smallest Member of a Sample. *Proceedings of the Cambridge Philosophical Society*, 24:180–190, 1928.

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

T. Gneiting and M. Katzfuss. Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151, 2014.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

I. J. Good. Rational decisions. *Journal of the Royal Statistical Society*, 14:107–114, 1952.

T. S. Gubareva and B. I. Gartsman. Estimating distribution parameters of extreme hydrometeorological characteristics by lmoment method. *Water Resources*, 37(5): 437–445, 2010.

E. J. Gumbel. Floods estimated by probability methods. *Engineering News-Record*, (134), 1945.

V. Hasselblad. Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64(328):1459–1471, 1969.

J. E. Heffernan and A. G. Stephenson. *ismev: An Introduction to Statistical Modeling of Extreme Values*, 2016. URL `https://CRAN.R-project.org/package=ismev`. R package version 1.41.

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

J. M. Landwehr, N. C. Matalas, and J. R. Wallis. Probability weighted moments compared with some traditional techniques in estimating gumbel parameters and quantiles. *Water Resources Research*, 15(5):1055–1064, 1979.

M. Lang, T. Ouarda, and B. Bobe. Towards operational guidelines for over-threshold modeling. *Journal of Hydrology*, 225(34):103 – 117, 1999.

S. C. Larson. The shrinkage of the coefficient of multiple correlation. *J. Edic. Psychol.*, 22:45–55, 1931.

H. A. Loaiciga and R. B. Leipnik. Analysis of extreme hydrologic events with gumbel distributions: marginal and additive cases. *Stochastic Environmental Research and Risk Assessment*, 13(4):251–259, 1999.

Lovdata. *Dam safety regulation (Damsikkerhetsforskriften)*, 2009. URL `https://lovdata.no/dokument/SF/forskrift/2009-12-18-1600`. hefte 14.

Lovdata. *Byggteknisk forskrift (TEK10)*, 2010. URL `https://lovdata.no/dokument/SF/forskrift/2010-03-26-489?q=tek%2010`. hefte 5.

H. Madsen, P. Rasmussen, and D. Rosbjerg. 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, 1997.

F. J. Marques, C. A. Coelho, and M. de Carvalho. On the distribution of linear combinations of independent gumbel random variables. *Statistics and Computing*, 25(3):683–701, 2015.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Wiley, 1996.

X.-L. Meng and S. Pedlow. Em: A bibliographic review with missing articles. pages 24–27, 1992.

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

F. Mosteller and J. Tukey. Data analysis, including statistics. In G. Lindzey and E. Aronson, editors, *Revised Handbook of Social Psychology*, volume 2, pages 80–203. Addison Wesley, 1968.

S. Nadarajah. Exact distribution of the linear combination of p gumbel random variables. *International Journal of Computer Mathematics*, 85(9):1355–1362, 2008.

J. Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3:119–131, 1975.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL `https://www.R-project.org/`.

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

D. Rosbjerg, H. Madsen, and P. F. Rasmussen. Prediction in partial duration series with generalized pareto-distributed exceedances. *Water Resources Research*, 28 (11):3001–3010, 1992.

F. Rossi, M. Fiorentino, and P. Versace. Two-component extreme value distribution for flood frequency analysis. *Water Resources Research*, 20(7):847–856, 1984.

W. Seidel, K. Mosler, and M. Alker. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52(3):481–487, 2000.

J.-Y. Shin, J.-H. Heo, C. Jeong, and T. Lee. Meta-heuristic maximum likelihood parameter estimation of the mixture normal distribution for hydro-meteorological variables. *Stochastic Environmental Research and Risk Assessment*, 28(2):347–358, 2014.

G. H. Steinbakk, T. L. Thorarinsdottir, T. Reitan, L. Schlichting, S. Hølleland, and K. Engeland. Propagation of rating curve uncertainty in design flood estimation. *Water Resources Research*, 52:6897–6915, 2016.

S. Steinius, P. Glad, T. Wang, and T. Væringstad. *Veileder for flomberegninger i små uregulerte felt*. NVE, 2015. Veileder nr. 7/2015.

P. Waylen and M.-k. Woo. Prediction of annual floods generated by mixed processes. *Water Resources Research*, 18(4):1283–1286, 1982.

D. Wilson, H. Hisdal, and D. Lawrence. Has streamflow changed in the nordic countries? recent trends and comparisons to hydrological projections. *Journal of Hydrology*, 394:334–346, 2010.

C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.