Høgskulen på Vestlandet

Master Thesis in Climate Change Management

GE4-304

Predefinert informasjon

Startdato:	10-05-2018 09:00	**Termin:**	2018 VÅR
Sluttdato:	28-05-2018 14:00	**Vurderingsform:**	Norsk 6-trinnsskala (A-F)
Eksamensform:	Master's thesis		
SIS-kode:	203 GE4-304 1 MA 2018 VÅR		
Intern sensor:	Thorben Dunse		

Deltakar

Namn:	Simen Norheim
Kandidatnr.:	610
HVL-id:	131265@hvl.no

Gruppe

Gruppenamn:	Enmannsgruppe
Gruppenummer:	7
Andre medlemmer i gruppa:	Deltakaren har levert inn i enkeltmannsgruppe

Flood frequency analysis: Comparing annual maximum series and peak over threshold

A case study for Norway

Simen Norheim

Master Thesis in Climate Change Management

Department of Environmental Sciences, Faculty of Engineering and Science

WESTERN NORWAY UNIVERSITY OF APPLIED SCIENCES

Sogndal
[May, 2018]

Western Norway University of Applied Sciences

Flood frequency analysis: Comparing annual maximum series and peak over threshold

A case study for Norway

Master thesis in Climate Change Management

Author: Simen Norheim	Author sign.

Thesis submitted: Spring 2018	Open/confidential thesis
Main Supervisor: Florian Kobierska Baffie, HVL Co-supervisors: Kolbjørn Engeland, NVE	
Keywords: Flood frequency analysis, Annual maximum series, Peak over threshold, Extreme value distributions, Goodness-of-fit	Number of pages: 48 Appendix: 1 Sogndal, 28.05, 2018 Place/Date/year

This thesis is a part of the master's program in Climate Change Management (Planlegging for klimaendringer) at the Department of Environmental Sciences, Faculty of Engineering and Science at the Western Norway University of Applied Sciences. The author(s) is responsible for the methods used, the results that are presented and the conclusions in the thesis.

Preface

This study is the final thesis of the *Climate change management* masters course at Høgskulen på Vestlandet (HVL). I would like to thank my supervisors Florian Kobierska Baffie at HVL and Kolbjørn Engeland at the Norwegian Water Resource and Energy Directorate (NVE) for providing the data used in this study and for great help and guidance through the course of this study. I would also like to thank the NVE Hydrological modelling department for helpful comments and for providing me a workplace where while writing my thesis.

Abstract

Flood frequency analysis (FFA) is used to estimate the frequency of flood events and the magnitude of extreme flood events expected to occur every T years. In many countries it is common practice to conduct an FFA using annual maximum series (AMS) as opposed to peak over threshold (POT). This is because extraction of AMS data is fairly easy, while extraction of POT data requires selection of threshold and ensuring peak independence. Another recurrent discussion in the literature is about the selection of extreme value distributions to use in an FFA. For AMS data I used the generalized extreme values (GEV) and Gumbel (GUM) distributions as these are commonly used AMS distributions. For POT data I fitted the generalized pareto (GP) and the exponential (EXP) distributions as these are the POT counterparts of the GEV and GUM distributions. The models have been compared in reliability, predictive ability and stability. The measure of each model's performance in these respects has been estimated using various tests such as goodness-of-fit tests and scoring rules. The models based on POT data were found to both fit and predict better than the AMS models. This discrepancy in performance increases as the average number of events per year in the POT datasets increases. The difference in performance between distributions fitted to the same data is much smaller. The 2-parameter GUM and EXP distributions show better predictive ability and much less variance when estimating design floods for large recurrence intervals. Models based on POT data were also found to be best suited for catchments where precipitation is a major contributor to flood events. Whereas AMS and POT models performed similarly for catchments where flood events are caused by exclusively meltwater. In essence this study found that utilization of POT over AMS data substantially increase the performance of FFA models, but for some regions AMS models will perform similarly.

Samandrag på norsk

Flomfrekvensanalyse brukes til å estimere frekvensen av flom hendelser og størrelsen på ekstreme hendelser som forventes å skje hvert T år. I mange land er det vanlig å gjennomføre en flomfrekvensanalyse ved bruk av årlig maksimalverdi (AMS) til fordel for flommer over terskel (POT). Det er fordi uttak av AMS data er relativt lett, mens uttak av POT data krever valg av terskel og forsikring at toppene er uavhengige av hverandre. Et annet hyppig diskusjonstema i faglitteraturen er valg av fordeling å bruke i analysen. Til AMS data har jeg tilpasset generalisert ekstrem verdi- (GEV) og Gumbel (GUM) fordelingene ettersom disse er hyppig brukte AMS fordelinger. Til POT data har jeg tilpasset generalisert pareto- (GP) og eksponentiell (EXP) fordelingene siden disse er POT motpartene til GEV og GUM fordelingen. Modellene har blitt sammenlignet med henhold til pålitelighet, prediksjons egenskaper og stabilitet. Hver modells ytelse i henhold til disse målene har blitt estimert ved bruk av forskjellige tester. Modellene basert på POT data viste både bedre tilpassing prediksjons egenskaper enn AMS modellene. Denne forskjellen i ytelse øker når gjennomsnittlig antall flommer per år i POT datasettet øker. Forskjellen i ytelse fra modeller tilpasset samme data er mye mindre. 2-parameter fordelingene GUM og EXP viser bedre prediksjons egenskaper samt mye mindre variasjon når dimensjonerende flom beregnes for høye gjentaksintervaller. Modeller basert på POT data viste seg også best egnet for nedbørsfelt hvor nedbør bidrar mest til flomhendelser. AMS og POT modeller er omtrent like gode for nedbørsfelt hvor alle større flommer er snøsmelteflommer. Dette studie har funnet at bruk av POT til fordel for AMS data gir vesentlig bedre modeller i flomfrekvensanalyse, men for noen regioner vil AMS og POT modeller gi lignende resultater.

Table of contents

Figure list

1 Introduction

Design flood estimation is an important tool for those working with management of rivers and riparian areas, this is particularly the case for those working with flood mitigation. A design flood estimate is given as discharge (m^3), but can be translated to local stage levels (m) and velocity (m/s) using hydraulic models. Depending on the mitigative work in question one might use different estimates, whereas an area planner might be concerned with stage levels, an engineer might consider erosion as a function of velocity instead. Further, the accuracy of these estimates is a highly important aspect of this, particularly for those working on projects with potentially very high risk of infrastructural damage and injuries, i.e. dam construction. Accuracy of design flood estimate is not just important for this due to the risk of failure, but also because once constructed they are expensive to alter to accommodate new estimates should it be established that previous estimates were too modest. Investigation of the accuracy of design flood estimates will remain highly important as we see a water cycle intensification (Huntington 2006).

In response to flood hazards, countries have different standards/regulations for construction of flood mitigation measures. Floods are often described by the term return period (T), in practice a flood magnitude with return period T will occur or be exceeded on average once every T-years. Safety standards and regulations relating to flood dangers varies from nation to nation, but a variety of buildings and infrastructure is often set to safety classes of 100- or 200-year floods. According to Norwegian building regulations buildings and infrastructure falls under one of the following classes 20- (F1), 200- (F2) and 1000- (F3) year floods. Which class a given building falls under depends on the potential impact of a flood. Critical buildings such as hospitals or other first responder fall under class F3 (1000 year), housing or other buildings that would likely result in injuries if flooded is class F2 (200 year) while uninhabited garages and the like falls under F1 (20 year) (2017). Dams have two safety classes 500- and 1000-year return periods (Lovdata 2009).

Considering the gauged record of most river and waterways is shorter than 100 years (or even shorter than 50), design flood discharges beyond the gauged record often requires extrapolation from the gauged data. With short record lengths extrapolation to design flood estimates may happen far beyond the largest event on record, thus the estimates are uncertain and may vary highly depending on the method (Nagy et al. 2017). Flood frequency analysis (FFA) entails fitting probability distributions to the observed data; this in turn enables estimation of design flood values for any return period of interest.

Instances where a flood frequency analysis is desired for an ungauged river, a few methods have commonly been employed. One method is the prediction of ungauged basins (PUB), discussed in depth in (Hrachowitz et al. 2013). The PUB method investigates the correlation between different hydrological properties of a gauged catchment (such as various discharge values) and the catchments climatic and spatial/geological properties. The correlation between these properties is then applied to an ungauged river with known climatic and spatial/geological properties to predict discharge (Chen et al. 2006). Another method, referred to as regional frequency analysis estimates design flood values based on adjacent, gauged rivers. Regional frequency analysis considers basins determined to be hydrologically homogeneous to the catchment of interest, hydrological homogeneous regions often refer to contiguous catchments. The data gathered from the gauged river is then adapted to the characteristics of the ungauged basin (Ouarda et al. 2001). In this study we only consider gauged catchments.

Two different data types are used in flood frequency analysis, the annual maximum series (AMS) data and the peaks over threshold (POT) data. The annual maximum series take into consideration the highest discharge level of a river for each year. This can be the hydrological year, typically ending when it is common for precipitation to fall as snow (since this water will not drain until the following year), but can also be considered in calendar year. Either way, this means that for n years the dataset will be composed of n data points. The advantage of using AMS data is that selecting the highest discharges per year is a simple task and ensuring that floods are independent of other floods in the dataset is only relevant to investigate if the floods coincide around the turn of the year. Which type of year one considers (i.e. water- or calendar year) could impact the importance of ensuring independence of events, especially if climatic factors vary and may thus cause mild periods at times previously below freezing. Problems with the AMS approach is that only one event is considered per year, an accurate FFA would then require a station be active for 30-50 years if we use the recommendations of (Midttømme et al. 2011, Castellarin et al. 2012) for utilization of 2 and 3 parameter distributions. Finally, only selecting one event per year could result in some observations being of insignificant magnitude, this can also be problematic for catchments that are highly affected by aperiodic events like El Niño Southern Oscillation.

Peak over threshold series consider independent flood peaks that exceed a given threshold. In practice this usually means that a POT data set is comprised of more data points than years of gauging for the

station. Many studies have discussed criteria for what constitutes an independent flood event. Beard (1974) suggests flood peaks be separated by five days plus the natural logarithm of the area, additionally the discharge should drop by 75% between the peaks, Ball et al. (2016) suggests this is viable for catchments larger than 1000km^2. The independence criterion presented by Beard (1974) is cited in (USWRC 1982) among the guidelines for flood frequency analysis however, suggests independence criterion be decided based on the studied rivers characteristics. Cunnane (1979) introduced an independence criterion that the discharge between the peaks dropped ⅔ from the lower peak, and that the events occur no closer to each other than thrice the time of rise of the first event. While a POT dataset is recommended, it often include the potentially very time-consuming work of selecting independent peaks (Keast and Ellison 2013). Several studies have previously looked at the performance of peaks over threshold against annual maximum data sets. Tavares and Da Silva (1983) claims the POT dataset perform better, showing a smaller variance than the AMS once the dataset is compiled of at least 2 peaks per year. Cunnane (1973) suggests the peaks over threshold dataset is superior with 1.65 peaks per year. Bezak et al. (2014) found their peaks over threshold dataset to be better with even just 1 event per year, but it performed best when there were on average 8 peaks per year.

Suggested probability distributions for flood frequency analysis can usually be found in guidelines for different countries or areas. Castellarin et al. (2012) summarize the guidelines for countries within Europe, while the United States Water Resource Council (USWRC 1982) provides guidelines in the US through bulletin 17b, these have later been improved by (Stedinger and Griffis 2008, 2011) for flood frequency analysis in the USA. Ball et al. (2016) and the Canadian Standard Association (CSA 2012) provide guidelines for Australia and Canada respectively (Das et al. 2013). These four guidelines tell us that mainly three different distributions are suggested utilized for annual maximum series:

– Generalized extreme values (GEV) in Australia, Austria, Cyprus, Germany, France, Italy, Lithuania, Slovakia, Spain.

– Gumbel (GUM) in Canada, Finland, Greece and Norway.

– Log-Pearson III (LP3) in Australia, Lithuania, Poland, Slovenia and the USA.

For peak over threshold series the recommended probability distributions are generally the same as for annual maximum series. For several of the countries it is not specified different distribution recommendations for annual maximum and peak over threshold series, while some countries only specify their recommended distributions for annual maximum series. Often studies search for the best fitting distributions, but some adopt a well-established distribution and suggest alterations to make it more universally fitting. Stedinger and Griffis (2011) suggests keeping the log-Pearson III distribution and adding new parameters to account for the non-stationary nature of today's watersheds with regards to urbanization and climate change. Countries suggesting different distributions could reflect on how the distributions perform for different types of rivers. Many other studies have assessed how different distributions perform, Nguyen et al. (2017) found that for annual maximum rainfall analysis the generalized normal (GNO), Pearson type III (PE3) and GEV were the most fitting distributions, they recommended using the GEV distribution "due to its more solid theoretical basis". In their study, they also found that the heavily suggested 2-parameter Gumbel distribution, showed weaker in descriptive capabilities, but solid predictive ability. On the other hand, the five parameter Wakeby (WAK) distribution, outperformed all other distributions in descriptive score, but was among the worst performers in predictive ability. They suggested this was due to the five parameters of the distribution, a high number of parameters allows it to mimic watersheds with high confidence, but when predicting, many parameters can make the model to rigid, often over- or underestimating (Nguyen et al. 2017). Nagy et al. (2017) conducts a flood frequency analysis for both annual maximum and peak over threshold series, they found that Generalized Pareto (GP), Pearson type III, Log Pearson III and General extreme values all produce satisfying results, of these the GEV distribution usually performs the worst. Zalina et al. (2002) and Blain and Meschiatti (2014) test for several distributions and find the best fit to be with the GEV distribution. Blain and Meschiatti (2014) also found satisfying results using the kappa distribution, they did however not use the LP3 distribution which has been found satisfactory in several other studies. Zalina et al. (2002) tried for LP3 but found it consistently overestimating in their study. There have also been several studies looking at flood data from areas that is associated with great variety of climate and weather. Gubareva (2011) studied rivers in central Europe (Austria and France), Siberia, Taiwan and Hawaii, testing for 4 different distributions (PE3, GEV, GP and 3LN). She finds the best fitting distribution to be GEV for the majority of rivers studied, but explain that for certain rivers the GP show better performance. She suggests that extreme events in regions prone to a severe floods exhibit power type behavior (Gubareva 2011), similar to what suggested in other studies (Kidson and Richards 2005). Abida and Ellouze (2008) and Rahman et al. (2013) study rivers for Tunisia and Australia

respectively, using a variety of distributions. Abida and Ellouze (2008) concludes that the GNO distribution is the best fit for northern Tunisia while both the GEV and the GNO performs well for central/southern Tunisia. Rahman et al. (2013) conducts a flood frequency analysis in Australia and presents a table showing the top 4 best fitting distributions for each state. There the LP3 distribution score highest in the most states, with four different distributions showing the best fit in the seven states (LP3 in three states, GEV in two, GP in one and GUM in one). The study concludes that the best performers, all suggested utilized for future flood frequency analysis studies, are LP3, GEV and GP.

Despite many studies, no consensus is reached on a universally superior method for flood frequency analysis, nor has a basis been established for which data type to use. Thus, constructing the best methodology for frequency analysis of hydrological extremes involves investigating different combinations of probability distributions, data type and parameter estimation methods. This study will look into differences in performance related to choice of probability distributions and data type.

While many have searched for a universal distribution it is also argued that this ignores specific characteristics of a given catchment, and that FFA should be done for a given catchment using the distribution that proves most fitting to the given catchments through goodness of fit tests and tests investigating their forecasting ability (Gneiting and Raftery 2007, Nagy et al. 2017).

The goals of this study will be to:

- Assess which data type is best suited for a flood frequency analysis, annual maximum series (AMS) or peak over threshold (POT)
- Assess the how well different distributions perform for AMS and POT data
- Look at how the flood generating process (FGP) affects the points above

2 Methods

2.1 Data

The dataset used in this study contains data from 529 stations in Norway. Selections of stations to consider here was based on record length and presence of both AMS and POT data. The required record length was set to 30 years because this is recommended for flood frequency analysis using 2-parameter distributions, while at least 50 years of data is recommended for frequency analysis using 3-parameter distributions (Midttømme et al. 2011). Of the 529 stations 271 satisfied these prerequisites and were

used in this study. Discharge levels from all stations are given as daily average peak while data points from recent years also includes momentary peaks, the daily average peaks were used in this study.

The data used in this study is taken from NVE's hydrological database Hydra II. The AMS data from Hydra II is assumed to be comprised of independent events since only a single event is considered per year. POT datasets on the other hand demands more careful selection. First a threshold has to be set, to define above which limit is an event considered significant for this dataset. This varies from station to station, relating to what flood generating process (FGP) is present. In the Hydra II dataset this was set to the 98- and 95 percentiles, only data from the former used in this study. Because several peaks can be counted per year it is not a given that two events are independent of each other an independence criterion has to be set. For this data the independence criterion was adopted from (Cunnane 1979), stating that a peak can only be considered independent from a former peak if the discharge between the peaks dropped ⅔ from the smaller flood event, and if the peaks occur temporally at least thrice the time to rise (of the first event) apart. The length of POT series relative to the AMS series for the stations used can be seen in Figure 1.

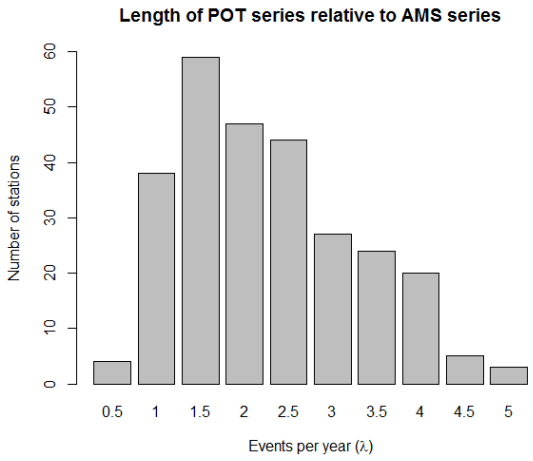

Figure 1: Length of POT series relative to the AMS series from the same station. Each value on the x-axis includes series with ± 0.25 λ (i.e. value one on the x-axis includes all stations with a λ of 0.75-1.25).

Flood events occur as a result of several factors i.e. heavy rainfall, a particular big influx of meltwater or dam breaks. Any flood event may be caused by any one of these processes or a combination of these. Any particular flood event may be attributed to these processes, while a catchment's characteristics

(such as steepness yearly precipitation etc.) may discern which of the initiating factors is most prevalent for said catchment. An FGP value has been assigned to most catchments in (Engeland et al. 2016) and many singular events have also been assigned a FGP value. The FGP value ranges from 0-1 and concerns how much meltwater and rainwater contribute to floods. A flood resulting from only meltwater will have FGP = 0 while one resulting from only rainwater will have FGP = 1. The FGP values used in this study is the average FGP of all events in a time series. A more detailed description of the data and data selection procedure can be found in (Engeland et al. 2016).

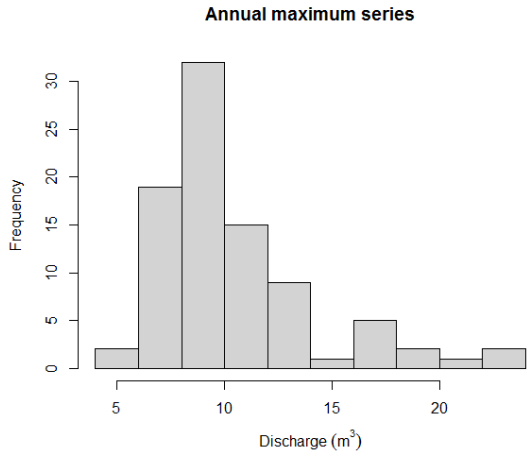

Figure 2: An annual maximum series from a catchment where the flood generating process has the value 1 (rainwater).

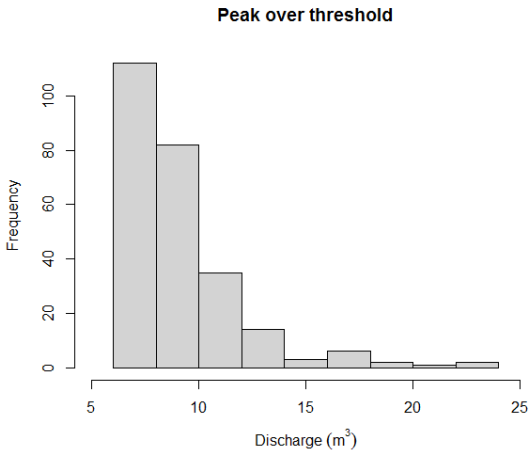

Figure 3: A peak over threshold timeseries from a catchment where the flood generating process has the value 1 (rainwater)

Figure 2 andFigure 3 shows the timeseries for AMS and POT for station 91.2 respectively. Station 91.2 has a flood generating process of 0.92 meaning precipitation contributes highly to the flood events in that catchment. When the flood generating process is mostly attributed to rainwater it is likely that several significantly large floods happen every year, thus a POT dataset is longer than AMS as shown in Figure 2 andFigure 3. A map showing the flood generating processes of each catchment is presented in Figure 4.

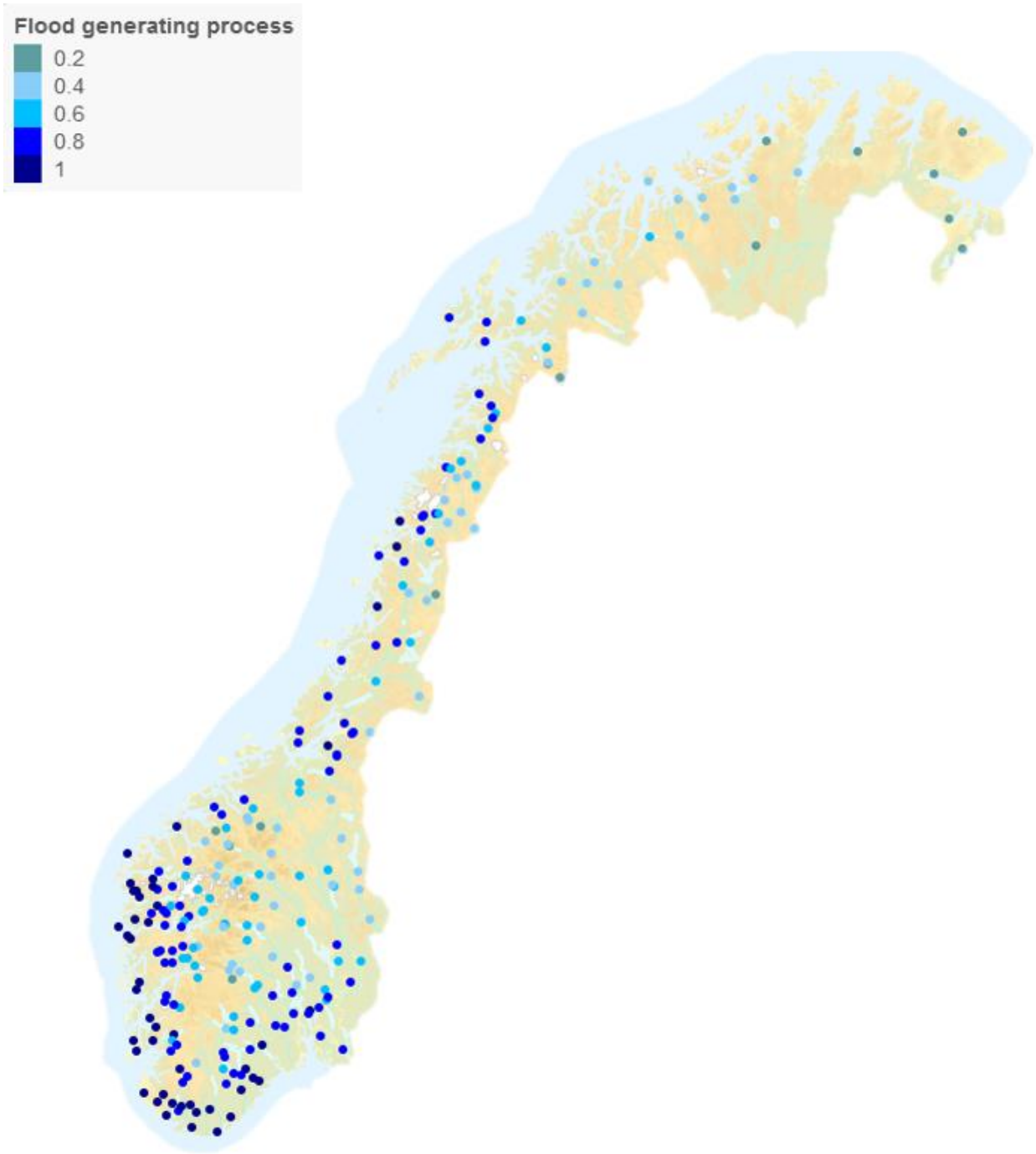

Figure 4: Average flood generating process of each station. An FGP=0 mean events are derived solely from meltwater, while FGP=1 mean events are derived entirely from precipitation. Each color corresponds to an FGP value between the value stated in the legend and the one stated above i.e. stations marked with deep blue has an FGP of 0.8-1.

2.2 Distribution selection

Distributions fitted in this study was selected based on occurrences in the literature. Two distributions (one 2-parameter and one 3-parameter) was selected for fitting to AMS data while another pair of distributions was selected for fitting to POT data. Based on use in the literature and recommendations in Castellarin et al. (2012), the choice fell on Gumbel and GEV for AMS data (2-parameter and 3-parameter respectively) and EXP and GP for POT data (2-parameter and 3-parameter respectively). In theory these are the distributions that follows a general extreme value distribution for AMS and POT data.

2.3 Annual maximum series

The basis for extreme value distributions originates from (Fisher and Tippett 1928). Stating that when a sample is taken from a parent distribution of independent and identically distributed data, and the max and min values are chosen, the distribution of the maximum and minimum will limit itself to two main types. These were later merged into the Generalized extreme value distribution by Von Mises (1936). The cumulative distribution function (CDF) for GEV can be written as:

$$D(x) = \begin{cases} exp\left\{-\left[1 - k\left(\frac{x-\xi}{\alpha}\right)\right]^{\frac{1}{k}}\right\} & for\ k\ \neq 0 \\ exp\left\{-\left(-\frac{x-\xi}{\alpha}\right)\right\} & for\ k = 0 \end{cases}$$

Equation 1.1

Here x is a given the event, measured in discharge. ξ is the location parameter, deciding the shift or placement of the curve. The scale parameter α explains the dispersion of the data and the shape parameter *k* defines the shape of the curve. If $k \neq 0$ you get the distribution function of the 3-parameter GEV distribution, if $k=0$ you instead get a 2-parameter version of the GEV called the Gumbel distribution. There are two other special cases of the GEV distribution, if $k<0$ a Frechet distribution is given, while $k>0$ yields the Weibull distribution. This is because (Von Mises 1936) combines the three different distributions from (Fisher and Tippett 1928) into one distribution.

Prediction of design flood values is the goal of flood frequency analysis. For any return period *T*, there is a related return level QT meaning the flood size (m^3) of a flood with the exceedance probability of 1/*T*

every year. For the estimation of return levels, the cumulative distribution function can be altered like this:

$$1 - \frac{1}{T} = D(x)$$

Equation 1.2

$$x_T = \begin{cases} m + \frac{\alpha}{k}\left\{1 - \left[-log\left(1 - \frac{1}{T}\right)\right]^k\right\} & for\ k \neq 0 \\ m - \alpha\ log\left[-log\left(1 - \frac{1}{T}\right)\right] & for\ k = 0 \end{cases}$$

Equation 1.3

Where *T* is the chosen return period for which one wish to estimate design flood values. This extrapolation of the data is used to determine design flood levels for return periods substantially larger than the length of the data series.

2.4 Peak over threshold series

The probability distributions applied to POT series in this study are the Generalized pareto (GP) and the Exponential (EXP). The GP distribution was introduced by Pickands III (1975) as an alternative that describes POT distributions where the GEV special cases describe AMS distributions. The CDF of the GP distribution can be written as such:

$$S(x) = \begin{cases} 1 - \left[1 - k\left(\frac{x - v}{\alpha}\right)\right]^{\frac{1}{k}} & for\ k\ \neq 0 \\ 1 - exp\left(-\frac{x - v}{\alpha}\right) & for\ k = 0 \end{cases}$$

Equation 2.1

Where x is a given event measured in discharge, m is the location parameter, α is the scale parameter and k is the shape parameter. The GP distribution is given when $k \neq 0$, GP distributed POT data implies that annual maximum data be GEV distributed. As a special case of the GP distribution, when k is absent or $k = 0$ the exponential distribution is given. Where GP implies GEV distribution, EXP implies Gumbel

distribution of annual maximum data (Lang et al. 1999). Estimation of design flood values for any return level using the GP and EXP distributions is done adapting their cumulative distribution functions like this:

$$1 - \frac{1}{T} = exp\{-\lambda(1 - S(x))\}$$

Equation 2.2

$$x_T = \begin{cases} \frac{\alpha}{k}\left\{1 - \frac{\left[-log\left(1 - \frac{1}{T}\right)\right]}{\lambda}\right\}^k + v & for\ k \neq 0 \\ m - \alpha\ log\left[-\frac{1}{\lambda} log\left(1 - \frac{1}{T}\right)\right] & for\ k = 0 \end{cases}$$

Equation 2.3

Where ƛ pertains to the average number of exceedances above the threshold u per year on record. In this study we have substituted away the threshold u for an estimated location parameter v.

2.5 L-moments

Parameter estimation in this study was done using the L-moments approach. Since the unification of techniques into the L-moments technique by Hosking (2009), the technique has gained wide recognition among scholars. The L-moments approach estimates location ($ƛ_1$), scale ($ƛ_2$) and the skewness (τ_3) for the parent distribution (Hosking 2009). This parameter estimation technique was chosen because it utilizes linear transformations of the data, this means that the estimations are less affected by outliers in the data. L-moments are thus more robust and can boast better stability than other traditional moments techniques. This was thoroughly discussed (Hosking 2009), where L-moments for several distributions are given. Equations for the L-moments used to estimate distribution parameters in this study can be found in Appendix 1.

2.6 Validation

In this study I wanted to obtain the predictive performance of our models. In addition I wanted to estimate the predictive performance of models originating from both AMS and POT datasets. Garavaglia et al. (2011) refers to two main benchmarks when comparing frequency analysis models, reliability and stability. The reliability of a model considers how similar the estimated model is to the parent

distribution. Because the parent distribution is unknown, the model has to be validated on observed data to convey the reliability of the model. The second benchmark, stability refers to the models ability to generate similar results when estimated on different sub-sets of data. While both reliability and stability is important in a model, reliability should be considered first. Stability of a model is important to ensure that the model is not prone to being biased by outliers. Because flood frequency analysis is often used in construction of safety measures, a biased model could entail incorrect construction regulation. However, it is the reliability that explains how well a model predicts observations, consequently the reliability is paramount in selecting a model. Stability can be used to identify the best model when several models show similar reliabilities (Renard et al. 2013).

Testing and ranking the predictive performance of the distributions can be done many ways. A goodness of fit test analyzes how well the model fit the parent distribution, but when combined with cross validation to ensure out of sample testing, it can also evaluate the predictive ability of the model (Steyerberg et al. 2010). There are also other scoring rule methods that can be employed to rank models originating from differing datasets.

2.7 Goodness-of-fit

A goodness-of-fit test can be used to assess the reliability of a model or how well a distribution fits to the empirical data. There are several different goodness-of-fit tests, with several different test statistics. They measure the discrepancy in the modeled values versus the empirical values. The tests decide limits within which the observations must fall for the sample to be drawn from a population with a given distribution. The Anderson-Darling test measures its test statistic based on the average deviation of the fitted distribution from the unknown parent distribution. The Anderson-Darling distribution additionally contains a function that gives more weight to observations in the tails (Choulakian and Stephens 2001, Thas and Ottoy 2003). The Kolmogorov-Smirnov test on the other hand calculates the maximum deviation between the fitted distribution and the parent distribution (Justel et al. 1997, Del Barrio et al. 2000). Figure 5 shows an example of a distribution fitted to observed data.

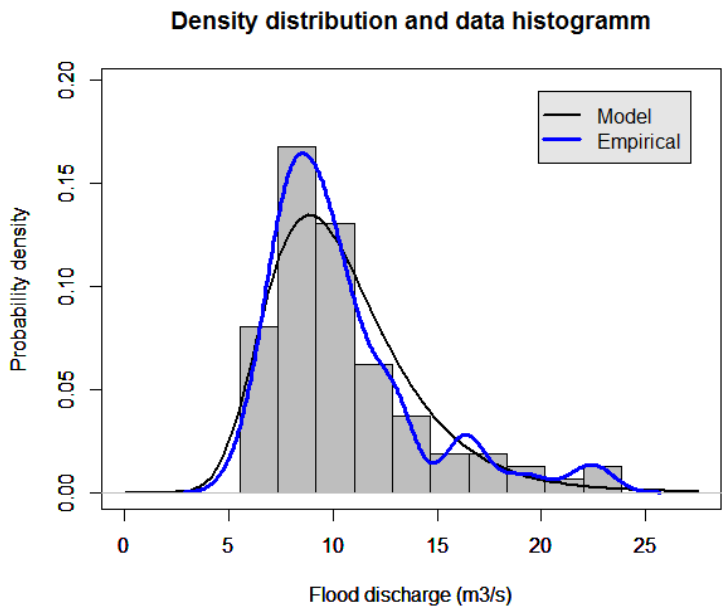

Figure 5: Visual presentation of goodness-of-fit. Empirical (blue line) is the true distribution of the observed data, while the black line is the model distribution.

2.8 Scoring methods

When conducting a flood frequency analysis, it is often desired to get the predictive performance of a model, given design flood estimates or quantiles. This can be achieved using scoring rules such as Brier score (Brier 1950), or the quantile score. The Brier score presented in Brier (1950) is a scoring rule where the disparity between predictions and observations are measured, this then gives a score from zero to two where zero is perfect forecast while 2 is the worst possible (Brier 1950, Steyerberg et al. 2010). The half Brier score can be used to test for how well a model predicts exceedances over a given threshold or quantile, it is called half Brier score because, in the original equation the score ranges from zero to two while in the half it ranges from zero to one. It is given in (Gneiting and Raftery 2007) as:

$$BS\ (F, y) = \big(F(y) - \mathbb{1}(y \geq x)\big)^2$$

Equation 3.1

Where F(y) is the calculated probability of event y exceeding the given threshold x.

Quantile score can also be utilized to test for predictive ability in models. The quantile score will yield the weighted error of predictions relating to observations, and is given in (Gneiting and Raftery 2007) as:

$$QS(F, y) = (F^{-1}(y) - y)(\mathbb{1}\{F^{-1}(y) \geq y\} - x)$$

Equation 3.2

2.9 Cross validation

For models to show adequate predictive ability, the estimated model must be tested on an out-of-sample dataset. The out-of-sample dataset can be from a different but similar population, but when this is not an option it is often obtained by dividing the sample into two parts. The training or estimation dataset, used to estimate the model, and the test or validation dataset used to validate the model. It is desired to estimate models on a large portion of the sample, such that the estimated model is not significantly different than the model we would obtain from estimation using the entire sample. This is to ensure that the model is not biased to different directions depending on the training data. However, at the same time a sufficient amount of out-of-sample data must be available for validation, to inhibit excessive variance in the validation data.

This is often achieved using cross validation. There are several different types of cross validation, but the general idea is to withhold one part of the dataset and estimate the model on the remaining n parts. The estimated model is then validated on the part of the population that was not used for model estimation. This is useful when one population cannot be used to predict another population. Among these methods of cross validation, k-fold cross validation is common due to its low computational cost compared to other cross validation methods (Fushiki 2011). In k-fold cross validation the data is divided into k folds of equal size. The data must be randomized if the order of data is consequential (e.g. time series of non-stationary data), it is common practice to randomize data regardless of data type. Each fold is then used once for validation. Below it is explained how this was done in this study, this was done for every station:

- AMS: The data was sampled randomly and divided into 10 equal folds.

- POT: All events from the same year were grouped together because we compare annual exceedance probability, these blocks were then sampled randomly and divided into folds of equal size.
- Parameters were estimated using nine folds while the last fold was withheld and used to validate the model. The validation of the model was done with goodness-of-fit tests and the Brier and quantile score tests.
- This was repeated 10 times so that every fold was used as the validation set once.
- The test scores were then averaged out.

2.10 Programming

Statistical analyses for this study was done in the open source coding program R (R Core Team 2017). The coding language is specialized for statistical use and its libraries are continuously expanded and updated by its users. The base language is quite similar to other coding languages, but specialized packages are readily available at github or r-cran. These packages can be specialized for specific biological or hydrological analyses or be broader, focusing on data manipulation and organization. The coding done for this study used both packages with broader and narrower scopes.

With a methodology similar to that of (Baffie et al. 2017) the main coding done here has been done using the packages developed during their study ,"fitdistrib" and "FlomKart" (Baffie 2016a, b). Fitdistrib is a package which contains functions for fitting extreme value probability distributions to data. FlomKart was used to assess the goodness of fit of those distributions fitted using fitdistrib. In turn these packages call on functions from other packages, the most prominent of which is the evd and nsRFA packages (Stephenson 2002, Viglione 2014). For analysis of the AMS distributions existing functions in the packages were used, for analysis of the EXP and GP distributions I had to expand the existing functions and supplement the package with some new functions. My coding can be found here.

2.11 Stability

Bootstrapping is a method for measuring the variability of an estimate. The general idea is that whenever a model is estimated, the estimation is done on a portion of data, which itself may be average

or extreme to either side. To account for this a method as presented by Diaconis and Efron (1983) was designed. When bootstrapping the data of length n is resampled with replacements, creating a new artificial dataset comprised of n points where the same point may be repeated as many as n times. The statistic of interest is then calculated from the resampled data. This process is repeated many times, in this study it was repeated 1000 times but it may be repeated many more times. Confidence intervals can be decided and from those the variability of the statistic can be obtained (Felsenstein 1985).

The statistic estimated through bootstrapping in this study was not dimensionless and could therefore not be compared to each other. To estimate the stability of the distribution across all stations, the coefficient of variation was calculated from the bootstrapping results.

3 Results

To establish which flood frequency analysis method of those used here is the most appropriate one, I evaluated performance at each station based on several tests. For reliability of the model, i.e. how well the model can describe the data, I employed the Anderson-Darling and Kolmogorov-Smirnov goodness-of-fit tests. To assess how accurately the models predict design flood events the Brier and quantile scoring rules were considered. Lastly, the stability of the models was tested. The stability of the models can be used to infer the best model given that performance in reliability and predictive ability are similar. The plots presented here uses a smoothing function to better discern trends while ignoring smaller variations or noise. Figure 9 shows each point with the smooth graph over, the points are not shown in other plots to make the plots easier to read.

3.1 Goodness-of-fit

The Kolmogorov-Smirnov and the Anderson-Darling tests statistics for each station is the average test statistic derived from a ten-fold cross validation. For the goodness of fit tests, the GEV and GUM distributions have been fitted to AMS data while the EXP and GP distributions have been fitted to POT data.

Table 1 gives the average test statistic for each distribution across all stations. The best fitting distribution according to each test is shown with bold text, both the AD and KS test yield GP as the best fitting distribution. With the other distribution fitted to POT data yielding the next lowest test statistics in both tests. The relative discrepancy between POT and AMS data in the Anderson-Darling test is much lower than in the Kolmogorov-Smirnov test. The AD test statistic is 6-10% larger for the AMS models than the POT models, while the KS test statistic for AMS models is almost 50% larger than that of the POT models.

Table 1: The average goodness-of-fit tests scores for each distribution. The bold number signifies the lowest/best score for the given tests.

	Anderson-Darling	*Kolmogorov-Smirnov*
EXP	1.0873	0.2736
GEV	1.1993	0.3947
GP	**1.0839**	**0.2646**
GUM	1.1438	0.3953

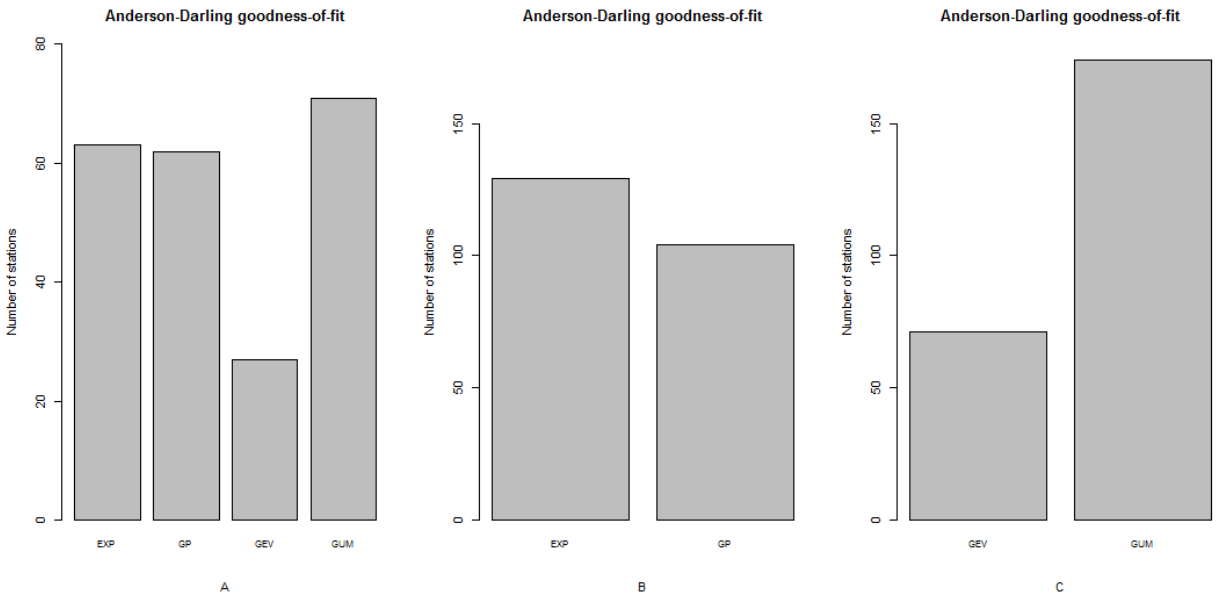

Figure 6:The number of stations for which each distribution showed the best fit according to the Anderson-Darling test. Figure A shows the number of stations for which each distribution showed the best fit, while figures B and C show the same but only compares models based on the same dataset.

Comparing Figure 6 to table 1 shows that while the GP distribution shows the best fit across all stations it is only the best fitting distribution for about 60 stations. The EXP yield similar results while the GUM shows best fit for about 70 stations (when all distributions are compared). When only comparing models based on the same data the 2-parameter distributions does best for both AMS and POT. Among the POT models both show best fit for a large number of stations, while for the AMS models the GUM distribution show best fit for the vast majority of stations.

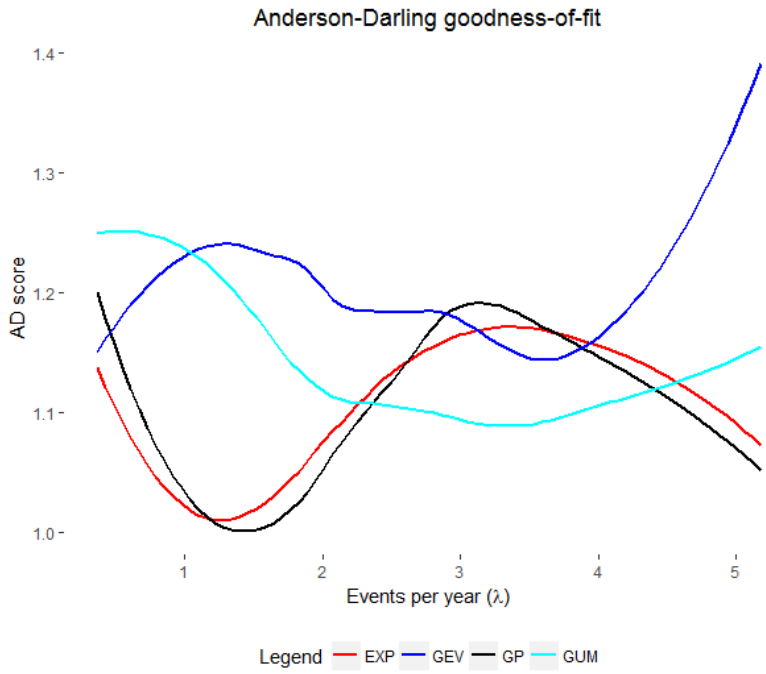

Figure 7: When the Anderson-Darling score is plotted against the number of events per year in the POT data set for a given catchment, the graphs vary quite a bit.

Figure 7 shows the Anderson-Darling test statistic plotted against events per year (λ) for each distribution. Figure 7 shows no noticeable trends for either distribution. Rather all distributions undulate in a wave like fashion as λ increases. The distributions based on POT data and those based on AMS data seem to undulate opposed to each other, when the POT distributions perform well, the AMS distributions perform poorly and vice versa. The undulation of the POT distributions is quite similar with smaller differences while the undulation of the AMS distribution is more disjointed from each other. Table 1 shows that the two POT distributions yield the best fit to their parent distribution, this is not easily distinguished in Figure 7.

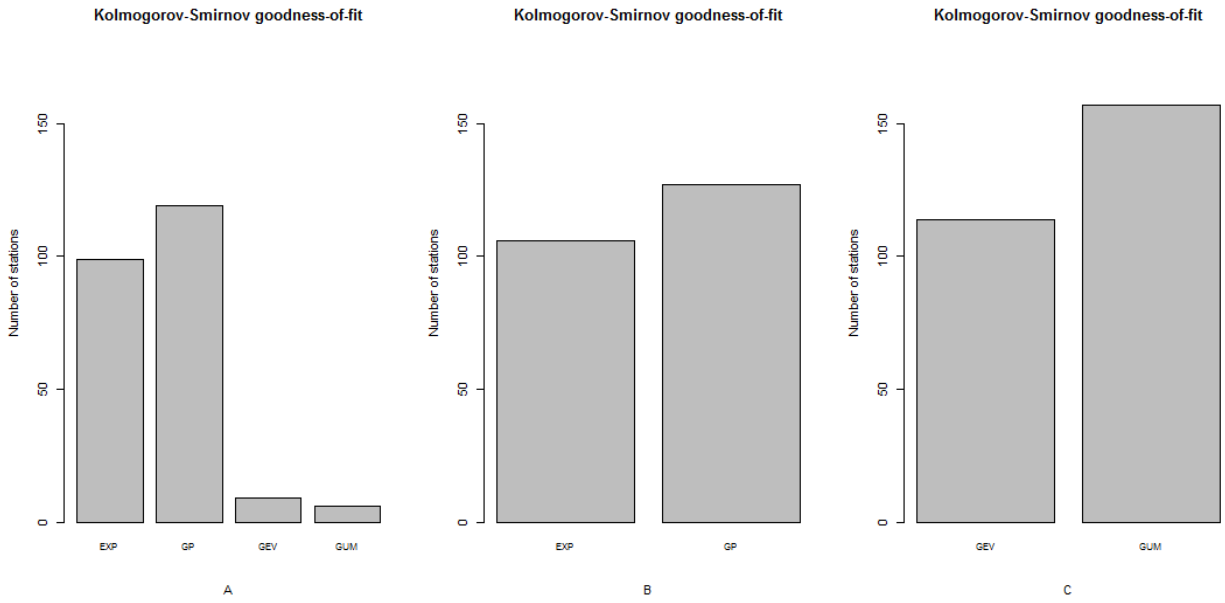

Figure 8: The number of stations for which each distribution fits the best, according to the Kolmogorov-Smirnov test. Figure A shows the number of stations for which each distribution showed the best fit, while figures B and C show the same but only compares models based on the same dataset.

Where the Anderson-Darling test shows quite varying results with all distributions fitting best to a number of stations, the Kolmogorov-Smirnov test clearly favors the POT distributions according to Figure 8 A. GP and EXP shows the best fit for 250 of the 271 stations and similarly the POT distributions show the best average fit across all stations (table 1). When models derived from the same data are compared the GP distribution outperforms the EXP distribution as opposed to what seen in Figure 6. Among the AMS models the GUM shows best fit for the most stations according to the AD and KS test.

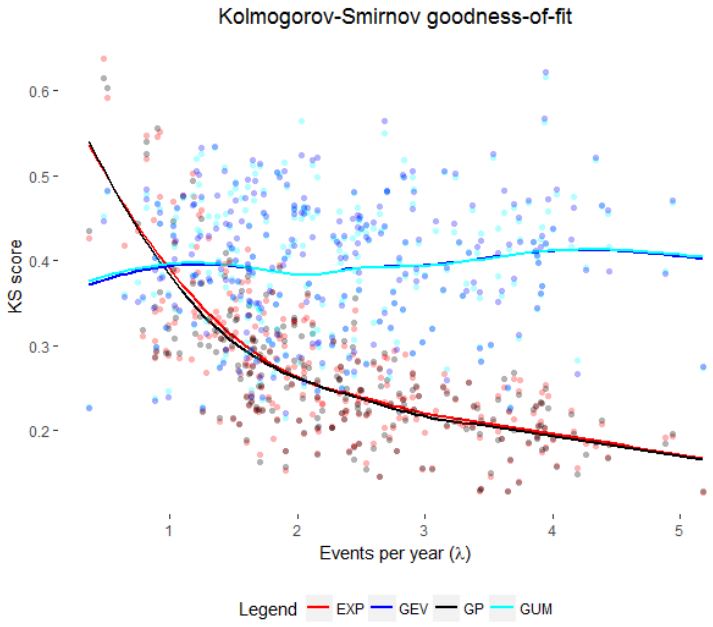

Figure 9: Presents an example of what the smoothing function does.

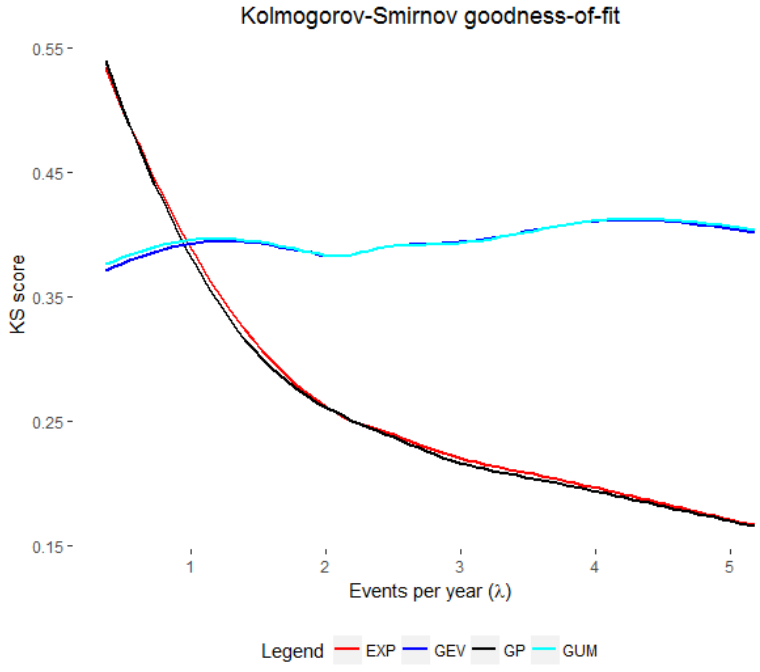

Figure 10: Kolmogorov-Smirnov score plotted against events per year in the POT data set. The distributions fitted to AMS data shows no noticeable trend while distributions fitted to POT data starts dropping rapidly and then flattens out towards the right-hand side of the plot.

Unlike what is seen for the AD test in Figure 7, Figure 10 shows the POT models to have a clear trend towards better fit as λ increases. AMS data performs better than the POT data when λ<0.97, after which the POT distributions fit better. The disparity between the two data types continues to grow as λ

increases, however, at a decreasing pace. The difference between the distributions fitted to the same data is very small.

The Kolmogorov-Smirnov test statistic as a function of flood generating process is presented in Figure 11. The AMS models can be seen to show very little variation in the KS score as the FGP changes. The POT models on the other hand, show a clear trend towards better fit as the flood generating process moves towards 1 (precipitation floods). There is negligible difference in fit between the AMS models, while for the POT models the GP show a slightly better fit at all FGP values.

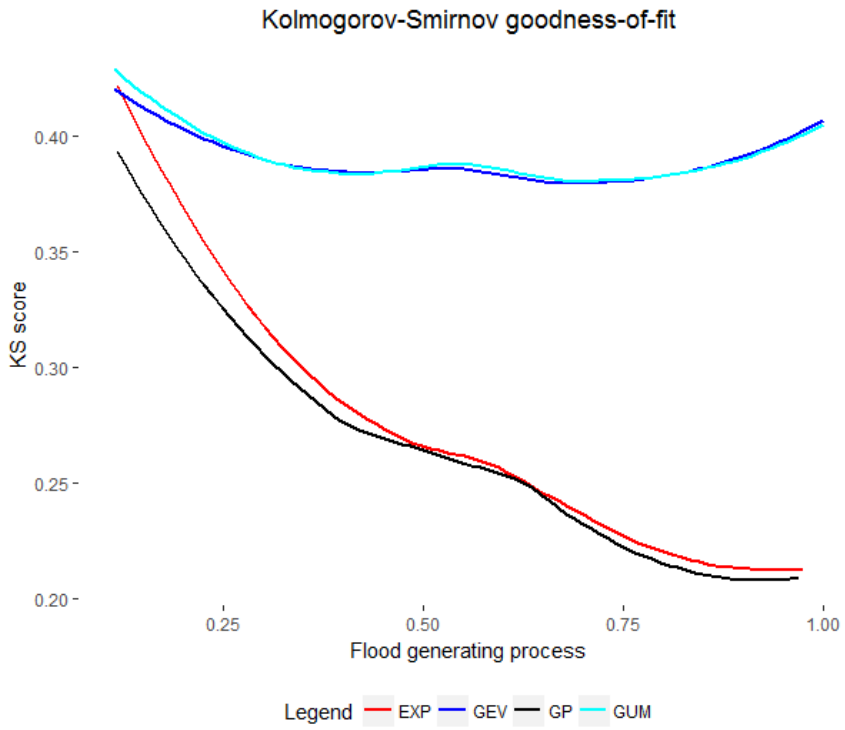

Figure 11: Kolmogorov-Smirnov goodness-of-fit test statistic as a function of FGP

Figure 12 and Figure 13 shows the best fitting distribution according to the AD and KS for each station plotted on a map. This enables recognition of spatial differences in test performances.

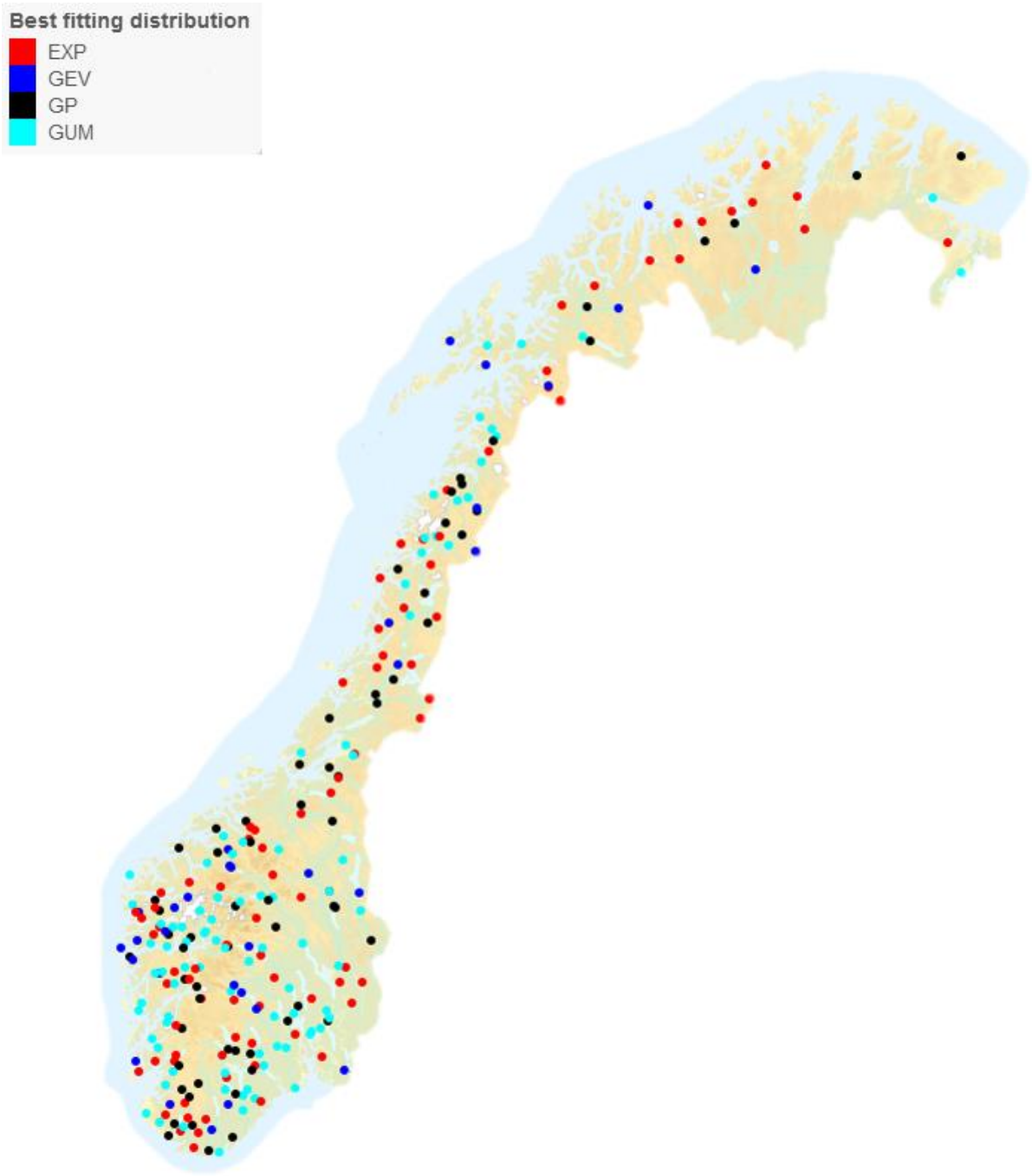

Figure 12: Distribution with the lowest (best) Anderson-Darling test score for all stations

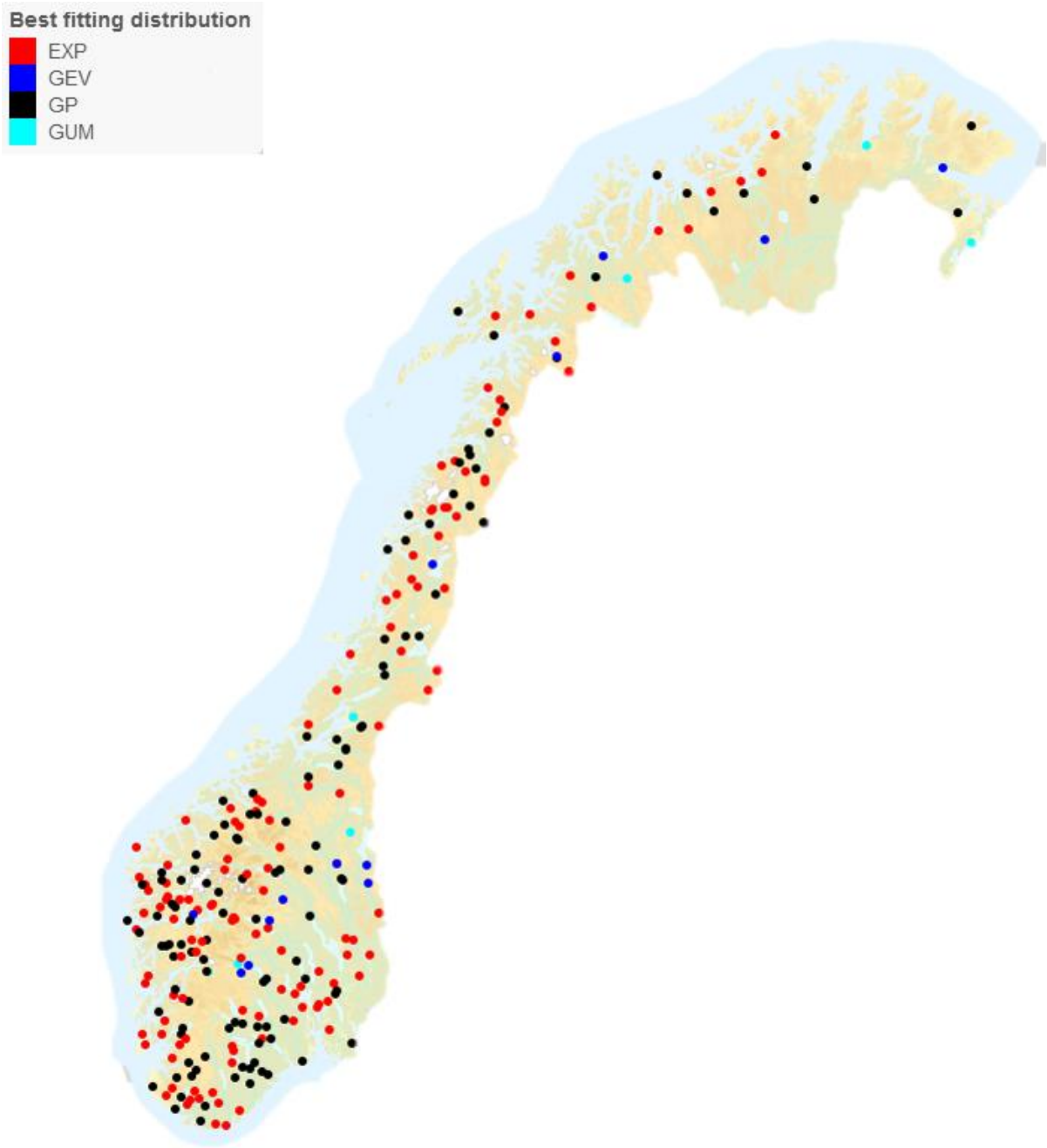

Figure 13: Distribution with the lowest (best) Kolmogorov-Smirnov test score for all stations

Figure 12 mapping the performance of distributions according to the Anderson-Darling test shows no obvious spatial trends. Figure 13 showing the Kolmogorov-Smirnov test prefers the EXP and GP distributions for the vast majority of Norway. Almost all stations that prefer the GUM and GEV distributions can be found in inland areas or in the most northern part of the country.

3.2 Predictive performance

The distributions were compared according to their scores in tests designed to evaluate predictive accuracy. Scores of models validated on AMS data are not truly comparable to models validated on POT data. The scores tells us how well the models predict the observed events they are validated on. Thus, for these tests a third type of model was introduced, fitting the EXP and GP distributions to POT data and then validating on AMS data from the same station. The Brier and quantile scoring methods require return levels and annual exceedance probability for the return period we consider. The design flood return levels (m^3) and the annual exceedance probability were calculated for each distribution at each station. Thresholds used for calculating the scores are thus distinct for each distribution. The equations 1.3 and 2.3 explained earlier were used to estimate return levels for several return periods. Most presented scores are for a 20-year return period, but Table 3 gives average scores for several return periods.

In table 2 the Brier and quantile score for a 20-year return period is given for each distribution, these scores are the average across all stations. The EXP can be seen to yield the best score in both tests with the GP distribution yielding the second-best scores. The models estimated on POT and validated on AMS data show that for the Brier score the models estimated on POT data perform better than those estimated and validated on AMS data. The quantile score shows the models estimated and validated on AMS data perform slightly better than the models estimated on POT and validated on AMS data.

Table 2: The average scores for each distribution per the Brier scoring method and the quantile scoring method. GP AMS and EXP AMS are models estimated on POT data and validated on AMS data. The bold number signifies the lowest/best score for the given tests.

	GP	*EXP*	*GEV*	*GUM*	*GP AMS*	*EXP AMS*
Brier score	0,0321	**0,0282**	0,0545	0,0522	0,0356	0,0325
Quantile score	0,0493	**0,0486**	0,0524	0,0521	0,0549	0,0546

The total number of stations where each model yields the best quantile- and brier score for a 20-year return period can be seen in Figure 14 and Figure 15 respectively. The two models that score best for

the most stations in the quantile score are the two POT models, with the EXP model yielding the best score for about twice as many stations as the GP model. Both models estimated and validated on AMS data score better than the models estimated on POT data and validated on AMS data. For the Brier score the four models estimated on POT data yield the best scores for almost all stations, with the EXP AMS model scoring best. The other models estimated on POT data also perform well. When we compare the models validated on AMS data the GUM model yields the best quantile score in almost 50% of the stations (Figure 14 B), the EXP AMS and GEV models perform similarly. For the Brier score the models estimated on POT data outperform the AMS based models for about 80% of the stations (Figure 15 B), with the EXP AMS performing best.

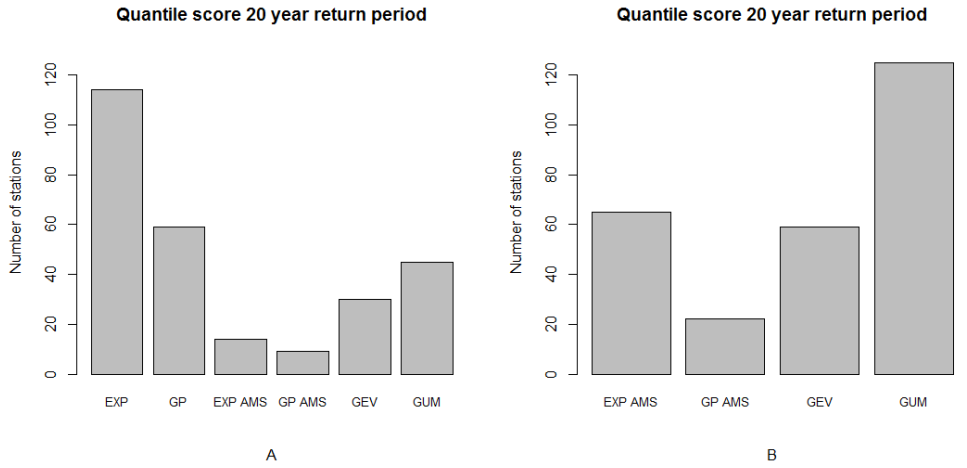

Figure 14: Shows how many stations each model performs best for according to the quantile scoring method. In plot A every model is shown together, those validated on AMS and POT data. Plot B only show the models validated on AMS data as this makes them comparable.

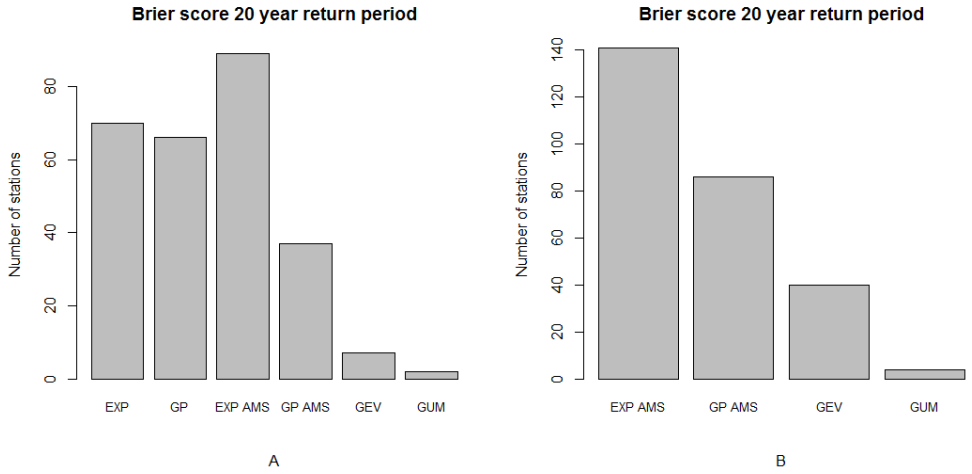

Figure 15: Shows how many stations each model performs best for according to the Brier scoring method. In plot A every model is shown together, those validated on AMS and POT data. Plot B only show the models validated on AMS data as this makes them comparable.

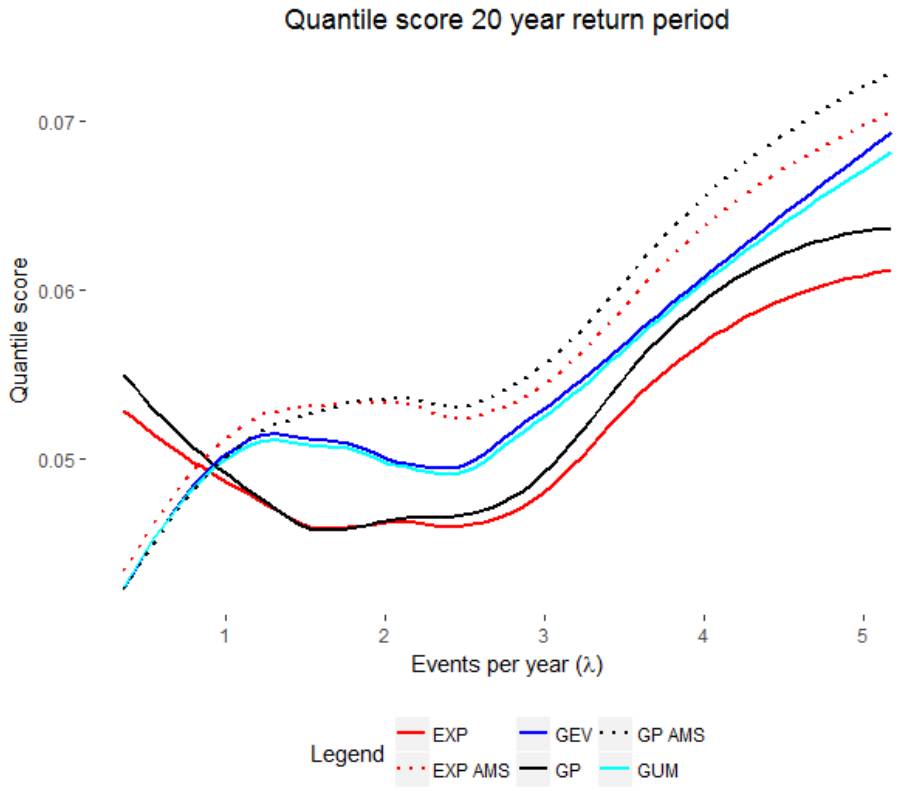

Figure 16: Quantile score as a function of flood per year in the POT data. Models validated on AMS data perform better until POT data exceeds around 1.1-1.2 events per year. After 2.5-3 events per year all models experience an uptick.

Figure 16 show that models estimated and validated on POT data prove themselves superior for most catchments, with the exclusion of catchments whose POT data consists of less than 0.9 events per year. Differences between models estimated on AMS data is negligible, while the two POT based models start to deviate slightly from each other after the number of events per year exceeds 2.2.

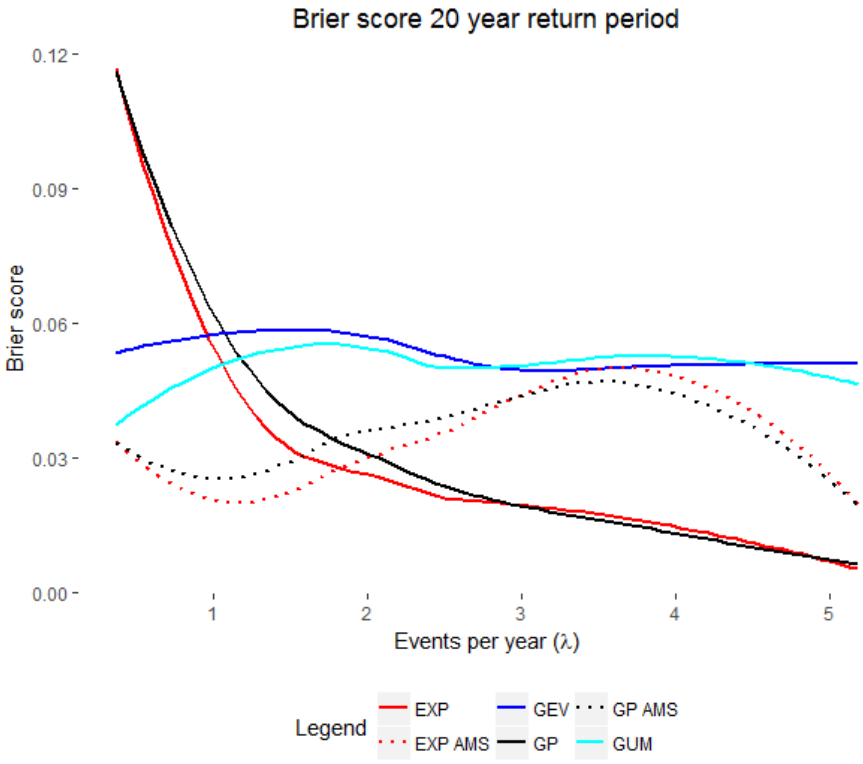

Figure 17: Brier score for a 20-year return period plotted against events per year in the POT data.

The Brier score shows that the models estimated on POT data score the best for almost every station. When the POT data consists of less than two events per year the models estimated on POT data and validated on AMS data scores the best. Like in Figure 10 the GEV and GUM distribution show no noticeable trend, while GP and EXP (estimated and validated on POT) shows a decreasing Brier score as λ increases. Towards the right-hand side of the plot the GP and EXP graphs flatten. The models estimated on POT data and validated on AMS undulate similarly to the graphs seen in Figure 7.

Table 3 the EXP distribution can be seen to perform best on average in both the Brier and the quantile scoring method for all return periods calculated between 20- and 200-year.

Table 3: Each distributions average Brier and Quantile score for return periods 20-, 50-, 100- and 200-year.

	BS 20	*BS 50*	*BS 100*	*BS 200*	*QS 20*	*QS 50*	*QS 100*	*QS 200*
EXP	**0.0282**	**0.0120**	**0.0071**	**0.0042**	**0.0486**	**0.0256**	**0.0152**	**0.0089**
GP	0.0321	0.0172	0.0106	0.0077	0.0493	0.0265	0.0161	0.0099
GEV	0.0545	0.0269	0.0176	0.0116	0.0524	0.0288	0.0179	0.0111
GUM	0.0522	0.0234	0.0135	0.0084	0.0521	0.0279	0.0171	0.0104
EXP AMS	0.0325	0.0145	0.0084	0.0059	0.0546	0.0287	0.0173	0.0103
GP AMS	0.0356	0.0183	0.0117	0.0089	0.0549	0.0293	0.0181	0.0113

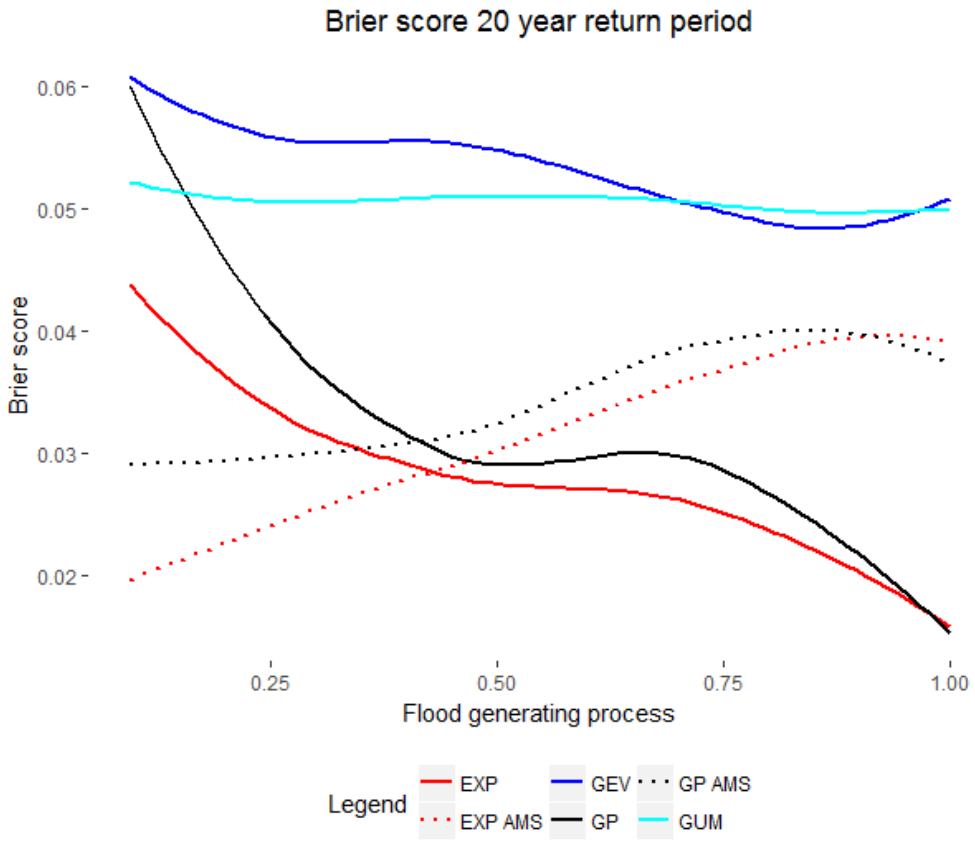

Figure 18:Brier score for a 20-year return period for all models plotted against the flood generating process of the catchments (0 means all meltwater while 1 means all rainwater).

Flood generating process is an important characteristic of a catchment in regards to FFAs, the FGP of each station can be seen in Figure 4. In Figure 18 it is shown that the POT models become increasingly accurate when the flood generating process moves from meltwater to rainwater (0-1). The AMS models perform similarly regardless of flood generating process, with the GEV model showing a slight decreasing trend. Among the models validated on AMS data those estimated on POT data score better for all FGP values.

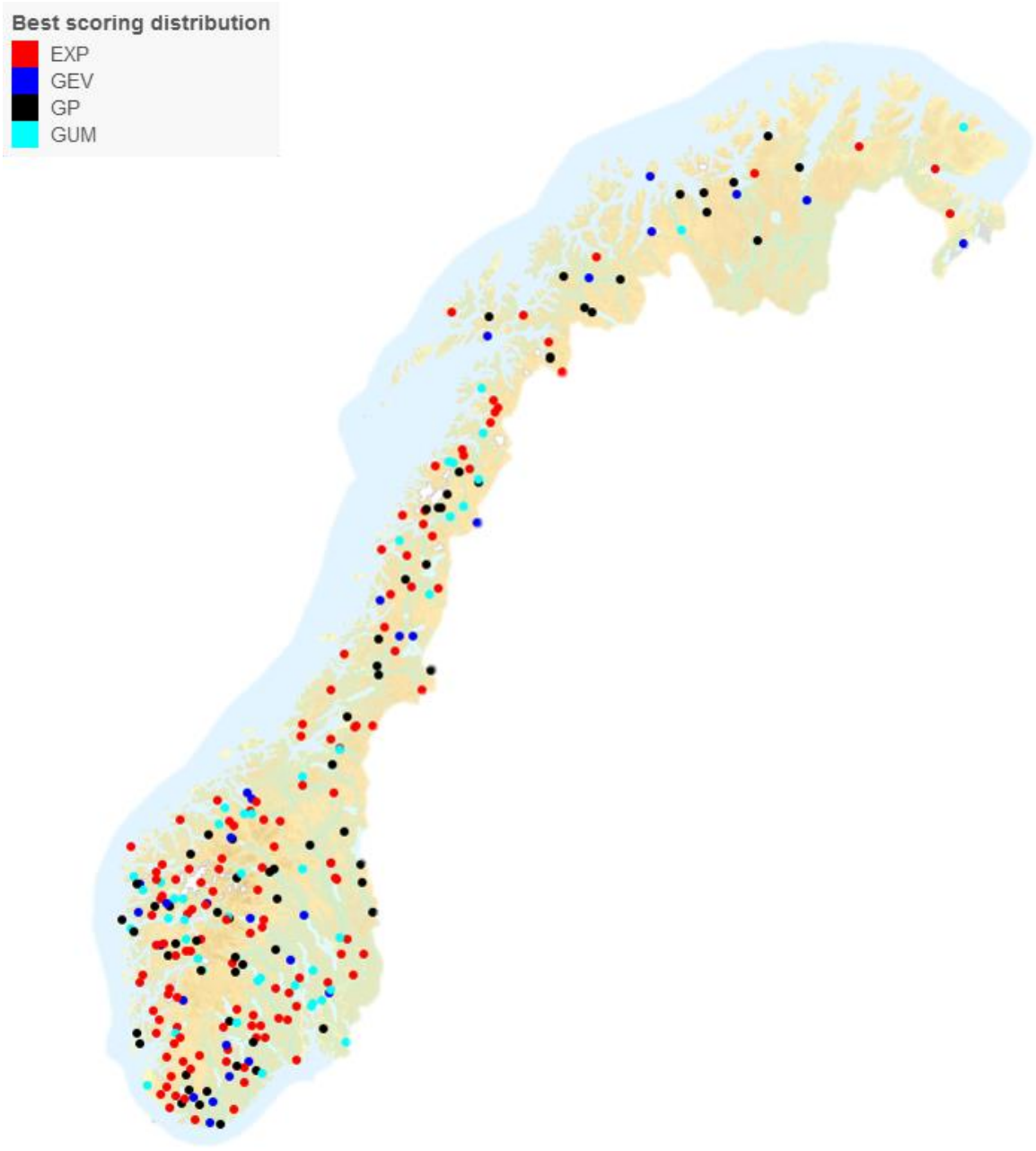

Figure 19: Best scoring distributions according to quantile score for 20-year return period.

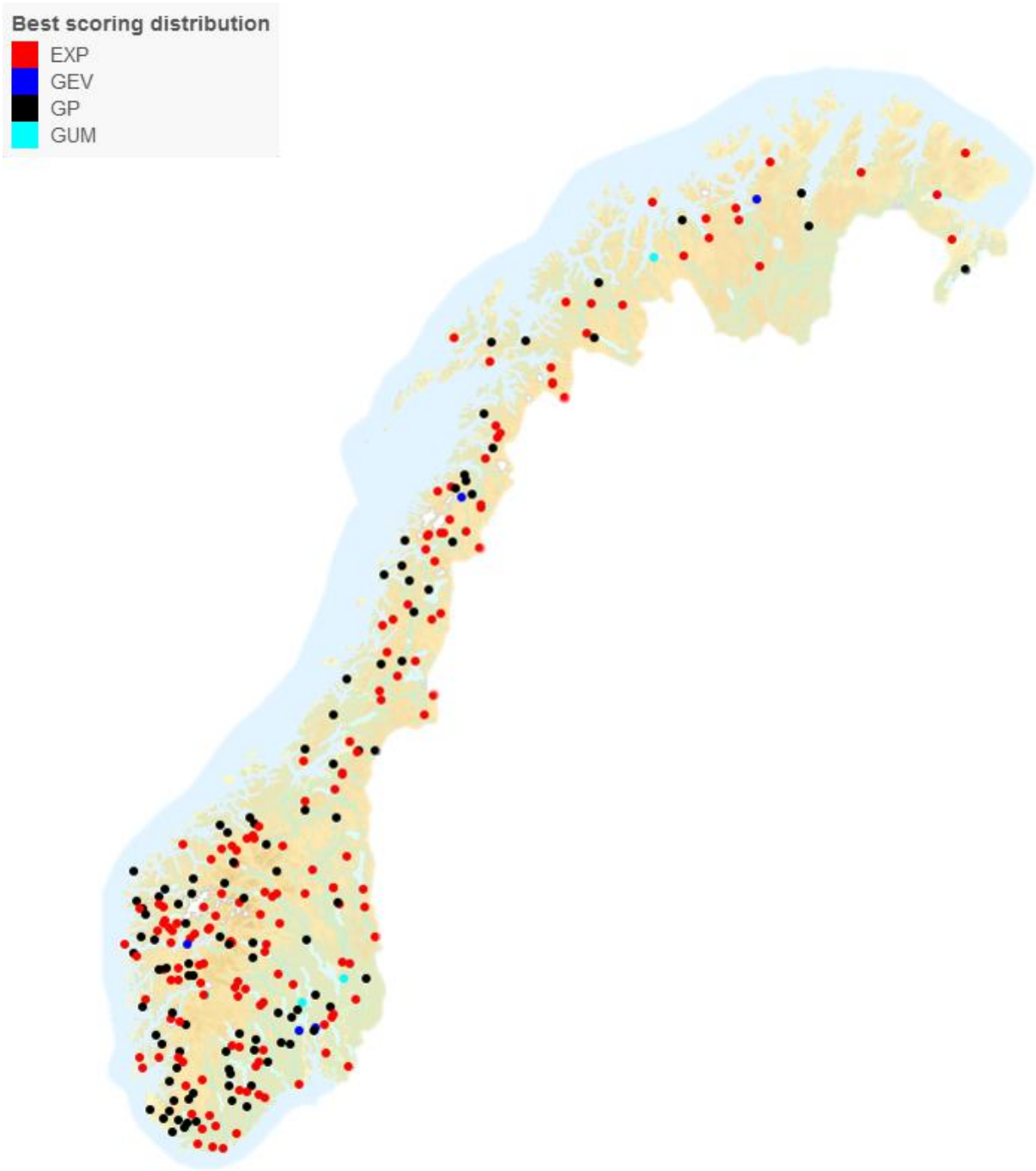

Figure 20: Best scoring distributions according to Brier score for a 20-year return period.

Figure 19 and Figure 20 shows the best performing distribution in the quantile and Brier score respectively. There is not made a distinction between models fitted to the same distribution in Figure 19

andFigure 20 (i.e. EXP and EXP AMS are both marked as red). For the quantile score (Figure 19) the AMS distributions is found to perform best in a few areas around Oslo, Sogndal and in the southern and northern most parts of the country. For the Brier score AMS distributions can mostly be seen close to Oslo in the eastern part of Norway

3.3 Stability

In a flood frequency analysis reliability and predictive ability are the most important aspects, consequently these results were presented first. Stability of the models can help assess which model is superior if differences in descriptive and predictive are negligible.

Estimating the stability of the distributions was done by measuring the design flood values for each distribution through bootstrapping. The coefficient of variation was calculated from the bootstrapping results because an estimation of stability must be dimensionless.

Figure 21 shows the dispersion of the coefficient of variation (CV) for every station for a 20- and 200-year return period. The 3-parameter distributions, GEV and GP shows the largest dispersion of CV values. The mean of the 3-parameter distributions is also slightly higher than the two 2-parameter distributions (EXP and GUM). When 200-year return period design floods are estimated the two 3-parameter distributions show significantly larger variation than the 2-parameter distributions. The EXP distribution yields the lowest total dispersion of CV values for both return periods. It also yields the lowest mean coefficient of variation at around 6%and 7.5% for 20- and 200-year return period respectively. The GUM distributions show higher means at around 7 and 8% for 20- and 200-year return periods, while the 3-parameter distributions have a CV at around 7.5% and 14.5% for 20- and 200- year respectively. The GP distribution has the highest CV for both return periods. The thickness in the x-direction signify the frequency at which a given value occurs.

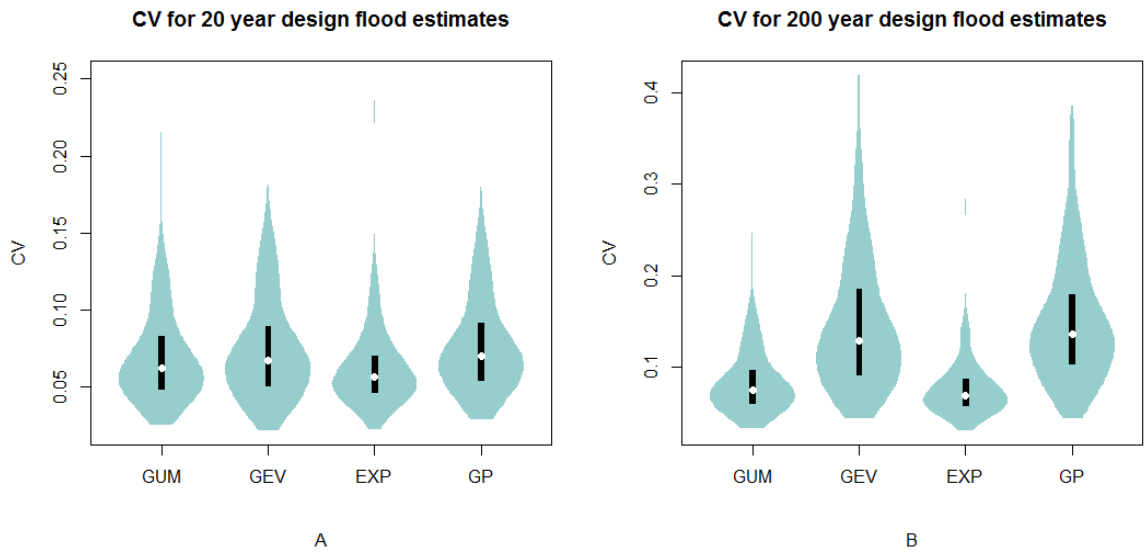

Figure 21: Violin plot of the coefficient of variation of distributions among stations. Plot A shows the coefficient of variation for a 20-year return period, while plot B shows the coefficient of variation for a 200-year return period. The white dot shows the mean value, while the black box spans from the 25-75% quartile range.

4 Discussion

Here I will discuss the method and results of this study and compare it to other studies. I will also discuss potential errors within the results and what I consider to be inherent flaws in the data.

4.1 Limitations of the data

A potential limitation of distributions commonly used in flood frequency analysis is that they don't consider long term trends. These distributions assume the data to be stationary, i.e. that climatic factors and the spatial properties of the catchment have remained the same throughout the data gathering period. The models also assume that the same characteristics remain unchanged for whatever recurrence interval we extrapolate to. On this Bayazit (2015) stated that we should not even assume that any future trend is equal to an observed trend. Gumbel (1941) eluded to the problem of assuming stationarity, stating that these methods must assume stationarity of the data and that these methods will become unreliable the data is non-stationary. More recently Stedinger and Griffis (2011) suggested adding a parameter to a commonly used distribution to account for non-stationarity. Leclerc and Ouarda (2007) suggested a non-stationary parameter for all distributions used in an FFA, and that selection of distribution should also consider accuracy of non-stationarity. With gauged records often being short

quantifying and extrapolation of observed trends can be highly difficult (Bayazit 2015). Yilmaz et al. (2014) found an increasing trend for extreme precipitation events in Melbourne, Australia. However, models accounting for non-stationarity did not perform better than those assuming stationarity. The authors argued that more studies spanning larger areas should be conducted on the matter. Such studies must be done with care as one should be careful in assuming observed trends in one area are applicable to another similar area (Huntington 2006).

The threshold set when selecting events for the POT data was an arbitrary one, set to the 98th percentile (Engeland et al. 2016). While this may be suitable for some catchments, other catchments may suffer from such a threshold because it neglects to account for catchments characteristics (Lang et al. 1999). Rosbjerg et al. (1992) suggests that a threshold should be determined based on catchment properties. This is highly time-consuming work and with the lack of a universal POT data collection framework automation of the process for large areas is understandable.

Measuring discharge accurately is difficult and hard to automate, therefore a discharge level is calculated from a measure of the river stage level. As the river bed changes constantly with erosion and deposition of materials the true stage level of a river is not necessarily known. If the river cross section is unknown the model must assume this shape, which can affect the accuracy of the data (Saleh et al. 2013). This uncertainty has not been accounted for in this study.

4.2 Methods

Originally the EXP and GP distributions are respectively 1- and 2-parameter distributions. Both lack a location parameter as this is substituted by the selected threshold. It's common practice to also estimate the location parameter, thus treating them as 2- and 3-parameter distributions (Engeland 2005). For this study I ran the EXP and GP model twice, once treating them as 1- and 2-parameter distributions adding the threshold as the "location" parameter, and once as 2- and 3-parameter distributions estimating the location parameter. Models where all parameters were estimated proved more stable than those where the location parameter was left out. Results presented here are derived from the 2- and 3-parameter versions of the distributions as these models showed better stability.

Midttømme et al. (2011) suggests that at least 30 years of data is required to conduct an FFA using 2-parameter distributions. Stations selected for this study consequently had two requirements, to be comprised of at least 30 years of data and provide both AMS and POT data. For three stations this means that the length of the POT data set was extremely low, 18 data points for the shortest set. The length of these data sets can prove problematic when doing a ten-fold cross validation. When the validation data set is very short the test scores are more likely to show instability. If two of the larger events are selected into the same validation set it would be biased towards large values. At the same time another of the validation sets might be biased towards lower values due to a lack of larger values. Both these cases would yield a rather poor goodness-of-fit, and the mean test score for the ten iterations would be worse. A solution to this could be selecting stations with more than 50 years of data, as suggested for using 3-parameter distributions (Midttømme et al. 2011), but this would have decreased the amount of stations tested from 271 to 100.

4.3 Results

In this study I use two goodness-of-fit tests, the Anderson-Darling and the Kolmogorov-Smirnov. The Kolmogorov-Smirnov test shows stability producing similar results regardless of which data is used for validation. The Anderson-Darling test on the other hand does not. As seen in Figure 7 all distributions show an undulating behavior. Closer investigation of the points where the undulating is most pronounced shows that the varying results are caused by outliers. Data from some stations contained one or two flood events that were much larger than the rest of the events. When these larger events are used for validation the test statistic calculated shows a poor goodness-of-fit. For some POT data sets, these one or two largest events occurred in the same year. The pooling of events (see section 2.9) occurring in the same year together then enhanced this problem. The AMS data had another issue causing variance in reliability estimates. With AMS data consisting of single largest event occurring each year, the AMS dataset sometimes includes events much smaller than the second or third smallest event. When these events are used for validation it biases the results in the same way a particularly large event would.

Considering the Anderson-Darling test yields unstable results, less weight should be given to results derived from the AD test than the KS test. All tests except the AD suggest models estimated on POT data

are superior in reliability, stability and forecasting accuracy. While the average goodness-of-fit score of GP is better than the EXP's score, the number of stations for which each distribution shows best fit is quite similar. This means that while GP on average fits better, it might be more appropriate to evaluate distributions based on performance on a case by case basis (Abida and Ellouze 2008, Gubareva 2011, Rahman et al. 2013). The 3-parameter distributions did not yield numerical results for all stations. The GEV distribution lacks AD scores for 25 stations, while GP lacks both AD and KS scores for 38 stations. This happens when the estimation data contains almost all lower values and the validation set contains particularly large values. The goodness-of-fit test then does not yield a numerical result, this occurs once for the EXP distribution and no times for the GUM distribution. It is uncertain how much this lack of values affects the results. While the average test statistic could have been calculated by taking out the non-numeric values (i.e. calculated the average score based on 8 or 9 iterations rather than 10) this would have biased the results and was thus not done. That these stations were left out of the results also biases the results, but the bias caused by this was deemed less consequential. Due to the nature of this issue (models with a small location parameter validated on the larger values), if these stations had given proper test statistics it is probable that the results might be more heavily favoring the 2-parameter distributions.

In Figure 16 the quantile score of all models show an uptick as λ exceeds 3. When considering the outliers in that area, I found that data from some stations with $\lambda>3$ contain a few very large events. Because the quantile score puts more weight on large events these events likely biased the test results causing the uptick in quantile score for all models.

There seems to be a clear advantage in using the peak over threshold data when conducting a flood frequency analysis. According to both the KS test and the Brier score test, having more data points to estimate and validate the models results in better performance. When $\lambda>1.08$ the POT models perform better than the AMS models. Similar results have been found in other studies (Cunnane 1973, Madsen 1996, Bezak et al. 2014). Both Brier and KS test shows continuously better performance as number of events per year increases. The improvement in performance of POT models happens at a progressively slower pace as λ increases, similar to what found by others (Bezak et al. 2014, Nagy et al. 2017). When selecting a threshold, one must weigh the amount of work of selecting peaks versus the improvement in performance a lower threshold would gain the analysis. When selecting a threshold, one should also

consider the desired number of events per year as a suggestion rather than a requirement. This is because some catchments would have to include many smaller events to compile a data set of λ events per year. One of the advantages of the POT data is that all events may be of consequential size, whereas some floods contained in the AMS data may be very small. This was found to cause some instability for the AMS models, but was not investigated further so its importance is uncertain. The POT data set from rivers which experience one major event per year, will likely be similar to the AMS data from the same river if one neglects events of insignificant magnitude.

As the λ of each station is different and the results are made using a smoothing function of scores from all stations, increased accuracy and fit with increased λ can only be inferred. For example, the performance of POT data with λ=5 in this study is an estimation based on the performance of a few stations. While the results suggest that an increase in λ gives a certain increase in performance, this will depend on catchment characteristics and will thus not be true for all catchments.

Between the models estimated on AMS data, the GUM distribution scores slightly better for all tests except for the KS test. The GUM distribution also shows better stability. Between the POT models the GP fits the best according to the goodness-of-fit tests, while the EXP distribution yields better scores in predictive ability. Additionally, even if the GP yield better fit overall EXP shows best fit for the most stations (Figure 8) according to the KS test. The EXP distribution has the best average score and scores best for the most stations for both Brier and quantile score, the relative disparity in Brier and quantile score increase for longer return periods. Similar results were obtained by Blanchet et al. (2015), they argued that the GP distribution overfit the data and that the GP distribution gives less accurate estimates for large recurrence intervals due to its heavy tail. The two most important aspects of the models, reliability or predictive ability show the EXP to be performing better. Conducting a flood frequency analysis using the EXP distribution also has the advantage of requiring a shorter time series than using a 3-parameter distribution would (Midttømme et al. 2011). Consequently, more stations should yield reliable design flood estimates. If an FFA is done over an area with varying regional characteristics, using several distributions should be considered. Changes in model performance from region to region can then be observed and more accurate local design flood estimates can be made.

Flood frequency analysis allows us to extrapolate a design flood value for a recurrence interval that we very rarely see. When we estimate a 200-year flood there is often very few observations that help ensuring accuracy of these estimations. As the recurrence interval estimated for increases so will the variance in estimations. This increase in variance is because the estimates often times are dependent on a few observations, sometimes these observations correspond to much lower recurrence intervals. Nagy et al. (2017) suggested that robust estimates of return levels require the recurrence interval less than two times the length of the data series. Stability of the models is inferred by calculating the coefficient of variation from return level estimates. The coefficient of variation suggests that the 2-parameter GUM and EXP distributions are the most stable distributions, similar to what was found in other studies (Baffie et al. 2017, Nagy et al. 2017).

Figure 11 and Figure 18 suggests that the AMS and POT models perform similarly for catchments with a flood generating process dominated by meltwater (FGP=0). With a more precipitation dominated flood generating process the POT models perform increasingly well. AMS models perform about as well for FGP≈1 as they did for FPG≈0. It is unsurprising that AMS and POT models perform similarly for catchments with low FGP values, because a catchment with FGP close to zero will likely produce similar AMS and POT datasets. Whereas a catchment with FGP ≈ 1 may experience several flood events per year. POT data from such catchments can easily contain more events of significant magnitude than the AMS data from the same catchment. Figure 18 also shows that the EXP distribution is favored over GP in the Brier score for all flood generating processes, apart from catchments where floods are almost completely derived from precipitation.

5 Conclusions

A few frequently discussed topics regarding flood frequency analysis have been considered here using gauged data from stations in all parts of Norway. Models have been estimated based on AMS and POT data and then been evaluated based on reliability predictive ability and stability.

The AD and KS goodness-of-fit tests show the POT based models as a better fit. The Anderson-Darling suggests the POT models fit slightly better and the Kolmogorov-Smirnov test suggests POT models fit

better if λ>0.97 with the disparity in fit increasing as λ increases. Similar results are found in the scoring methods. The Brier and quantile scoring methods suggests POT based models are superior when λ> 1.08 and 0.9 respectively. Per today it is common practice in Norway to use AMS data. This is understandable considering lack of consensus and thus unclear guidelines in threshold selection independence criteria. There was found that utilization of POT over AMS data can substantially increase the performance of models

Two distributions were fitted each to AMS and POT data, GEV/GUM and EXP/GP respectively. Both goodness-of-fit tests claims differences between distributions fitted to the same data is quite small. With the exception of GEV and GUM in the AD test (which shows instability). The Brier and quantile scoring rules suggests the 2-parameter distributions perform increasingly better (relative to the 3-parameter distributions) as the recurrence interval we estimate for increases. The 2-parameter distributions outperformed the 3-parameter distribution in predictive ability and stability while showing similar reliability. For models validated on AMS data the EXP AMS is the most accurate predictive distribution.

Both the choice of data and distribution are affected by the flood generating processes of the area for which an FFA is conducted. When floods are derived from mostly rainwater there is little advantage to using POT data, but the more precipitation contributes to flood events the larger advantage the POT data has. The GP fits slightly better for any FGP value, while the EXP shows better predictive ability than GP for almost all FGP values, this discrepancy is larger for low FGP values.

The selection of data series should be evaluated depending on flood generating processes of the catchments in question. For catchments where precipitation is important POT data should be used, AMS data can be used if the FGP is close to zero. Choice of distribution should be based on performance indexes.

6 Reference

2017. Byggteknisk forskrift (TEK17).*in* K. o. moderniseringsdepartementet, editor. Norske Stat.

Abida, H., and M. Ellouze. 2008. Probability distribution of flood flows in Tunisia. Hydrology and Earth System Sciences **12**:703-714.
Baffie, F. K. 2016a. fitdistrib: A list a functions to fit data to probability distributions.
Baffie, F. K. 2016b. FlomKart: Set of functions to estimate the performance of various methods for flood frequency analysis.
Baffie, F. K., K. Engeland, and T. Thorarinsdottir. 2017. Evaluation of design flood estimates - a case study for Norway. Hydrology Research.
Ball, J., M. Babister, R. Nathan, W. Weeks, E. Weinmann, M. Retallick, and I. Testoni. 2016. Australian Rainfall and Runoff: A Guide to Flood Estimation. Commonwealth of Australia.
Bayazit, M. 2015. Nonstationarity of Hydrological Records and Recent Trends in Trend Analysis: A State-of-the-art Review. Environmental Processes-an International Journal **2**:527-542.
Beard, L. R. 1974. Flood flow frequency techniques. Technical Report. CRWR.
Bezak, N., M. Brilly, and M. Sraj. 2014. Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques **59**:959-977.
Blain, G. C., and M. C. Meschiatti. 2014. Using multi-parameters distributions to assess the probability of occurrence of extreme rainfall data. Revista Brasileira De Engenharia Agricola E Ambiental **18**:307-313.
Blanchet, J., J. Touati, D. Lawrence, F. Garavaglia, and E. Paquet. 2015. Evaluation of a compound distribution based on weather pattern subsampling for extreme rainfall in Norway. Natural Hazards and Earth System Sciences **15**:2653-2667.
Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. Monthey Weather Review **78**:1-3.
Castellarin, A., S. Kohnova, L. Gaál, A. Fleig, J. L. Salinas, A. Toumazis, T. R. Kjeldsen, and N. Macdonald. 2012. *in* E. C. i. S. a. Technology, editor. Centre of Ecology and Hydrology.
Chen, Y. D., G. Huang, Q. Shao, and C.-y. Xu. 2006. Regional analysis of low flow using L-moments for Dongjiang basin, South China. Hydrological sciences journal **51**:1051-1064.
Choulakian, V., and M. A. Stephens. 2001. Goodness-of-fit tests for the generalized Pareto distribution. Technometrics **43**:478-484.
CSA. 2012. *in* C. S. association, editor.
Cunnane, C. 1973. A particular comparison of annual maxima and partial duration series methods of flood frequency prediction. Journal of Hydrology **18**:257-271.
Cunnane, C. 1979. A note on the Poisson assumption in partial duration series models. Water Resources Research **15**:489-494.
Das, S., N. Millington, and S. Simonovic. 2013. Distribution choice for the assessment of design rainfall for the city of London (Ontario, Canada) under climate change. Canadian Journal of Civil Engineering **40**:121-129.
Del Barrio, E., J. A. Cuesta-Albertos, C. Matrán, S. Csörgö, C. M. Cuadras, T. de Wet, E. Giné, R. Lockhart, A. Munk, and W. Stute. 2000. Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. Test **9**:1-96.
Diaconis, P., and B. Efron. 1983. Computer-intensive methods in statistics. Scientific American **248**:116-131.
Engeland, K. 2005. A short introduction to extreme value theory.*in* K. Engeland, editor.
Engeland, K., L. Schlichting, F. Randen, K. S. Nordtun, T. Reitan, T. Wang, E. Holmqvist, A. Voksø, and V. Eide. 2016. Flomdata.*in* NVE, editor. NVE.
Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783-791.
Fisher, R. A., and L. H. C. Tippett. 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Pages 180-190 *in* Mathematical Proceedings of the Cambridge Philosophical Society. Cambridge University Press.
Fushiki, T. 2011. Estimation of prediction error by using K-fold cross-validation. Statistics and Computing **21**:137-146.
Garavaglia, F., M. Lang, E. Paquet, J. Gailhard, R. Garçon, and B. Renard. 2011. Reliability and robustness of rainfall compound distribution model based on weather pattern sub-sampling. Hydrology and Earth System Sciences Discussions **15**:p. 519-p. 532.
Gneiting, T., and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**:359-378.
Gubareva, T. S. 2011. Types of probability distributions in the evaluation of extreme floods. Water Resources **38**:962-971.
Gumbel, E. J. 1941. The return period of flood flows. The annals of mathematical statistics **12**:163-190.
Hosking, J. R. 2009. L-Moments. Wiley StatsRef: Statistics Reference Online.
Hrachowitz, M., H. Savenije, G. Blöschl, J. McDonnell, M. Sivapalan, J. Pomeroy, B. Arheimer, T. Blume, M. Clark, and U. Ehret. 2013. A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrological sciences journal **58**:1198-1255.
Huntington, T. G. 2006. Evidence for intensification of the global water cycle: review and synthesis. Journal of Hydrology **319**:83-95.
Justel, A., D. Peña, and R. Zamar. 1997. A multivariate Kolmogorov-Smirnov test of goodness of fit. Statistics & Probability Letters **35**:251-259.
Keast, D., and J. Ellison. 2013. Magnitude Frequency Analysis of Small Floods Using the Annual and Partial Series. Water **5**:1816-1829.
Kidson, R., and K. S. Richards. 2005. Flood frequency analysis: assumptions and alternatives. Progress in Physical Geography **29**:392-410.

Lang, M., T. Ouarda, and B. Bobée. 1999. Towards operational guidelines for over-threshold modeling. Journal of Hydrology **225**:103-117.
Leclerc, M., and T. B. Ouarda. 2007. Non-stationary regional flood frequency analysis at ungauged sites. Journal of Hydrology **343**:254-265.
Lovdata. 2009. Forskrift om Sikkerhet ved vassdragsanlegg (damsikkerhetsforskriften).*in* O.-o. energidepartementet, editor.
Madsen, H. 1996. At-site and regional modelling of extreme hydrologic events. Technical University of Denmark, Department of Hydrodynamics and Water Resources.
Midttømme, G. H. P., L. E

Holmqvist, E, Ø. Nøtsund, H. Hisdal, and R. Sivertsgård. 2011. Retningslinjer for flomberegninger.*in* NVE, editor. NVE.
Nagy, B. K., M. Mohssen, and K. F. D. Hughey. 2017. Flood frequency analysis for a braided river catchment in New Zealand: Comparing annual maximum and partial duration series with varying record lengths. Journal of Hydrology **547**:365-374.
Nguyen, T. H., S. El Outayek, S. H. Lim, and V. T. V. Nguyen. 2017. A systematic approach to selecting the best probability models for annual maximum rainfalls - A case study using data in Ontario (Canada). Journal of Hydrology **553**:49-58.
Ouarda, T. B., C. Girard, G. S. Cavadias, and B. Bobée. 2001. Regional flood frequency estimation with canonical correlation analysis. Journal of Hydrology **254**:157-173.
Pickands III, J. 1975. Statistical inference using extreme order statistics. the Annals of Statistics:119-131.
R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.
Rahman, A. S., A. Rahman, M. A. Zaman, K. Haddad, A. Ahsan, and M. Imteaz. 2013. A study on selection of probability distributions for at-site flood frequency analysis in Australia. Natural Hazards **69**:1803-1813.
Renard, B., K. Kochanek, M. Lang, F. Garavaglia, E. Paquet, L. Neppel, K. Najib, J. Carreau, P. Arnaud, and Y. Aubert. 2013. Data-based comparison of frequency analysis methods: A general framework. Water Resources Research **49**:825-843.
Rosbjerg, D., H. Madsen, and P. F. Rasmussen. 1992. Prediction in partial duration series with generalized pareto-distributed exceedances. Water Resources Research **28**:3001-3010.
Saleh, F., A. Ducharne, N. Flipo, L. Oudin, and E. Ledoux. 2013. Impact of river bed morphology on discharge and water levels simulated by a 1D Saint-Venant hydraulic model at regional scale. Journal of Hydrology **476**:169-177.
Stedinger, J. R., and V. W. Griffis. 2008. Flood frequency analysis in the United States: Time to update. Journal of Hydrologic Engineering **13**:199-204.
Stedinger, J. R., and V. W. Griffis. 2011. Getting From Here to Where? Flood Frequency Analysis and Climate. Journal of the American Water Resources Association **47**:506-513.
Stephenson, A. G. 2002. evd: Extreme Value Distributions.
Steyerberg, E. W., A. J. Vickers, N. R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. J. Pencina, and M. W. Kattan. 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.) **21**:128.
Tavares, L. V., and J. E. Da Silva. 1983. Partial duration series method revisited. Journal of Hydrology **64**:1-14.
Thas, O., and J.-P. Ottoy. 2003. Some generalizations of the Anderson–Darling statistic. Statistics & Probability Letters **64**:255-261.
USWRC. 1982. Guidelines for Determining Flood Flow Frequency.*in* U. S. W. R. Council, editor. Hydrological Communications.
Viglione, A. 2014. nsRFA: Non-supervised Regional Frequency Analysis.
Von Mises, R. 1936. La distribution de la plus grande de n valeurs. Rev. math. Union interbalcanique **1**.
Yilmaz, A., I. Hossain, and B. Perera. 2014. Effect of climate change and variability on extreme rainfall intensity-frequency-duration relationships: a case study of Melbourne. Hydrology and Earth System Sciences **18**:4065.
Zalina, M. D., M. N. M. Desa, V. T. V. Nguyen, and A. H. M. Kassim. 2002. Selecting a probability distribution for extreme rainfall series in Malaysia. Water Science and Technology **45**:63-68.

7 Appendix

Appendix 1:L-moments for the distributions used in this study. Found in (Hosking 2009)

Distribution	*F(x)*	*L-moments*
Exponential	$x = \xi - \alpha \log(1 - F)$	L-location: $\lambda_1 = \xi + \alpha$ L-scale: $\lambda_2 = \frac{1}{2}\alpha$
Generalized extreme values	$x = \xi + \frac{\alpha}{k}\{1 - (-\log F)^k\}$	L-location: $\lambda_1 = \xi + \frac{\alpha}{k}\{1 - \Gamma(1+k)\}$ L-scale: $\lambda_2 = \frac{\alpha}{k}(1 - 2^{-k})\Gamma(1+k)$ L-skewness: $\tau_3 = 2\frac{1-3^{-k}}{1-2^{-k}} - 3$
Generalized pareto	$x = \xi + \frac{\alpha}{k}\{1 - (1-F)^k\}$	L-location: $\lambda_1 = \xi + \frac{\alpha}{1+k}$ L-scale: $\lambda_2 = \frac{\alpha}{(1+k)(2+k)}$ L-skewness: $\tau_3 = \frac{1-k}{3+k}$
Gumbel	$x = \xi - \alpha \log(-\log F)$	L-location: $\lambda_1 = \xi + \gamma\alpha$ L-scale: $\lambda_2 = \alpha \log 2$