Stockholms universitet

Return period estimation for wind storm losses in Norway

Oscar Lindberg

Masteruppsats i försäkringsmatematik
Master Thesis in Actuarial Mathematics

Masteruppsats 2021:3
Försäkringsmatematik
Juni 2021

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2021:3**
http://www.math.su.se

Return period estimation for wind storm losses in Norway

Oscar Lindberg*

June 2021

Abstract

The insurance risk for natural catastrophes is heavily determined by extreme loss events occurring with low frequency. For insurance providers, this makes probability estimates for individual catastrophe events a crucial part of inference. In this thesis, we analyze historical wind storm loss data from the Norwegian Natural Perils Pool, in order to model the probability of extreme losses occurring from storm events. To do this, we use a Peaks over threshold model, in which the excess losses for storm events above a selected cost threshold follows a Generalized Pareto distribution. We also have access to a catastrophe model that simulates storm events and is used to estimate the probability of extreme losses. We compare this catastrophe model to a maximum likelihood estimated Peaks over threshold model, and also consider a Bayesian estimation of the Generalized Pareto distribution, in which prior uncertainty is based on the catastrophe model. Estimation uncertainty is evaluated using profile likelihood methods, a Bootstrap analysis, and through the posterior distribution for the Bayesian model.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: oscar.lindberg97@gmail.com. Supervisor: Filip Lindskog.

Acknowledgements

I would like to thank my supervisor Filip Lindskog at Stockholm university, for his support and advice, and to my external supervisor at Guy Carpenter, Robert Stenlund, for providing me with this challenge and helping me throughout the project.

Contents

1 Introduction

In this thesis, we explore a few ways of modeling future insurance losses for wind storms in Norway. Using storm data from the Norwegian Natural Perils Pool, where individual claims have been aggregated into storm events, the main focus of this thesis is to estimate the probability distribution of the largest annual wind storm loss. When performing risk assessment for natural perils, a large focus is put on estimating how often a catastrophic storm is likely to occur. A common source of analysis is through return periods, and for a fixed time interval, the return period of an event can be defined as the reciprocal of the probability of the event occurring in the time interval, multiplied by the length of the time interval. For example, if the most costly storm event next year has a 10% probability of exceeding a specific loss, this loss is said to have a 10 year return period, and is denoted as the 10 year return level for losses.

Although the Norwegian Natural Perils Pool covers several types of natural disasters, this thesis will only model wind storms. The reason for focusing on windstorms is because they have caused the highest damage among natural disasters in Norway. While there is often some ambiguity in the precise definitions of different catastrophe types, wind storms are typically defined as extra-tropical cyclones that generates strong surface winds. An extra-tropical cyclone being an area characterized by having lower atmospheric pressure than its surrounding areas.

Having a decent assessment of the probability distribution of large storms is of great importance, given the extreme amount of damage a storm can bring. The scarcity of historical data makes inference on extreme storm losses less reliable. To approach this problem, extreme value theory, an area in statistics that deals with the estimation of rare events, can be a useful tool for inference. Extreme value theory is applied in a wide variety of fields, from finance, where the interest could be in large insurance losses or fluctuations of financial investments, to environmental processes, where the interest may lie in estimating the occurrence of large floods or extreme wind speeds. A fundamental part of extreme value theory is the extremal types theorem, which can be used to motivate the Peaks over threshold model. By selecting a threshold for a data set, this model assumes a Generalized Pareto distribution to the excess values above the threshold.

When assessing the probability of large insurance costs from storm events, another approach is to use what is known as a catastrophe model. A catastrophe model uses methods from several domains, such as meteorology and engineering. This approach assesses the probability of large storm losses by generating a large number of synthetic storm events, estimating what each event would cost, and the probability of each event occurring. Such a model is meant to extrapolate, sometimes far beyond historical data, and gain a better understanding of extreme event scenarios. Through a reinsurance broking agency called Guy

Carpenter, we have access to a catastrophe model for wind storms in Norway.

1.1 Objectives

In this thesis, we will model the probability distribution of wind storm event insurance losses in Norway, and the distribution of the annual maximum storm event loss, using a Peaks over threshold model. The probability distribution for the annual maximum storm event loss will be evaluated using estimates of return levels for losses on a yearly scale. The Peaks over threshold model involves fitting a Generalized Pareto distribution to excess losses, and we will use both maximum likelihood and a Bayesian method to estimate parameters of the Generalized Pareto model, in which prior uncertainty is based on the catastrophe model. We will assess uncertainty in the Generalized Pareto distribution parameters, quantiles of storm losses, and yearly return levels of losses. This will be done using profile likelihood, a Bootstrap simulation, and by evaluating the posterior distribution of the Bayesian model. The interest being to compare how uncertainty in the model differs when using profile likelihood, Bootstrap, and the Bayesian method.

2 Background

2.1 Norwegian Natural Perils Pool

In 1980, the Norwegian Natural Perils Pool was founded, which is governed by the Natural Perils insurance act, and connected insurance against fire to natural perils. This means any property in Norway ensured against fire is also automatically covered against natural perils through the Norwegian Natural Perils pool. The purpose of establishing this insurance pool was to provide compensation for natural peril caused damage, as well as aiding in preventive measures against natural perils. The perils (meaning causes of damages) covered by this insurance pool are

- Storm
- Flood
- Sea surge
- Land slide
- Earthquake
- volcano eruption

In figure 1, we see that over half of historical losses for the Norwegian pool are caused by wind storms. In this plot, we have excluded earthquakes and volcano eruptions, as they make up less than 0.1 % of historical loss.

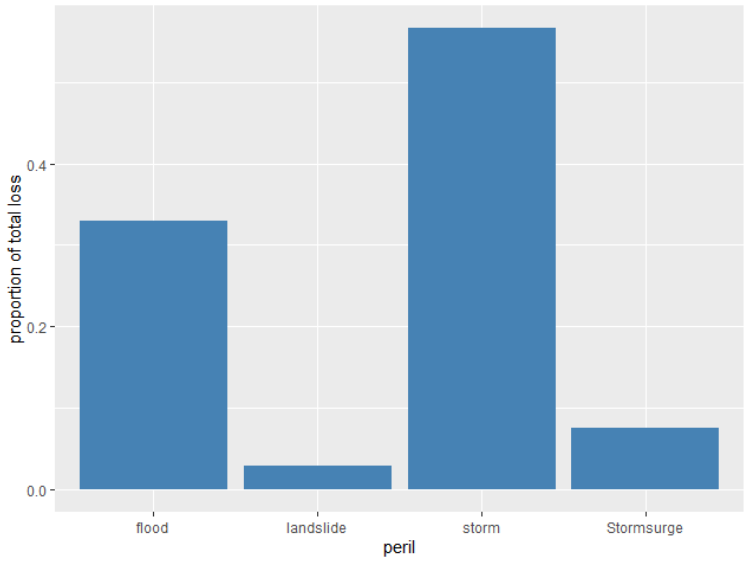

Figure 1: Proportion of historical storm loss by peril

When assessing the risk of a natural peril, it is not only important to consider historical weather and climate information, as factors such as urbanization and building standards can have a great affect on the risks faced by a region. For Norway, the long and rugged coastline is exposed to storms and stretches out to about 25 000 km (which includes several Fjords and islands). The varied topography leads to Norway having steep and fast flowing rivers. Flooding occurs in almost every region of Norway, with the central south region being especially vulnerable to flooding, due to large inland rivers.

To get a visual sense of how the regions of Norway has been affected, we show in figure 2 the sum of the number of claims related to wind storms 1980-2020 for each county. There has been a region reform in Norway 2014-2020 where counties (as well as municipalities) have been merged together, and this plot shows the county split 2019, after counties "Nord-Trøndelag" and "Syd Trøndelag" merged.

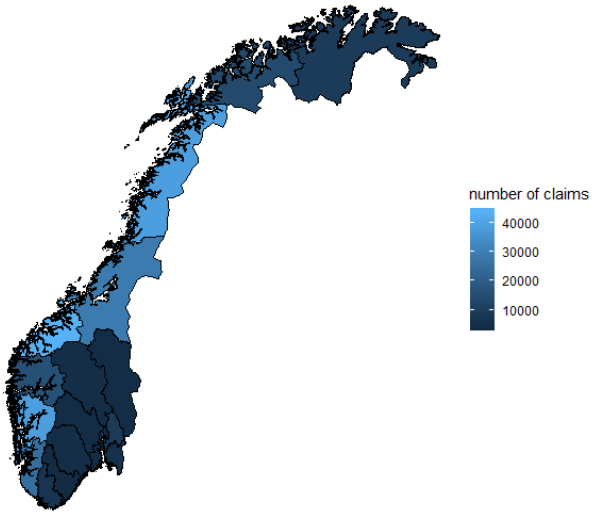

Figure 2: Total number of storm claims 1980-2020 by county

During the last 40 years, the largest storm loss for the insurance pool was the "New years day" storm (Nyttårstormen) in 1992. The worst damage was caused in the north west region of the country, as we see in figure 3, where we show the number of claims caused during this storm. The meteorological institute of Norway stated that a storm of that magnitude is expected to occur less than once every 200 years.[Meteorologisk institutt, 2017] It is however important to distinguish between storm return periods from an extreme weather perspective, versus a financial loss perspective, as they can differ greatly.

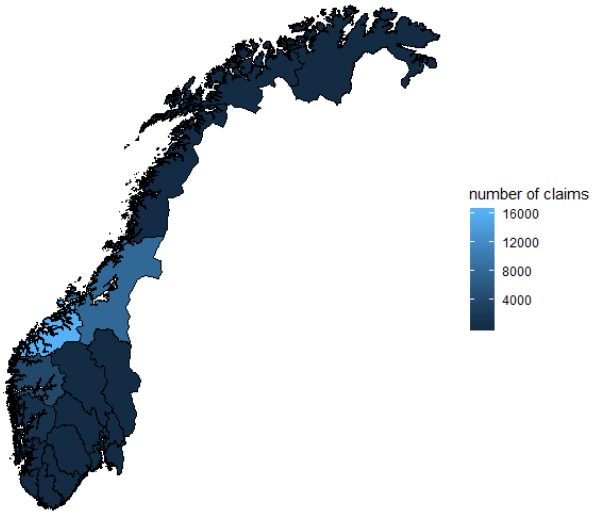

Figure 3: Number of claims during the New year day storm 1992 by county

Another large storm was Dagmar, which occurred on Christmas day in 2011 and impacted the north western and inner parts of the country, as well as Sweden and Finland. The third largest wind storm was in 2015, which was named Nina, and damaged Norway, Sweden and Denmark. The most recent large event was the landslide in Gjerdrum in December of 2020, which is in the south east part of Norway.

2.2 Reinsurance

A large part of insurance regulation is about solvency, making sure that an insurer has the ability to reimburse the people they have obliged to cover. To make sure an insurance provider has enough assets to cover their risks, solvency capital requirements are calculated on a regular basis, which are used to estimate how much capital an insurance provider needs to hold in order to remain solvent with at least a 99.5% probability. This can also be phrased in terms of return periods, where the solvency capital requirement is meant to cover a once in a 200 year scenario.

If an insurance provider is reluctant to expose themselves to a financial risk above a certain threshold, they may diversify their risk by paying a premium to a reinsurance provider to assist with reimbursements for more extreme scenarios. This reinsurance coverage could for example be proportional, in which the reinsurer would pay a set share of the total claim cost. Another type is the so called excess of loss, which is often defined by a lower and upper cost threshold. The reinsurer would then be obliged to pay the excess loss (i.e. the amount by

which the cost exceeds the threshold) above the lower threshold up until the claim cost reaches the upper threshold. The market for natural perils insurance is characterized by low frequency, high loss events, which makes the demand for reinsurance especially high given the extraordinary potential loss that can occur for single events.

2.3 Catastrophe models

The purpose of a catastrophe (CAT) model is to estimate loss from extreme, wide-impact events (that are termed catastrophes), while the loss estimated is usually financial. In practice, catastrophe models are used in several domains, such as reinsurance, where they are used in pricing and structuring of reinsurance contracts. They are also used in the calculations of capital requirements for non-life insurance. An important distinction is that a catastrophe model does not attempt to predict natural catastrophes, its focus is only to estimate probabilities for different catastrophe scenarios.

In [Mitchell-Wallace et al., 2017], the general theory and application of catastrophe models is described, which we will go through in very summarized terms below. A catastrophe model is often divided into the following 4 components

- Exposure
- Hazard
- Vulnerability
- Financial

The exposure component specifies the risks covered by the model. This information is specified as a database containing the exposure values covered by the model, split by area and type of risk. If we for example only model buildings, we would for each building category and location calculate the number of buildings, the total sum insured, and deductible information. The total sum insured can be divided into building sum insured and content sum insured, where the building sum insured is the value of the physical structure (including walls, floor, roof) and the content sum insured is the value of the content (the items inside) of the building. These would be the most important features of the risks, and are typically denoted as primary characteristics of the exposures. Beyond this, secondary characteristics, which are not as important as the primary ones, may for example include the age and number of stories of the buildings.

As a catastrophe model generates simulated storms, the hazard component combines the information from simulated storms (such as peak wind speeds for each location) with the exposure data to estimate the hazard impact of the storm for each location. This information is used to define event footprints, which are meant to reflect the relative intensity of a hazard for a given storm.

The vulnerability component estimates how the risk of a location would respond to predicted hazard conditions. These estimates are mainly based on engineering studies or past experiences. For earthquakes, the Peak Ground Acceleration measure is commonly used to estimate the damage a given type of earthquake may bring to an area.

Once we have estimated the damage done by a potential storm, the financial component then applies the financial and insurance terms to each storm, such as the policy conditions and deductibles. The final output is then an event loss table, which is an estimate of the financial cost for each individual storm event generated, with an associated probability for each event. The information is expressed both in terms of Ground-Up loss, meaning before we apply deductibles and other relevant financial terms, and Gross loss, which is the loss after considering these financial terms. This event loss table is used to estimate several risk measures, including the exceedance probability of the maximum storm loss per year, and the annual average loss expected to occur.

2.4 Market database

Part of the work behind this thesis has been to estimate the total sum insured for the risks covered by the Norwegian Natural Perils Pool, which is used in the CAT models, and also provides underwriting information for the insurance Pool. As approximately all buildings in Norway are covered by this insurance pool, we chose to use public building characteristics information from Statistics Norway (SSB) and pricing information of buildings from Finance Norway (FNO), to estimate the total value of properties for every municipality in Norway. In figure 4, we display the total sum insured by every county in Norway, where the darker counties indicate a higher total sum insured.

In figure 5, we also show the total sum insured, split by the building (or occupancy) types used in the CAT model. The occupancy type "Residential - General" includes separate houses, which as we see comprises roughly a third of the total sum insured.

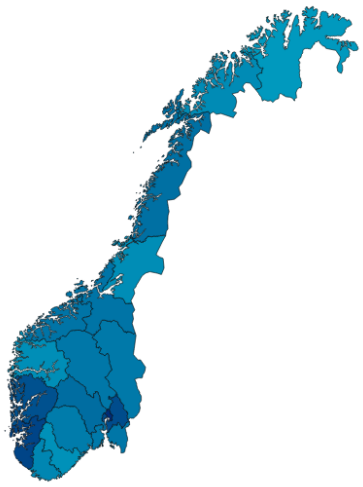

Figure 4: Total sum insured by county in Norway

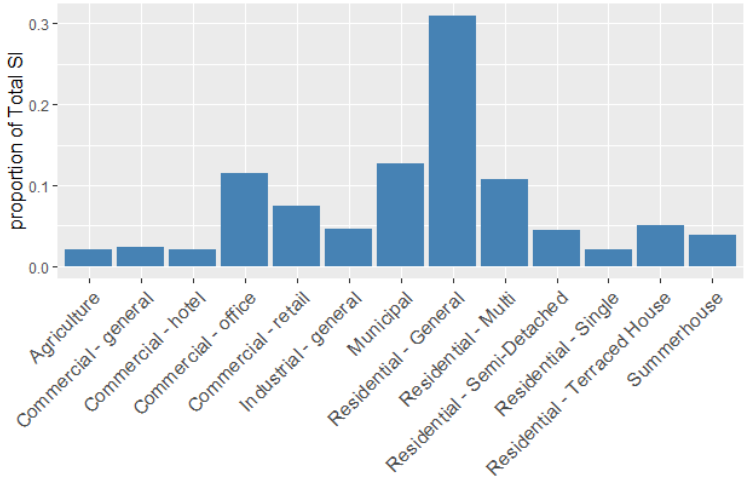

Figure 5: Total sum insured by occupancy type

2.5 Motivation for a Bayesian model

In a statistical model, a Bayesian method treats the unknown parameters as random variables, then formulating what is known as a prior distribution for the parameters, this prior being based on initial beliefs of uncertainty before observing the data. This process can be a useful way of incorporating different

sources of information into the analysis, something that is especially practical when data is scarce. Taking into account both prior uncertainty and the likelihood of observed data, a Bayesian model puts its focus on the posterior distribution of the unknown parameters, which reflects uncertainty in the model after observing the data. As a Bayesian model specifies uncertainty in estimates directly using the posterior, it has the advantage of not having to rely on asymptotic theory which can be unsuitable for data with low sample size. The main cost of using this method is the subjective nature of specifying a prior distribution, and while the effect the prior has on the model typically decreases with sample size, a prior specified without much consideration can lead to very poor estimates when data is scarce.

In this analysis, we want to make inference about insurance losses for future wind storms in Norway, and have access to over 40 years of historical storms. Beyond that, we have a catastrophe model, which could be described as a type of scenario analysis in that it simulates future losses by generating synthetic storm events. A Bayesian model that translates properties of the CAT model into a prior distribution for the Generalized Pareto distribution could potentially reduce uncertainty in the probability of extreme losses. This Bayesian model may also act as a compromise between two methods of inference, one based purely on historical losses, and the other based on assessing loss probabilities by combining market exposure information with research on extreme weather scenarios.

3 Storm data

We will be analyzing data containing storm claims in Norway between 1980-2020. The individual claims have been aggregated into storm events losses, leading to 83 storm events. To define a storm event, the so called hour clause is used, in which a storm event size is defined by the sum of claims that enter in a 72-hour window. While we have not set a definitive limit, only 72-hour windows with a sufficiently high insurance cost gets treated as events in our data. This window clause definition has a very important consequence when signing reinsurance, for example, if two days with extreme weather happen next to each other, the reinsurance compensation may depend widely on whether we count this as one event or two.

In figure 6 we show the storm event losses (inflation adjusted, and transformed for anonymity) over time, where we see that the cost differences between the three most costly storms (Nyttårstormen, Dagmar and Nina) are extremely large, followed by more dense distribution of losses.

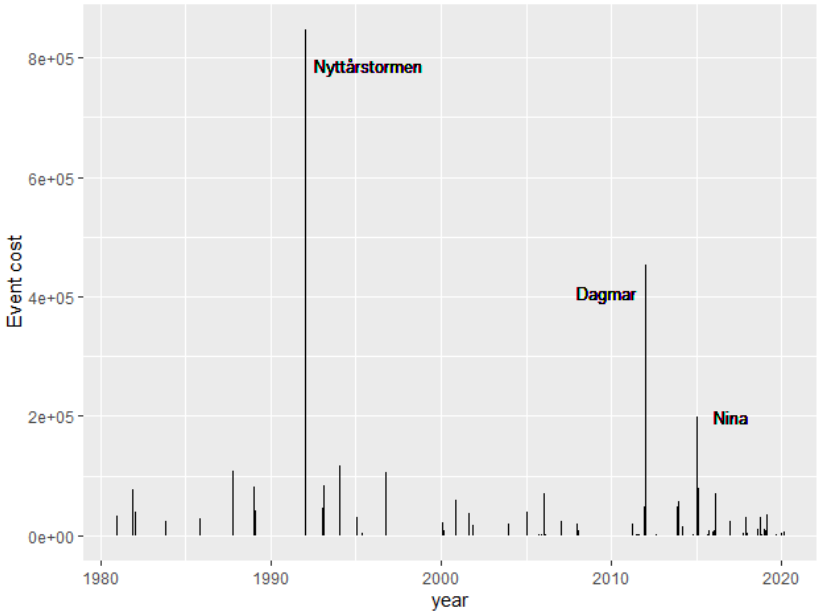

Figure 6: Storm event losses over time

We have over 40 years of data, and during that time, there has been a great change in the value of properties, and the number of properties in each region in Norway. When comparing storm costs from different years, it is important to adjust for such observable factors that impact claim costs in order to see the "true" economic impact of each storm. This is also important for the statistical model we will use, as it will assume all data is from the same distribution, which means there should not be a trend over time for the data being modelled. We have used an average of two inflation measures, the consumer price index and building price index to get storm losses from different years on the same pricing level. We have also adjusted for portfolio changes by scaling costs based on the total sum insured for the given period of which a storm occurred.

4 Theory

In the following section, we will go through some of the basics of extreme value theory, see how it motivates using the Generalized Pareto distribution for modeling extreme data using the Peaks over threshold method, and how to estimate yearly return levels for extreme events. We will also go through some basic aspects of modeling the frequency of counting processes, how to estimate parameter uncertainty using likelihood ratio and Bootstrap methods, and some theory behind Bayesian modeling.

4.1 Extreme value theory

We will primarily use the notation of [Coles et al., 2001, Ch. 3-4] in order to describe extreme value models.

Extreme value theory deals with trying to model the probability of extreme events, the foundation for many extreme value models are built upon the extremal types theorem, also known as the Fisher-Tippet Theorem, which is about a certain probability distribution convergence related to the maximum of random variables. We will describe this theorem briefly below.

For a set of independent identically distributed continuous random variables $X_1, .., X_n$ with cumulative density function F, we will denote
$M_n = \max(X_1, .., X_n)$. The cumulative density function for the maximum of these variables is

$$P(M_n \leq x) = P(X_1 \leq x, .., X_n \leq x) = \prod_{i=1}^{n} P(X_i \leq n) = F(x)^n. \tag{1}$$

The extremal types theorem states that if there exists sequences $\{a_n\}, \{b_n\}$ such that

$$P(\frac{M_n - b_n}{a_n} \leq z) \to G(z) \text{ as } n \to \infty \tag{2}$$

then the linearly transformed maximum $M_n^* = \frac{M_n - b_n}{a_n}$ will converge in distribution as $n \to \infty$ to one of the following three forms.

$$\begin{aligned} &I : G(z) = \exp(-\exp(-\frac{z-b}{a})), -\infty < z < \infty \\ &II : G(z) = \begin{cases} 0, z \leq b \\ exp(-(\frac{z-b}{a})^{-\alpha}), z > b \end{cases} \\ &III : G(z) = \begin{cases} \exp(-(-(\frac{z-b}{a})^{-\alpha})), z > b \\ 1, z > b \end{cases} \end{aligned} \tag{3}$$

for parameters $a > 0, b$ and in case II,III $\alpha > 0$. These distributions are known as extreme value distributions, with types I, II and III known as the Gumbel, Frechet and Weibull families respectively. In practice, the condition of 2 is met for nearly all continuous distributions. The proof of this theorem, which is quite extensive, can be found in
[Pickands III et al., 1975]. These three distribution families can also be generalized into what is known as the generalized extreme value distribution, which takes the form

$$G(z) = \exp(-(1 + \xi(\frac{z-\mu}{\sigma}))^{-1/\xi}), 1 + \xi(z - \mu)/\sigma > 0 \tag{4}$$

Here, the type I,II and III case earlier described corresponds to

$$\begin{cases} \xi \to 0, \text{ (Gumbel)} \\ \xi > 0, \text{ (Frechet)} \\ \xi < 0, \text{ (Weibull)} \end{cases}$$

If a distribution, F, has the property of equation (2), it is said to be in the maximal domain of attraction, sometimes expressed as $F \in \text{MDA}(G)$, where G refers to which of the three types the distribution belongs to. The main way of differentiating between the three family types is by their tail behaviour, often expressed through the survival function $(1 - G(z))$, which specifies the probability of exceeding a given value for distribution G. The type I case, $\xi < 0$ corresponds to distributions having a finite upper end point, while $\xi \to 0$ has a survival function that decreases exponentially for large values, such that the tail is not very heavy. The case where $\xi > 0$ means the survival function decreases at a sub exponential rate (i.e. very slowly), which means the right tail of the distribution will be heavier.

We show in figure 7 one example of how the scaled maximum of random variables may converge. In this simulation, we have generated 500 blocks of data, with each block containing 500 generated values of independent Pareto random variables. Scaling the maximum of Pareto random variables by appropriate stabilizing sequences, which turns out to be $a_n = F^{-1}(1 - 1/n), b_n = 0$ (we show this choice holds up in Appendix A.3), the scale transformed maximum $\frac{M_n}{a_n}$ converges to a Frechet distribution, which was one of the three types of Generalized extreme value distributions. The red curve in figure 7 is the probability density function for the Frechet distribution, which is the derivative of the cumulative distribution function in equation (4),

$$f(x) = \frac{1}{\sigma}(1 + \xi\frac{x - \mu}{\sigma})^{-(1/\xi+1)} \cdot \exp(-(1 + \xi\frac{x - \mu}{\sigma})^{-1/\xi}).$$

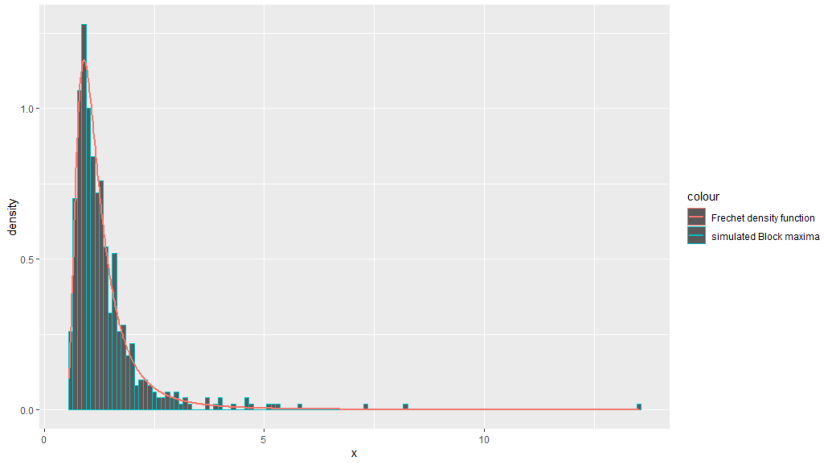

Figure 7: Simulated block maxima for Pareto variables

This convergence is the basis behind the so called block maxima method, which splits a data set into blocks, often by time measures (such as monthly observations), and measures the maximum of each block. In this method, you would fit a generalized extreme value distribution to the maximum observed values. The motivation behind this model is that if the distribution follows the maximal domain for attraction, then for large values of n,

$$P(M_N^* < z) \approx G(z)$$
$$P(M_N < z) = P(M_n^* < z \cdot a_n + b_n) \approx G(z \cdot a_n + b_n) = G^*(z)$$

where G^* is also member of the Generalized extreme value family of distributions. This means we do not need to know the normalizing sequences a_n and b_n in practice in order to use a Generalized extreme value model for the data, we only need to know they exist. It is important to specify the blocks such that there is enough data within each block for the GEV distribution to be appropriate, while at the same time having enough blocks to fit a sample with. One important drawback of the block maxima method is the loss of information you get by only including the maximum value for each block. This ignores any other extreme values that was not the largest within its block from the model.

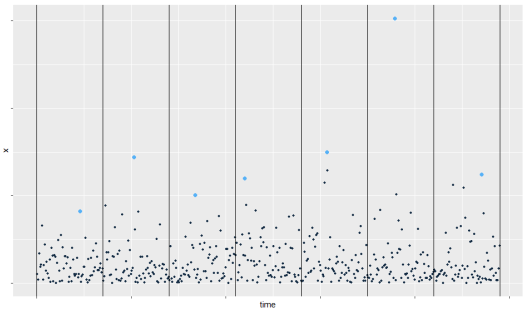

Figure 8: Block maxima illustration example with 7 blocks

4.1.1 Peaks over threshold

The peaks over threshold method puts it focus on modeling the tail distribution of data by only considering data above a certain threshold (i.e. $P(X|X > u)$, for some threshold u). If a distribution is in the maximal domain of attraction, one can show that for a sufficiently large threshold u, the conditional excess distribution is approximately

$$P(X - u \leq y|X > u) \approx H(y),$$

where $H(y)$ is the cumulative distribution function for the Generalized Pareto distribution, which can be defined as

$$H(y) = \begin{cases} 1 - (1 + \xi\frac{y}{\sigma})^{-1/\xi}, \xi \neq 0 \\ 1 - exp(\frac{y}{\sigma}), \xi = 0 \end{cases} , y \in \begin{cases} [0, \infty], \xi \geq 0 \\ [0, -\sigma/\xi], \xi < 0 \end{cases} \tag{5}$$

with scale parameter $\sigma > 0$ and shape parameter ξ, we will denote this distribution as $GP(\xi, \sigma)$.

To understand why this approximation is suitable, we can first note that from the extremal types theorem, if a distribution X is in the maximal domain of attraction, we have an approximation for large values of n,

$$F_X^n(z) \approx \exp(-(1 + \xi(\frac{z - \mu}{\sigma}))^{-1/\xi})$$

for parameters μ, σ, ξ. Taking the logarithm of both sides of this equation yields

$$n \cdot \log(F_X(z)) \approx -(1 + \xi(\frac{z - \mu}{\sigma}))^{-1/\xi}.$$

For large values of z, we can use a first order Taylor expansion of $\log(F_X(z)) \approx -(1 - F_X(z))$. Using this approximation, and dividing by $-n$ on both sides, we achieve

$$P(X > u) = 1 - F_X(u) \approx \frac{1}{n}(-(1 + \xi(\frac{u - \mu}{\sigma}))^{-1/\xi}),$$

which finally gives us

$$p(X > u + y|X > u) = \frac{p(X > u + y)}{p(X > u)} \approx \frac{\frac{1}{n}(-(1 + \xi(\frac{u+y-\mu}{\sigma}))^{-1/\xi})}{\frac{1}{n}(-(1 + \xi(\frac{u-\mu}{\sigma}))^{-1/\xi})} =$$
$$(1 + \frac{\xi(u + y - \mu)/\sigma}{1 + \xi(u - \mu)/\sigma})^{-1/\xi} = (1 + \frac{\xi y}{\tilde{\sigma}})^{-1/\xi}, \tag{6}$$
$$\tilde{\sigma} = \sigma + \xi(u - \mu).$$

We may note that the shape parameter ξ is the same for the generalized extreme value distribution and this Generalized Pareto distribution.

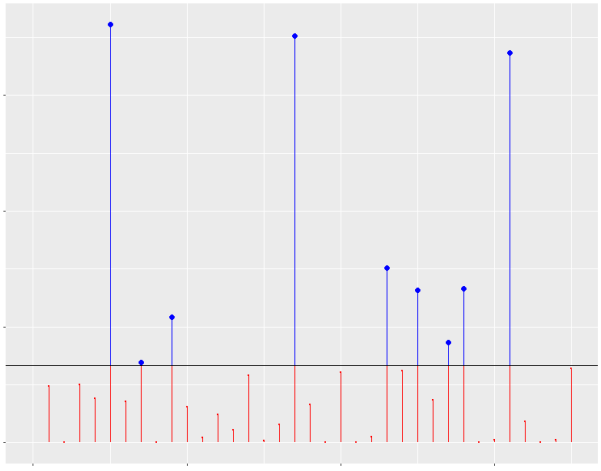

Figure 9: Peaks over threshold illustration

The probability density function for a Generalized Pareto distribution $Y \sim GP(\xi, \sigma)$ is equal to

$$f_Y(y) = \frac{\partial}{\partial y} H(y) = \frac{1}{\sigma} \cdot (1 + \xi \frac{y}{\sigma})^{-(1+1/\xi)}, \xi \neq 0,$$

which means the log likelihood for a set of independent data points $y_1, .., y_n$ with distribution Y is

$$l(y_1, .., y_n|\xi, \sigma) = \log(\prod_{i=1}^{n} f_Y(y_i)) = \sum_{i=1}^{n} \log(\frac{1}{\sigma} \cdot (1 + \xi \frac{y_i}{\sigma})^{-(1+1/\xi)}) =$$
$$- (n \log(\sigma) + (1 + 1/\xi) \sum_{i=1}^{n} \log(1 + \xi y_i/\sigma)) \tag{7}$$

Numerical methods can be used to optimize the log likelihood, and thus obtaining the maximum likelihood estimator

$$(\hat{\xi}_{ml}, \hat{\sigma}_{ml}) = \underset{\xi,\sigma}{\operatorname{argmax}}\ l(y_1, .., y_n|\xi, \sigma)$$

For different probability levels p, we want to find the values for which $P(Y < q_p) = p$, which can be obtained by using the inverse cumulative distribution function, $H^{-1}(y)$. By this definition, $H^{-1}(p)$ defines the level p quantiles of the Generalized Pareto distribution, since $P(Y < H^{-1}(p)) = H(H^{-1}(p)) = p$. This inverse function can be obtained through the following steps

$$\begin{aligned}
1 - H(y) &= (1 + \xi \frac{y}{\sigma})^{-1/\xi} \\
(1 - H(y))^{-\xi} &= 1 + \xi \frac{y}{\sigma} \\
\frac{\sigma}{\xi}((1 - H(y))^{-\xi} - 1) &= y
\end{aligned}$$

From these calculations, we receive the level p quantile of the generalized Pareto distribution

$$H^{-1}(y) = \frac{\sigma}{\xi}((1 - y)^{-\xi} - 1). \tag{8}$$

If we would like to sample from the Generalized Pareto distribution, we may first generate a uniform random variable $U \sim \text{unif}(0, 1)$ between 0 and 1. Using the symmetric property of the uniform distribution $1 - U \sim \text{unif}(0, 1)$, along with the transformation theorem for random variables, we obtain

$$Y = H^{-1}(1 - U) = \frac{\sigma}{\xi}((U)^{-\xi} - 1) \sim GP(\xi, \sigma).$$

since

$$P(H^{-1}(1 - U) \leq y) = P(1 - U < H(y)) = H(y).$$

We can also use this property in order to find the mean of the Generalized Pareto distribution.

$$\begin{aligned}
&U \sim \text{unif}(0, 1) \\
&Y = \frac{\sigma}{\xi}((U)^{-\xi} - 1) \sim GP(\xi, \sigma) \\
&E[U^{-\xi}] = \int_0^1 x^{-\xi} dx = [-x^{-\xi-1}/(\xi - 1)]_{x=0}^{x=1} = \frac{-1}{\xi - 1} = \frac{1}{1 - \xi}, \xi < 1 \\
&E[Y] = E[\sigma(U^{-\xi} - 1)/\xi] = \frac{\sigma}{\xi}(E[U^{-\xi}] - 1) = \frac{\sigma}{\xi}(\frac{1}{1 - \xi} - 1) = \\
&\frac{\sigma}{\xi}(\frac{\xi}{1 - \xi}) = \frac{\sigma}{1 - \xi}, \xi < 1
\end{aligned}$$

The mean of the distribution is only finite for $\xi < 1$, which can be problematic for certain situations, since in many practical cases, we know the true mean of the distribution to be finite.

Another important property is that if $Y \sim GP(\xi, \sigma)$, then $Y - u|Y > u \sim GP(\xi, \sigma + \xi u), u > 0$, meaning that the excess distribution for higher thresholds follows the same distributional form, which is a property

unique to the Generalized Pareto distribution. The property can be shown by calculating the conditional probability

$$
\begin{aligned}
&P(Y-u>y|Y>u) = P(Y>y+u|Y>u) = \frac{P(Y>y+u|Y>u)}{P(Y>u)} = \\
&\frac{(1+\xi\frac{y+u}{\sigma})^{-1/\xi}}{(1+\xi\frac{u}{\sigma})^{-1/\xi}} = (\frac{(1+\xi\frac{y+u}{\sigma})}{(1+\xi\frac{u}{\sigma})})^{-1/\xi} = \\
&(1+\frac{\xi\frac{y}{\sigma}}{(1+\xi\frac{u}{\sigma})})^{-1/\xi} = (1+\xi\frac{y}{(\sigma+\xi u)})^{-1/\xi} = (1+\xi\frac{y}{\tilde{\sigma}})^{-1/\xi} \\
&\tilde{\sigma} = \sigma+\xi u
\end{aligned}
\tag{9}
$$

Selecting the threshold value u for the Peaks over threshold model is a classic example of the bias-variance trade off, as selecting a higher threshold should lead to a better approximation of the Generalized Pareto distribution, thus reducing bias. Meanwhile, increasing the threshold also reduces the number of data points for estimation, giving higher variance in estimation.

One visual guide to choosing an appropriate threshold, is to plot the estimated parameters for different choices of thresholds. In this approach, one plots the estimate for ξ, and the transformed scale parameter $\sigma_0 = \sigma + \xi u$, as they should both in theory be the same for different thresholds, a consequence of equation (9). One would then choose a threshold where the estimates for σ_0 and ξ look stable around a small region of thresholds.

Another method of determining a suitable threshold is by using a certain linearity property of quantiles, using that
$Y = X - u|X > u \sim GP(\xi, \sigma + \xi(u-\mu))$, the level p quantile of the conditional excess would be

$$H_Y^{-1}(p) = \frac{\sigma+\xi(u-\mu)}{\xi}((1-p))^{-\xi} - 1)$$

which is a function linear in the threshold u. An approach to determine a suitable threshold is then to choose a probability level p, and for different choices of thresholds, plot the quantiles $H_Y^{-1}(p)$ against the threshold, and find the lowest point for which you find a stable linear trend, with some random variation.

4.2 Frequency

As we are interested in the frequency of extreme storm losses, we need to assume a model for storm frequency.

When modeling observed count data over time, it is common to use models relating to Poisson processes. A Poisson process $\{N(t) \geq 0\}$, which represents the total number of events up to (and including) time t, is characterized by a hazard rate function $\lambda(t)$. If we define the cumulative hazard rate function as

$M(t) = \int_0^t \lambda(s)ds$, a Poisson process has the property

$$N(t_2) - N(t_1) \sim po(M(t_2) - M(t_1)), t_1 < t_2.$$

A second condition for Poisson processes is independent increments, meaning $N(t_4) - N(t_3)$ is independent of $N(t_2) - N(t_1)$, for all time intervals $t_1 < t_2 \leq t_3 < t_4$. If the rate function is constant, meaning $\lambda(t) = \lambda, \forall t \geq 0$, it is known as a homogeneous Poisson process, such that $\int_{t_1}^{t_2} \lambda(s)ds = \lambda \cdot (t_2 - t_1)$. In that case $N(t) \sim po(\lambda t)$, and the number of events increases linearly in expectation. If we have yearly data, the number of events each year would follow the same Poisson distribution. This would mean the variance and the expected number of events per year would be the same, an assumption that is often violated for many counting processes, where the variance tends to be larger than the mean. This type of violation is called overdispersion, and can be accounted for by using a frequency model where the variance is larger then the mean, such as the Negative Binomial distribution.

4.3 Distribution of annual maximum event

If we have N number of events per year, with i.i.d. values $x_1, .., x_N$ independent of N. Assuming N is known, the maximum value $M_N = \max(x_1, .., x_N)$, where we set $M_0 = 0$, will have a cumulative distribution function of

$$P(M_N \leq x | N = n) = P(x_1 \leq x, .., x_n \leq x) = F(x)^n, n = 0, 1, 2, ..$$

If the number of events, N, is also random, we can still get the marginal probability distribution for M_N by first using the relation between the conditional distribution $P(M_N|N)$ and the joint probability $P(M_N, N)$ as

$$P(M_n \leq x, N = n) = P(M_N \leq x | N = n)P(N = n).$$

Next, to get the marginal distribution of M_N, we sum the joint distribution of $P(M_N, N)$ over all possible values of N, to achieve

$$\begin{aligned} F_{M_n}(x) = P(M_N \leq x) &= \sum_{n=0}^{\infty} P(M_n \leq x | N = n)P(N = n) \\ &= \sum_{n=0}^{\infty} F(x)^n P(N = n) = E[F(x)^N] = \phi_N(F(x)) \end{aligned} \tag{10}$$

where the last expectation is with respect to N. The expression $\phi_N(z) = E[z^N]$ is the probability generating function for the number of events, and since $F(x)^n \in [0, 1]$ this generating function needs to exist for values between 0 and 1 in order for 10 to hold. If we know the inverse of the function in (10), we can express the T year return level for individual storm losses as

$$F_{M_N}^{-1}(1 - 1/T).$$

Another important quantity often analyzed for storm events is the occurrence excess probability curve, which is the probability of the maximum storm losses exceeding a given value, meaning

$$p(M_N > x) = 1 - \phi_N(F(x)).$$

Let us show an example, where we generate 1000 "years" of data, where the number of events each year follow a negative binomial distribution $N \sim Nbin(r, p)$ with r=5 and p =0.5 such that the mean is $E[N] = \frac{p \cdot r}{1-p} = 5$, and the event sizes each year follow a Gamma distribution $x_1, .., x_N \sim Ga(\alpha, \beta)$ with mean $\frac{\alpha}{\beta} = 100000$ and standard deviation $\sqrt{\frac{\alpha}{\beta^2}} = 14142$. The drawings are shown in figure 10, where the curve is the theoretical probability density function for M_N. The density function is obtained by taking the derivative of the CDF in equation (10), using the chain rule, i.e

$$\frac{\partial \phi_N(F(x))}{\partial x} = \frac{\partial \phi_N(F(x))}{\partial F(x)} \frac{\partial F(x)}{\partial x},$$

where $F(x)$ is the Gamma cumulative distribution function, and

$$\phi_N(z) = (\frac{1-p}{1-pz})^r, |z| < \frac{1}{p}$$

is the probability generating function for the negative binomial distribution, which has a derivative of

$$\frac{\partial \phi_N(z)}{\partial z} = \frac{pr}{1-pz}(\frac{p-1}{pz-1})^r.$$

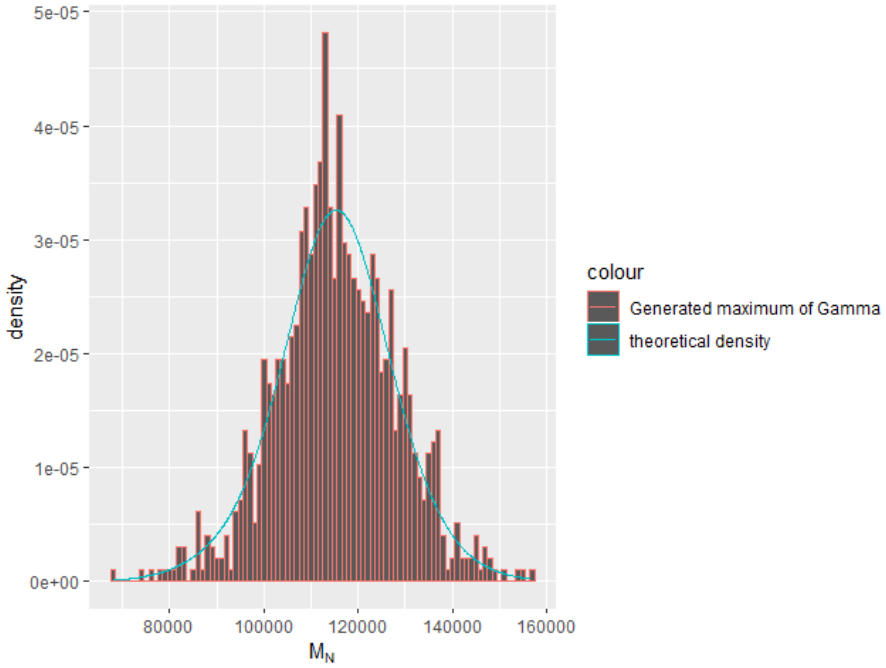

Figure 10: Drawings of maximum of Gamma variables

4.4 Likelihood ratio inference

If we have data points $x_1, .., x_n$, i.i.d. realizations from a statistical model with parameters $\theta = (\theta_1, .., \theta_d)$, we may denote the log likelihood function as $l(\theta)$, and the maximum likelihood estimator as $\hat{\theta}_{ml}$. One common source of inference for the parameters is the deviance function:

$$D(\theta) = -2(l(\theta) - l(\hat{\theta}_{ml})).$$

For a parameter estimate $\hat{\theta}$, the deviance function $D(\hat{\theta})$ is a measure of how much less likely this parameter estimate is in relation to the maximum likelihood estimator. If the true parameter of the model is θ_0, it is possible to show that under suitable regularity conditions, for large values of n, the deviance of the true parameter has an approximate distribution

$$D(\theta_0) \sim \chi^2(d).$$

This property can be used to create a $100 \cdot (1 - \alpha)$ % confidence region for θ by

$$C_\alpha(\theta) : \{\theta : D(\theta) < \chi^2_{1-\alpha}(d)\}.$$

where $\chi^2_{1-\alpha}(d)$ denotes the $(1 - \alpha)$ quantile of the "Chi Squared" distribution with d degrees of freedom.

[Coles et al., 2001, p.35] When the number of parameters, d, is large, this χ^2 approximation becomes less accurate.

Sometimes we interested in one particular parameter of the model, $\theta_k, k \in \{1, .., d\}$. We may denote the other parameters as $\theta_{-k} = \{\theta_i, i \neq k\}$, giving two components $\theta = (\theta_k, \theta_{-k})$. To make inference for θ_k, we may use the profile log likelihood function, which is defined as

$$l_p(\theta_k) = \max_{\theta_{-k}} l(\theta_k, \theta_{-k}).$$

For a given value of θ_k, the likelihood is optimized with respect to every other parameter θ_{-k}, and the profile log likelihood is then the highest log likelihood, conditioned on a parameter value for θ_k. If the true parameter is $\theta_{0,k}$, it is possible to show that for large values of n, we have an approximation

$$2 \cdot (l(\hat{\theta}_{ml}) - l_p(\theta_{0,k})) \sim \chi^2(1),$$

and this is used to create a $100 \cdot (1 - \alpha)$ % confidence interval for θ_k by

$$C_\alpha(\theta_k) = \{\theta_k : 2 \cdot (l(\hat{\theta}_{ml}) - l_p(\theta_k)) < \chi^2_{1-\alpha}(1)\}.$$

The inequality can also be rewritten as $l_p(\theta_k) > l(\hat{\theta}_{ml}) - \chi^2_{1-\alpha}(1)/2$. Figure 11 shows an example of how the profile log likelihood for a parameter could look like, where the dark horizontal line is equal to $l(\hat{\theta}_{ml}) - \chi^2_{0.95}(1)/2$, and the blue line is the 95 % confidence interval for θ_k, the values for which $l_p(\theta_k) > l(\hat{\theta}_{ml}) - \chi^2_{0.95}(1)/2$.

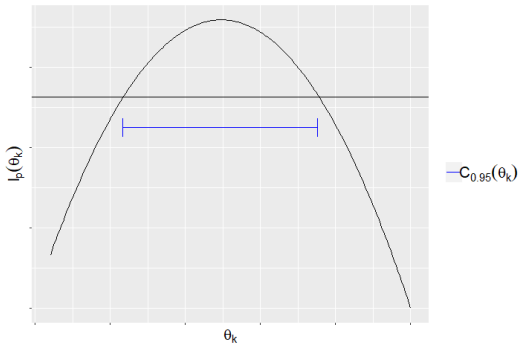

Figure 11: Profile log likelihood example

We may also use the likelihood ratio for some combination of the parameters, such as the quantiles, which for the Generalized Pareto distribution and a probability level p can be written as

$$q_p = F^{-1}(p) = \frac{\sigma}{\xi}((1-p)^{-\xi} - 1).$$

For a given he value of p, the quantile itself can be treated as a model parameter, and we may rewrite the relationship between the quantile and the other parameters as

$$\sigma = \frac{q_p \xi}{((1-p)^{-\xi} - 1)}.$$

We may then obtain the profile likelihood for quantiles by considering different choices for q_p, inserting the above expression for σ back into the Generalized Pareto log likelihood, and optimizing the likelihood with respect to ξ.

In [Coles et al., 2001], as an alternative to Wald based intervals, they argue a likelihood ratio based confidence interval method may be preferred when attempting to estimate confidence intervals for quantiles of the generalized Pareto distribution. The reason being that this likelihood ratio approach often give confidence intervals that are not symmetrical around the maximum likelihood estimator, instead giving a skewed interval that is often a better representation of the uncertainty of extreme quantiles.

4.5 Bootstrap methods

Having observed an independent sample $x_1, .., x_n$ from an unknown distribution X with $F(x) = P(X \leq x)$, the Bootstrap method aims to make inference on a population characteristic θ, which can be expressed as a parameter. If we have an estimator of this parameter $\hat{\theta}$ as a function of the data, we are often interested in how certain this estimate is. If we could draw new samples from the true probability distribution F, each of size n, we could potentially learn about distribution properties of the estimator $\hat{\theta}$, such as its standard deviation. In reality, we have limited data, and do not know the true distribution F. Instead, Bootstrap methods proposes using an approximation, $\hat{F}$, of the true distribution based on the data. By drawing new samples from $\hat{F}$ (each of size n usually), B number of times, and in each drawing calculating $\hat{\theta}$, we would obtain a Bootstrap generated set of estimators

$$\hat{\theta}^{(1)}, .., \hat{\theta}^{(B)}.$$

This is then used as an approximation of the sampling distribution of the estimator $\hat{\theta}$. In this procedure, there are two general ways of approximating F, which brings us to parametric versus non-parametric Bootstrap. In a non-parametric Bootstrap, we would assume equal probability to each observed value, $\frac{1}{n}$, and use the empirical distribution of the data as an approximation $\hat{F}$, where the empirical distribution for the data is:

$$\hat{F}(x) = \frac{\sum_{i=1}^{n} I(x_i \leq x)}{n}$$

$$\text{where } I(x_i < x) = \begin{cases} 1, & \text{if } x_i \leq x \\ 0, & \text{otherwise} \end{cases}$$

From the strong law of large numbers, the empirical CDF estimator converges towards the theoretical CDF as the number of samples goes to infinity

$$\lim_{n\to\infty} \hat{F}(x) \to F(x) \quad \forall x.$$

We can generate a Bootstrap sample by repeatedly drawing values randomly from the observed data points, with replacement, and calculating $\hat{\theta}$ for each resample.

In a parametric Bootstrap, we assume a parametric model for the data and estimate the parameters based on the original data. We then repeatedly draw new samples from this parametric distribution, and calculate $\hat{\theta}$ for each drawing. Whether using parametric or non-parametric Bootstrap, the end result is meant to be the same, a sample $\hat{\theta}^{(1)}, .., \hat{\theta}^{(B)}$, assumed to reflect the distribution of the estimator.

If we would like to estimate a confidence interval for θ using Bootstrap, one common and simple method is to directly use the quantiles of $\hat{\theta}^{(1)}, .., \hat{\theta}^{(B)}$, which is called the percentile method. For example, if we wanted a two sided 95% confidence interval, and had $B = 1000$ Bootstrap samples, we would take the 25_{th} and the 975_{th} smallest value of $\hat{\theta}^{(1)}, .., \hat{\theta}^{(B)}$ as end points of the confidence interval. A more detailed description of Bootstrap is given in [Efron and Tibshirani, 1986].

4.6 Bayesian models

When estimating parameters for a statistical model, maximum likelihood techniques aims to find the parameters that give the highest probability to the observed data. However, sometimes it is useful to consider external information beyond the data set to further reduce uncertainty. One way to put some weight to external information in a statistical model, is by using a Bayesian model, which assumes a prior distribution for the parameters of the model. Unlike maximum likelihood, where we usually assume θ (the "true parameter" of the model) to be an unknown constant, a Bayesian model views θ as a random variable. When we have additional information we would like to take into account, it is sometimes useful to incorporate this information into the prior distribution. The general Bayesian philosophy is to choose the prior for a parameter θ based on your belief about θ before observing the data, commonly articulated using a distributional family for convenience. This prior might for example be based on initial constraints for the data regarding quantiles or the mean.

An vital part of Bayesian theory has to do with proportionality. For instance, let us say $f(x)$ is a probability density function, and there is another function

$$g(x) = kf(x) \propto f(x) \; \forall x$$

where $k \neq 0$ is a constant, and the $\propto$ sign signifies that the functions are proportional to each other. If we then know that a random variable Y, has a density

function proportional to $g(x)$, then Y will follow the probability distribution uniquely defined by $f(x)$.

To describe Bayesian models, we will mainly use the theory of
[Carlin and Louis, 2008, Ch. 2-3]. If you have observed a set of i.i.d. data points $x = (x_1, .., x_n)^{'}$ from a continuous distribution with parameters $\theta = (\theta_1, .., \theta_p)^{'}$, and likelihood $p(x|\theta)$, a Bayesian model assumes a prior distribution $p(\theta)$ for the unknown parameters. Inference is then made using what is known as the posterior distribution, which, using Bayes formula can be expressed as

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta).$$

The posterior distribution can be interpreted as the distribution of the unknown parameters after taking the observed data into account, which becomes the main source of inference in a Bayesian model. The denominator, $p(x)$, is the marginal distribution of the data, which for a given data set is seen as a normalizing constant since it does not change with θ. We then say that the posterior distribution is proportional to the likelihood and the prior of the parameters.

If we denote Θ as the space of possible values for θ, and a future data point as $\tilde{x}$, predictions for a future value $\tilde{x}$ is made by using the posterior predictive distribution

$$p(\tilde{x}|x, \theta) = \int_{\Theta} p(\tilde{x}|\theta)p(\theta|x)d\theta.$$

If however, we want to make point estimates for the parameters, one could use the posterior mean

$$E[\theta|x] = \int_{\Theta} \theta p(\theta|x)d\theta.$$

Another choice is the posterior mode, which can be seen as the Bayesian equivalent to the maximum likelihood estimate, being that it optimizes the posterior:

$$\hat{\theta}_{mode} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(\theta|x) = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(x|\theta)p(\theta)$$

When comparing the posterior mean versus mode, neither of them paints a full picture of what parameter to use. Instead, it is recommended to inspect the full posterior distribution when considering different parameters.

One important aspect of a Bayesian model is to consider how much weight the prior gets in the model, sometimes expressed in terms of data-prior dominance. The likelihood function scales with the number of data points, meaning that the likelihood will have more influence as the number of data points grows, resulting in the prior having less effect on the posterior distribution. In an extreme value model, there is usually a low sample size for the likelihood, such that the data will sometimes fail to "dominate the prior". If we choose a restrictive prior for a parameter, this may then have an extremely large effect on the posterior.

If there is too much influence from the prior, it means putting more weight to (sometimes highly subjective) judgements, where as we generally want the data to "speak for itself" in order to make the most reliable inference. One way to make sure the prior does not get too much weight is by choosing a so called "vague" prior, which does not give a large preference to a certain region of the parameter space, thus remaining more impartial before observing data. For example, a uniform prior will have constant density, meaning we regard each possible value of the parameters as equally likely before observing data.

4.6.1 Markov chain Monte Carlo sampling

Say we have a Bayesian model with a parameter θ, with posterior density $p(\theta|y)$. A common case is that the posterior has an unrecognizable distribution we do not know how to integrate, and a common approach is then to use Markov Chain Monte Carlo (MCMC) methods in order to sample from $p(\theta|y)$. This is achieved by creating a sequence $\theta^{(1)}, \theta^{(2)}, ..$ that will converge to a sample from the posterior distribution $p(\theta|y)$, once enough iterations have been created. An MCMC chain will have the following Markovian property

$$p(\theta^{(t+1)}|\theta^{(t)}, \theta^{(t-1)}, .., \theta^{(1)}) = p(\theta^{(t+1)}|\theta^{(t)}).$$

This property means given the current state of the process $\theta^{(t)}$, the probability distribution of the next state of the process $\theta^{(t+1)}$, is independent of any past states. To create such a process that will converge to the posterior distribution of interest, there are a number of algorithms that can be used. One of these algorithms is called Slice sampling, which we will use in this thesis.

We will start by explaining the Slice sampling method in the one parameter case $\dim(\theta) = 1$. To use this method, we need to be able to evaluate the posterior density up to a proportional constant,

$$p(\theta|x) \propto h(\theta) = p(x|\theta) \cdot p(\theta).$$

For simplicity in notation, the proportional density $h()$ is written only as a function of θ, as the data, x, is the same throughout the process. The idea of a slice sampler is to use an auxiliary variable U to create an extended joint target distribution

$$p(\theta, U) \propto \begin{cases} 1, 0 < U < h(\theta) \\ 0, \text{otherwise} \end{cases}$$

If this can be achieved, the marginal distribution for θ (although technically still conditioned on the data), will be

$$\int_0^{h(\theta} p(\theta, u)du \propto h(\theta)$$

which is the posterior we are interested in sampling from. If for a given value of θ, we draw a uniform variable $U|\theta \sim \text{unif}(0, h(\theta))$, this uniform variable will have conditional density

$$p(u|\theta) = \frac{1}{h(\theta)} \cdot I(0 < u < h(\theta)).$$

Now the joint distribution of $p(\theta, U)$ is proportional to

$$p(\theta, u) \propto p(u|\theta) \cdot h(\theta) \propto 1 \cdot I(0 < u < h(\theta)).$$

The conditional density of $\theta|U$ is then proportional to the joint density above, meaning

$$p(\theta|u) = \frac{p(\theta, u)}{p(u)} \propto p(\theta, u) \propto I(0 < u < h(\theta)).$$

This is true since for a fixed value of u, the value $p(u)$ does not depend on θ. This means the conditional variable of $\theta|u$ is uniform on the space $S_U = \{\theta : 0 < U < h(\theta)\}$. In [Neal, 2003], they discuss how to find the region S_U in more detail.

To generate values from $p(\theta, U|y)$, we can start with an initial value $\theta^{(0)}$, and generate a sequence, where in step $t = 1, .., T$, we

$$\begin{cases} \text{step 1: Draw } U|\theta^{(t)} \sim \text{unif}(0, h(\theta^{(t)})), \\ \text{step 2: Draw } \theta^*|U \sim \text{unif}(\theta : U \leq h(\theta)) \\ \text{step 3: set } \theta^{(t+1)} = \theta^* \end{cases}$$

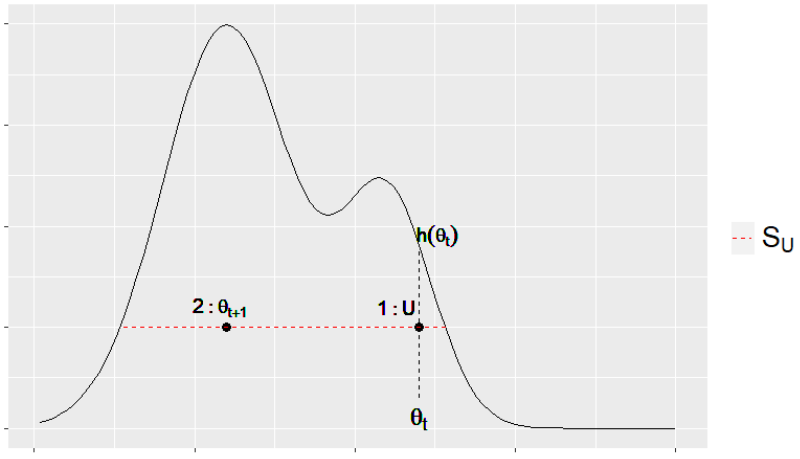

Figure 12: Illustration of Slice sampling iteration steps, starting with θ_t

If we have two parameters θ_1, θ_2, the Slice sampling algorithm can be used in a similar fashion, with the each iteration of the algorithm now being

$$\text{step 1:} \begin{cases} \text{Draw } U|\theta_1^{(t)}, \theta_2^{(t)} \sim \text{unif}(0, h(\theta_1^{(t)}, \theta_2^{(t)})), \\ \text{Draw } \theta_1^{(t+1)} \text{ from } \theta_1^*|U \sim \text{unif}(\theta_1 : U \leq h(\theta_1^{(t)}, \theta_2^{(t)})) \end{cases}$$

$$\text{step 2:} \begin{cases} \text{Draw } U|\theta_1^{(t+1)}, \theta_2^{(t)} \sim \text{unif}(0, h(\theta_1^{(t+1)}, \theta_2^{(t)}), \\ \text{Draw } \theta_2^{(t+1)} \text{ from } \theta_2^*|U \sim \text{unif}(\theta_2 : U \leq h(\theta_1^{(t+1)}, \theta_2^{(t)}) \end{cases}$$

4.6.2 Convergence diagnostics for MCMC algorithms

When running an MCMC chain such as the Slice sampling method to evaluate the posterior distribution, it is standard practice to have a so called "burn-in" period, which assumes that the values produced by the chain after the burn in period has converged to the stationary distribution. In this context, we use a more practical and slightly vague definition by saying that an MCMC algorithm has converged at time t_1 if the samples produced beyond time t_1 can be safely assumed to come from the stationary distribution. We may check if there are signs of failure to converge by running several parallel MCMC chains, and assessing (for example, visually) whether each chain is stable in terms of mean and variance over time, and whether the different chains have the same distribution.

One popular method is called the Gelman-Rubin statistic, introduced in [Gelman et al., 1992]. This method assesses convergence by running several parallel chains, each with different initial values, discarding the samples from the selected burn in period, and comparing the within chain variance and between chain for each parameter sampled. For a parameter θ, we denote the sample mean and variance of chain m as $\bar{\theta}_m$ and $\hat{\sigma}_m^2$ respectively, and the mean of all chain values as $\bar{\theta} = \frac{1}{m}\sum_{m=1}^{M} \bar{\theta}_m$. We then measure

$$B = \frac{N}{M-1}\sum_{i=1}^{M}(\bar{\theta}_m - \bar{\theta})^2$$

$$W = \frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}_m^2.$$

where B is known as the between chain variance, and W is the average within chain variance. The final statistic of interest is called the scale reduction factor, which can be written as

$$\sqrt{R} = \sqrt{\frac{\frac{N-1}{N}W + \frac{M+1}{M \cdot N}B}{W} \cdot \frac{df}{(df-2)}}$$

$$= \sqrt{\frac{N-1}{N} + \frac{M+1}{M \cdot N}\frac{B}{W} \cdot \frac{df}{(df-2)}}$$

where *df* is the degrees of freedom from a t distribution fit to the posterior of θ. If $\sqrt{R}$ is far from 1, it is a sign that the MCMC algorithm has yet to converge. It should however be noted that this test can fail to detect convergence failures in many cases.

4.6.3 Bayesian credibility interval for parameters

The posterior of a Bayesian model can be used in order to estimate the probability that a parameter falls within a certain interval. A $100 \cdot (1-\alpha)$ % credible set for a parameter θ can be defined as a subset C_α of possible values for θ for which,

$$(1-\alpha) = \int_{C_\alpha} p(\theta|y)d\theta.$$

The probability of θ being in the credible set C_α, given the observed data, is then α. If we have a model with p parameters $\theta = (\theta_1, .., \theta_p)$ and have generated a posterior sample of size m , $\theta^{(1)}, .., \theta^{(m)}$, we may estimate a $100 \cdot (1-\alpha)$ % credibility interval for parameter i, θ_i, by the observed $\alpha/2$ and $(1-\alpha/2)$ quantile of the posterior sample $\theta_i^{(1)}, .., \theta_i^{(m)}$.

This method can also be extended to another statistic that is a function of the parameters. For example, if the level p quantile of the model distribution q_p is some function of the parameters, we may estimate the posterior distribution of the quantile as $q_p^{(1)}, .., q_p^{(m)}$ by inserting the expression of the quantile for each value of θ in the chain. We may then construct a credibility interval for q_p using the same method. We can also use the mean of this sample as an estimate of the posterior mean for q_p.

5 Model

5.1 Threshold selection

The first step in our model is to select a threshold for our Peaks over threshold model. We will use two visual approaches in selecting the threshold value. In figure 13, we plot estimated parameters for ξ and $\sigma_0 = \sigma + \xi u$ against different choices of thresholds, u, which should be stable around an appropriate choice of threshold. We see that both ξ and σ_0 appear mostly stable around a threshold value of 30 000.

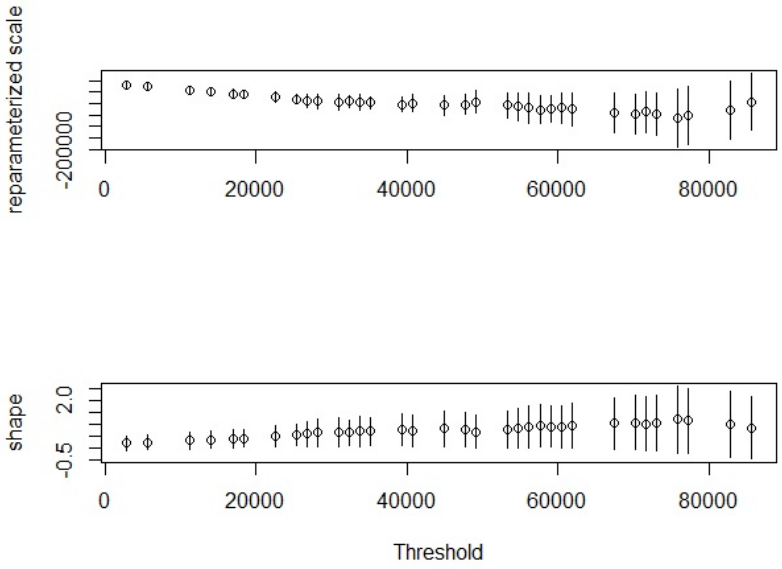

Figure 13: parameter stability plot

Next, we plot the estimated quantiles $H_Y^{-1}(p)$, for probability level $p = 0.9$, against threshold choices, which should be linear in the threshold, with some random variation. This is shown in figure 14, where the vertical line represents our final choice of threshold, which was 28570. Choosing this threshold left us with 29 observations (or peaks) over the threshold.

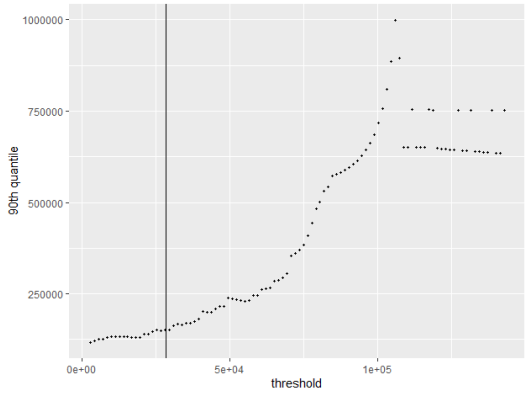

Figure 14: quantile stability plot

5.2 Storm frequency

We assume in our model that the number of storms per year that exceed the threshold follows a Poisson distribution.

$$N_u \sim po(\lambda)$$
$$P(N_u = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, ..$$

Having observed only 29 storms exceeding the threshold in a 41 year period, it can be difficult to check whether the model is appropriate. We tested for trends in the number of large storms per year using a log linear Poisson GLM, which can be expressed as $\log(E[N_u]) = \beta_0 + \beta t$, where t is the year observed. The p-value for testing $H_0 : \beta_1 = 0$ was 0.73, meaning we see no evidence of a trend. Below, in figure 15, we display the cumulative number of large storms over time, and compare it to the expected number of storms according to the Poisson assumption, which should increase linearly.

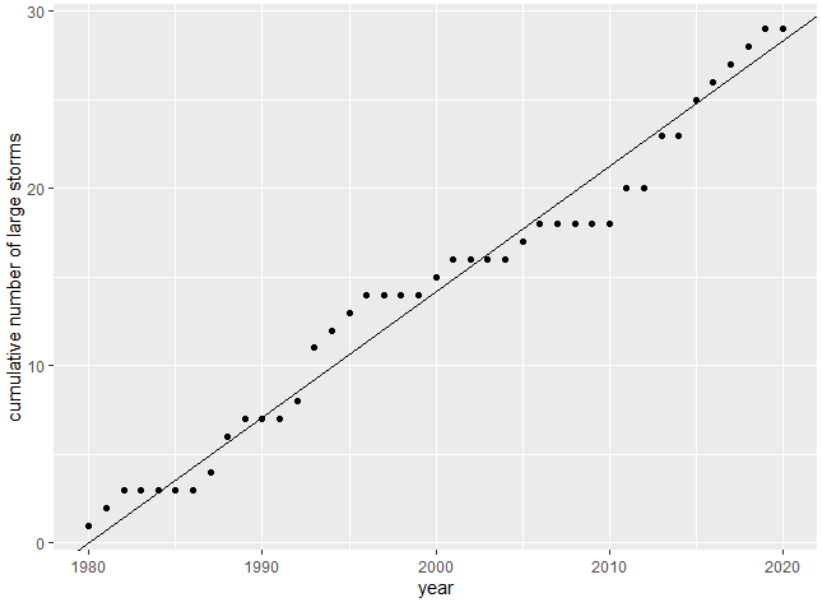

Figure 15: cumulative number of large storms over time

The cumulative number of large storms seem to follow this line pretty well, indicating that our Poisson model could be a decent assumption. For a Poisson distribution, maximum likelihood estimator of the rate is the mean. The maximum likelihood estimate for the rate of this Poisson distribution is the average number of large storms per year over time, which was

$$\hat{\lambda} = \frac{29}{41} \approx 0.71.$$

A Poisson model assumes that the mean equals the variance, and for our data, the number of large storms per year showed a sample variance of 0.6621, which is fairly close to the sample mean.

In figure 16, we show what months the 29 largest storms occurred, where we see that almost all wind storms occurred during the winter half of the year. This trend means the homogeneous Poisson assumption we make is not accurate on a monthly scale, as winter months would have much higher frequency than summer months. However, this distinction becomes less important to consider when aggregating the data to a yearly basis.

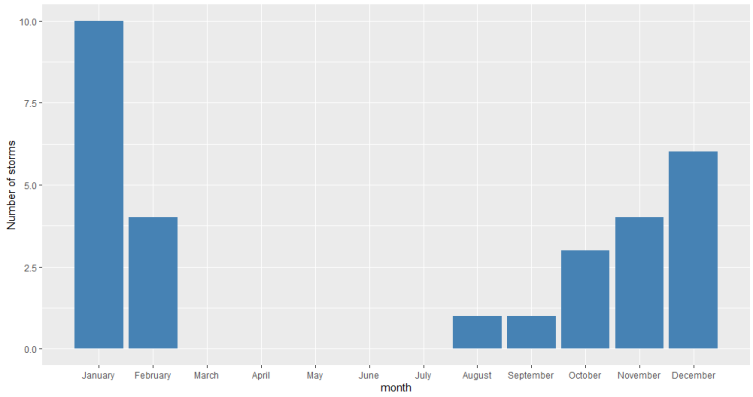

Figure 16: Number of historical storms by month

5.3 Quantiles and Return levels

In our model, we will assume that each year, N_u, the number of storm events where the loss is higher than our threshold u, follows a Poisson distribution, $N_u \sim po(\lambda)$. If X is the distribution of a storm loss, we will assume the conditional distribution $Y = X - u | X > u \sim GP(\xi, \sigma)$, and that the excess losses each year $Y_1, .., Y_{N_u}$ are independent. We are interested in estimating the quantiles of the conditional storm loss distribution $X|X > u$, which for a probability level p, can be obtained by shifting the assumed Generalized Pareto distribution by the threshold u, meaning

$$q_p = u + H^{-1}(p) = u + \frac{\sigma}{\xi}((1-p)^{-\xi} - 1). \tag{11}$$

For example, $q_{0.5}$ would answer what is the median loss, given that the loss has exceeded the threshold u.

If we denote $M_{N_u} = \max(Y_1 + u, .., Y_{N_u} + u)$ as the annual maximum storm loss, we are also interested in the T year return level for the annual maximum loss, which can be written as

$$z_{1-1/T} = \{x : P(M_{N_u} > x) = 1/T\}.$$

For example, $z_{1-1/100}$, the 100 year return level would be the value for which the annual maximum loss has a 1% probability of exceeding. Technically, our definition of the annual maximum loss has failed to include losses below the threshold. However, as we are only concerned with modeling sufficiently costly scenarios, the distribution of the losses below the threshold become less important to the analysis. To find an expression for the yearly return level, we can use the probability generating function of the Poisson distribution

$$\phi_{N_u}(z) = E[z^{N_u}] = exp(\lambda(z-1))$$

Putting the expression above and the Generalized Pareto distribution function into equation (10), we obtain

$$\begin{aligned} &F_{M_{N_u}}(y+u) = P(M_{N_u} - u < y) = exp(\lambda(H(y)-1)) \\ &= exp(\lambda(1-(1+\xi\frac{y}{\sigma})^{-1/\xi}-1)) = exp(-\lambda(1+\xi\frac{y}{\sigma})^{-1/\xi}). \end{aligned} \tag{12}$$

The expression in equation (12) can also be rewritten as a CDF of an extreme value distribution, as shown in [Rootzén and Tajvidi, 1997, p.76].

To get the inverse of the probability function function for M_{N_u}, a few calculation steps can be used

$$\begin{aligned} F_{M_{N_u}}(y) &= exp(-\lambda(1+\xi\frac{y-u}{\sigma})^{-1/\xi}) \\ log(F_{M_{N_u}}(y)) &= -\lambda(1+\xi\frac{y-u}{\sigma})^{-1/\xi} \\ \frac{-log(F_{M_{N_u}}(y))}{\lambda} &= (1+\xi\frac{y+u}{\sigma})^{-1/\xi} \\ (\frac{-log(F_{M_{N_u}}(y))}{\lambda})^{-\xi} &= (1+\xi\frac{y-u}{\sigma}) \\ \frac{\sigma}{\xi}((\frac{-log(F_{M_{N_u}}(y))}{\lambda})^{-\xi}-1) &= y-u \end{aligned}$$

The level p quantiles of the annual maximum loss will be

$$z_p = F^{-1}_{M_{N_u}}(p) = u + \frac{\sigma}{\xi}((\frac{-log(p)}{\lambda})^{-\xi}-1). \tag{13}$$

If we want the T year return level, which is the inverse of the annual maximum exceedance probability, this is equal to

$$z_{1-1/T} = u + \frac{\sigma}{\xi}((\frac{-log(1/(1-T))}{\lambda})^{-\xi}-1). \tag{14}$$

The yearly return level in this model has three unknown parameters, λ, ξ, σ, and for a given estimate of these and time length T, we can plug these into equation 14 as an estimate of the T year return level.

5.4 Likelihood ratio intervals

For the Generalize Pareto distribution, we have two unknown parameters, and can achieve a 2 dimensional joint 95% confidence region, by considering each set of parameters where

$$C_{\xi,\sigma} = \{\xi, \sigma; D(\xi, \sigma) < \chi^2_{0.95}(2) \approx 5.99\}.$$

To find this region, we will use a simple numerical approach. We will consider a large 2 dimensional grid of possible values for (ξ, σ), with each coordinate being equally spaced out. The step size for the distance between coordinates were 0.0075 for ξ, and 0.5 % of the maximum likelihood estimator for σ. We then measure the log likelihood for each coordinate of (ξ, σ), and see which coordinates have

$$D(\xi, \sigma) < \chi^2_{0.95}(2),$$

setting this as the 95% confidence region for (ξ, σ).

We will also use profile likelihood functions to obtain 95% confidence intervals for ξ and σ, and quantiles for the storm losses q_p, and yearly return level $z_{1-1/T}$. For simplicity, we will treat the maximum likelihood estimator for the Poisson parameter λ as true, such that we can measure the profile likelihood for the T year return level by only using the Generalized Pareto likelihood.

As the T year return level in our model can be written as a function of λ, ξ, σ, we can rewrite the relation and solve the expression for the scale parameter σ.

$$\begin{aligned}
z_{1-1/T} &= u + \frac{\sigma}{\xi}((\frac{-\log(1/(1-T))}{\lambda})^{-\xi} - 1) \\
\frac{z_{1-1/T} - u}{\sigma} &= \frac{1}{\xi}((\frac{-\log(1/(1-T))}{\lambda})^{-\xi} - 1) \\
\sigma &= \frac{\xi(z_{1-1/T} - u)}{(\frac{-\log(1/(1-T))}{\lambda})^{-\xi} - 1}
\end{aligned}$$

For a given value of $z_{1-1/T}$, we can plug the above expression for σ into the log likelihood, and obtain the profile likelihood by optimizing with respect to ξ, as we will only consider $\hat{\lambda}_{ml}$ for the frequency parameter.

For the quantiles q_p, we will consider probability levels $p = 0.5, 0.9, 0.95, 0.99$, while for $z_{1-1/T}$, we will consider $T = 2, 3, .., 200$ yearly return periods.

5.5 Bootstrap

Another way we will judge model uncertainty is through a parametric Bootstrap. Using the maximum likelihood estimates from the Generalized Pareto distribution, we will sample new data from this model. Specifically, we will in step i of the process:

- draw 29 new excess losses $\tilde{y}_1, .., \tilde{y}_{29} \sim GP(\hat{\xi}_{ml}, \hat{\sigma}_{ml})$.

- re estimate the parameters of the Generalized Pareto distribution based on the new generated data to obtain $\theta^{(i)} = (\xi^{(i)}, \sigma^{(i)})$

We will run 50 000 steps of this process, to obtain 50000 sequences of parameter estimates $\theta^{(1)}, .., \theta^{(50000)}$, which will use as an approximation of the joint distribution of the parameter estimates of (ξ, σ). We will also insert the simulated values of ξ and σ into equation (11) as a Bootstrap sample of q_p, the level p quantile of the loss distribution $X|X > u$. The same will be done for the T year return level, where we insert the simulated parameters into equation (14), and always use the maximum likelihood estimator $\hat{\lambda}_{ml}$ for the Poisson rate parameter.

From the Bootstrap samples of $\xi, \sigma, q_p, z_{1-1/T}$, we will summarize these in terms of sample means, and 95% confidence intervals using the percentile method. Technically, the sample mean of the Bootstrap might be less interesting, since this Bootstrap method is generally meant to estimate uncertainty in parameters, however, this can still give some perspective to see how the mean of the Bootstrap compares to the confidence interval boundaries for each parameter.

As the sample size of our original data is only 29, there are significant limitations to accuracy of this method. The assumptions behind this method is that the maximum likelihood estimator is a reliable estimator for the true distribution, even though this estimator is known not to be very stable for this sample size, especially the shape parameter ξ. Despite this drawback, we deem the parametric Bootstrap to be more appropriate than a non-parametric Bootstrap, as the non-parametric Bootstrap would mean sampling thousands of times from only 29 data points of a very heavy tailed distribution, which is less reflective of the randomness we are trying to capture. One paper which explored Bootstrap methods for extreme value models was [Kyselỳ, 2008], which found that parametric bootstrap approaches were preferred to non parametric based on their simulated examples with sample sizes n=20, 40, 60, and 100. One of their main arguments being that the non-parametric Bootstrap gave too narrow confidence intervals for parameters and quantile estimates, thus underestimating the uncertainty in the model. This finding was also most notable for small sample sizes and heavy tailed distributions.

5.6 Bayesian model

As an alternative to maximum likelihood, we will consider a Bayesian estimation of the Generalized Pareto distribution to the storm loss excesses. For the prior distribution, we will first reparameterize to $\nu = \log(\sigma)$. Since the function $g(x) = \log(x)$ has a unique inverse for every $x > 0$, this is known as a one-to-one transformation, $\sigma \to \log(\sigma) = \nu$, such that the maximum likelihood estimator is invariant to the transformation $\hat{\nu}_{ml} = \log(\hat{\sigma}_{ml})$. The parameter transformation frees us to choose a prior on the full real axis $\nu \in [-\infty, \infty]$, and still follow the parameter restriction $exp(\nu) = \sigma \in [0, \infty]$.

To convert properties of the CAT model into a prior distribution, we will use a Bootstrap method. The catastrophe model brings a large set of simulated storm losses, let us denote these by $\tilde{x}_1, \tilde{x}_2, ...$ If there are m of these generated storms that exceeds our selected threshold, u, we can denote the simulated excess storm losses as $\tilde{y}_1, .., \tilde{y}_m$. Our prior specification procedure will be as follows.

- Draw a Bootstrap sample of size 29 from $\tilde{y}_1, .., \tilde{y}_m$, denoted $\tilde{y}^{(1)}, .., \tilde{y}^{(29)}$
- Fit a maximum likelihood GP distribution to the Bootstrap sample $\tilde{y}^{(1)}, .., \tilde{y}^{(29)}$, to obtain parameters ξ^*, σ^*
- Repeat this process 50 000 times to obtain Bootstrap samples of $\xi^{(1)}, .., \xi^{(50000)}$ and $\sigma^{(1)}, .., \sigma^{(50000)}$
- Fit a Gaussian distribution using maximum likelihood to

$$\xi^{(1)}, .., \xi^{(50000)} \sim N(\mu_\xi, \tau_\xi^2)$$
$$log(\sigma^{(1)}), .., log(\sigma^{(50000)}) = \nu^{(1)}, .., \nu^{(50000)} \sim N(\mu_\nu, \tau_\nu^2)$$

- Use this fit to define hyperparameters for the prior

$$\xi \sim N(\mu_\xi, \tau_\xi^2)$$
$$log(\sigma) = \nu \sim N(\mu_\nu, \tau_\nu^2)$$

The synthetic storm losses from the CAT model are meant to estimate the underlying distribution of potential losses that may occur. By drawing random resamples from this synthetic data, and fitting a Generalized Pareto distribution to each drawing to obtain Bootstrap samples of ξ and σ, we hope to capture the prior uncertainty in these parameters, without referencing the observed historical losses. By only drawing 29 points (same as the observed number of data) in each iteration, the Bootstrap samples will vary a lot, making the prior less restrictive.

In Appendix A.1, we go through the details of specifying the prior, as well as comparing the CAT model to historical losses, and how a Generalized Pareto distribution fits to the CAT model losses. Since $log(\sigma) \sim N(\mu_\nu, \tau_\nu^2)$, we can conclude that $\sigma \sim \text{LogN}(\mu_\nu, \tau_\nu^2)$, meaning the scale parameter follows a Lognormal distribution. We will for simplicity assume independence between ξ and σ, such that we will get a joint prior density of

$$\begin{aligned} p(\xi, \sigma) &= p(\xi) \cdot p(\sigma) \\ &= \frac{1}{\sqrt{2\pi\tau_\xi^2}} exp(-\frac{(\xi - \mu_\xi)^2}{2\tau_\xi^2}) \\ &\cdot \frac{1}{\sigma\sqrt{2\pi\tau_\nu^2}} exp(-\frac{(log(\sigma) - \mu_\nu)^2}{2\tau_\nu^2}). \end{aligned}$$

The joint posterior distribution will now have density proportional to

$$
\begin{aligned}
&p(\xi, \sigma|y) \propto p(y|\xi, \sigma)p(\xi, \sigma) \\
&= \prod_{i=1}^{29} \frac{1}{\sigma} \cdot (1 + \xi \frac{y_i}{\sigma})^{-(1+1/\xi)} \\
&\cdot \frac{1}{\sqrt{2\pi\tau_\xi^2}} exp(-\frac{(\xi - \mu_\xi)^2}{2\tau_\xi^2}) \\
&\cdot \frac{1}{\sigma\sqrt{2\pi\tau_\nu^2}} exp(-\frac{(log(\sigma) - \mu_\nu)^2}{2\tau_\nu^2}).
\end{aligned}
$$

As this joint posterior is not a recognisable distribution we know how to integrate and sample from, we will use the Slice sampling method to obtain posterior chains for ξ and σ. We will use a burn in period of 4000, and run 5 chains, each of size 10000 after the burn in period. The initial values $\xi^{(0)}, \sigma^{(0)}$ will be set to different parts of the parameter space for the five chains. This is implemented using an R package called Runjags.

5.6.1 Posterior estimates

From the Slice sampling, we receive a drawn sample of $(\theta^{(1)}, .., \theta^{(50000)})$, where $\theta^{(i)} = (\xi^{(i)}, \sigma^{(i)})$, this sample is an estimate of drawings from the posterior distribution $p(\xi, \sigma|x)$. Realizations from the posterior of a quantile q_p can be obtained by substituting simulated values of $\theta^{(i)}$ in equation (11).

If the Poisson rate parameter λ was known, we could also obtain posterior samples of the T year return level by substituting drawings of $\theta^{(i)}$ into equation (14). In our model, we will treat the maximum likelihood estimator for λ as true in order to do this, which is a convenient assumption used in order to estimate uncertainty in return levels caused by uncertainty in ξ and σ.

The estimated posterior samples of $\xi, \sigma, q_p, z_{1-1/T}$ will be summarized in terms of posterior means, and 95% credibility intervals using the percentile method.

5.6.2 Bayesian simulation example

To get a sense of how a Bayesian Generalized Pareto model fits to data when using Slice sampling, we have fitted the model to simulated data drawn from the Generalized Pareto distribution for a few different scenarios. We have varied the sample size by $n = 30, 60, 100$, and varied the shape parameter $\xi = 0.4, 0.7$, but kept the scale parameter $\sigma = exp(5)$ the same for each scenario, meaning we have 6 scenarios. For the MCMC simulations, we have used a burn in period of 4000, and generated 3 chains, each of size 10000. In each scenario, we have chosen the prior means equal to the true parameter. In some sense, this can be seen as the optimal circumstances to fit this model, where we know the model is correctly specified, with independent identically distributed data, and the prior

mean is equal to to be the actual parameter we are trying to estimate.

We have chosen Gaussian priors for ξ and $\nu = log(\sigma)$, and set the variance of ξ to $1/2$, and the prior variance for ν to 1, in each of the sample sizes. This way, we can see how the prior influence changes with sample size, and for two different choices of ξ.

In figure 17, we show the generated posterior samples when the true shape parameter is $\xi = 0.4$. We can clearly see how the perceived "vagueness" of the priors depends on the sample size, as the prior appears much flatter in the bottom figures, where $n = 100$ and the posterior has a more narrow distribution. We also see that for sample size $n = 30$, the posterior mode (the peak of the posterior) for ξ and ν looks almost like an equal trade off between the prior mean and the maximum likelihood estimator. However, the posterior mean of the shape parameter seems to be higher than the maximum likelihood estimator in each case. With sample size 100, the posterior mode is a lot closer to the maximum likelihood estimator, a consequence of the receding influence of the prior for higher sample sizes.

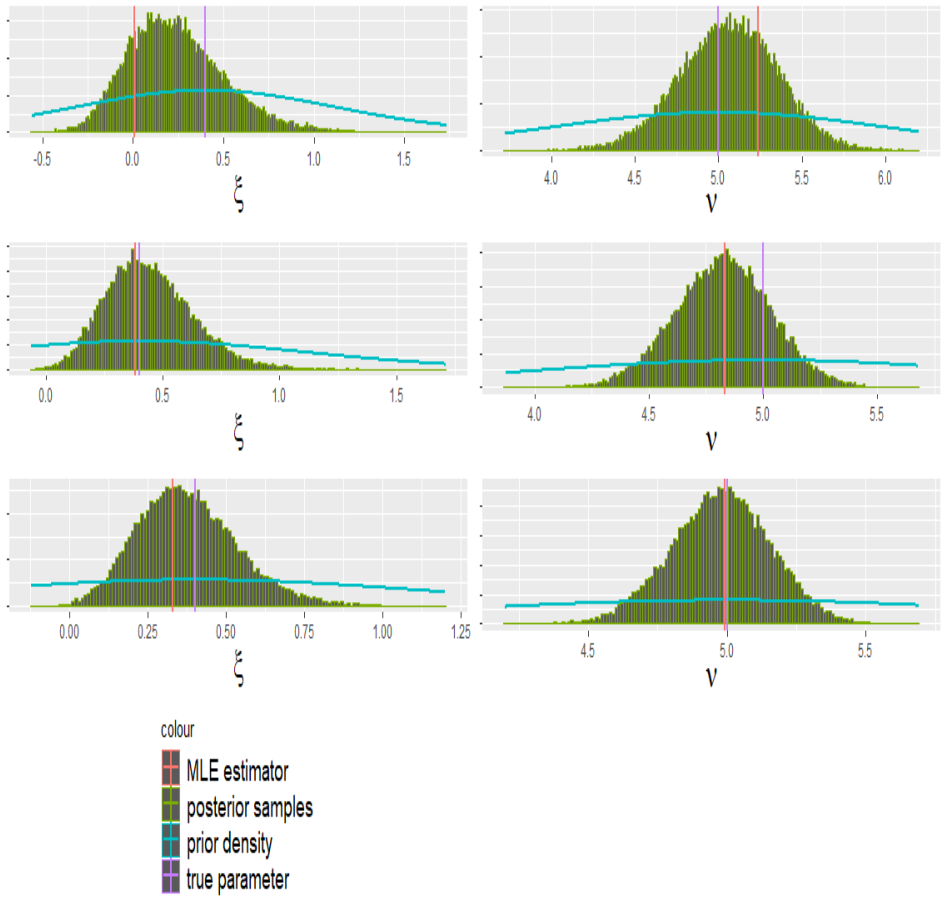

Figure 17: MCMC simulations $\xi = 0.4$, n=30,60,100

Next, we show the corresponding simulations for the more heavy tailed case where $\xi = 0.7$. We may note that in both figure 17 and 18, the posterior for ξ is noticeably asymmetrical, with the right tail being at least slightly heavier in each case.

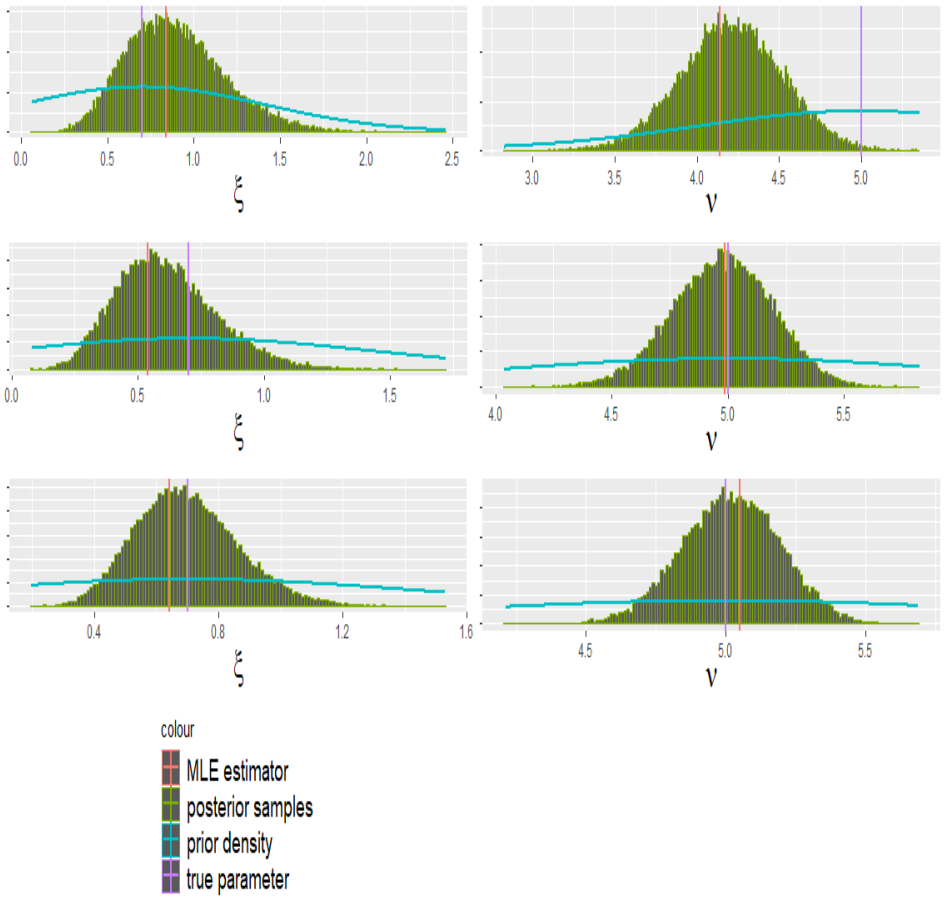

Figure 18: MCMC simulations $\xi = 0.7$, n=30,60,100

We may see how estimates of the r observation return level looks in each of the six scenarios. The r observational return level is the level that is expected to be exceeded once every r observations, which may be seen as the $1 - 1/r$ probability quantile. For the Generalized Pareto distribution, this is

$$x_r = \frac{\sigma}{\xi}((\frac{1}{r})^{-\xi} - 1).$$

In figure 19 and 20, we show the estimates of the r observation return levels by using the posterior means of the MCMC samples as point estimates for ξ and σ in each of the 6 simulated data scenarios. This is compared to using the maximum likelihood estimator and the true theoretical return levels. We see that in each of the 6 scenarios, the green curve is above the red curve, which means using the Bayesian posterior mean is more conservative than the maximum likelihood estimate of the return levels for these examples. This is

likely due to the posterior mean of ξ being higher than the maximum likelihood estimator in all cases. The theoretical return levels (blue curve) is also higher than the maximum likelihood estimator in each plot, most notably in the case $n = 30, \xi = 0.4$, indicating how the high parameter uncertainty at times can cause very naive estimates of return levels.

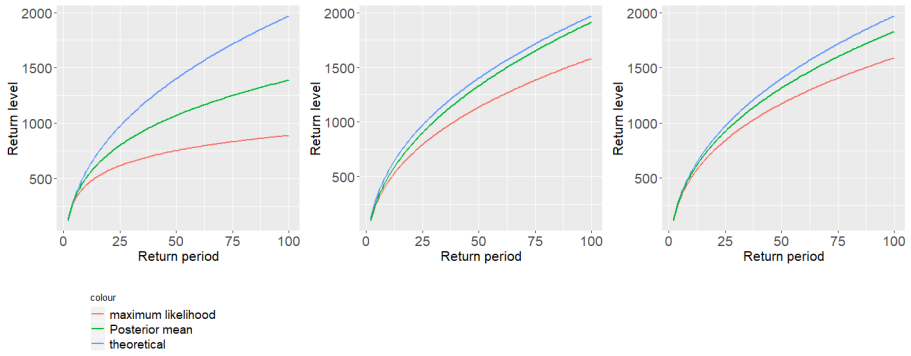

Figure 19: Return period estimates for $\xi = 0.4$, n=30,60,100

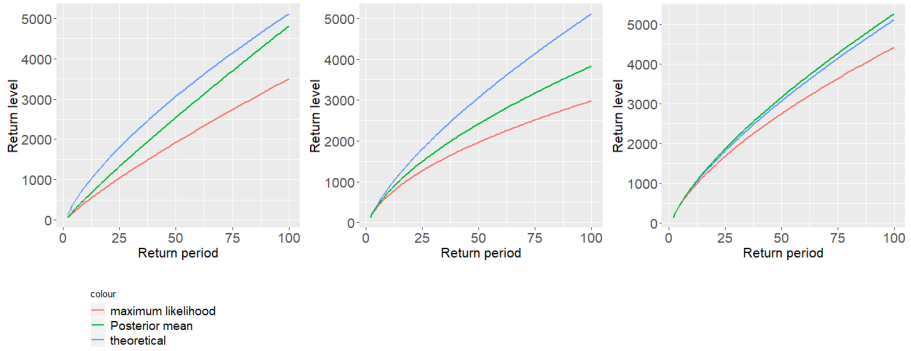

Figure 20: Return period estimates for $\xi = 0.7$, n=30,60,100

6 Results

6.1 parameter estimates

We will start by showing the parameter estimates for the Generalized Pareto distribution using the different methods. For the maximum likelihood estimator, the confidence interval shown was based on the χ^2 assumption of the profile likelihood. For the Bootstrap and Bayesian method, we had generated samples

for both ξ and σ, and will display the sample mean as the point estimate, and the empirical 2.5 % and 97.5 % quantile as interval boundaries.

point estimate	lower bound	upper bound	interval length	estimation
0.67	0.24	1.43	1.19	MLE
0.58	-0.02	1.12	1.14	Bootstrap
0.567	0.234	0.961	0.726	Bayesian

Table 1: Point estimates for ξ with 95% confidence (credibility) intervals

point estimate	lower bound	upper bound	interval length	estimation
27417	15130	53184	38054	MLE
31018	15276	56598	41322	Bootstrap
56627	35923	85099	49177	Bayesian

Table 2: Point estimates for σ with 95% confidence (credibility) intervals

In table 1, we see that the point estimates for ξ differ between the methods, as the Bootstrap and Bayesian are both lower than the maximum likelihood estimator. The interval lengths for ξ are highest when using the profile likelihood confidence interval, and lowest for the Bayesian credibility interval, where the upper bound is also lower than 1. In table 2, we see that the scale parameter from the Bayesian model is extremely high compared to the maximum likelihood, which was caused by the prior for $\nu = log(\sigma)$ having a much higher mean then the maximum likelihood estimator.

To assess how each estimation compares to historical data, we have plotted the empirical storm cost quantiles against the modeled quantiles from the Generalized Pareto distribution, using the point estimates from the three methods shown in tables 1 and 2. In figure 21, we see that the maximum likelihood and Bootstrap estimates are almost the same, while the Bayesian quantiles are generally higher than the empirical ones, with the exception of the two largest storms, which fall close to the line. This is because the posterior mean for σ was much higher than the maximum likelihood, resulting in higher quantile estimates.

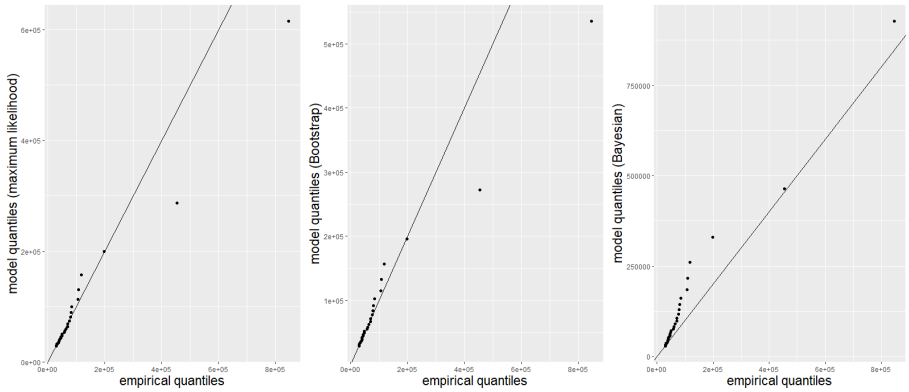

Figure 21: quantile-quantile plots

In figure 22, we show the parameter region for ξ and σ where the deviance $D(\xi, \sigma)$ is smaller then $\chi^2_{0.95}(2)$ as an approximation of a 95 % confidence region for the joint distribution of ξ and σ.

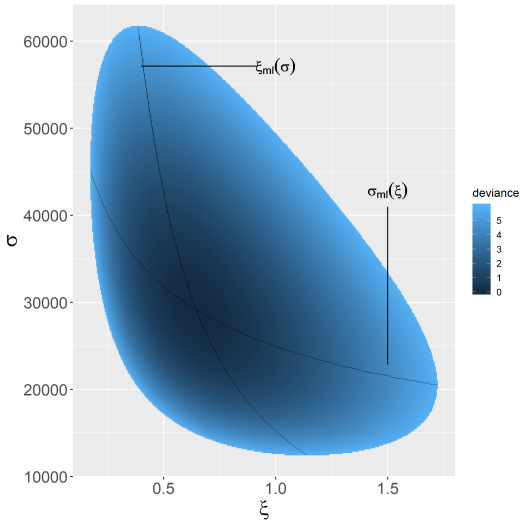

Figure 22: Likelihood ratio 95 % confidence region for GPD parameters

In figure 22, the colour is the deviance from the maximum likelihood $l(\hat{\sigma}_{ml}, \hat{\xi}_{ml})$, where a darker area means a lower deviance, and thus higher likelihood. We can note that $\xi > 0$ for all values in the confidence region, which emphasises the heavy tailed property of the loss distribution. The region is very large, which is because of the high parameter uncertainty from the low sample size. We have

also included the following two curves:

$$\sigma_{ml}(\xi) = \underset{\sigma}{\text{argmax}}\ l(\xi, \sigma)$$
$$\xi_{ml}(\sigma) = \underset{\xi}{\text{argmax}}\ l(\xi, \sigma)$$

The function $\sigma_{ml}(\xi)$ is for a given value of ξ, the value for σ with the highest likelihood, while $\xi_{ml}(\sigma)$ is the estimator of ξ with the highest likelihood, conditioned on σ. These two curves then cross at the optimal point for both parameters, the maximum likelihood estimator. Since the quantiles are increasing in both ξ and σ, there is a negative dependence between the parameter estimates, causing the negative slope of the two curves in the figure.

In figure 23, we show the estimated parameters generated from the Bootstrap for all 50 000 sequences. We can see that the estimates for σ seems to have a slight upward bias, while ξ seems to have a slight downward bias. And unlike the "deviance confidence region", this Bootstrap has values where $\xi \leq 0$.

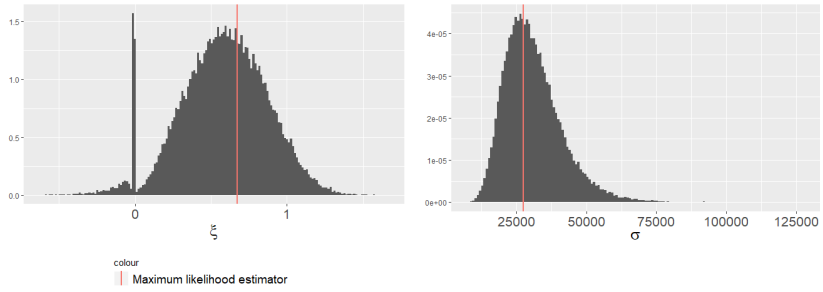

Figure 23: Bootstrap parameters

In figure 34 in Appendix A.2, we display the 5 MCMC chains for both parameter ξ and σ, and see that the mean and variance for each parameter appear stable, both within each chain and between the 5 chains. We have also displayed the autocorrelation function for each chain, which for a parameter θ estimates the correlation between $\theta^{(t)}$ and $\theta^{(t-l)}$ in the chain for lag $l = 1, 2, ...$ This is shown in figure 35, where we see that the first lag autocorrelation seems to vary around 0.25 to 0.3 for the different chains, for both parameters. The second lag autocorrelation is below (but still close to) 0.1 in most chains, with higher order correlations being close to 0. The Gelman-Rubin statistic gave scale reduction factors for ξ and σ respectively that were very close to 1. This means we do not find any clear evidence of convergence issues for the MCMC sampling.

In figure 24, we show the posterior samples for the Bayesian model when using Gaussian priors, with reparametrization $\nu = log(\sigma)$. There is a very large distance between the prior and the maximum likelihood for ν, where we see that the posterior is almost halfway in between the two. This difference becomes

much more extreme when considering $\sigma = exp(\nu)$. The shape parameter ξ on the other hand has a prior mean close to the maximum likelihood estimator, with lower variance than the corresponding Bootstrap samples for ξ. Although, unlike the previous Bayesian fits to simulated data, the posterior mean for ξ is below the maximum likelihood estimator and the prior mean, which may be a sign that the shape parameter is being "pushed downward" as a compensation for the scale parameter being restricted to larger estimates.

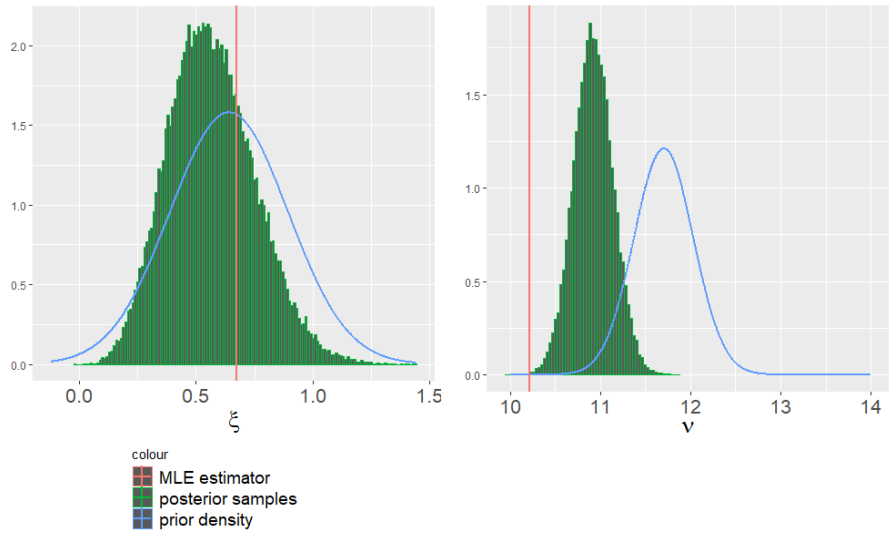

Figure 24: Posterior samples of $\xi, \nu = log(\sigma)$, when using Gaussian priors

6.2 Quantile uncertainty

In figure 25, we show as an example what the 99_{th} quantile of the storm loss distribution is for each of the parameters where the deviance, $D(\xi, \sigma)$, is smaller than the $\chi^2_{0.95}(1)$ threshold. By 99_{th} quantile, we mean $q_{0.99} = u + F_Y^{-1}(0.99)$, where Y denotes the GP distribution for the excesses losses.

We may notice that there is more contrast in colour as we move along the shape parameter ξ , then the scale parameter σ, signifying how the tail distribution is largely driven by ξ.

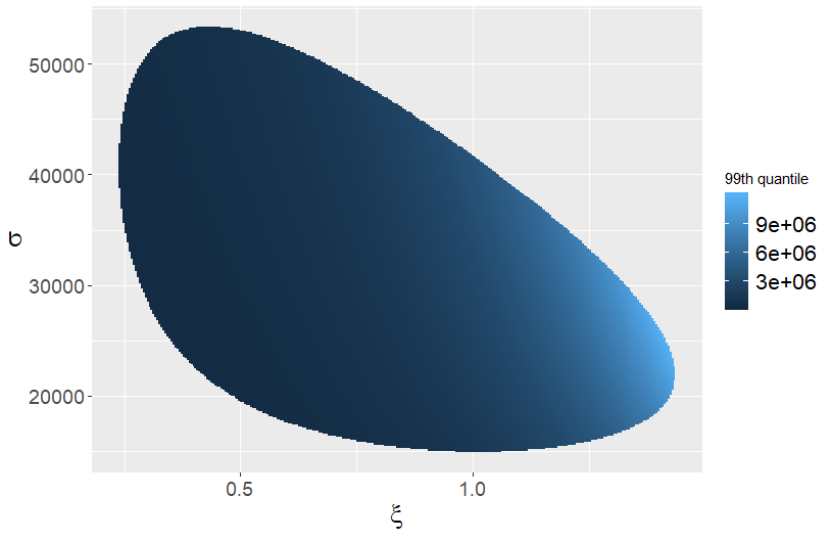

Figure 25: 99th quantile of storm loss distribution, for $D(\xi, \sigma) < \chi^2_{0.95} = 3.84$

In figure 26, we show the profile likelihood $l_p(q_{99})$ of the same quantile, where the blue line shows the 95 % confidence interval for q_{99}.

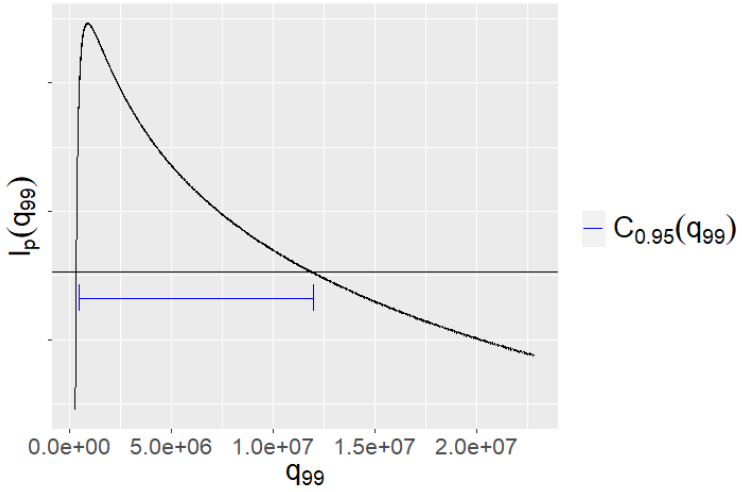

Figure 26: Profile log likelihood for 99th quantile of storm loss distribution

In figure 27, we compare quantile uncertainty for probability levels $p = 0.5, 0.9, 0.95, 0.99$ using 95% confidence intervals based on profile likelihood, Bootstrap, and 95% credibility intervals for the Bayesian model. The point estimates within each interval is the maximum likelihood estimate for the pro-

file likelihood, and the mean quantile for the Bootstrap and Bayesian plots. The upper left figure refers to estimates of the median loss among the losses that exceed the threshold. There we see that the Bootstrap and profile likelihood estimation is close to identical, while the Bayesian model has a higher median, a difference caused by the much higher scale parameter for the Bayesian model. For probability level 0.99, which can be interpreted as a once in a 100 observation storm loss, we see that the profile likelihood interval is much wider and skewed compared to the Bootstrap and Bayesian case. The reason for the Bayesian interval being much more narrow is that it rejects the extremely heavy tailed scenarios where $\xi > 1$ to a higher degree than the profile likelihood does.

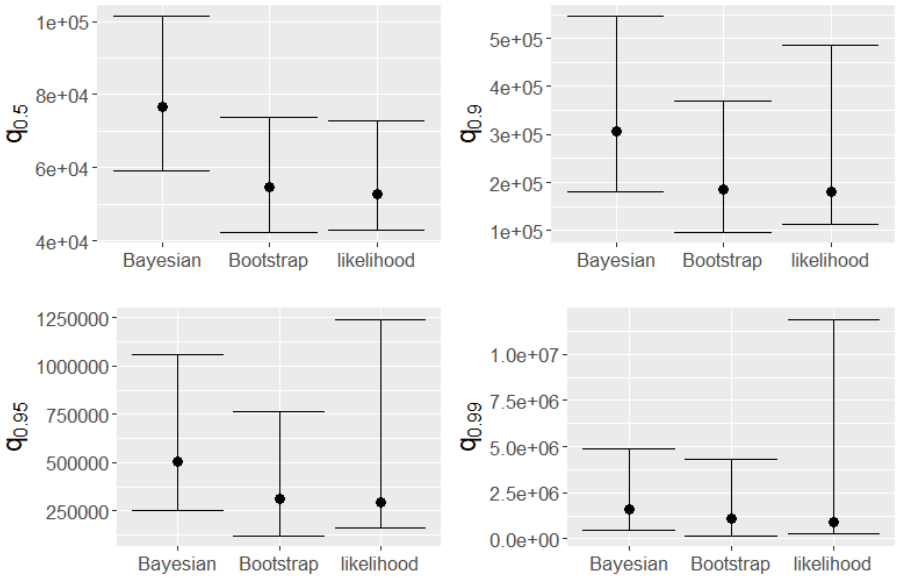

Figure 27: estimated quantiles of GP distribution with 95 % confidence (credibility) intervals

6.3 return period estimates

In figure 28, we show the estimated return levels, $z_{1-1/T}$ (for periods of T=2,3,..,200 years), where we have added a 95 % confidence interval based on the profile likelihood of the return level, and see that the upper bound of this interval is extremely large, as the maximum likelihood estimator is almost at the bottom of the interval. This is a result of the shape parameter ξ being very unstable, as we saw in the confidence interval for ξ in table 1.

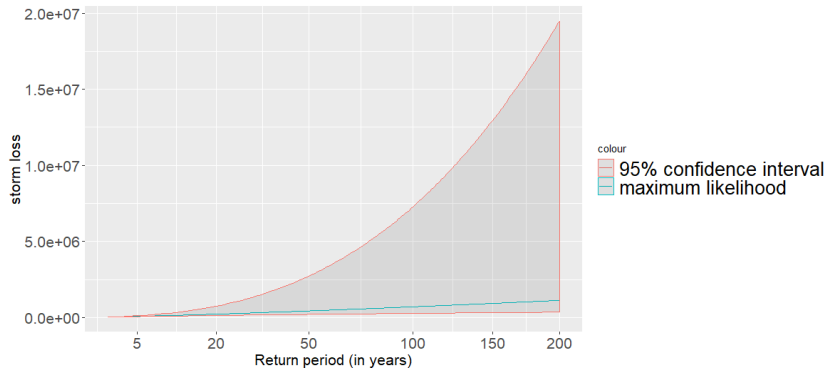

Figure 28: Return period with estimated 95% confidence interval based on profile likelihood

In figure 29, we show the return level estimates for storm losses based on the Bootstrapping of return levels, where the "Bootstrap mean" curve is the average return level for each given return period. For the 200 year return level, the upper bound of the confidence interval is roughly two and a half the size of the mean estimate, which is a lot more narrow than the profile likelihood approach, but still a fairly wide interval. The mean Bootstrap is close to the maximum likelihood return levels, with the maximum likelihood giving slightly lower return levels for high return periods.

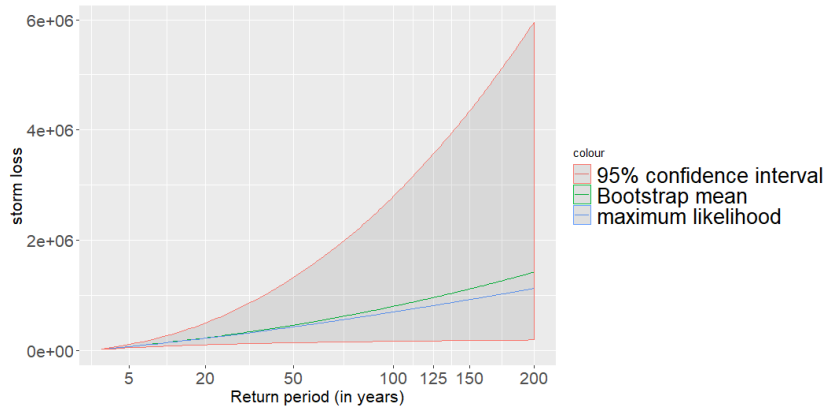

Figure 29: Return period Bootstrap with estimated 95% confidence interval

In figure 30, we compare return level estimates using the Bayesian model to ones estimated by maximum likelihood and the CAT model. We may observe that the Bayesian posterior mean for return levels is always in between the CAT model and the maximum likelihood estimate, and closer to the maximum likelihood. This is a result that might seem expected, given that the prior for

the Bayesian model is based on the CAT model, but this result is not always the case, as we saw in the return period plots for simulated data in figure 19 and 20. The 95% credibility interval is similar to the Bootstrap interval, and also much more narrow compared to the profile likelihood case, which is because the shape parameter has a much lower variance under the Bayesian model, and thus rejects the extremely heavy tailed scenarios with $\xi > 1$ that the profile likelihood method still deemed plausible.

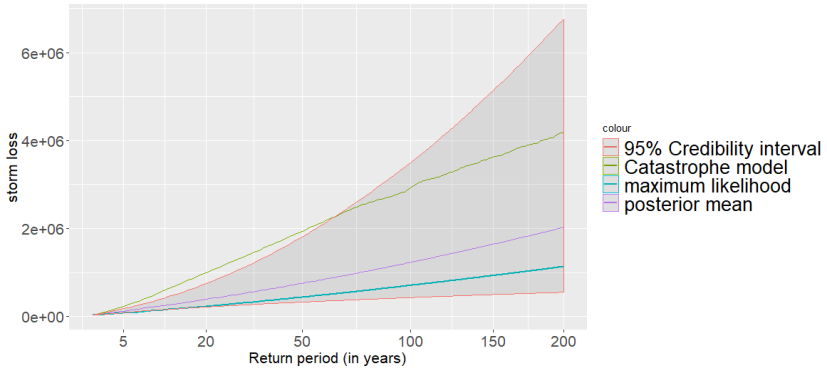

Figure 30: Return period using Bayesian model

7 Discussion

In this thesis, we have looked at a few methods of estimating the Peaks over threshold model for wind storm losses in Norway. As the choice of threshold left only 29 observations from a very heavy tailed distribution, there was a large parameter uncertainty in the analysis. The Bootstrap showed narrower confidence intervals for parameter and quantiles compared to profile likelihood intervals.

The CAT model had much more conservative estimates for return periods compared to the maximum likelihood return periods from the Peaks over threshold model. The Bayesian Peaks over threshold estimation gave return level estimates that became a trade off between the CAT model and maximum likelihood estimated model. The confidence intervals for return levels were extremely skewed when using profile likelihood, because of instability in shape parameter estimates ξ. The credibility intervals for high quantiles and return levels were much more narrow than the profile likelihood confidence intervals, a result of the posterior samples of ξ having very few values above 1. It is unclear whether the reduction in tail uncertainty for the Bayesian model was well founded, or if the prior restrictions provided a false sense of stability for the model.

One way of extending this analysis would be to explore different choices of thresholds in the model. Another important analysis would be to consider

other choices of incorporating prior information to the Bayesian model, as well as comparing the Slice sampling algorithm to other forms of MCMC methods, such as Metropolis-Hastings.

References

[Carlin and Louis, 2008] Carlin, B. P. and Louis, T. A. (2008). *Bayesian methods for data analysis*. CRC Press.

[Coles et al., 2001] Coles, S., Bawa, J., Trenner, L., and Dorazio, P. (2001). *An introduction to statistical modeling of extreme values*, volume 208. Springer.

[Efron and Tibshirani, 1986] Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, pages 54–75.

[Gelman et al., 1992] Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

[Kyselỳ, 2008] Kyselỳ, J. (2008). A cautionary note on the use of nonparametric bootstrap for estimating uncertainties in extreme-value models. *Journal of Applied Meteorology and Climatology*, 47(12):3236–3251.

[Meteorologisk institutt, 2017] Meteorologisk institutt (2017). 25 år sidan den historiske nyttårsorkanen. `https://www.met.no/nyhetsarkiv/25-ar-siden-den-historiske-nyttarsorkanen/`. [Online; accessed 12-April-2021].

[Mitchell-Wallace et al., 2017] Mitchell-Wallace, K., Jones, M., Hillier, J., and Foote, M. (2017). *Natural catastrophe risk management and modelling: A practitioner's guide*. John Wiley & Sons.

[Neal, 2003] Neal, R. M. (2003). Slice sampling. *Annals of statistics*, pages 705–741.

[Pickands III et al., 1975] Pickands III, J. et al. (1975). Statistical inference using extreme order statistics. *Annals of statistics*, 3(1):119–131.

[Rootzén and Tajvidi, 1997] Rootzén, H. and Tajvidi, N. (1997). Extreme value statistics and wind storm losses: a case study. *Scandinavian Actuarial Journal*, 1997(1):70–94.

Appendix A Appendix

A.1 Prior specification for Bayesian model

When using the Bayesian Generalized Pareto model, we have constructed Gaussian priors

$\xi \sim N(\mu_\xi, \tau_\xi^2), \nu = log(\sigma) \sim N(\mu_\nu, \tau_\nu^2)$. The four hyperparameters $\mu_\xi, \tau_\xi^2, \mu_\nu, \tau_\nu^2$ were based on a Bootstrap analysis of the CAT model storm costs. When comparing historical storms to the CAT model generated storm costs, there is a noticeable deviation between the distributions, as seen in figure 31, where the CAT model has a conservative probability of large storms occurring compared to historical losses. It is unclear if this is a sign of bias from the CAT model, or just a consequence of low sample size from the historical storms.

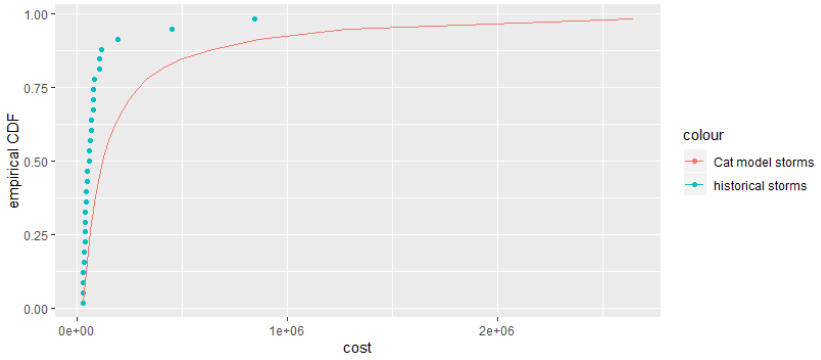

Figure 31: Empirical CDF versus CAT model probability distribution

In figure 32, we show the Bootstrap drawings of ξ and $\nu = log(\sigma)$. These were obtained by repeatedly drawing 29 random samples (with replacement) from the CAT model generated storms, and for each drawing, fitting a Generalized Pareto distribution to the drawn sample by maximum likelihood, and saving $\hat{\xi}_{ml}, \hat{\nu}_{ml} = log(\hat{\sigma}_{ml})$ as Bootstrap samples. These drawings were repeated 50 000 times, so we get $\xi^{(1)}, .., \xi^{(50000)}$, and $\nu^{(1)}, .., \nu^{(50000)}$ as basis for the prior.

In figure 32, we display histograms of the 50000 Bootstrap samples together with the Gaussian fits that we then use as priors for the Bayesian models. We see that for ν, the Gaussian fit matches the Bootstrap drawings almost perfectly, indicating that if our use of Bootstrap is appropriate to determine initial parameter uncertainty, the Gaussian prior is a good choice for ν. The ξ Bootstraps are slightly less decently fitted by a Gaussian distribution, but still captures the distribution adequately.

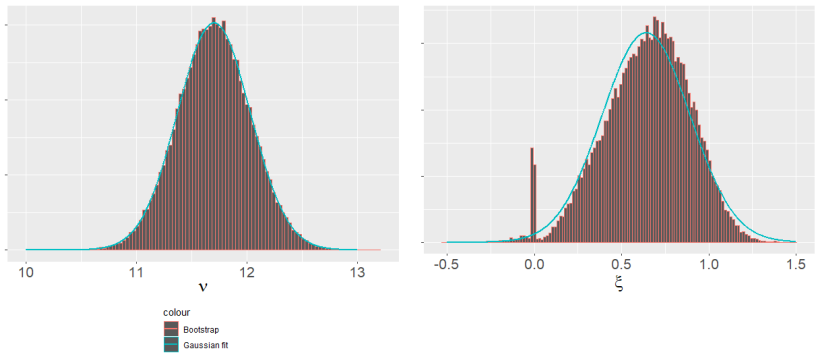

Figure 32: Histogram of Bootstrap drawings from CAT model with Gaussian fit

We may also see what a Generalized Pareto distribution fit to the CAT model generated storm losses would look like, which is shown in figure 33, where we plot the empirical quantiles of the CAT model losses against the corresponding Generalized Pareto quantiles. We have divided the quantile plot into two figures, where the left plot shows all quantiles up to probability level 0.983, and the right figure shows the higher quantiles. The fit gave a positive shape parameter estimate that was smaller than 1. We see in the left plot that the Generalized Pareto distribution captures this data very well, with the right plot showing that the tail is lower for the CAT model in relation to the Generalized Pareto fit.

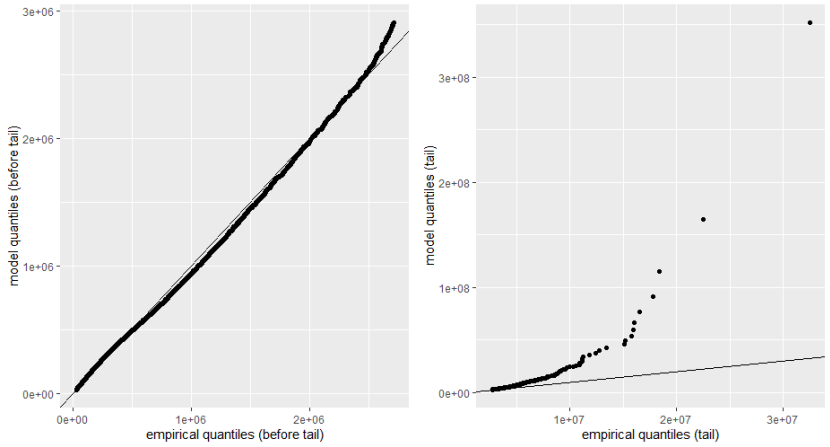

Figure 33: Quantile-Quantile plot for Generalized Pareto fit to CAT model generated storm losses

A.2 MCMC diagnostic

From our Bayesian model, we display in figure 34 the 5 MCMC chains for ξ and σ respectively. These 5 chains each were intitated with different starting values.

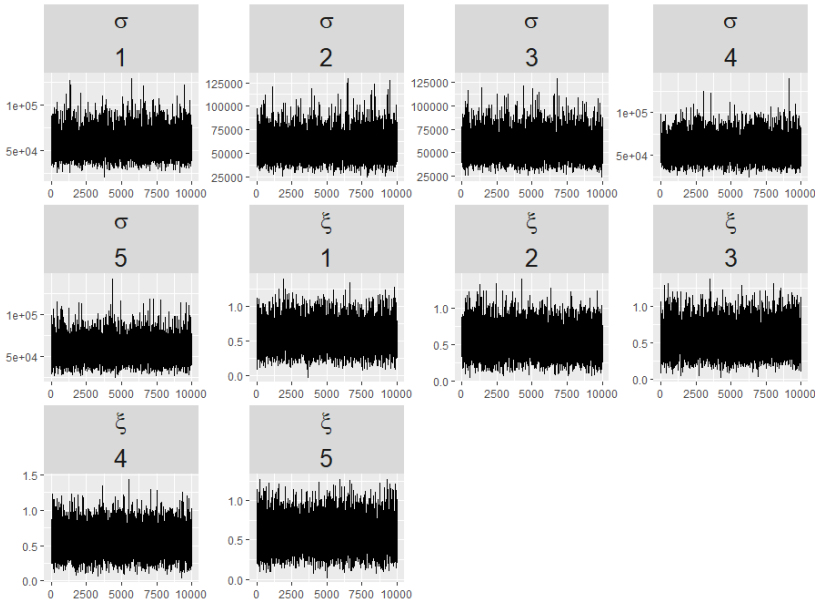

Figure 34: MCMC chains from Bayesian model

In figure 35, we display the estimated autocorrelation function for each chain, meaning the sample correlation between $\xi^{(i)}, \xi^{(i-l)}$, as well as $\sigma^{(i)}, \sigma^{(i-l)}$, for lags $l = 1, .., 20$.

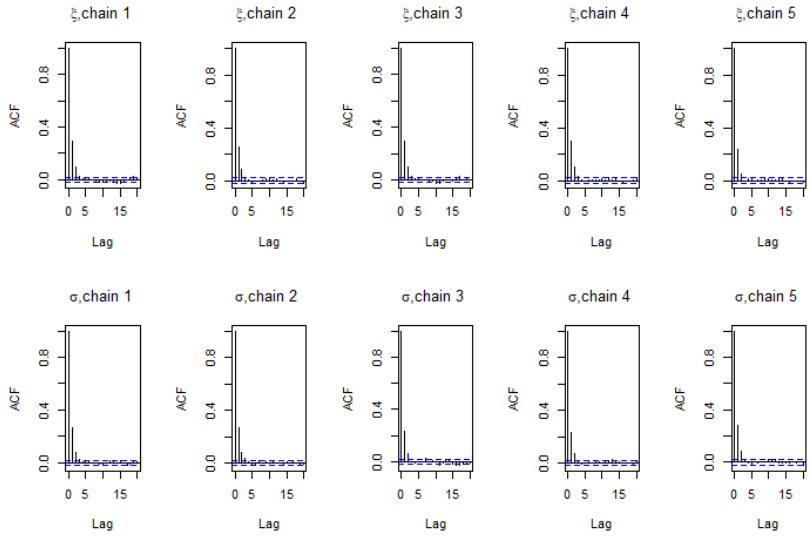

Figure 35: Autocorrelation from MCMC chains

A.3 Maximum of independent Pareto random variables

If $X_1, ..X_n \sim Pa(\alpha, x_m)$ are independent, then the CDF for these random vaiables is

$$F_{X_i}(x) = (1 - (\frac{x_m}{x_\cdot})^\alpha), x > x_m$$

while the inverse CDF is

$$F_{X_i}^{-1}(x) = \frac{x_m}{(1-x)^{1/\alpha}}.$$

If we are interested in the distribution of $M_n = \max(x_1, .., x_n),$, and use that $p(M_n < x) = F_{X_i}(x)^n$, we can use $a_n = F^{-1}(1 - \frac{1}{n})$ as a (non-random) scaling sequence for M_n. Then we see that the distribution of the scaled maximum $M_n^* = \frac{M_n}{a_n}$ and see that

$$\begin{aligned} p(M_n^* < x) &= p(\frac{M_n}{a_n} < x) = p(M_n < x \cdot a_n) \\ &= F(x \cdot a_n)^n = (1 - (\frac{x_m}{x \cdot a_n})^\alpha)^n \\ &= (1 - (\frac{x_m}{x\frac{x_m}{(1/n)^{1/\alpha}}})^\alpha)^n = (1 - (\frac{(1/n)^{1/\alpha}}{x})^\alpha)^n \\ &= (1 - \frac{1}{x^\alpha n})^n \to e^{-\frac{1}{x^\alpha}} = e^{-x^{-\alpha}}, n \to \infty \end{aligned} \tag{15}$$

This means that the maximum M_n scaled by a_n converges to a Frechet distribution, as defined in equation (4).