Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Marion Helen Røed

Spatial Extreme Value Modelling of Sea Level Data from the Oslo Fjord

Master's thesis in Applied Physics and Mathematics
Supervisor: Sara Martino
July 2021

NTNU
Norwegian University of
Science and Technology

Marion Helen Røed

Spatial Extreme Value Modelling of Sea Level Data from the Oslo Fjord

Master's thesis in Applied Physics and Mathematics
Supervisor: Sara Martino
July 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

This thesis aims to model the occurrence of extreme sea levels in the Oslo fjord using a spatial model. This is done for the purpose of obtaining information about the probability of flooding, which is important for the creation of flooding charts which is used by for example insurance and construction companies to make informed decisions. To address this problem, a Bayesian hierarchical model was proposed. The generalized extreme value model (GEV) was applied to yearly maximum sea level values from the Oslo fjord. The parameters of the GEV model were assumed to be either constant or slowly varying in space. The parameters that were assumed to vary were described with Gaussian random fields. The mean value of the random fields was assumed to be depending on spatial covariates that were expected to explain the sea level. The variance of the random fields was described with a Matern correlation function. It was found that a spatial model lead to improved results at monitored locations compared to a univariate model, and it was also found that the improvements seemed to be mostly caused by having a common shape parameter of the GEV distribution. Spatial interpolation was also tried in this analysis, however it was not feasible with the amount of sea level data available in the area. Future research could try to study a larger portion of the Norwegian coast and it could also see if it is possible to solve the problems related to limited data and spatial interpolation.

Sammendrag

Målet for denne oppgaven var å modellere ekstreme havnivå i Oslofjordområdet ved hjelp av en romlig modell. Modeller for ekstreme havnivå blir brukt blant annet til å lage flomkart. Flomkart er nyttig for eksempel for forsikringselskaper og ved planlegging av veier og bygninger. En Bayesisk hierarkisk modell ble brukt i denne oppgaven. Den generaliserte ekstremverdi-modellen (GEV) ble anvendt på årlig maksimum havnivå fra Oslofjorden. Parameterne til GEV modellen ble antatt å variere sakte eller å være konstante i området. Parameterne som varierte ble beskrevet ved hjelp av Gaussiske tilfeldige felt. Feltene hadde gjennom-snittsverdi avhengig av kovariater som forklarte variasjon i havnivå og varians beskrevet av en Matern korrelasjonsfunksjon. Resultatene av analysen antydet at den romlige modellen førte til forbedringer sammenlignet med diskrete modeller på stasjonene. Det virket som om forbedringene for det meste var på grunn av felles formparameter i GEV modellen. Romlig interpolasjon ble også prøvd, men det ble funnet at resultatene ikke var brukbare i praksis grunnet veldig brede usikkerhetsintervall. Usikkerhetsintervallene var brede på grunn av at kun begrensede mengder data var tilgjengelig. Fremtidige undersøkelser kan for eksempel prøve å analysere større deler av den norske fjorden og også prøve å løse problemer relatert til romlig interpolasjon og begrensede mengder tilgjengelig data.

Preface

This thesis completes my masters degree in Applied Physics and Mathematics at the Norwegian University of Science and Technology (NTNU).

I would like to thank my supervisor Sara Martino for guidance and support throughout the work with the thesis. I would also like to thank Hilde Sande Borck and her colleges at Kartverket for providing the data used in this thesis as well as answering questions about the topic at hand.

Contents

1 Introduction

Flooding of coastal areas is a large problem. We need reliable models so that we can specify the risk of flooding at any given coastal location. This is a point of interest to for example insurance companies and for the planning of construction work. The risk of flooding is usually described with return levels and associated return periods. The return level for the m year return period is the sea level that is expected to be exceeded every m years.

The main goal in this project is to use a spatial model to specify the risk of flooding. It is interesting to study a spatial model, as opposed to a univariate model, as the spatial model makes it possible to share information across locations. The spatial model is especially useful since lack of data is a big challenge when modelling extremes. Extremes are by definition rare, which means that we do not have excessive amount of data even though we have hourly measurements of sea level data collected over a period of many years. The number of locations being low is also problematic, as we need the data to have a spatial structure to be able to estimate the spatial parameters.

We base our analysis mainly on a spatial analysis of extreme sea levels on the Canadian coast that has been performed by Beck et al. (2020). The most notable difference between their analysis and this analysis is that we choose to not separate the sea level into a surge and a tide part. The separation of the sea level is done due to the sea level being a combination of a deterministic part, that is the tide, and a stochastic part based on meteorological conditions (Coles and Tawn, 2005). The choice of studying the undivided sea level was made based on a preliminary analysis that showed that it was very difficult to recover the sea level, which is the point of interest, after the analysis was performed (Røed, 2021). The preliminary analysis also showed that there did not seem to be a large difference between the results obtained by studying the undivided sea level and the surge and tide separately when studying the data from Oslo (Røed, 2021). The stations studied in this analysis are located in close proximity to Oslo, so we assume similar behaviour. The choice of studying the undivided sea level might be an oversimplification, however we found that it was necessary due to the difficulties related to recovering the sea level.

Another difference between the analysis by Beck et al. (2020) and this analysis is that we study Norwegian data, more specifically data from the Oslo fjord. Skjong et al. (2013) did also study Norwegian sea level data, however they performed univariate analysis. The choice of studying just the Oslo fjord was made due to the behaviour of the sea level changing drastically along the Norwegian coast line, and it is therefore difficult to model the entire coast line at once. The eastern coast line is dominated by the tide, and the tide level increases towards the north, while the western coast line is dominated by the surge (Skjong et al., 2013).

We follow Beck et al. (2020) and use the generalized extreme value (GEV) model in our analysis. There are two groups of extreme value models, there are block maxima and threshold exceedance models (Coles, 2001). The GEV model is a block maxima model, which means that the extreme values are found by dividing the data into blocks and taking the maximum of each block. The main problem with this approach is that it can be quite wasteful, only the largest value in the block is used even if there are multiple large values inside the block. Coles

and Tawn (2005) use the point process (PP) model, which is a threshold exceedance model, in their analysis. The problem with this model and any other threshold exceedance model is that hand tuning is necessary. The threshold needs to be chosen, and this is usually done by manually inspecting plots (Coles, 2001). This means that a great deal of manual work is needed to fit a model based on many locations, which we want to do in our spatial model.

The GEV model is stationary in time, which means that the year in which an extreme occurred is not taken into account in the model. The assumption of stationarity might be slightly unrealistic due to for example climate change, however a non-stationary model would be quite difficult to fit to the data since there is not a large amount of data available (Dyrrdal et al., 2015)

We follow Cooley et al. (2007) and assume that the GEV distributions are independent rather than linked with a copula. Beck et al. (2020) argue against this approach, stating that it is overly simplistic. Bracken et al. (2016) and Beck et al. (2020) use Gaussian and student t copulas respectively in their analyses. We decide to follow the simpler method due to having limited amounts of data, which makes it difficult to estimate dependencies.

Cooley et al. (2007), Bracken et al. (2016) and Dyrrdal et al. (2015) study the occurrence of extreme precipitation rather than sea levels. The methods used are however very similar to those used by Beck et al. (2020). The three papers model spatial extremes using hierarchical Bayesian models. Dyrrdal et al. (2015) and Bracken et al. (2016) use the GEV model while Cooley et al. (2007) use a threshold exceedance model. There does not seem to be as many papers using similar methods for dealing with spatial modelling of extreme sea levels.

The data used in the analysis are presented in Section 2. In this section we introduce the sea level data from discrete locations and the covariate data available in the entire Oslo fjord. The covariate data is used to try to explain the spatial behaviour of the sea level. In Section 3 the model is presented, the method used for estimating the parameters of the model is explained, we also talk about the application of the estimated parameters and the challenges involved when one wants to check the model. We perform a simulation study in Section 4. We analyse simulated data from 15, 4 and 8 locations. The simulation study based on 15 locations is done for the purpose of checking the model. The simulation study based on 4 locations is done to find out which results can be expected from the analysis of the real data, since only data from 4 locations are available. The simulation study based on 8 locations is done to check the expected improvements of obtaining some additional data. In section 5 we study the sea level data. In this section we look at a univariate model, a model where the shape parameter of the GEV model is common across locations, while the rest are independent in space and lastly we apply the spatial model to the data. The results are discussed in Section 6 and we make a summary in Section 7.

2 Problem and Data

In this section the problem studied in this thesis will be described. The problem at hand is to study the probability of the occurrence of extreme sea levels. We are in particular interested in being able to specify return levels with associated return periods. That is we want to be able to say that the sea level is expected to exceed x cm every m years. The structure of this section is as follows: First we will talk about why it is necessary to study the problem and give a general introduction of the problem. Next we will talk about the structure of the sea level data and how we choose to handle this structure. Lastly, in the two subsections, the sea level data and the covariate data will be described in detail.

Studying extreme sea levels is a topic of interest due to the risk of flooding of different areas being important knowledge for construction work, to homeowners and to insurance companies. Long periods of sea level data are needed to be able to address extreme sea levels. This is due to extremes being rare and to be able to comment on the probability of different levels of extremes a certain amount of extremes need to have occurred in the data set. This means that having for example 50 years of hourly measurements of the sea level is not an excessive amount of data in the extreme setting. In this project we are interested in studying extreme sea levels in a spatial setting. In practice this means that we can borrow information from other locations in the area, meaning that it will be easier to specify probabilities of extremes at each location. In the spatial setting we can also try to obtain probabilities of extremes in areas with no sea level data.

The sea level has a complex structure that depends on meteorological conditions and the tide (Coles and Tawn, 2005). The eastern coast of Norway is surge dominant (Skjong et al., 2013), this means that the amplitude of the sea level fluctuations due to the tide are small compared to the amplitude of the sea level fluctuations related to meteorological conditions. The tide part of the sea level is deterministic, we can explain exactly what the tide is currently, what it has been in the past and what it will be in the future. The tide part of the sea level is not interesting for stochastic analysis, since it is known. However, the sea level cannot be divided into a sum of a tide dependent part and a part that depends on meteorological conditions, because there is dependencies between the the two parts (Coles and Tawn, 2005). It was found in the preliminary work that it is difficult to handle the dependency (Røed, 2021). When the sea level data is divided into parts, then it is challenging to recover the sea level, which is the physical phenomenon of interest. It was also found in the preliminary work that there was not a huge difference between the results obtained by studying the sea level data and the surge data in the cases when it was possible to recover the sea level (Røed, 2021). Thus the undivided sea level data is studied in this project. This is done for the sake of simplicity and for the sake of being certain that the results are accurate and comparable with the physical sea level.

2.1 Sea level Data

Hourly sea level measurements in the area of interest, that is the Oslo fjord, was made available by Kartverket (Kartverket, 2021). Kartverket collects sea level data from numerous locations

along the Norwegian coast. Some of the locations are temporary stations, that is the sea level data are collected only for a short period of time, while other locations are permanent stations. The permanent stations are the ones that are of most interest in this project, due to the need for data from a longer period of time to be able to say something about the probability of extreme occurrences. There are 24 permanent stations in Norway and 4 of them are located in the Oslo fjord, the permanent stations are shown in Figure 1. The sea level data used in this project consist of hourly measurements of the sea level at the four locations: Helgeroa, Oscarsborg, Oslo and Viker. The four locations are shown in Figure 2. The data series have length 77 years, 56 years, 101 years and 30 years for Helgeroa, Oscarsborg, Oslo and Viker respectively.

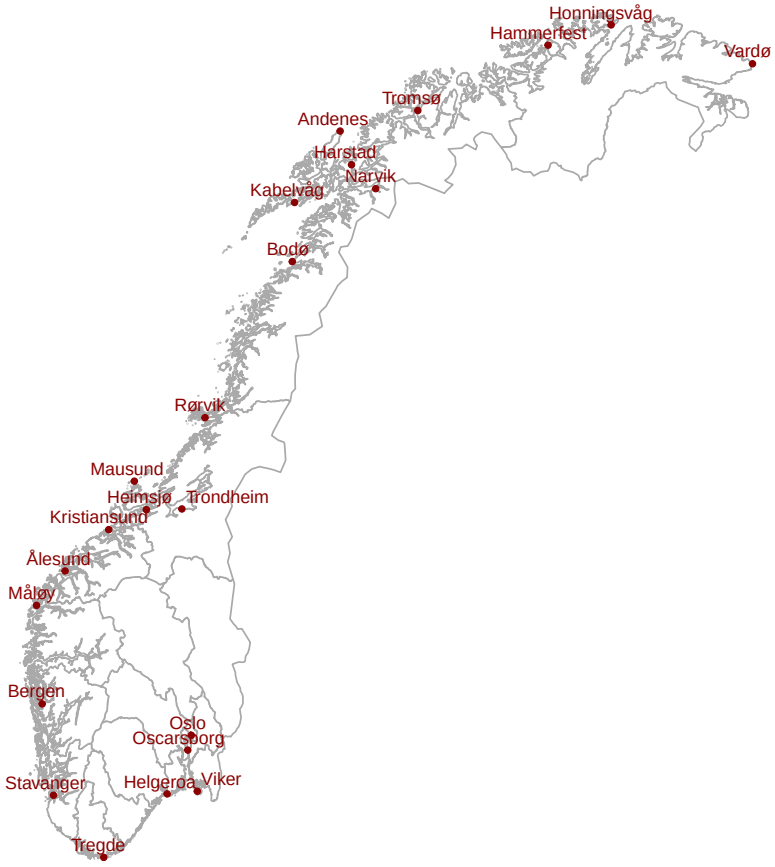


Figure 1: The figure shows the permanent stations in Norway where sea level data are collected hourly.

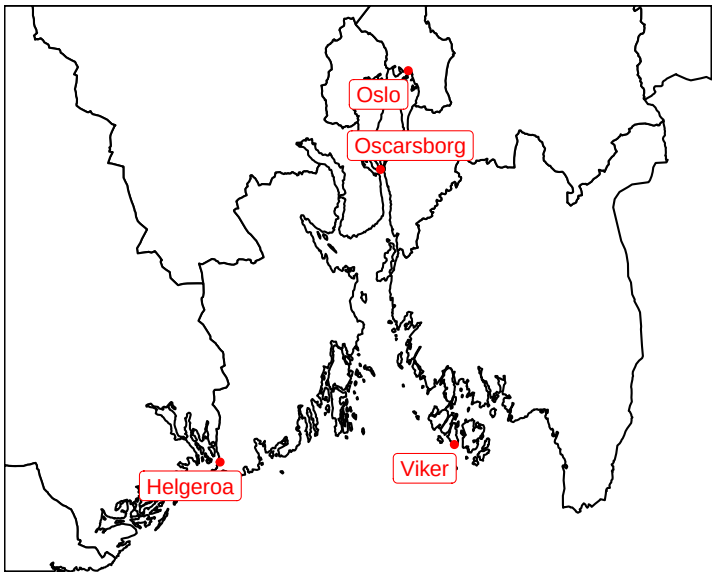


Figure 2: The figure shows the four locations in the Oslo fjord where there is available sea level data for long periods of time. The locations are from top to bottom: Oslo, Oscarsborg, Helgeroa to the left and Viker to the right.

Most of the sea level data in Norway that has been collected for a long period of time has a trend. This trend is caused by the land slowly rising, which happens because the land was pushed into the mantel by the weight of the ice in the last ice age (Skjong et al., 2013). The speed that the land rises differs around the country, we will estimate the trend with quadratic regression and then remove the trend from the data, this approach was chosen based on personal communication with Kartverket. We estimate the trend at location l using the model

$$y_l(t) = \beta_{0,l} + \beta_{1,l} \cdot t + \beta_{2,l} \cdot t^2 + {\varepsilon_l}^2(t) \tag{1}$$

Where $y_l(t)$ is the sea level at location l and t is time measured in years from 0 to number of years available in the different data sets. The estimated β values are shown in Table 1.

		Estimated parameters	standard error	p-value
Helgeroa	β_0	11.9	0.1	$< 2 \cdot 10^{-16}$
	β_1	-0.21	$4 \cdot 10^{-3}$	$< 2 \cdot 10^{-16}$
	β_2	$8 \cdot 10^{-4}$	$4 \cdot 10^{-5}$	$< 2 \cdot 10^{-16}$
Oscarsborg	β_0	16.4	0.1	$< 2 \cdot 10^{-16}$
	β_1	-0.52	0.01	$< 2 \cdot 10^{-16}$
	β_2	$4 \cdot 10^{-3}$	$1 \cdot 10^{-4}$	$< 2 \cdot 10^{-16}$
Oslo	β_0	34.4	0.1	$< 2 \cdot 10^{-16}$
	β_1	-0.56	$3 \cdot 10^{-3}$	$< 2 \cdot 10^{-16}$
	β_2	$2 \cdot 10^{-3}$	$3 \cdot 10^{-5}$	$< 2 \cdot 10^{-16}$
Viker	β_0	0.4	0.1	$< 2 \cdot 10^{-16}$
	β_1	0.008	0.021	$< 2 \cdot 10^{-16}$
	β_2	$-1 \cdot 10^{-3}$	$7 \cdot 10^{-4}$	$< 2 \cdot 10^{-16}$

Table 1: Parameter values for the quadratic regression used to remove the trend from the data sets.

The original and the de-trended time series are shown in Figure 3. All the locations except Viker had a decreasing trend, which is what we expect the land rising to produce. Viker on the other hand has a slightly increasing trend. The trend could mean that either the land does not rise as much at this location, or since the time series is quite short, it is also possible that the effect of global warming is slightly larger than that of the land rising for this time period.

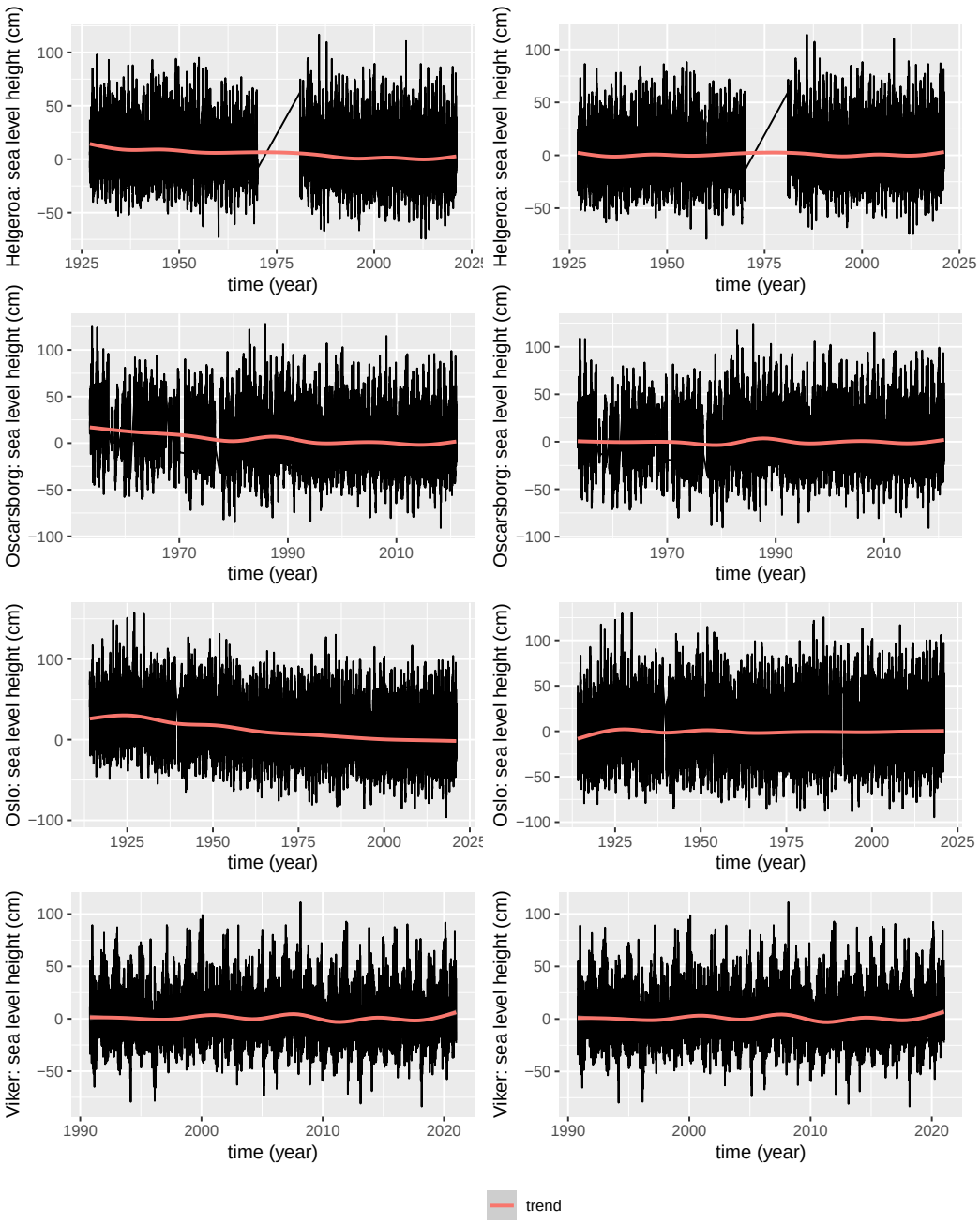


Figure 3: The figure shows the data from the four locations: Helgeroa, Oscarsborg, Oslo and Viker. The data to the left has a trend while the data to the right have been centred with regression. Note that the x-axis differs for the locations.

The time series presented in Figure 3 have several missing values. This is not a problem for fitting Equation (1), which is used to recover the long period trend. However, missing data

can be problematic when one wants to look at yearly maxima. In fact the more missing data there is in a year, the larger the chance that the yearly maximum value is missing (Beck et al., 2020) (Skjong et al., 2013). We want the yearly maximum to be available for every year in the data set due to these values being the ones that we use to fit the model. We will for this reason remove any year from the data sets for which a substantial amount of data are missing. We have chosen to set the limit to three months. The years that are missing more than three months for the four locations are presented in Table 2.

Location	Removed years
Helgeroa	1941, 1942, 1960, 1962, 1963, 1968, 2021
Oscarsborg	1953, 1957. 1958, 1959, 1961, 1964, 1967, 1968, 1976, 1977, 1990, 2021
Oslo	1914, 1915, 1972, 1974, 1991, 2021.
Viker	1990, 2021.

Table 2: The years that have been removed from the data sets due to the data missing for these years exceeding 3 months.

The yearly maximum data values for the four locations are presented in Figure 4. One can see that Helgeroa, Oscarsborg and Oslo all have their maximum value in 1987, whereas Viker does not have any data from this year. The maximum value differs the most from the other values for Oslo and Oscarsborg.

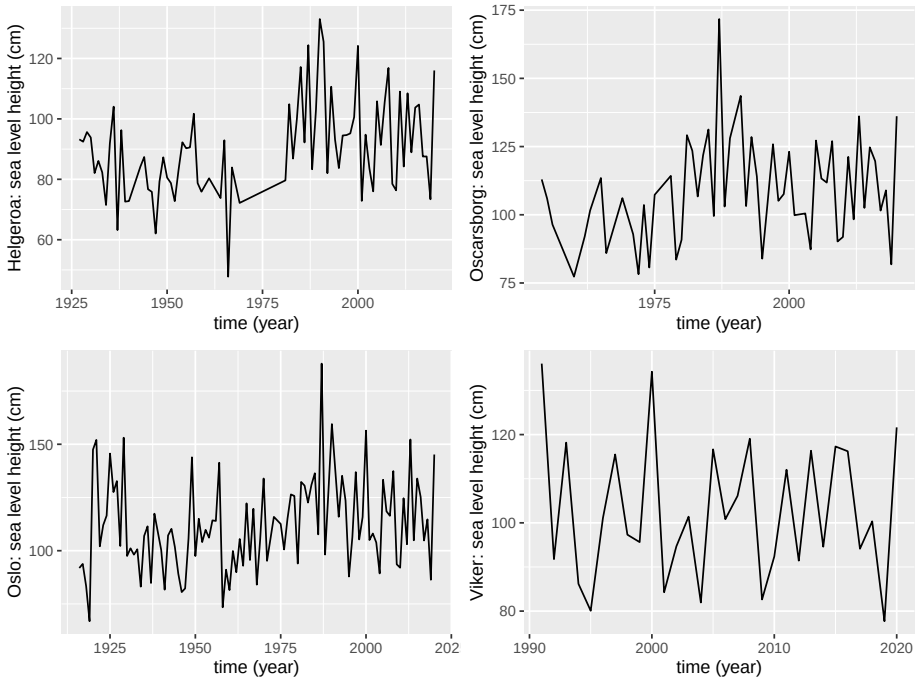


Figure 4: The figure shows the yearly maximum data from the four locations: Helgeroa, Oscarsborg, Oslo and Viker.

It is expected that we observe similar behaviour for the data sets as the locations are in close proximity. This means that spatial dependence between the locations is likely present in the data. The Figure 5 shows the annual maximum sea level values of coinciding years of the four locations plotted against each other. One can see that there seems to be quite strong dependence between all the locations. Cooley et al. (2007) state that spatial dependence of extremes is not well understood, but that the advantages of spatial analysis should outweigh potential negative effects of spatial dependence. Another potential issue is dependence between the annual maxima sea levels, however Cooley et al. (2007) state that maxima of stationary series still converge to a GEV as long as the dependence is short range.

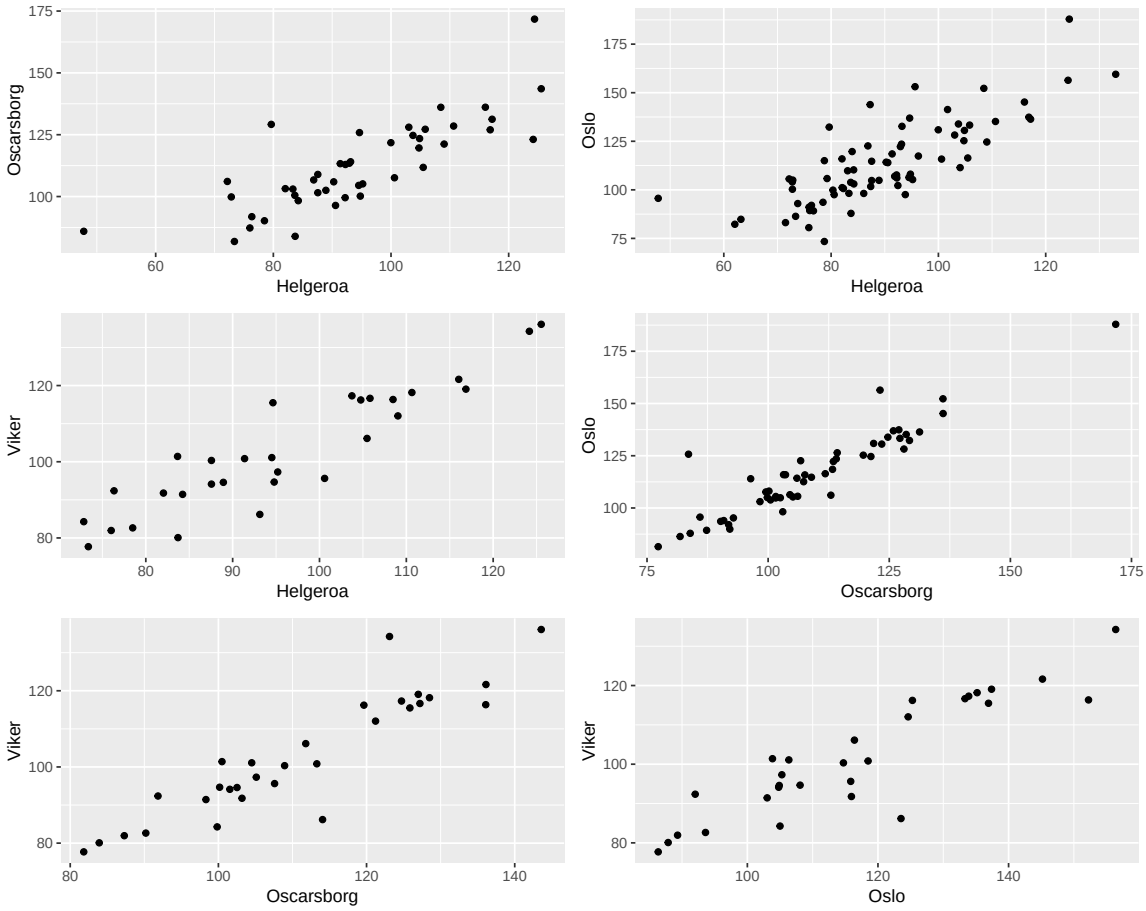


Figure 5: The figure shows the annual maximum sea level values of coinciding years of the four locations plotted against each other.

2.2 Covariate Data

There are three different covariates used in this project. The covariates were chosen based on personal communication with Kartverket. The covariates are expected to have some relation

to the sea level. The covariates chosen are atmospheric pressure at sea level, wind speed and tide levels. The pressure levels and the wind speed are distributed by the Copernicus program (Bell et al., 2021), (Hersbach et al., 2021). The Copernicus program does atmospheric reanalysis. The data distributed by the Copernicus program has resolution $0.25^\circ \times 0.25^\circ$ and the time coverage is hourly values from 1950 to 2020. The tide levels are distributed by Geonorge (2021). The tide data are based on harmonic analysis of sea level measurements along the coast. The tide data has resolution $0.01^\circ \times 0.005^\circ$, and is based on data from 1996 to 2014.

Atmospheric pressure at sea level is expected to influence the sea level: we expect high sea levels when the pressure is low and low sea levels when the pressure is high (Beck et al., 2020). The reasoning behind this expectation is that high pressure should push the ocean downwards and low pressure should give the ocean room to expand. We follow Beck et al. (2020) and compute the mean annual minimum pressure. That is we find the minimum pressure level for each year and take the mean of these values. The covariate values are shown in Figure 6. The data are available on a grid that is 6×11 and there are hourly measurements available in a period of 70 years. The grid nodes are quite large, which means that the covariate values are identical for quite a large area.

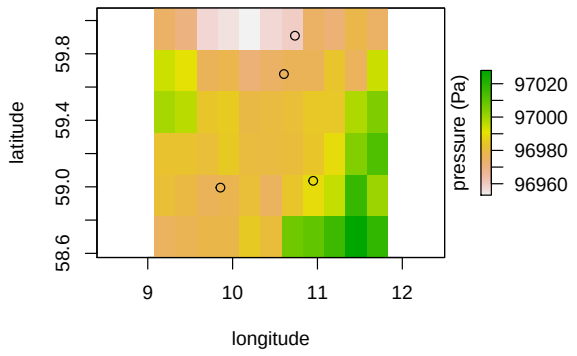


Figure 6: The plot shows the mean annual minimum sea level pressure calculated from data between the years 1950 and 2020. The pressure is measured in Pascal. The four locations Helgeroa, Oscarsborg, Oslo and Viker are displayed as circles.

Strong winds are associated with high sea levels (Hieronymus et al., 2018). We consider here the mean annual maximum wind speed as a possible covariate. That is the maximum wind speed is found for every year and the mean of those values is studied. The wind data are available as a north-south component, directed towards the north, and an east-west component, directed towards the east, these are shown in Figure 7. The wind data comes from the same model as the sea level pressure data and has the same resolution.

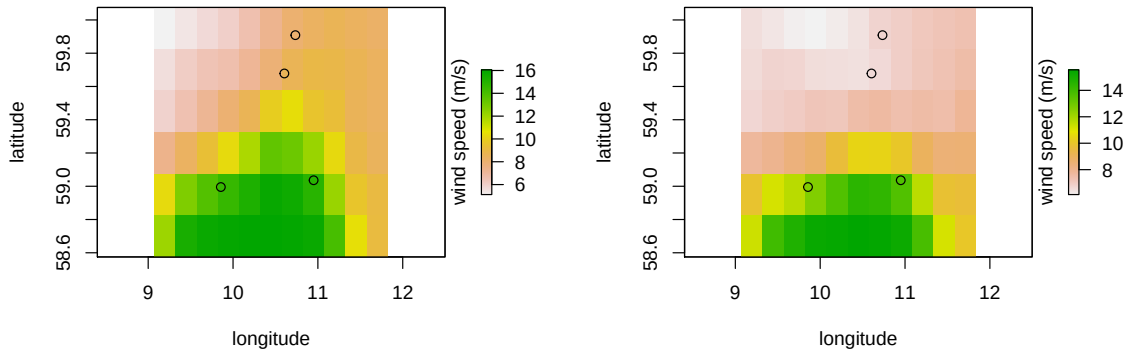


Figure 7: The plot shows the north-south and the east-west components of the wind speed covariate data, that is the mean annual maximum wind speed calculated from data between the years 1950 and 2020 in northern and eastern direction. The wind speed is measured in meters per second. The four locations Helgeroa, Oscarsborg, Oslo and Viker are displayed as circles.

The tide level covariate is the measurement of the height of the mean sea level above the lowest astronomical tide from 1996 to 2014, these measurements are hopefully reasonable approximations for height of highest astronomical tide above mean sea level (T. Taskjelle, personal communication). Figure 8 shows the tide covariate data. The grid resolution is in this case 281×229, which means that there are substantially more values than for the other two covariates.

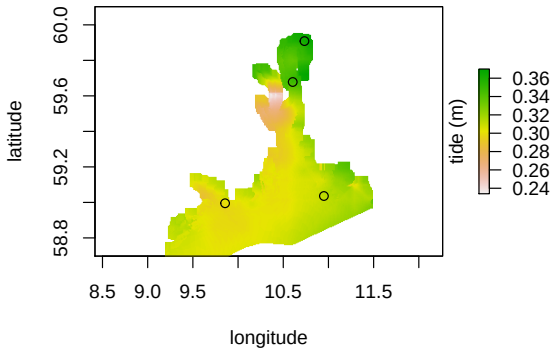


Figure 8: The plot shows the tide covariate data, that is the height of the mean sea level above the lowest astronomical tide from 1996 to 2014. The four locations Helgeroa, Oscarsborg, Oslo and Viker are displayed as circles.

The covariates presented in Figures 6, 7 and 8 have quite different scale and range. The

covariates have been scaled so that they exist on $[0, 1]$ in the following analysis for the purpose of making the covariates comparable.

3 Methods

In this section the spatial extreme value model used to explain the probability of extreme sea levels is presented. The second topic covered in this section is parameter estimation techniques, in particular Markov Chain Monte Carlo (MCMC) and the program STAN used to perform MCMC. Next we will talk about the application of the parameter estimates, in particular the calculation of return levels with associated return periods and the calculation of parameter values in areas where no data are available. The last topic covered in this section is the evaluation of the fit of the model.

We work in a Bayesian framework and use the Bayes rule to define a model for the parameters given the data. The Bayes rule is defined as

$$p(x|z) = \frac{p(z|x)p(x)}{p(z)}. \tag{2}$$

The Bayes rule can be simplified to $p(x|z) \propto p(z|x)p(x)$ when z represents the data, due to $p(z)$ being a constant. With the use of the Bayes rule and the identity $p(y, x) = p(y|x)p(x)$ we get

$$p(\mathbf{q}, \mathbf{s}, \xi, \eta|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{q}, \mathbf{s}, \xi) \cdot p(\mathbf{q}|\eta) \cdot p(\mathbf{s}|\eta) \cdot p(\xi, \eta), \tag{3}$$

where $p(\mathbf{z}|\mathbf{q}, \mathbf{s}, \xi) = \prod_{i=1}^{l} \prod_{j=1}^{k_i} \text{GEV}(z_{ij}|q_i, s_i, \xi)$, l is the number of locations and k_i is the number of data z available in location i. $p(\mathbf{q}|\eta)$ and $p(\mathbf{s}|\eta)$ represents the Gaussian fields with parameters included in η. $p(\xi, \eta)$ represents the probability distributions of the hyper-parameters of the model. In Bayesian inference the parameters of the likelihood function are associated with probability distributions (Givens and Hoeting, 2013). This means that the parameters of the model are treated as random variables. A big advantage of Bayesian analysis is that there is more room for using common sense when interpreting results, for example one can say that a credible interval has a large probability of containing the parameter value, rather than having to talk about repeated trials as one does in frequentist statistics (Gelman et al., 2018). Gelman et al. (2018) states that Bayesian analysis can be divided into three steps: The first step is setting up the full probability model, that is the joint model of parameters and data. The second step is conditioning on the observed data, that is finding the posterior distribution, that is distribution of the parameters given the observed data. The last step is evaluating the fit of the model and check how well the model fits the data. The first step is handled in Section 3.1. The second step involves calculating the distributions of the parameters of the model given the data, this is done using MCMC and STAN, which is discussed in more detail in Section 3.2. The last step involves checking the model and this step is discussed in more detail in Section 3.4.

3.1 Model

The most common model for extremes is called the generalized extreme value (GEV) model. The GEV model is based on block maxima data, which are found by dividing the data into blocks and taking the maximum of each block (Coles, 2001). The sets of block maxima data converge to the GEV model when the block size increases towards infinity(Coles, 2001). This

means that the theorem describing the GEV model can be seen as the extreme value analogue to the central limit theorem (Coles, 2001). A problem related to using block maxima is that it is very inefficient since only one data point is kept per block. Moreover, the shape parameter can be quite difficult to estimate and the uncertainty of the parameter can be quite large when the data series are short (Dyrrdal et al., 2015), this is the case in our project. The spatial model can be advantageous since assuming that the parameters of the model are either constant over a region or slowly varying in space makes it possible to share information between locations and thus achieve a more robust inference. We assume that the parameters that slowly vary in space can be described with Gaussian random fields based on some covariates that are available on the entire field and a correlation function based on distances between locations of interest.

In the following two subsections first the GEV model is presented, including a reparametrization of the GEV model. Then the spatial model is presented, including the hierarchical model for the parameters given the data.

3.1.1 The GEV Model

The GEV model is presented in Theorem 1, which is also known as the extremal types theorem (Coles, 2001). The following theory about the GEV model is based on the introductory book on extreme value modelling by Coles (2001). The theorem is the basis for all extreme value analysis. The extremal types theorem is the extreme value analysis analogue to the central limit theorem. When the block maxima $M_n = \max\{X_1, ..., X_n\}$ are scaled with sequences of constants, if the distribution function is non-degenerate, then it is a member of the GEV family.

Theorem 1 *If there exist sequences of constants* $\{a_n > 0\}$ *and* $\{b_n\}$ *such that*

$$Pr\Big\{\frac{M_n - b_n}{a_n} \leq z\Big\} \to G(z), \text{ as } n \to \infty,$$

where G *is a non-degenerate distribution function, then* G *is a member of the Generalized Extreme Value (GEV) family*

$$GEV(z|\mu, \sigma, \xi) = \exp\Big\{-\Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\}, \tag{4}$$

which is defined on $\{z|1 + \frac{\xi(z-\mu)}{\sigma} > 0\}$*, where* $\mu \in \mathbb{R}$*,* $\xi \in \mathbb{R}$ *and* $\sigma > 0$*.*

The sequences $\{a_n\}$ and $\{b_n\}$ in Theorem 1 are unknown, however we do not have to find these sequences since

$$\begin{aligned} Pr\Big\{\frac{M_n - b_n}{a_n} \leq z\Big\} &\approx G(z) \\ Pr\{M_n \leq z\} &\approx G\Big(\frac{z - b_n}{a_n}\Big) = G^*(z), \end{aligned} \tag{5}$$

where the function $G^*(z)$ is also a member of the GEV family, however with different parameters μ, σ and ξ.

The parameters of the GEV model are the location μ, scale σ and shape ξ. The shape parameter decides the tail behaviour of the GEV distribution. If the parameter ξ is zero, positive or negative then the GEV distribution function is the Gumbel, Fréchet or Weibull distribution functions respectively. The Fréchet distribution has a finite lower end-point and the Weibull distribution has a finite upper end-point. The Gumbel distribution has lighter tail than the Fréchet distribution as the tail decays exponentially rather than polynomially. The upper tail of the distribution is the point of interest in extreme value modelling, the finite upper end-point of the Weibull distribution could lead to some non-physical results as an upper limit usually does not exist in practical applications. The GEV probability distribution functions for ξ negative, zero and positive are shown in Figure 9.

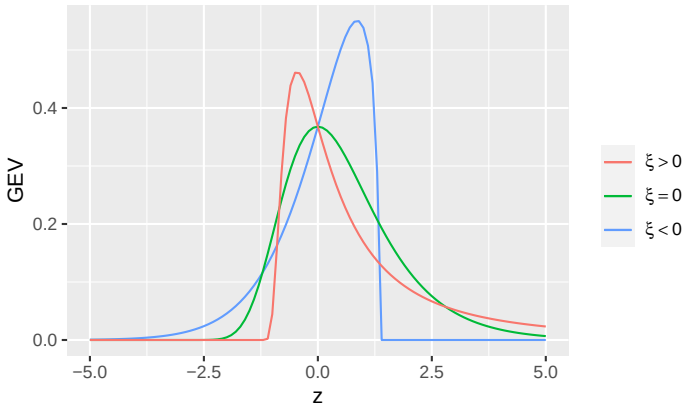


Figure 9: The GEV probability density function with $\mu = 0$, $\sigma = 1$ and $\xi = \{-0.75, 0, 0.75\}$.

The location and scale parameters μ and σ have no simple interpretation in terms of the phenomenon under study. We adopt the reparametrization of the GEV distribution proposed by Castro-Camilo et al. (2021), for the purpose of obtaining parameters that are easier to interpret. The reparametrization substitute μ and σ with the parameter q_α linked with a quantile and s_β linked with the difference between two quantiles. The biggest advantage of the reparametrization is that the parameters get a physical meaning, which means that it is easier to assign priors and the parameters are easier to interpret.

The parameter q_α can be defined as

$$\Pr\{Z \leq q_\alpha\} = \alpha, \tag{6}$$

where the location parameter is the quantile q_α related to the probability α. The spread parameter s_β can be defined as

$$s_\beta = q_{1-\beta/2} - q_{\beta/2}, \tag{7}$$

where $q_{1-\beta/2}$ and $q_{\beta/2}$ are quantiles related to the probabilities $1 - \beta/2$ and $\beta/2$. The spread parameter is therefore a representation of the width of the distribution.

The GEV distribution described by Equation (4) is a cumulative density function which means that the following relation holds

$$\mathrm{GEV}(z|\mu,\sigma,\xi) = \Pr\{Z \leq z\}. \tag{8}$$

The following equation for the quantile q_α of the GEV distribution can be derived from Equations (4), (6) and (8).

$$\begin{aligned} \mathrm{GEV}(q_\alpha|\mu,\sigma,\xi) &= \Pr\{Z \leq q_\alpha\} = \alpha \\ \exp\Big\{ -\Big[1+\xi\Big(\frac{q_\alpha-\mu}{\sigma}\Big)\Big]^{-1/\xi}\Big\} &= \alpha \\ \xi(\frac{q_\alpha-\mu}{\sigma}) &= ((-\log(\alpha))^{-\xi}-1) \\ q_\alpha &= \mu + \frac{\sigma}{\xi}(l_{\alpha,\xi}-1), \end{aligned} \tag{9}$$

where we define $l_{\gamma,\xi} = (-\log(\gamma))^{-\xi}$. The definition of s_β can be derived from Equations (7) and (9).

$$\begin{aligned} s_\beta &= q_{1-\beta/2} - q_{\beta/2} \\ &= \mu + \frac{\sigma}{\xi}(l_{1-\beta/2,\xi}-1) - \mu - \frac{\sigma}{\xi}(l_{\beta/2,\xi}-1) \\ &= \frac{\sigma}{\xi}(l_{1-\beta/2,\xi} - l_{\beta/2,\xi}) \end{aligned} \tag{10}$$

We also want functions for μ and σ given parameters q_α and s_β, for the purpose of defining the reparametrized GEV function. There is a one-to-one mapping between (μ,σ,ξ) and (q_α, s_β, ξ) (Castro-Camilo et al., 2021). Using Equations (9) and (10) we get

$$\sigma = \frac{s_\beta \xi}{(l_{1-\beta/2,\xi} - l_{\beta/2,\xi})} \tag{11}$$

and

$$\begin{aligned} \mu &= q_\alpha - \frac{\sigma}{\xi}(l_{\alpha,\xi}-1) \\ &= q_\alpha - \frac{s_\beta(l_{\alpha,\xi}-1)}{(l_{1-\beta/2,\xi} - l_{\beta/2,\xi})}. \end{aligned} \tag{12}$$

The reparametrized GEV distribution can then be found from Equations (4), (11) and (12).

$$\mathrm{GEV}(z|q_\alpha, s_\beta, \xi) = \exp\Big\{ -\Big[\Big(\frac{z-q_\alpha}{s_\beta(l_{1-\beta/2,\xi} - l_{\beta/2,\xi})^{-1}} + l_{\alpha,\xi}\Big)\Big]_+^{-1/\xi}\Big\}, \tag{13}$$

where $a_+ = \max(a, 0)$.

3.1.2 Spatial Model

The spatial model is defined for the purpose of improving the estimates at the different locations by borrowing information by nearby stations. The model can also be used to try to estimate the parameters in areas with no available data. For this purpose we assume that the parameters of the GEV distribution either can be explained by a Gaussian random field dependent on information available in the entire area or are constant over the field.

The shape parameter ξ is assumed to be constant in the entire area as Beck et al. (2020) argues that there is no statistical basis to support that the parameter varies spatially. We are also studying a small area with very few stations, which means that it might be best to have a simple shape parameter (Cooley et al., 2007). The parameter ξ is quite difficult to estimate due to the GEV distribution as a function of ξ being on a complicated form (Dyrrdal et al., 2015). A transformation of the parameter is used in the model that restricts the parameter to exist on the interval $[-0.5, 0.5]$, since the parameter is not expected to diverge far from 0. We have

$$\xi = a + (b - a) \cdot \frac{\exp\{\tilde{\xi}\}}{1 + \exp\{\tilde{\xi}\}}, \tag{14}$$

where $\tilde{\xi}$ is a hyperparameter of the model, $a = -0.5$ and $b = 0.5$.

The parameters q and $\ln s$ are assumed to vary spatially and are therefore explained with Gaussian random fields (Beck et al., 2020). We assume that the transformed parameter $\ln s$ is explained by a Gaussian random field rather than the parameter s, this choice is made due to the need of keeping the parameter s positive. The parameter s needs to be positive because it describes the width of the GEV distribution. We assume that the mean of the Gaussian fields are dependent on spatial covariates, while the variances are explained with the help of correlation functions dependent on distances between locations. We model $\mathbf{q}$ and $\mathbf{lns}$ as

$$\begin{aligned} q_{\alpha,i} &= \mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{q}} + u_i \\ \ln s_{\beta,i} &= \mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{s}} + w_i \end{aligned} \tag{15}$$

where $\mathbf{x}_i$ is the covariate vector at location i, it consists of an intercept and the four covariates described in Section 2.2. $\boldsymbol{\beta}_q$ and $\boldsymbol{\beta}_s$ are parameter vectors to be estimated that together with the covariates decides the mean of the parameters q and $\ln s$ in every point i. u_i and w_i are the values of the Gaussian random fields with mean 0 at location i. The Gaussian random field have specified covariance matrices $\boldsymbol{\Sigma}_q$ and $\boldsymbol{\Sigma}_s$. That is

$$\begin{aligned} \mathbf{u} &\sim \text{normal}(0, \boldsymbol{\Sigma}_q) \\ \Sigma_{q,ik} &= \tau_q^2 \cdot \Sigma(d_{ik}|\nu, \rho_q) \\ \mathbf{w} &\sim \text{normal}(0, \boldsymbol{\Sigma}_s) \\ \Sigma_{s,ik} &= \tau_s^2 \cdot \Sigma(d_{ik}|\nu, \rho_s) \end{aligned} \tag{16}$$

where d_{ik} is the distance between the two locations i and k, the parameters τ_q, τ_s are the standard deviation of the Gaussian random fields and $\Sigma(d|\nu, \rho)$ is the Matern correlation function with ranges ρ_q and ρ_s. The Matern correlation function has two parameters, a

measure of differentiability ν and a range parameter ρ that measures how fast the correlation of the random field decays with distance (Stein, 1999). The Matern correlation function is defined as

$$\Sigma(d|\nu,\rho) = \frac{2^{1-\nu}}{\Gamma(\nu)}\Big(\frac{(8\nu)^{1/2}d}{\rho}\Big)^{\nu}\mathcal{K}_{\nu}\Big(\frac{(8\nu)^{1/2}d}{\rho}\Big), \tag{17}$$

where $\mathcal{K}_{\nu}$ is the modified Bessel function of the second kind of order ν, d is the distance between two points and ρ is the distance for which the correlation function has value approximately equal to 0.13 (Lindgren and Rue, 2015). The measure of differentiability ν is difficult to estimate, and the parameter is therefore assumed to be 1 which is a common choice for the parameter (Lindgren et al., 2011). The correlation function is plotted using $\nu = 1$ for a series of different ranges in Figure 10.

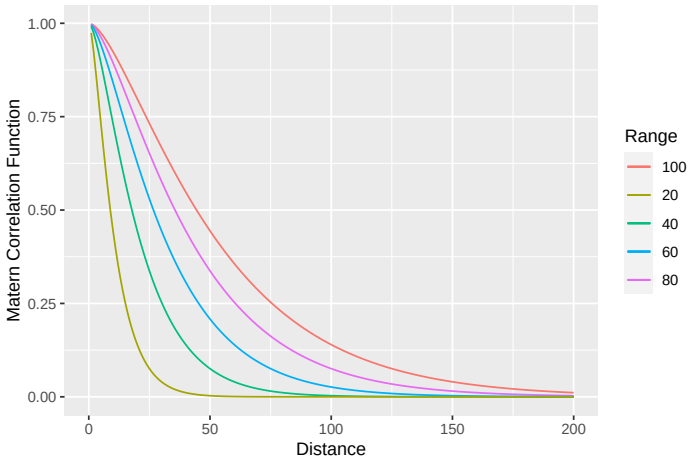


Figure 10: Matern correlation function using parameter $\nu = 1$ and range parameters 20, 40, 60, 80 and 100.

3.1.3 Specification of Priors

We need to assign prior distributions to the hyperparameters to fully specify the model in the Bayesian setting. The parameters that need priors are the regression parameters β_q and β_s, the standard deviations of the Gaussian random fields τ_q and τ_s, the ranges of the correlation functions ρ_q and ρ_s and the parameter $\tilde{\xi}$ used in the transformation of the shape parameter ξ. These prior distributions can be based on subjective belief, they can be based on previous data and analyses or they can be chosen to have very little impact on the full distribution. If there is not much prior information to base the priors on then one needs to make sure that the chosen priors does not have a strong effect on the posterior distribution (Givens and Hoeting, 2013). That is we do not want the posterior to be sensitive to the prior if it is not based on known information. One method one can use to make the posterior less sensitive is to choose priors that are quite wide with relatively low probabilities. However, one should not choose the priors too wide as this will cause the parameter estimation algorithm to run slowly (Stan Development Team, 2019). In our analysis we use slightly informative priors that is we want

the priors to guide the distribution somewhat, but not constrain it too much.

The prior of $\tilde{\xi}$ is chosen to be normally distributed with mean 0 and standard deviation 10.

$$\tilde{\xi} \sim \text{normal}(0, 10) \tag{18}$$

We can find the associated distribution function for ξ by applying the transformation presented in Equation (19) to the values of $\tilde{\xi}$.

$$\xi = a + (b - a) \cdot \frac{\exp\{\tilde{\xi}\}}{1 + \exp\{\tilde{\xi}\}}, \tag{19}$$

The resulting distribution function is presented in Figure 11.

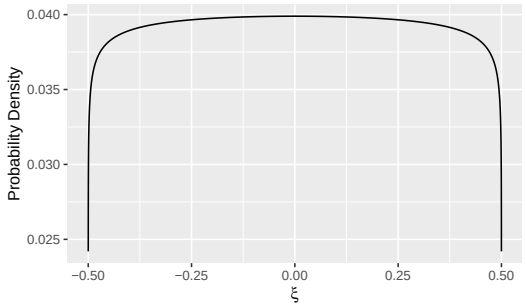


Figure 11: The distribution function of the parameter ξ associated with the prior of the parameter $\tilde{\xi}$.

The priors of $\boldsymbol{\beta}_q$ and $\boldsymbol{\beta}_s$ are chosen to be normally distributed with large standard deviations since we have very little information about which values to expect for the parameters.

$$\begin{aligned} \beta_q &\sim \text{normal}(0, 100) \\ \beta_s &\sim \text{normal}(0, 100) \end{aligned} \tag{20}$$

The priors of the standard deviation τ_q and τ_s are chosen to be gamma distributed with the mode at a low value while covering a decently large range of values. The reasoning behind this choice is that the standard deviation needs to be positive and we expect it to be reasonably small, due to the maximum sea levels not varying too much in size, however we do not have any tangible information describing which value to expect. The mean of the gamma distribution is 5 and the variance is 20. The priors are

$$\begin{aligned} \tau_q &\sim \text{gamma}(5/4, 1/4) \\ \tau_s &\sim \text{gamma}(5/4, 1/4). \end{aligned} \tag{21}$$

The priors of the ranges of the Matern correlation fields ρ_q and ρ_s are chosen to be gamma with modes at approximately half of the maximum distance between the locations studied and the distributions are chosen so that they cover the area slightly past the maximum distance

between locations (Sara Martino, personal communication). The maximum distance between locations is 113 km. The mean of the gamma distribution is 60 and the variance is 720. The priors are

$$\begin{aligned} \rho_q &\sim \text{gamma}(5, 1/12) \\ \rho_s &\sim \text{gamma}(5, 1/12) \end{aligned} \tag{22}$$

The standard deviation and range priors are shown in Figure 12.

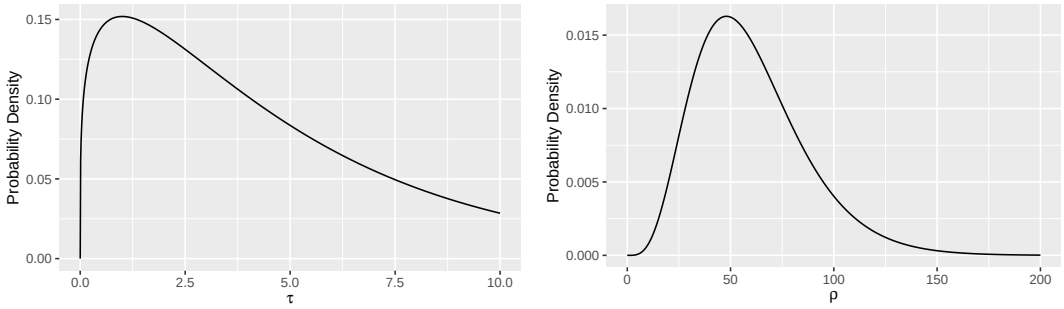


Figure 12: The priors of the standard deviation τ_q and τ_s on the left and the priors of the range ρ_q and ρ_s on the right

3.2 Parameter Estimation

Inference is in the Bayesian setting based on the posterior distribution of the model parameters given the data. The posterior distribution is in our case

$$\begin{aligned} p(\mathbf{q}, \ln \mathbf{s}, \tilde{\xi}, \beta_q, \beta_s, \tau_q, \tau_s, \rho_q, \rho_s | \mathbf{z}) \propto\ & p(\mathbf{z}|\mathbf{q}, \ln \mathbf{s}, \tilde{\xi}) \\ & \cdot p(\mathbf{q}|\beta_q, \tau_q, \rho_q) \\ & \cdot p(\ln \mathbf{s}|\beta_s, \tau_s, \rho_s) \\ & \cdot p(\tilde{\xi}) p(\beta_q) p(\beta_s) p(\tau_q) p(\tau_s) p(\rho_q) p(\rho_s). \end{aligned} \tag{23}$$

where the first line represents the data layer, which is described by the multivariate GEV distribution. The second and third line represents the Gaussian random fields describing the spatial behaviour of parameters q and $\ln s$. The fourth line represents the priors of the hyperparameters.

This distribution is not in closed form and cannot be analysed analytically. We are therefore using using Bayesian estimation, more specifically we are performing Markov Chain Monte Carlo (MCMC). A disadvantage about MCMC is that it is a computationally intensive approach. We are however using the package STAN in R, STAN is a program built for performing efficient MCMC.

3.2.1 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) is one of multiple methods that can be used to estimate the parameters of an arbitrary probability distribution function. MCMC is an iterative method for drawing parameter values from a distribution function. Parameter values are proposed based on previous iterations and rejected or accepted based on information from the probability distribution function (Gelman et al., 2018). The parameter estimates only depend on the most recent draw, which means the result is a Markov chain. The Markov Chains are expected to become accurate estimates of the posterior distribution over time, as long as the Markov process created has the posterior distribution as its stationary distribution. If the proposal distribution is chosen so that the Markov Chain is aperiodic and irreducible, then there is a unique limiting stationary distribution for the Markov Chain (Givens and Hoeting, 2013).

If the MCMC algorithm is to be efficient, then the proposal distribution needs to return values from the entire target distribution (Givens and Hoeting, 2013). If the proposal distribution is too wide, then the proposed values will be rejected frequently and if the proposal distribution is no narrow then the entire target distribution might not be explored. The resulting parameter chains in these two cases will be poor representations of the target distribution. Other criteria to follow when choosing the proposal distribution is to make sure that sampling is easy and that the proposals travel reasonable distances in parameter space (Gelman et al., 2018).

The resulting chains can be quite dependent on the starting value in the first iterations, it can therefore be a good idea to exclude a number of iterations in the beginning of the chains (Givens and Hoeting, 2013). This starting period is called the burn-in period. It can also be a good idea to start the algorithm from different starting points to ensure that it has converged. Another potential issue with iterative simulation is correlation between the samples (Gelman et al., 2018). This can be an issue since correlated samples contain less information than independent samples and the correlation might cause the algorithm to be inefficient.

3.2.2 Hamiltonian Monte Carlo (HMC)

The MCMC algorithm can be quite inefficient due to the proposal values being chosen randomly (Gelman et al., 2018). The situation can be improved with reparametrization and efficient jumping rules, however the problem remains for high-dimensional target distributions. Hamiltonian Monte Carlo (HMC) solves this problem by choosing the proposed steps based on the geometry of the target density function rather than them being chosen randomly (Betancourt, 2018). Hamiltonian Monte Carlo is based on Hamiltonian dynamics, which explains motions of particles in an energy field with potential energy based on the negative log posterior distribution and kinetic energy chosen to keep the particles within the desired field. The following description of HMC is based on the paper A Conceptual Introduction to Hamiltonian Monte Carlo by Betancourt (2018).

The HMC method is based on differential geometry, which is a complex subject (Betancourt, 2018). However, differential geometry also explains classical physics. This means that there are analogous physical systems to our probabilistic systems, which can give us some intuition.

A physical system that can be used as an analogy for the probabilistic system is the system of an object in orbit around a planet with a gravitational field. The reason for this system being analogous is the introduction of the typical set. The typical set is the region of the target density function that contains the significant contributions to expectations (Betancourt, 2018). This means that the typical set is the region of the density function, from which the proposal values should be picked, when the goal is an efficient MCMC algorithm. The typical set can be explained as an analogy to the region that the object in orbit needs to stay within to stay in orbit. If the object deviates from this area then it will either fall to earth or disappear into space. The typical set can be located by studying the density function in large dimensions. The expectation of a function $g(x)$ is given as

$$E(g(x)) = \int g(x)\pi(x)dx, \tag{24}$$

where $\pi(x)$ is the probability density function. The region close to the mode of the distribution has large values of the density function $\pi(x)$, however the volume of the region is very small when the dimensions of the probability distribution is large. This means that the contribution to the expectation is very small. The farther away from the mode the region is, the smaller the value of the density function $\pi(x)$, this means that the contribution to the expectation is very low for the region even though the volume of the region is very large.

We want to pick proposal values within the typical set. This can be achieved by explaining the system with Hamiltonian equations (Betancourt, 2018). The Hamiltonian is the analogue to energy, it can be divided into kinetic and potential energy. The potential energy is a function of the target distribution while the kinetic energy is a function of auxiliary momentum variables. The auxiliary momentum is introduced in the algorithm, it is important for keeping the object in orbit, without momentum the object would fall down to the planet, with to much momentum the object would disappear into space. The momentum is decided during the warm-up period (Stan Development Team, 2019).

The Hamiltonian and the Hamiltonian equations are defined by Betancourt (2018) and Stan Development Team (2019). The joint density of the auxiliary momentum variables p and the parameter variables q is given as

$$\pi(q, p) = \pi(p|q)\pi(q). \tag{25}$$

The joint density is used to define the Hamiltonian

$$\begin{aligned} H(q, p) &= -\log \pi(q, p) \\ &= -\log \pi(p|q) - \log \pi(q) \\ &= K(p, q) + V(q). \end{aligned} \tag{26}$$

Where $K(p, q) = -\log \pi(p|q)$ is called the kinetic energy and $V(q) = -\log \pi(q)$ is called the potential energy. The new proposal values are found by first drawing a momentum. The drawn momentum value and the current parameter values define the starting position of the

iteration. The vector field that explains the movement for the proposal value can be described by the Hamiltonian equations.

$$\begin{aligned}\frac{dq}{dt} &= \frac{\partial H}{\partial p} = \frac{\partial K}{\partial p} \\ \frac{dp}{dt} &= -\frac{\partial H}{\partial q} = -\frac{\partial K}{\partial q} - \frac{\partial V}{\partial q}.\end{aligned} \tag{27}$$

The Hamiltonian equations then need to be solved to specify the new parameter estimates.

3.2.3 STAN

Programming Hamilton Monte Carlo is a quite tedious process as quite a large amount of details need to be implemented (Gelman et al., 2018). However, HMC is already implemented in STAN, all one needs to do is specify the likelihood of the data and the probability density functions of all the parameters including the priors, the remainder of the process is automated.

A momentum needs to be chosen to be able to solve the Hamiltonian equations. In STAN this momentum is chosen to be independent of the current parameter values (Stan Development Team, 2019), the distribution of the momentum is chosen to be

$$p \sim MVN(0, M), \tag{28}$$

where MVN is the multivariate normal distribution and M is the Euclidean metric. The Hamiltonian equations can in this case also be written as

$$\begin{aligned}\frac{dq}{dt} &= \frac{\partial K}{\partial p} \\ &= \frac{d}{dp}\Big(- \log\Big((2\pi)^{-k/2}\det(M)^{-1/2}\exp\Big(- \frac{1}{2}p^T M^{-1} p\Big)\Big)\Big) \\ &= \frac{d}{dp}\Big(\frac{1}{2}p^T M^{-1} p\Big) \\ &= M^{-1}p \\ \frac{dp}{dt} &= -\frac{\partial V}{\partial q},\end{aligned} \tag{29}$$

due to the momentum being independent of current parameter values. These equations are solved in STAN using the leapfrog integrator, which gives stable results when solving Hamiltonian equations (Stan Development Team, 2019). The leapfrog integrator alternates between updating the position and the momentum for small time intervals ϵ. The leapfrog algorithm alternates between making half-step updates of the momentum and full-step updates of the position.

$$\begin{aligned}p &= p - \frac{\epsilon}{2}\frac{dV}{dq} \\ q &= q + \epsilon M^{-1} p \\ p &= p - \frac{\epsilon}{2}\frac{dV}{dq}\end{aligned} \tag{30}$$

The integrator has a global numerical error of ϵ^2 (Stan Development Team, 2019). A acceptance step is performed to account for the numerical error. The acceptance probability is given as

$$\min(1, \exp(H(p, q) - H(p^*, q^*))), \tag{31}$$

where (p^*, q^*) is the proposal and (p, q) is the initial value. If the step is accepted then the proposal value is used as initial value in the next iteration, if the step is not accepted then the previous initial value is reused.

There are three parameters that need to be specified in the HMC algorithm: the time step ϵ, the Euclidean metric M and the number of steps taken by the leapfrog algorithm L (Stan Development Team, 2019). STAN optimizes the time step ϵ so that the preferred acceptance rate is achieved, the Euclidean metric M is estimated from the warm-up iterations and the no-U-turn sampling algorithm (NUTS) is used to find the correct number of steps for the leapfrog algorithm L (Stan Development Team, 2019).

STAN runs four parallel Markov Chains for the purpose of checking convergence, if all four converge to the same parameter distributions then one can be quite sure that the algorithm has converged (Gelman et al., 2018). A chosen percentage of the iterations are discarded as a warm-up period due to these parameter estimates being dependent on initial values in addition to the initial iterations being used to specify the metric M.

3.2.4 No-U-Turn Sampling (NUTS)

The no-U-turn sampling algorithm (NUTS) is used to optimize the number of steps L taken in the HMC algorithm. If the number of steps L is too large then computation time is wasted and if the number of steps L is too small then the algorithm behaves similarly to a random walk algorithm (Hoffman and Gelman, 2014). The NUTS algorithm performs HMC steps both forwards and backwards in time making a binary tree, and in each step the tree depth is increased (Stan Development Team, 2019). The algorithm stops iterating when the chain of proposal values starts doubling back on itself or the maximum tree depth is reached (Hoffman and Gelman, 2014)(Stan Development Team, 2019). The algorithm does not allow u-turns, thus the name. The proposal value chosen is sampled with multinomial sampling with bias towards the later iterations of the NUTS algorithm, this is done for the purpose of obtaining longer steps (Stan Development Team, 2019).

The implementation of the NUTS algorithm, together with the automatic specification of the Euclidean metric and the time step size, makes the STAN program completely automated, that is no hand tuning of parameters is required.

3.3 Application of Parameters

The parameter estimates are used to specify the probability distribution function so that we can use the distribution to obtain properties of interest. The properties that are of most interest in extreme value analysis are return levels with associated return periods. We are in this project interested in being able to calculate the return levels at the locations: Helgeroa,

Oscarsborg, Oslo and Viker. We are also interested in trying to specify the return levels at any other location in the Oslo fjord.

In the first subsection return levels are defined and the procedure to obtain return level estimates is described. In the second subsection the procedure to obtain spatial parameter estimates from areas without available data is described.

3.3.1 Return Levels

The return level z and return period m are properties of interest due to their practical application as descriptions of the expected sea level exceedance z every m years. The return level and return period estimates can be obtained by studying the probability that the yearly maximum values are larger than z, that is $Pr\{M_n > z\} = p$, and the associated value of z (Coles, 2001). z is the return level and $m = \frac{1}{p}$ is the return period. The return period can be understood as the expected number of years between each exceedance of z. According to Theorem 1 we have that

$$Pr\Big\{M_n \leq z\Big\} \to \text{GEV}(z|\mu, \sigma, \xi) \tag{32}$$

as $n \to \infty$. Thus, we get the equation

$$1 - \frac{1}{m} \approx \text{GEV}(z|\mu, \sigma, \xi) \tag{33}$$

for the dependence between the return level z and the return period m, given the parameters μ, σ and ξ (Coles, 2001). We need to use the relation between parameters $\{\mu, \sigma\}$ and $\{q, s\}$ to find the estimates due to the reparametrization of the model. The reparametrization formulas are defined in Section 3.1.1.

The parameter estimates obtained from the MCMC algorithm are samples that can be used to estimate the probability density of the parameters. An estimation procedure for the return levels is needed since we have multiple values for each parameter. We first divide the parameter samples into sets based on iteration number, so that each parameter set contains one estimate per parameter. Then we choose a return period and solves Equation (33) for each parameter set. This gives us samples of the return level for the return period that we chose. The set of return level estimates can then be used to determine the approximate probability distribution for the return level, which means that we can find approximations for properties such as the median and the credible interval.

3.3.2 Spatial Analysis

A possible application of the parameter estimates is to try to find parameter estimates in areas where there are no available sea level data. If parameter estimates can be found then return level estimates can be found as well. The algorithm used for obtaining estimates for the parameter q of the GEV model is presented as Algorithm 1. The same algorithm can be

used to find estimates for $\ln s$ by replacing the parameters related to q with the equivalent parameters related to s. The algorithm is in practice conditional multivariate normality.

Algorithm 1: Spatial Interpolation Algorithm (Beck et al., 2020)

Result: Parameter estimates for q at new locations
Let $S = \{s_0, ..., s_l\}$ be the existing locations
Let $S^* = \{s_0^*, ..., s_k^*\}$ be new locations of interest
for *i from 1 to number of parameter samples* **do**
The joint distribution of $\mathbf{q} = (q_1, ..., q_{l+k})$ can be written as:

$$\mathbf{q}_{S\cup S^*} \sim \text{normal}_{l+k}(\mathbf{X}_{S\cup S^*}\boldsymbol{\beta}_{q_i}, \tau_{q_i}^2 \boldsymbol{\Sigma}_{\mathbf{q}_{S\cup S^*}}),$$

where $\mathbf{X}_{S\cup S^*}$ is the covariate matrix with 5 columns and $l+k$ rows. $\boldsymbol{\beta}_{q_i}$ and τ_{q_i} represents the i-th parameter samples obtained for $\boldsymbol{\beta}_q$ and τ_q. The correlation matrix is given as

$$\boldsymbol{\Sigma}_{\mathbf{q}_{S\cup S^*}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{q}_S} & \boldsymbol{\Sigma}_{\mathbf{q}_{S,S^*}} \\ \boldsymbol{\Sigma}_{\mathbf{q}_{S^*,S}} & \boldsymbol{\Sigma}_{\mathbf{q}_{S^*}} \end{bmatrix}$$

where the entries of the matrix are given by the Matern correlation function with range ρ_{q_i}. The matrix $\boldsymbol{\Sigma}_{\mathbf{q}_{S,S^*}}$ has l rows and k columns and the matrix $\boldsymbol{\Sigma}_{\mathbf{q}_{S^*,S}}$ is the transpose.
Sample parameters for the new locations of interest given the the parameter samples obtained for the existing locations. That is, sample $\mathbf{q}_{S_i^*}$ given $\mathbf{q}_{S_i}$:

$$\mathbf{q}_{S_i^*}|\mathbf{q}_{S_i} \sim \text{normal}_k(\hat{\boldsymbol{\mu}}, \tau_{q_i}^2 \hat{\boldsymbol{\Sigma}})$$

where

$$\hat{\mu} = X_{\mathbf{q}_{S^*}}\beta_{q_i} + \boldsymbol{\Sigma}_{\mathbf{q}_{S^*,S}}\boldsymbol{\Sigma}_{\mathbf{q}_{S,S}}^{-1}(\mathbf{q}_{S_i} - X_{\mathbf{q}_S}\beta_{q_i})$$

and

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{\mathbf{q}_{S^*,S^*}} - \boldsymbol{\Sigma}_{\mathbf{q}_{S^*,S}}\boldsymbol{\Sigma}_{\mathbf{q}_{S,S}}^{-1}\boldsymbol{\Sigma}_{\mathbf{q}_{S,S^*}}.$$

end

3.4 Evaluating the Fit of the Model

An important step in the process of fitting a model is to check how well the model corresponds with the data (Gelman et al., 2018). One does not expect statistical models to be perfectly true, however one can study the model to check if there are any deficiencies in the model that impact practical applications and whether the prior assumptions seem reasonable (Gelman et al., 2018). In this section we will talk about how we can check the model in our specific case and the challenges involved.

One needs to check how sensitive the model is to the assumptions made. The priors can contain assumptions as the priors can be non-informative, weakly informative or informative. If they are weakly informative or informative then it is important to check to what extent the

priors affect the parameter estimates.

The practical application of highest interest is return levels. It is possible to make empirical return level plots directly from the data. This means that the correctness of the return level plots can be checked. However, the maximum return period that we can make empirical return level plots for are limited to the number of years that there is available data. This means that the comparison is going to be quite difficult at some locations since the empirical return levels are only available for very short periods of time, this is the case especially for Viker.

We also need to try to check the spatial model. Checking the spatial model is quite challenging, especially when one tries to make estimates in areas where no data is available to perform empirical model checks. We check the model in the locations where we have data by comparing the resulting parameter estimates and return levels to the equivalent estimates obtained with the univariate model. If one wants to check how the model does with spatial prediction, then one method that could be used to check this is cross-validation. That is one can leave one location out of the analysis and check whether the return levels and parameter estimates can be correctly obtained by only using the remaining locations in the analysis. This is however not feasible in our analysis as we only have four locations, which is too few to perform cross-validation.

An additional method used in this project to check the model is simulation studies. For each model fitted with real data an equivalent model with known parameter values will be fitted so that one can see firstly that the model works and secondly the best possible outcome of the analysis.

4 Simulation Study

The simulation studies are performed for the purpose of checking how the model performs when the results can be compared with known parameter values. The simulation studies can also give us an idea of which results to expect from the analysis of the sea level data, which is much needed since we have quite limited options when it comes to model checking. Additional results that can be obtained from a simulation study that cannot be obtained from the real data are the results of a model based on more than four locations. It can in particular be interesting to see whether the accuracy of the estimates increases when additional locations are added to the model. Generally when simulation studies are performed the simulations are rerun numerous times, this is however not feasible in these analyses as each one takes quite a long time to run.

We will in this section present the results from simulation studies based on 15, 4 and 8 locations. The first study is a multivariate GEV model based on fifteen locations, which should show how the model would have performed if more data was available. We are in this study checking two different priors for the range parameter ρ. The range parameter has been given an informative prior due to the parameter being quite difficult to estimate and it is therefore interesting to check the effect of the informative prior on the rest of the results. The second study is a multivariate GEV model based on four locations, which should give us an idea of what to expect when applying the model to the sea level data. We simulate data based on the multivariate model for four locations and use the data to perform three different analyses. First we apply a univariate model to the data from each location, this analysis is done so that we can use the results to check the performance of the multivariate model. Then we apply a model that assumes that the shape parameter of the GEV distribution is common across the locations while the other two parameters are independent in space, this analysis is done due to the shape parameter being quite difficult to estimate, so it is interesting to try to use all the available data to estimate the parameter. In the last analysis we apply the multivariate model, which is the main point of interest, to the data. The last study is a multivariate GEV model based on eight locations, where four locations have very short series of extreme values. This study is done because obtaining new sea level data is very time consuming, as a year of data is equivalent to one data point in the model, it is therefore interesting to check the effect of shorter series.

4.1 Analysis of 15 Locations

The locations for the first simulation study are shown in Figure 13. The first four locations are identical to the locations with available sea level data and the additional eleven stations are added sequentially using an algorithm that maximizes the minimum distance to any of the stations already added. This algorithm was selected based on the assumption that if a new station were to be added in the area, then it would be placed as far from existing stations as possible. A shortcoming of this reasoning is that population density probably will be an important factor when choosing a new location.

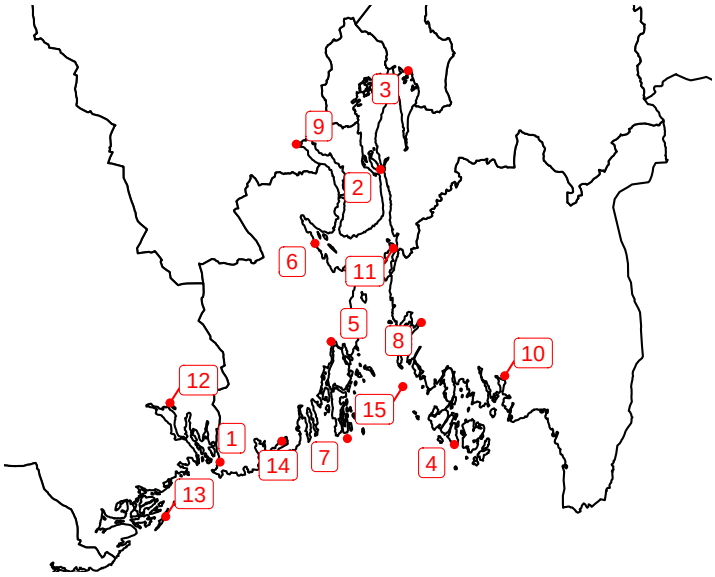


Figure 13: The locations used in the simulated multivariate GEV analysis for 15 stations. Station 1-4 are at the same locations as the stations with sea level data and the additional stations were chosen sequentially to be as far from existing stations as possible.

We restate the model defined in Section 3.1.2. We let z_{ij} be the j-th observed maximum sea level at location i. The model likelihood is

$$
\begin{aligned}
p(\mathbf{z}|\mathbf{q_\alpha}, \mathbf{\ln s_\beta}, \tilde{\xi}) &= \prod_{i=1}^{l}\prod_{j=1}^{n_i} \mathrm{GEV}(z_{ij}|q_{\alpha,i}, s_{\beta,i}, \xi) \\
\mathrm{GEV}(z_{ij}|q_{\alpha,i}, s_{\beta,i}, \xi) &= \exp\Big\{ -\Big[\Big(\frac{z_{ij}-q_{\alpha,i}}{s_{\beta,i}(l_{1-\beta/2,\xi}-l_{\beta/2,\xi})^{-1}} + l_{\alpha,\xi}\Big)\Big]_+^{-1/\xi}\Big\} \\
l_{\gamma,\xi} &= (-\log(\gamma))^{-\xi} \\
s_{\beta,i} &= \exp(\ln s_{\beta,i}) \\
\xi &= a + (b-a)\cdot\frac{\exp\{\tilde{\xi}\}}{1+\exp\{\tilde{\xi}\}}
\end{aligned}
\tag{34}
$$

where l is the number of locations and n_i is the number of data available in location i. q_α is the $\alpha-$quantile and s_β is the difference between the $1-\beta/2$ and $\beta/2$ quantiles. a and b are the lower and upper bounds of the shape parameter ξ. We set $a = 0$ and $b = 0.5$. We choose to look at the quantiles $\alpha = 0.5$ and $\beta = 0.05$ because the parameter q_α then represents the median while the parameter s_β represents the 95% credible interval of the GEV distribution.

Moreover we let

$$\begin{aligned} q_{\alpha,i} &= \mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{q} + u_i \\ \mathbf{u} &\sim \text{normal}(0, \boldsymbol{\Sigma}_q) \\ \Sigma_{q,ik} &= \tau_q^2 \cdot \Sigma(d_{ik}|\nu, \rho_q) \end{aligned} \tag{35}$$

$$\begin{aligned} \ln s_{\beta,i} &= \mathbf{x}_i^T \boldsymbol{\beta}_\mathbf{s} + w_i \\ \mathbf{w} &\sim \text{normal}(0, \boldsymbol{\Sigma}_s) \\ \Sigma_{s,ik} &= \tau_s^2 \cdot \Sigma(d_{ik}|\nu, \rho_s) \end{aligned}$$

where $\mathbf{x}_i$ is the covariate vector at location i consisting of an intercept and the covariates presented in Section 2.2. $\boldsymbol{\beta}_q$ and $\boldsymbol{\beta}_s$ are parameter vectors that together with the covariates decides the mean of the parameters q_α and $\ln s_\beta$ in every point i. u_i and w_i are the values of the Gaussian random fields at location i. $d_{ik} = |r_i - r_k|$ is the distance between two points in the Gaussian random field. The parameters τ_q, τ_s are the standard deviation of the Gaussian random fields and $\Sigma(d|\nu, \rho)$ is the Matern correlation function with ranges ρ_q and ρ_s. The parameter values that we have chosen are

Parameter	Value
ξ	-0.05
β_q	$90, 9, 12, -18, 15$
β_s	$4.3, 0.2, 0.1, -0.6, 0.2$
τ_q	0.3
τ_s	0.2
ρ_q	100
ρ_s	80

Table 3: The chosen parameter values for the simulation studies.

We sample simulated data to use in the analysis for each location i. The number of data points n_i for $i = 1, ..., 15$ is drawn uniformly between 50 and 100. In order to simulate data we first draw q_i and $\ln s_{\beta,i}$ $i = 1, ..., 15$ from Equation 35. Then for each station i we sample n_i data from the GEV distribution presented in Equation 34.

We use STAN to find samples of the parameters based on the posterior model, this model can be stated as follows

$$\begin{aligned} p(\mathbf{q}, \mathbf{ln\, s}, \tilde{\xi}, \beta_q, \beta_s, \tau_q, \tau_s, \rho_q, \rho_s|\mathbf{z}) \propto\ & p(\mathbf{z}|\mathbf{q}, \mathbf{ln\, s}, \tilde{\xi}) \\ & \cdot p(\mathbf{q}|\beta_q, \tau_q, \rho_q) \\ & \cdot p(\mathbf{ln\, s}|\beta_s, \tau_s, \rho_s) \\ & \cdot p(\tilde{\xi})p(\beta_q)p(\beta_s)p(\tau_q)p(\tau_s)p(\rho_q)p(\rho_s). \end{aligned} \tag{36}$$

The probability distributions representing the prior knowledge can be explained with the following equations

$$\begin{aligned}
\tilde{\xi} &\sim \text{normal}(0, 10) \\
\beta_q &\sim \text{normal}(0, 100) \\
\beta_s &\sim \text{normal}(0, 100) \\
\tau_q &\sim \text{gamma}(5/4, 1/4) \\
\tau_s &\sim \text{gamma}(5/4, 1/4) \\
\rho_q &\sim \text{gamma}(5, 1/12) \\
\rho_s &\sim \text{gamma}(5, 1/12)
\end{aligned} \tag{37}$$

4.1.1 Model Estimates

The obtained parameter estimates are shown in Figures 14, 15, 16, 17 and 18. One can see that all the histograms contain the chosen parameter values. The parameters that specify $\ln s$ in Figure 16 seem to have been estimated with more accuracy than the equivalent parameters for q in Figure 15. The Figures 14, 15 and 16 show that the priors are not noticeable except for the spatial parameters, which means that the priors should not have had much influence on the final parameter samples. The parameters of the Gaussian random fields are the ones that are expected to be hardest to estimate, due to the posterior distribution as a function of these parameters being on quite a complicated form. One can see that the standard deviations have deviated substantially from their priors while the ranges only have deviated slightly from their priors.

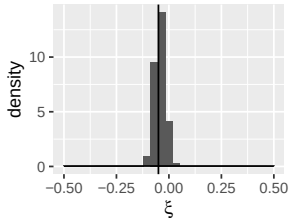


Figure 14: Histogram of the parameter ξ for 15 locations based on simulated data. The chosen parameter value is shown as a vertical line and the prior is also shown.

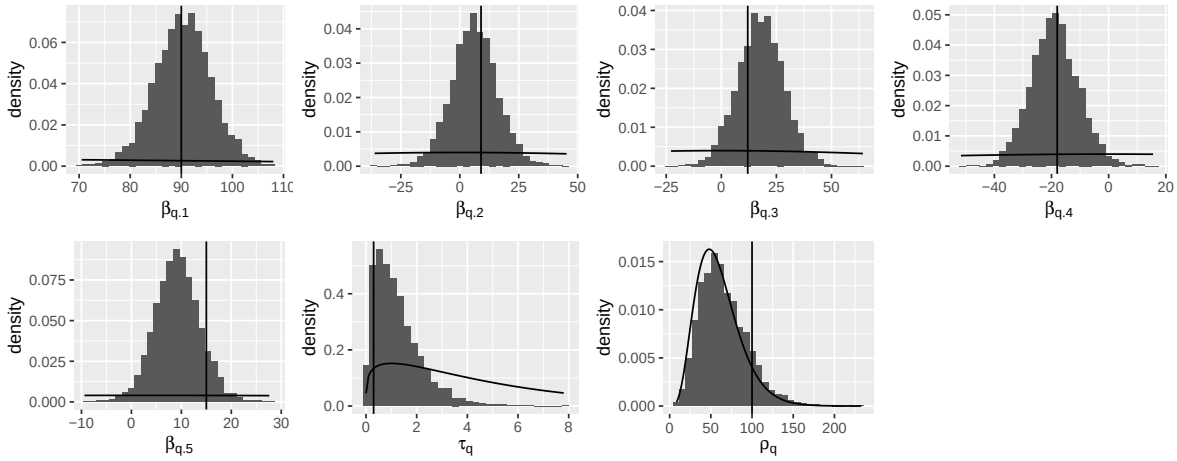


Figure 15: Histogram of the parameter estimates that specify the quantile q. The estimates are based on simulated data of 15 locations. The chosen parameter values are shown as vertical lines and the priors are also shown for all the parameters.

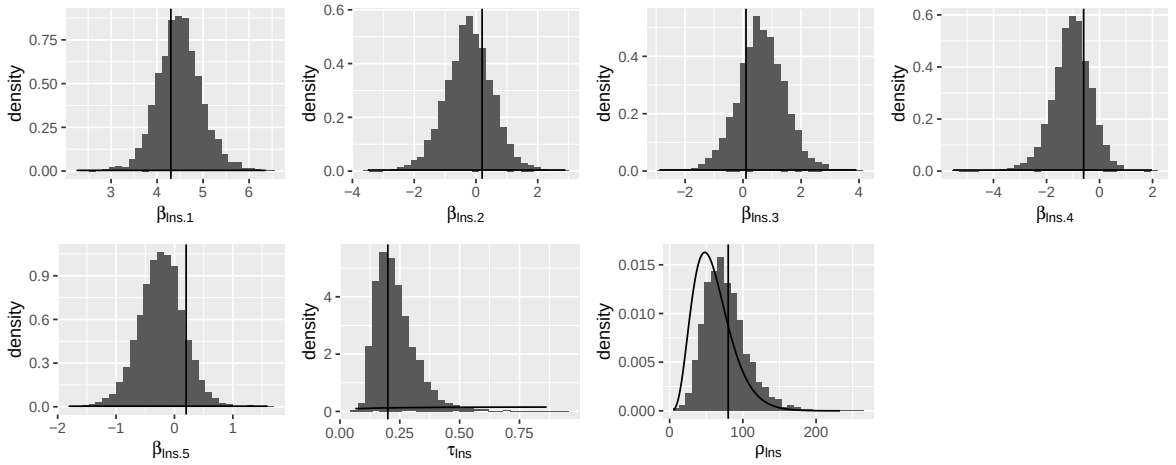


Figure 16: Histogram of the parameter estimates that specify $\ln s$. The estimates are based on simulated data of 15 locations. The chosen parameter values are shown as vertical lines and the priors are also shown for all the parameters.

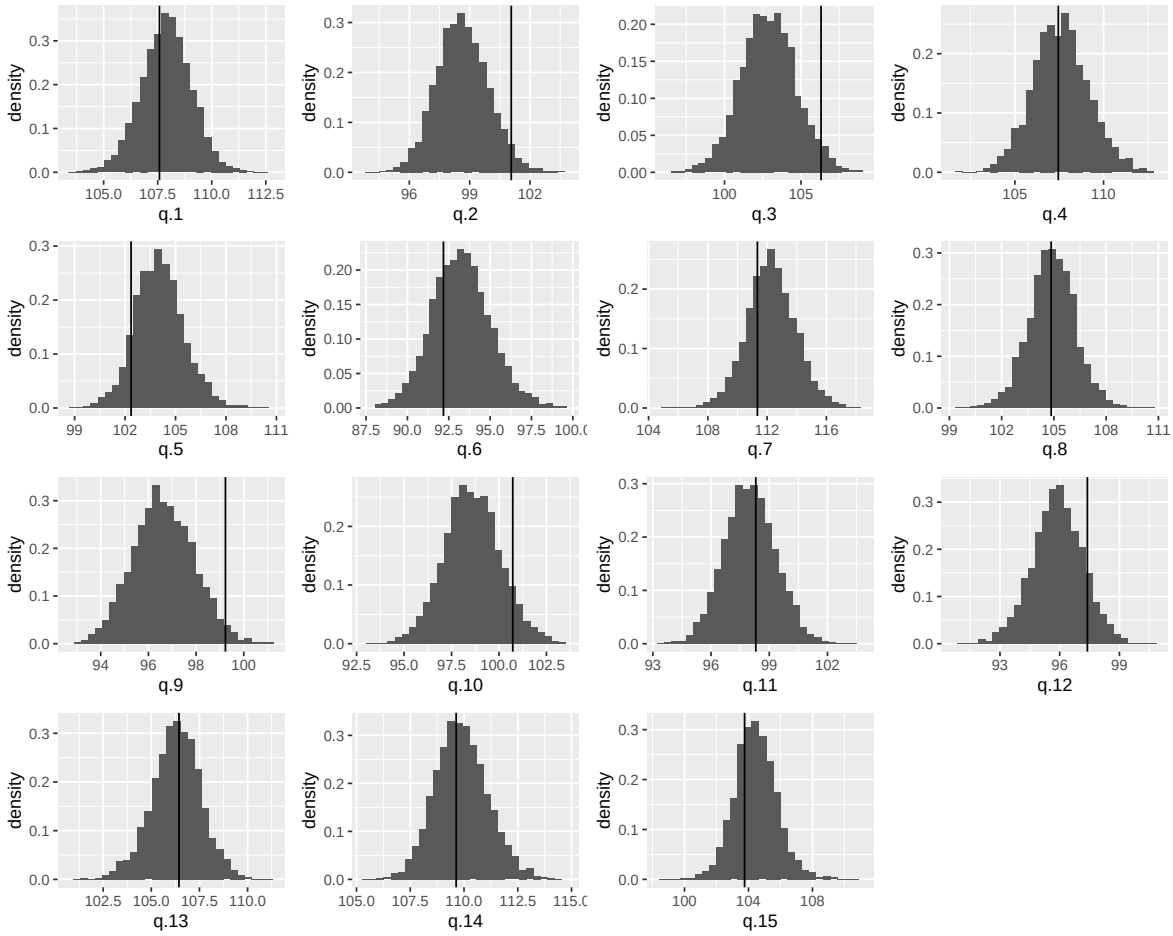


Figure 17: Histogram of the parameter estimates of q for 15 locations based on simulated data. The chosen parameter values are shown as vertical lines and the priors are also shown for all the parameters.

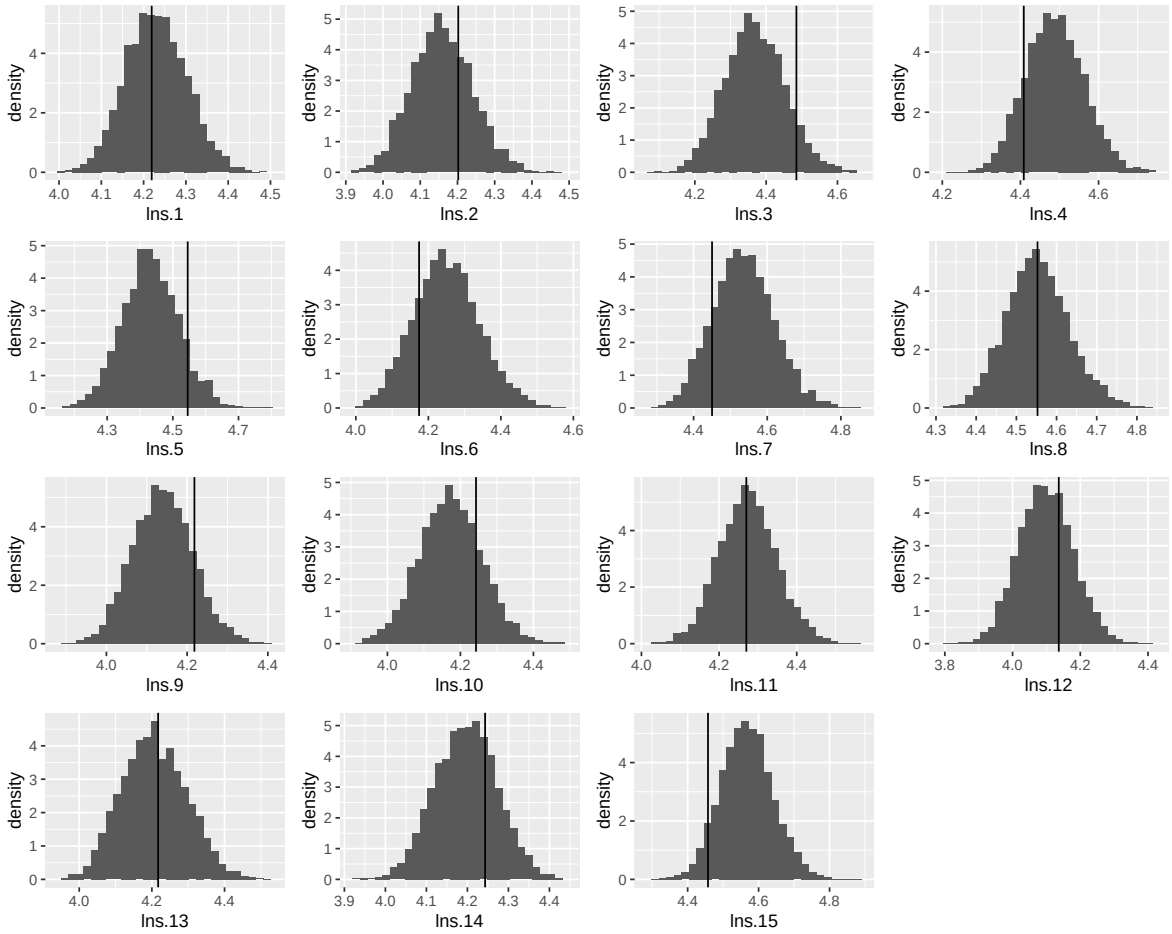


Figure 18: Histogram of the parameter estimates of $\ln s$ for 15 locations based on simulated data. The chosen parameter values are shown as vertical lines and the priors are also shown for all the parameters.

Return levels with associated return periods are properties of interest in this analysis, as they give us practical information about the expected frequency of the sea exceeding certain levels. The return levels and the methods used to obtain these are described in more detail in Section 3.3.1. Figure 19 shows the 95% credible interval and the median of the return levels plotted against the return periods for the 15 locations. The empirical return level values are also depicted. The empirical return levels are found by sorting the extreme data by size from largest to smallest $\{X_i, i \in [1, n]\}$ and the empirical return periods are $\{\frac{n}{i}, i \in [1, n]\}$. One can see that the uncertainty bounds have ranges that differ from about 25 cm to less than 50 cm. The empirical return level values are in most cases contained within the credible bounds, however some of the emirical return level points are clearly outside the bounds. This is not altogether unexpected due to the spatial fields leading to generalized results, which means that overall the fit of the model could improve while the model fit at some individual locations could worsen.

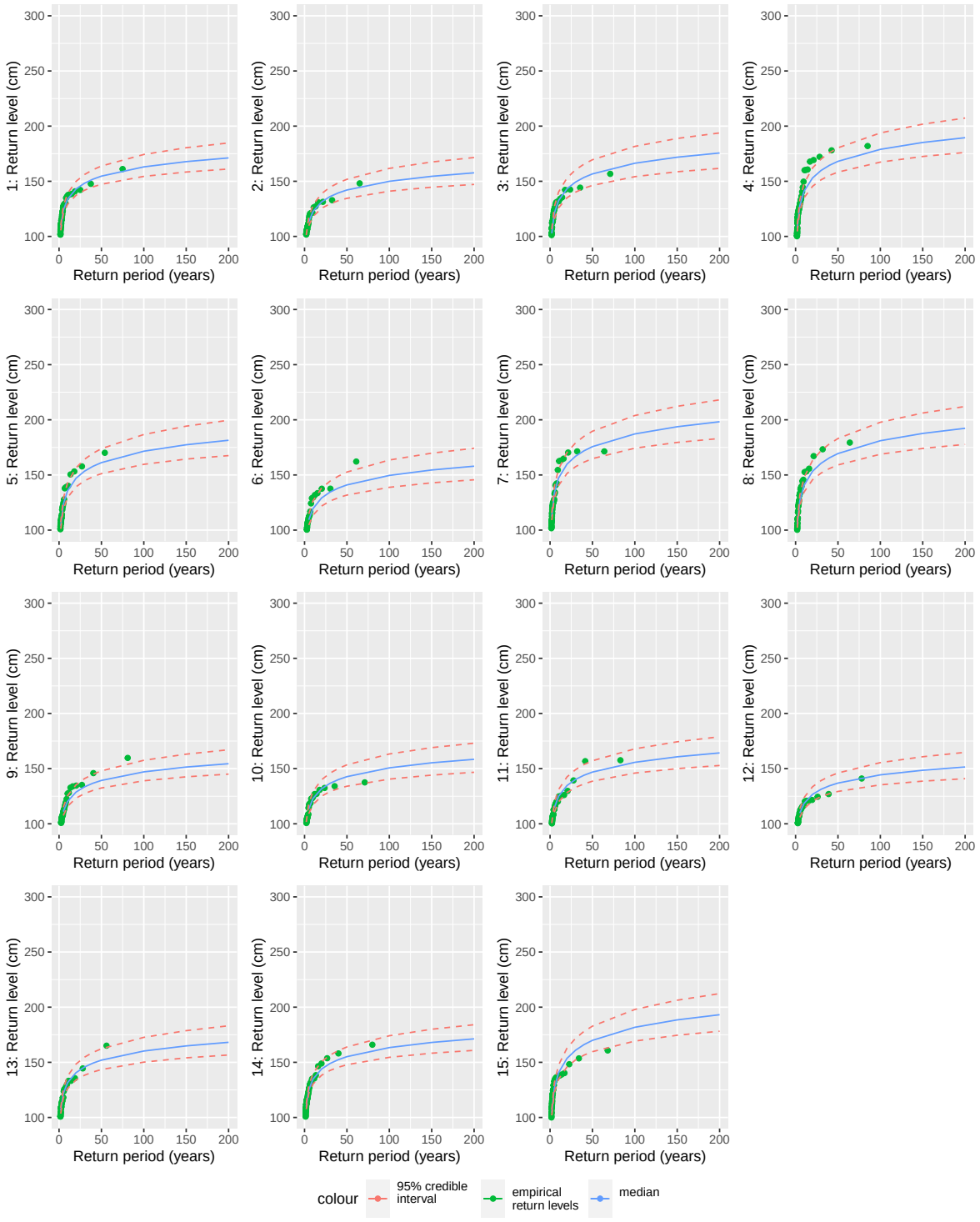


Figure 19: The return level plots of 15 locations based on simulated data. The plots contain the medians and the 95% credible intervals of the return levels plotted against the return periods. The empirical return levels based on the data are also shown.

Trying to extend the return level analysis to the entire Oslo fjord is another point of interest in this project. The procedure for obtaining estimates for the parameters q and $\ln s$ in areas where there are no data is explained in Section 3.3.2. The medians and the 95% credible intervals of the return levels associated with the 200 year return period for the entire Oslo fjord are presented in Figure 20. One can see that the largest difference between the lower and the upper bound is about 80 cm, however in most places the difference is closer to 40 cm. The location with the largest uncertainty bound is the extremely narrow area leading in to location 9. The results could indicate that having 15 stations in the area could give return levels with quite narrow uncertainty bounds in the majority of the area. However, one needs to keep in mind that this result is based on simulated data, real sea level data might not be as good of a fit to the model.

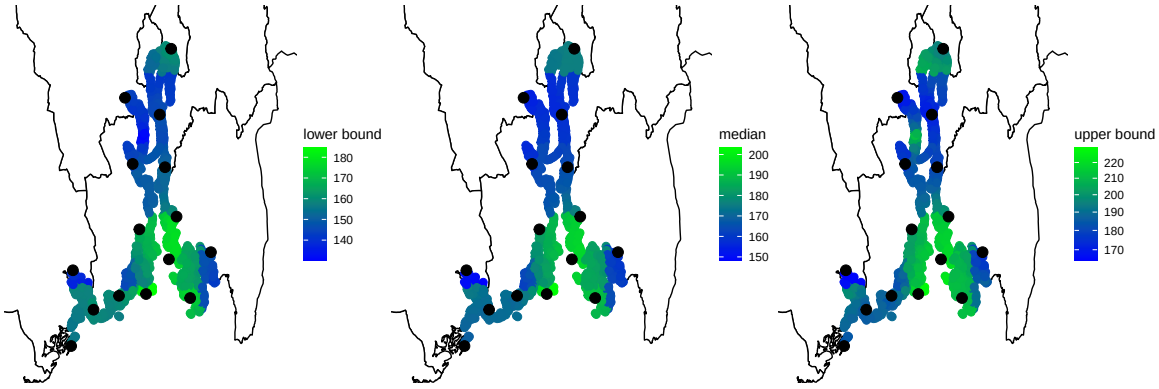


Figure 20: Maps of return levels in the entire Oslo fjord based on parameter estimates from 15 locations with simulated data. In the plots the lower bound of the 95% credible interval on the left, the median in the middle and the upper bound of the 95% credible interval on the right.

4.1.2 Sensitivity Analysis for the Prior of the Range Parameter

We perform a sensitivity analysis for the prior of the range parameter due to the prior being informative as this parameter being quite difficult to estimate. We fit the model once more, however this time a new prior for the range parameter is used while the rest of the model remain unchanged. The two priors are shown in Figure 21. The priors were chosen to be gamma distributed with a mean at approximately half of the maximum distance between the locations studied. The maximum distance between locations is 113 km. The initial prior for model 1 is gamma distributed with parameters 5 and 1/12, this means that the mean is 60 and the variance is 720. The new prior for model 2 is gamma distributed with parameters 1.5 and 1/40, this means that the mean remains 60, but the variance is changed to 2400. One can see that the mode of the prior has moved closer to zero and that the upper tail has become heavier.

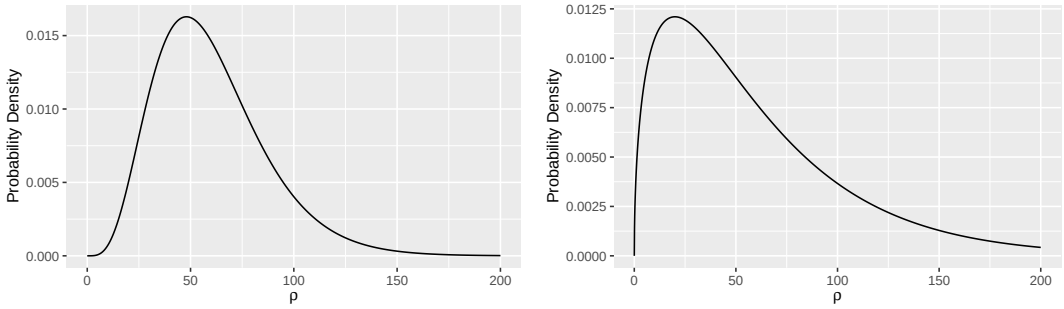


Figure 21: The first range prior is shown on the left. This prior is gamma distributed with parameters 5 and 1/12, this means that the mean is 60 and the variance is 720. The new range prior is shown on the right. This prior is gamma distributed with parameters 1.5 and 1/40, this means that the mean remains 60, but the variance is changed to 2400.

A comparison of the resulting parameter estimates and the results obtained in the parameter estimates in the previous section are shown in Tables 4, 5 and 6. The tables show the true values of the parameters together with the 95% credible bounds from the two analyses. The bounds seem to be practically unchanged. The largest difference between the two analyses observed in Table 4 is 0.81 when one excludes the range parameters. The largest differences in Tables 5 and 6 are 0.32 and 0.02 respectively. This result indicates that the informative priors for the range parameters do not seem to greatly influence the estimation of the rest of the parameters. This indicates that subjectively choosing a prior should be fine as long as it is chosen within a reasonable interval.

Parameter	True Value	Model 1		Model 2	
ξ	-0.05	-0.09	0.01	-0.09	0.01
$\beta_{q,1}$	90.00	79.15	101.18	78.67	101.35
$\beta_{q,2}$	9.00	-13.43	25.53	-13.89	25.58
$\beta_{q,3}$	12.00	-2.74	38.83	-3.08	39.64
$\beta_{q,4}$	-18.00	-35.17	-1.97	-34.78	-1.63
$\beta_{q,5}$	15.00	0.29	18.05	0.10	17.89
τ_q	0.30	0.10	3.63	0.08	3.92
ρ_q	100.00	21.79	130.96	6.09	217.08
$\beta_{s,1}$	4.30	3.55	5.47	3.53	5.53
$\beta_{s,2}$	0.20	-1.75	1.27	-1.72	1.25
$\beta_{s,3}$	0.10	-1.02	2.29	-0.97	2.28
$\beta_{s,4}$	-0.60	-2.58	0.30	-2.56	0.30
$\beta_{s,5}$	0.20	-0.99	0.54	-1.03	0.45
τ_s	0.20	0.12	0.46	0.11	0.63
ρ_s	80.00	31.13	142.83	23.56	252.53

Table 4: Comparison of the parameter estimates from the sensitivity analysis of the range parameter.

Parameter	True Value	Model 1		Model 2	
q_1	107.58	105.49	110.12	105.55	110.14
q_2	101.07	96.19	101.28	96.10	101.29
q_3	106.31	99.26	106.53	99.23	106.60
q_4	107.45	104.58	110.80	104.73	110.66
q_5	102.36	101.15	107.11	101.23	107.16
q_6	92.18	89.86	97.03	89.82	96.71
q_7	111.35	109.07	115.60	109.11	115.62
q_8	104.84	102.18	107.56	102.22	107.49
q_9	99.23	94.20	99.20	94.22	99.21
q_{10}	100.73	95.70	101.65	95.69	101.64
q_{11}	98.31	95.42	100.61	95.41	100.64
q_{12}	97.39	93.30	98.39	93.16	98.30
q_{13}	106.44	103.51	108.78	103.46	108.73
q_{14}	109.64	107.56	112.43	107.69	112.38
q_{15}	103.76	101.92	107.28	101.86	107.16

Table 5: Comparison of the parameter estimates of q from the sensitivity analysis of the range parameter.

Parameter	True Value	Model 1		Model 2	
$\ln s_1$	4.22	4.09	4.38	4.09	4.37
$\ln s_2$	4.20	4.00	4.33	4.00	4.32
$\ln s_3$	4.48	4.21	4.54	4.21	4.54
$\ln s_4$	4.41	4.35	4.64	4.35	4.63
$\ln s_5$	4.55	4.27	4.60	4.27	4.60
$\ln s_6$	4.17	4.08	4.44	4.08	4.43
$\ln s_7$	4.45	4.38	4.70	4.38	4.70
$\ln s_8$	4.55	4.41	4.72	4.41	4.70
$\ln s_9$	4.22	4.00	4.29	3.99	4.29
$\ln s_{10}$	4.24	4.01	4.35	4.01	4.36
$\ln s_{11}$	4.27	4.13	4.43	4.14	4.43
$\ln s_{12}$	4.14	3.95	4.26	3.94	4.25
$\ln s_{13}$	4.22	4.04	4.38	4.02	4.38
$\ln s_{14}$	4.24	4.05	4.34	4.05	4.35
$\ln s_{15}$	4.46	4.43	4.72	4.42	4.72

Table 6: Comparison of the parameter estimates of $\ln s$ from the sensitivity analysis of the range parameter.

4.2 Analysis of 4 Locations

We are performing a simulation study on four locations since there are only four locations in the Oslo fjord that have longer periods of sea level data. The results of this simulation study

should give us an idea of what to expect from the analysis of the sea level data. The four locations studied in this section are at the same locations as the stations with long series of sea level data. The locations are presented in Figure 22.

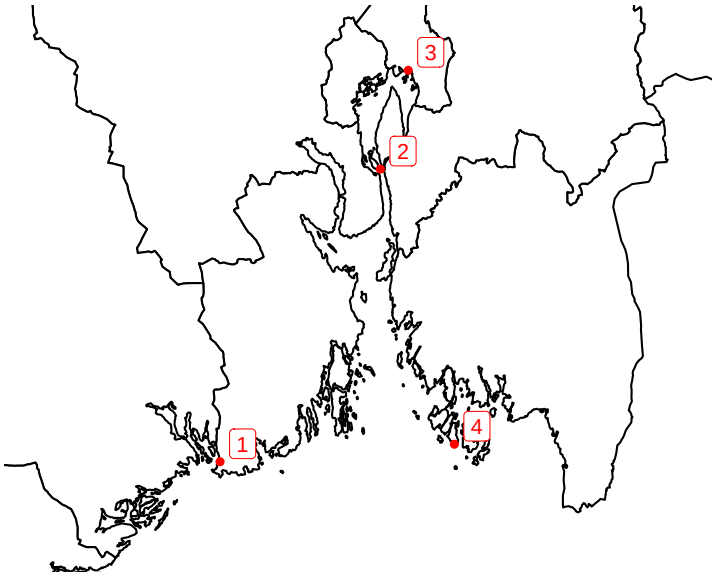


Figure 22: The locations used in the simulated multivariate GEV analysis for 4 stations. The stations are at the same locations as the stations with sea level data.

The data studied in this section are sampled according to the model presented in Section 4.1. In this experiment the model parameters are identical to those in Section 4.1, but both q, $\ln s$ and the data have been resampled. We will apply three different models to the data in this section. First we will perform a univariate analysis where the parameters of the GEV model are assumed to be independent in space. Next we will perform an analysis where the shape parameter is assumed to be constant in space, while the other parameters are assumed to be independent in space. The last analysis will apply the multivariate model to the data. A reason for performing an analysis with common parameter ξ, while the other parameters do not have spatial representations is that the parameter ξ is quite difficult to estimate and using data from multiple locations to estimate this parameter could be beneficial. We are also interested in studying the difference between the results from the last two analyses to see how much of a difference is caused by the spatial field.

4.2.1 Univariate Analysis

We perform a univariate analysis for the purpose of finding out whether the spatial model leads to improved results and if so how much of an improvement can be expected.

We are studying a simplified model in the univariate analysis. We let z_j be the j-th observed

maximum sea level at the location of interest. The model likelihood is

$$\begin{aligned} p(\mathbf{z}|q, \ln s, \tilde{\xi}) &= \prod_{j=1}^{k} \mathrm{GEV}(z_j|q, s, \xi) \\ \mathrm{GEV}(z_j|q, s, \xi) &= \exp\Big\{ - \Big[\Big(\frac{z_j - q}{s(l_{1-\beta/2,\xi} - l_{\beta/2,\xi})^{-1}} + l_{\alpha,\xi} \Big) \Big]_+^{-1/\xi} \Big\} \\ l_{\gamma,\xi} &= (-\log(\gamma))^{-\xi} \\ s &= \exp(\ln s) \\ \xi &= a + (b-a) \cdot \frac{\exp\{\tilde{\xi}\}}{1 + \exp\{\tilde{\xi}\}} \end{aligned} \tag{38}$$

where k is the number of data available at the location. q is the $\alpha-$quantile and s is the difference between the $1 - \beta/2$ and $\beta/2$ quantiles. a and b are the lower and upper bounds of the shape parameter ξ.

We fit the model using the data generated based on spatial representations of q and $\ln s$ and with common value for the parameter ξ across the spatial field. The values of $\alpha = 0.5$ and $\beta = 0.05$ were chosen so that the parameters q and s represent the median and the 95% credible interval of the GEV distribution respectively. We use STAN to find samples of the parameters based on the posterior model, this model can be stated as follows

$$\begin{aligned} p(q, \ln s, \tilde{\xi}|\mathbf{z}) \propto\ & p(\mathbf{z}|q, \ln s, \tilde{\xi}) \\ & \cdot p(\tilde{\xi}) p(q) p(\ln s). \end{aligned} \tag{39}$$

The probability distributions representing the prior knowledge can be explained with the following equations

$$\begin{aligned} \tilde{\xi} &\sim \mathrm{normal}(0, 10) \\ q &\sim \mathrm{normal}(100, 50) \\ \ln s &\sim \mathrm{normal}(0, 10) \end{aligned} \tag{40}$$

The resulting parameter estimates of the parameters ξ, q and $\ln s$ are presented in Figures 23, 24 and 25 respectively. The plots show the histograms of the parameter estimates for each location together with the parameter values and the priors. One can see that all the histograms cover the parameter value. The histograms are quite wide, this is not unexpected since the estimates are not based on excessive amounts of data. The priors are not notable in the plots, so they should have had little impact on the final parameter distributions. We will in the next section perform an analysis with a common value of the parameter ξ. One can see in Figure 23 that the histograms for the parameter estimates of ξ always cover the interval $[-0.2, 0.2]$, this means that a common parameter value probably will lie in this interval.

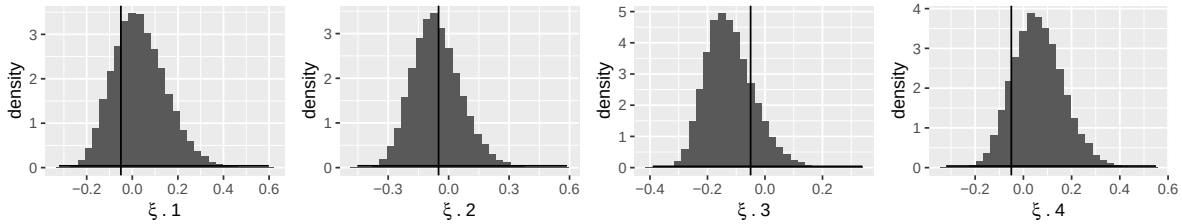


Figure 23: Histograms of the parameter estimates of ξ from a univariate analysis. The data used is the data generated for the multivariate analysis, for the purpose of checking whether a multivariate model leads to improvement. The chosen parameter values and the priors of the model are also depicted in the plots.

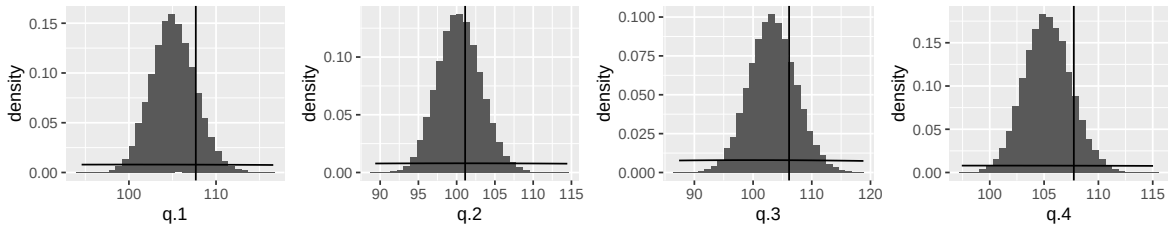


Figure 24: Histograms of the parameter estimates of q from a univariate analysis. The data used is the data generated for the multivariate analysis, for the purpose of checking whether a multivariate model leads to improvement. The chosen parameter values and the priors of the model are also depicted in the plots.

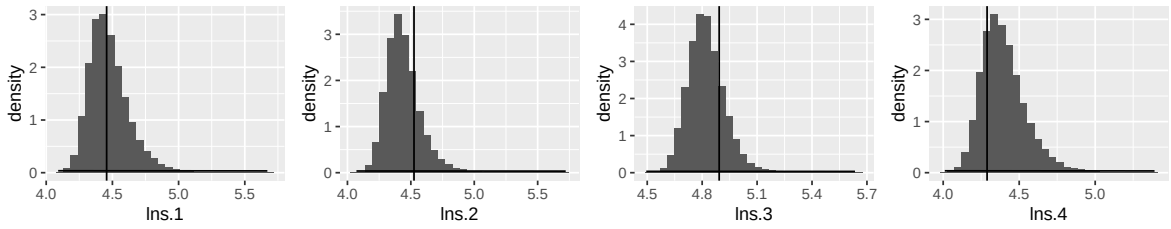


Figure 25: Histograms of the parameter estimates of $\ln s$ from a univariate analysis. The data used is the data generated for the multivariate analysis, for the purpose of checking whether a multivariate model leads to improvement. The chosen parameter values and the priors of the model are also depicted in the plots.

The return level plots of the univariate analysis are shown in Figure 26. The empirical return level estimates are included in the 95% credible intervals for all the locations, however one can see that the 95% credible intervals are very wide. This result indicates that it is difficult to use the results for specifying the risk of observing extremes, as the return level could vary quite drastically.

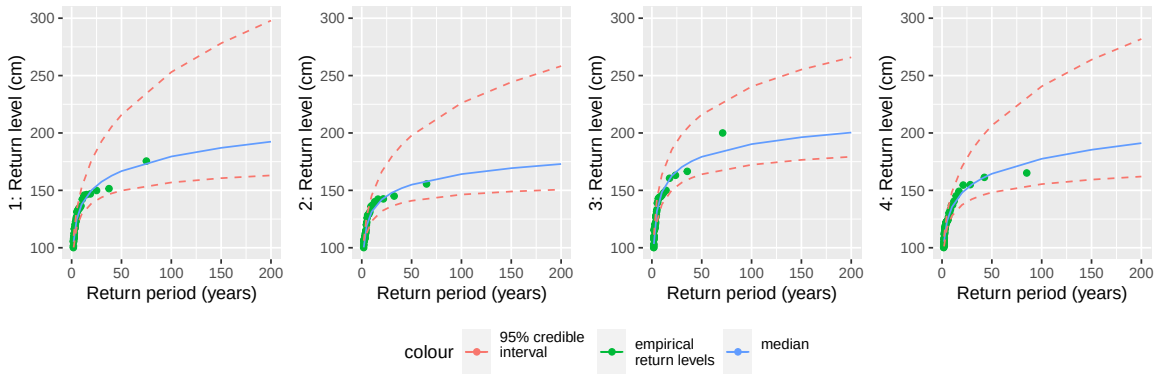


Figure 26: Return level plots for the univariate analysis of the four stations studied in the last section. The data used is the data generated for the multivariate analysis, for the purpose of checking whether a multivariate model leads to improvement.

4.2.2 Analysis with Common Parameter ξ for All Locations

We study a model where the parameter ξ is assumed to be common for the different locations while the parameters q and $\ln s$ are assumed to be independent in space. The purpose of this analysis is to investigate how much of a difference can be seen by choosing a common parameter ξ. We let z_{ij} be the j-th observed maximum sea level at location i. The model likelihood is

$$\begin{aligned} p(\mathbf{z}|\mathbf{q}, \ln \mathbf{s}, \tilde{\xi}) &= \prod_{i=1}^{l} \prod_{j=1}^{k_i} \mathrm{GEV}(z_{ij}|q_i, s_i, \xi) \\ \mathrm{GEV}(z_{ij}|q_i, s_i, \xi) &= \exp\Big\{ -\Big[\Big(\frac{z_{ij} - q_i}{s_i(l_{1-\beta/2,\xi} - l_{\beta/2,\xi})^{-1}} + l_{\alpha,\xi}\Big)\Big]_+^{-1/\xi}\Big\} \\ l_{\gamma,\xi} &= (-\log(\gamma))^{-\xi} \\ s_i &= \exp(\ln s_i) \\ \xi &= a + (b-a) \cdot \frac{\exp\{\tilde{\xi}\}}{1+\exp\{\tilde{\xi}\}} \end{aligned} \tag{41}$$

where l is the number of locations and k_i is the number of data z available in location i. q is the $\alpha-$quantile and s is the difference between the $1-\beta/2$ and the $\beta/2$ quantiles. a and b are the lower and upper bounds of the shape parameter ξ. The limits of the parameter ξ are chosen to be $a = -0.5$ and $b = 0.5$ and the quantiles α and β are chosen to be 0.5 and 0.05 respectively.

We fit the model using the data generated based on spatial representations of q and $\ln s$ and with common value for the parameter ξ across the spatial field. The posterior distribution is

$$\begin{aligned} p(\mathbf{q}, \ln \mathbf{s}, \tilde{\xi}|\mathbf{z}) &\propto p(\mathbf{z}|\mathbf{q}, \ln \mathbf{s}, \tilde{\xi}) \\ &\quad \cdot p(\tilde{\xi})p(\mathbf{q})p(\ln \mathbf{s}). \end{aligned} \tag{42}$$

The probability distributions representing the prior knowledge can be explained with the following equations

$$\begin{aligned} \tilde{\xi} &\sim \text{normal}(0, 10) \\ \mathbf{q} &\sim \text{normal}(100, 50) \\ \mathbf{ln\,s} &\sim \text{normal}(0, 10) \end{aligned} \tag{43}$$

The resulting parameter estimates of the parameters ξ, q and $\ln s$ are presented in Figures 27, 28 and 29 respectively. One can see in Figure 27 that the histogram of the parameter ξ is in the expected interval $[-0.2, 0.2]$. The histograms for the parameter q from the univariate analysis and this analysis are shown in Figures 24 and 28. The results are very similar, it is difficult to find any significant differences. The histograms for the parameter $\ln s$, shown in Figures 25 and 29, on the other hand are quite different. The shape of the histograms are very different and the upper bounds of the univariate histograms are much larger, the lower bounds are however quite similar.

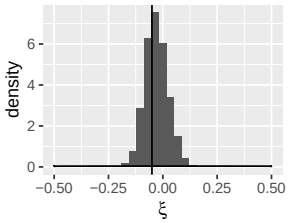


Figure 27: Histogram of the parameter estimate for ξ. The estimate is based on simulated data from 4 locations. The parameter ξ is assumed to be common for all locations. The chosen parameter value is depicted as a vertical line and the prior of the parameter is also added to the plot.

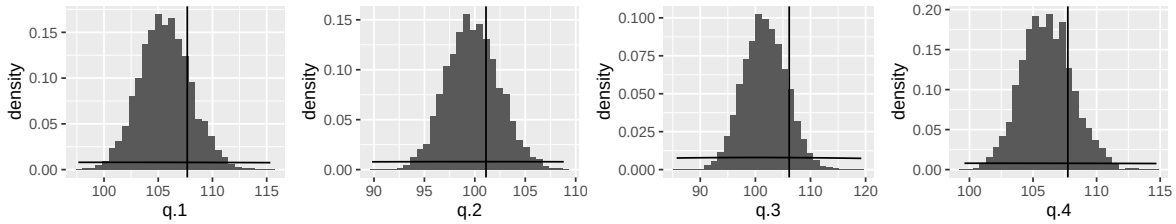


Figure 28: Histogram of the parameter estimates for q. The estimates are based on simulated data from 4 locations. The parameter q is not assumed to vary spatially. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

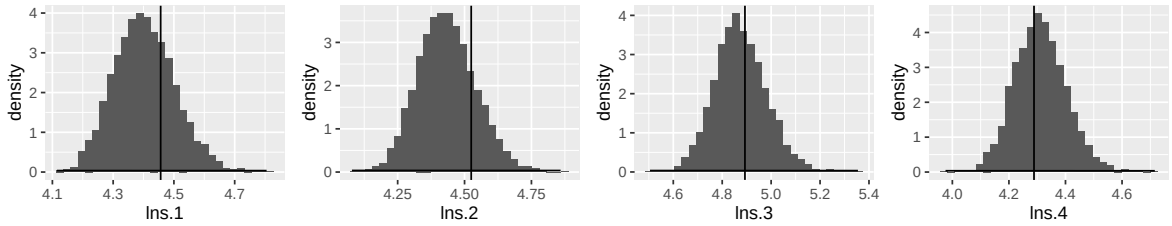


Figure 29: Histogram of the parameter estimates for $\ln s$. The estimates are based on simulated data from 4 locations. The parameter $\ln s$ is not assumed to vary spatially. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

The return level plots are shown in Figure 30. One can see that the empirical return level estimates are contained within the credible intervals for all the locations. We compare the results to the univariate case shown in Figure 26. One can see that the 95% credible intervals have become much narrower for all the locations except the third one, for which the credible intervals are close to unchanged.

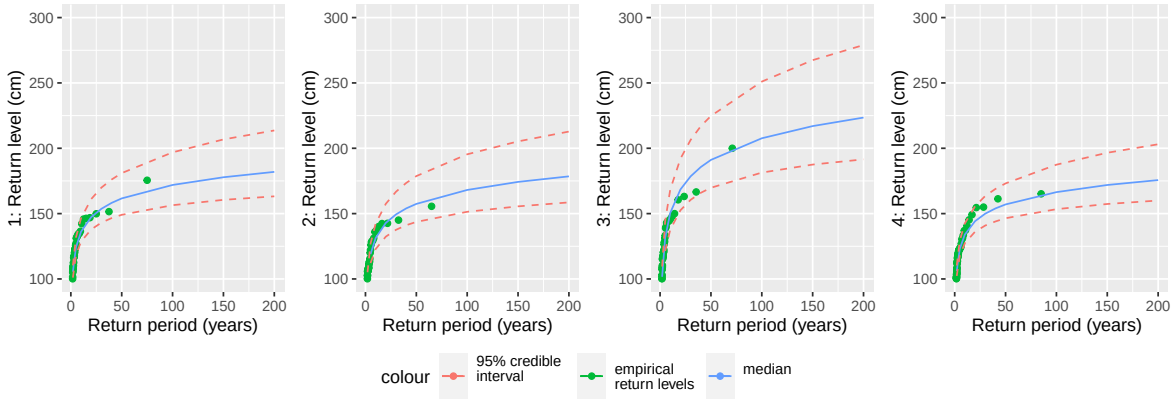


Figure 30: The return level plots of 4 locations based on simulated data. The model used assumes that the parameter ξ is common across all locations, while the parameters q and $\ln s$ do not vary in space. The plots contain the medians and the 95% credible intervals of the return levels plotted against the return periods. The empirical return levels based on the data are also shown.

4.2.3 Multivariate Analysis

We apply the multivariate model to the data. The model is the same as was presented in Section 4.1. The parameter estimates obtained in the analysis are presented in Figures 31, 32, 33, 34 and 35. One can see that the appointed parameter values are included in the histograms for all of the parameters. However, the parameter samples for the standard deviations and the ranges in Figures 32 and 33 have not deviated from the prior distributions, which means that the parameters have not been estimated based on the data. Another observation one can make from the plots are that the spread of the parameter samples are very wide. This means that the uncertainties of the estimates are very large, it is therefore reasonable to assume that

these uncertainties can lead to the uncertainties of derived entities to be considerable as well. This result is not unexpected since we only have four locations with data in quite a large area, so there is not excessive amount of information to use in the inference. The lack of data seem to be especially problematic for the spatial parameters, that is the standard deviation and the ranges of the spatial fields. We compare the histograms for ξ, q and $\ln s$ obtained in the last analysis, that is Figures 27, 28 and 29, to the results obtained in the multivariate analysis, that is Figures 31, 34 and 35. One can see that the histograms are close to identical in the two analyses, this indicates that the spatial fields does not seem to make much difference when we are studying four locations.

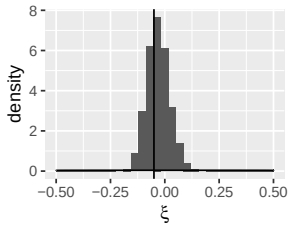


Figure 31: Histogram of the parameter estimate for ξ. The estimates are based on simulated data from 4 locations. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

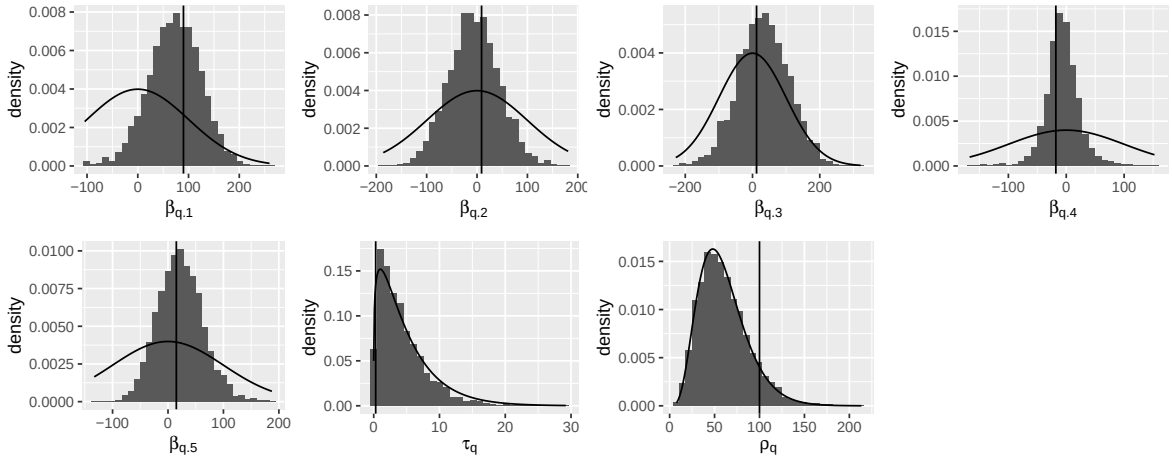


Figure 32: Histograms of the parameter estimates that describe the parameter q. The estimates are based on simulated data from 4 locations. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

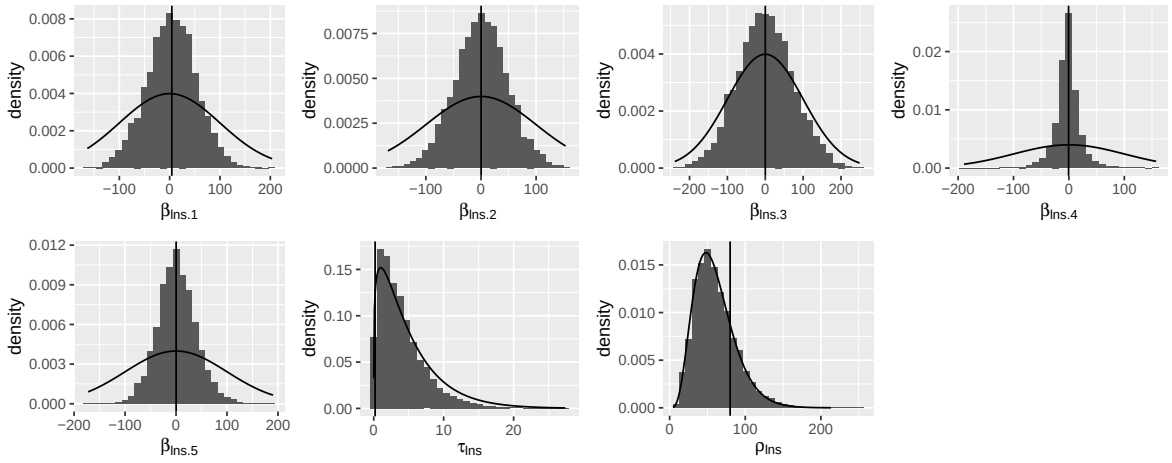


Figure 33: Histograms of the parameter estimates that describe the parameter $\ln s$. The estimates are based on simulated data from 4 locations. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

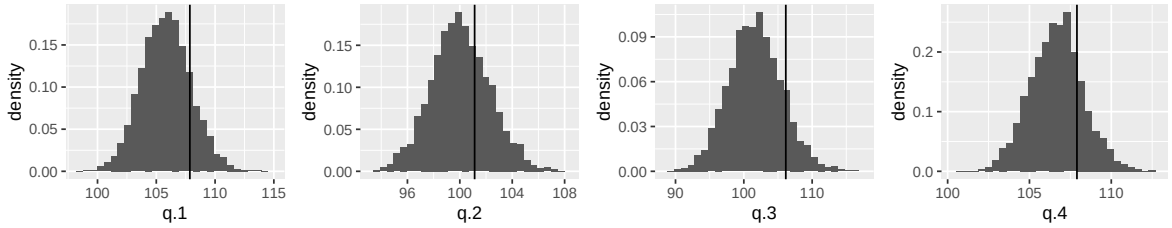


Figure 34: Histograms of the parameter estimates for **q** based on simulated data from 4 locations. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

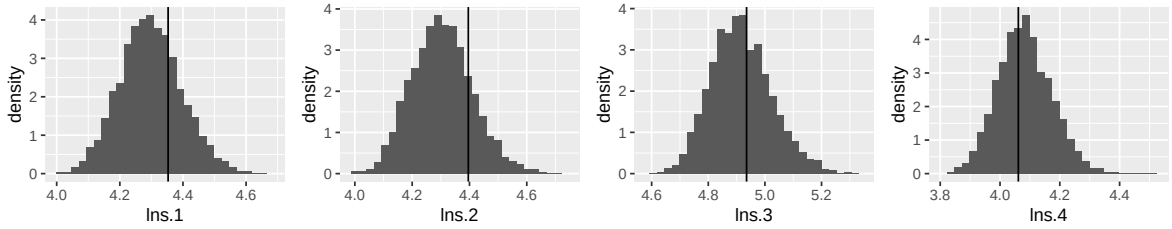


Figure 35: Histograms of the parameter estimates for $\mathbf{ln\,s}$ based on simulated data from 4 locations. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

The return level plots are shown in Figure 36. The figure shows the 95% credible interval and the median of the return levels plotted against the return periods for the four locations.

One can see that the spread of the uncertainty bounds vary from covering less than 50 cm to covering about 80 cm. The widest uncertainty is obtained by location 3 which is the location farthest to the north. One can see that the median of this return level plot seem to coincide with the last empirical return level value, but not with the rest of the empirical return level points. Locations 2 and 3 are quite close in vicinity so one could expect them to borrow information from each other and one would thus expect somewhat similar results for the two locations, this is however not the case. We compare the results to the other return level plots obtained in this section, that is the results presented in Figure 26 and 30. One can see that the return level plots from the analysis of the model with common ξ are close to identical to the return level plots obtained in this section. This result indicates that the improvements we see compared to the univariate model seem to be mostly caused by having a common parameter ξ. This result is not very surprising as the spatial parameters had not changed from their priors, in addition to the parameter ξ being quite difficult to estimate.

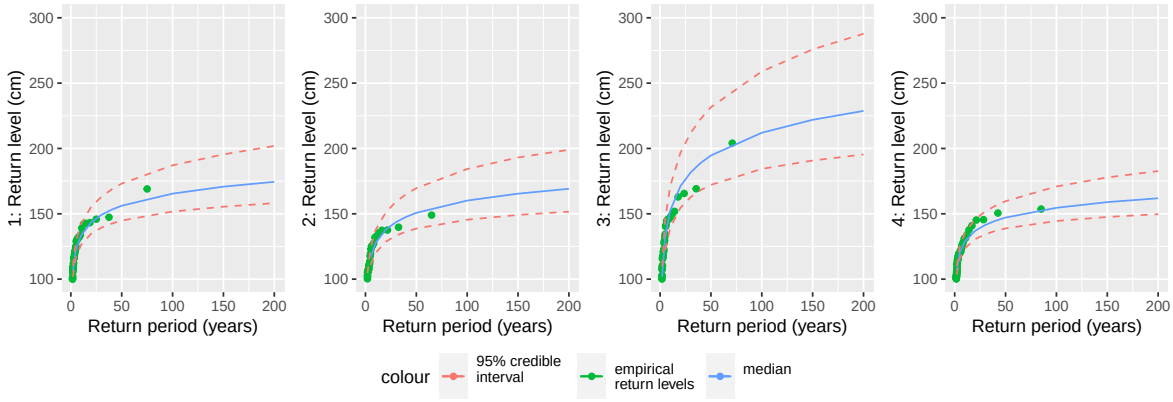


Figure 36: The return level plots of 4 locations based on simulated data. The plots contain the medians and the 95% credible intervals of the return levels plotted against the return periods. The empirical return levels based on the data are also shown.

Another point of interest is a spatial analysis of the return levels. That is we would like to try to obtain return levels in areas where there is not data available. The spatial map of the 95% credible interval and the median of the return levels associated with the 200 year return period are shown in Figure 37. One can see that the upper bound is extremely large in the areas where there are no available data. This result is expected since we have very few locations and quite large distances. One can see that both the upper and lower bounds are closer to the medians in areas in close proximity to the locations with data.

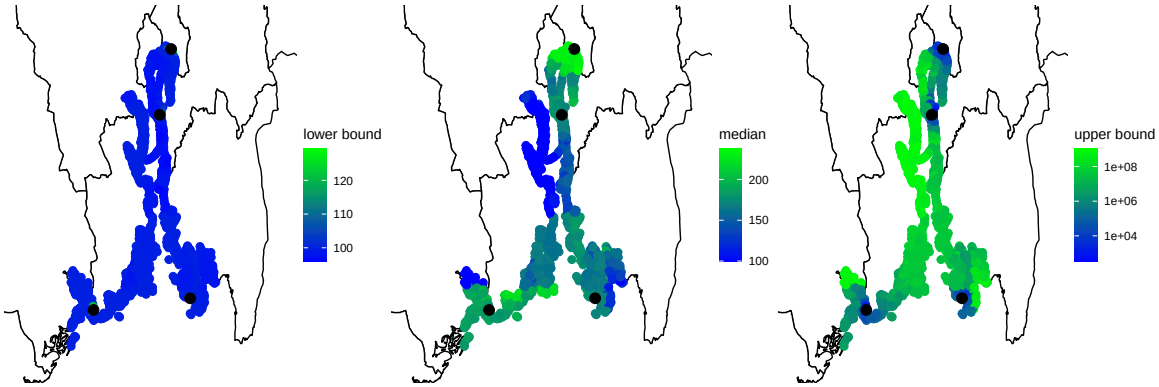


Figure 37: Maps of return levels in the entire Oslo fjord based on parameter estimates from 4 locations with simulated data. In the plots the lower bound of the 95% credible interval on the left, the median in the middle and the upper bound of the 95% credible interval on the right.

4.3 Analysis of 8 Locations, Where 4 Locations Have Shorter Data Series

We are performing a simulation study on 8 locations, where 4 locations have shorter data series. We are studying the stations that were examined in the simulation study of four locations together with four additional locations. The purpose of this study is to see if obtaining shorter data series in the Oslo fjord would lead to improvements. The short series do however need to have at least one year of data to be used in the GEV model, this means that obtaining additional data is very time consuming. The additional locations have in this case been chosen using the same method as was used when choosing the additional locations for the 15 location simulation study. That is every time a location is added the minimum distance to other locations is maximized, so that the location is as far as possible from any other station. The locations are plotted in Figure 13.

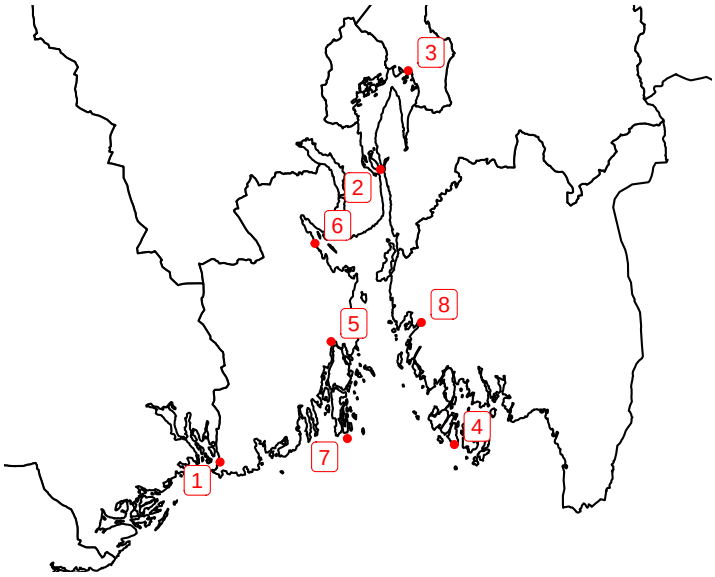


Figure 38: The locations used in the simulated multivariate GEV analysis for 8 stations. The 4 stations with lowest numbers are at the same locations as the stations with sea level data and the additional stations were chosen sequentially to be as far from existing stations as possible.

The model studied in this section is the model presented in Section 4.1. The number of data points for the first four locations are the same as in the previous analyses, while the number of data for the additional four locations is drawn uniformly between 5 and 15.

The parameter samples obtained in the analysis are presented in Figures 39, 40, 41, 42 and 43. One can see that all the histograms cover the chosen parameter values. We compare the results to the results obtained from the study of 4 locations in Section 4.2.3. One can see that the estimations of the parameter ξ in Figures 31 and 39 are practically unchanged. The estimates of the parameters used to specify q shown in Figures 32 and 40 show some improvements. The β parameters that specify the mean of spatial field for q have all been estimated with more certainty than in the case with 4 locations. The standard deviation and the range of the spatial field are on the other hand close to unchanged. The estimates of the parameters used to specify $\ln s$ shown in Figures 33 and 41 show improvements. The uncertainty of the β parameters have greatly improved, the standard deviation has deviated substantially from the prior and the range has deviated slightly from the prior.

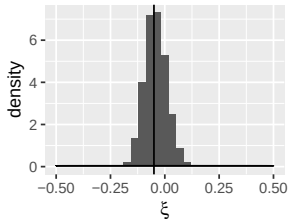


Figure 39: Histogram of the parameter estimate for ξ. The estimate is based on simulated data from 8 locations, where 4 of the locations have quite short data series. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

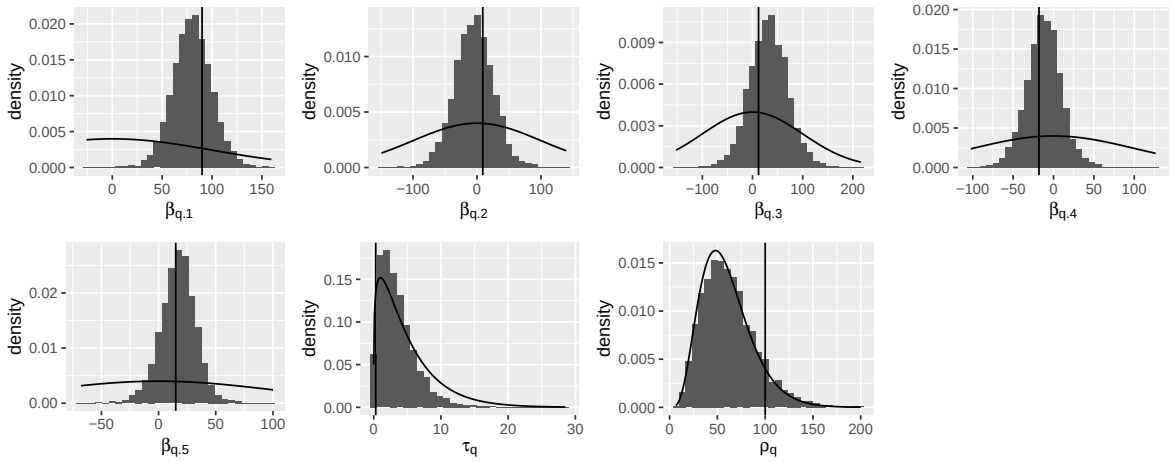


Figure 40: Histograms of the parameter estimates that describe the parameter q. The estimates are based on simulated data from 8 locations, where 4 locations have quite short data series. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

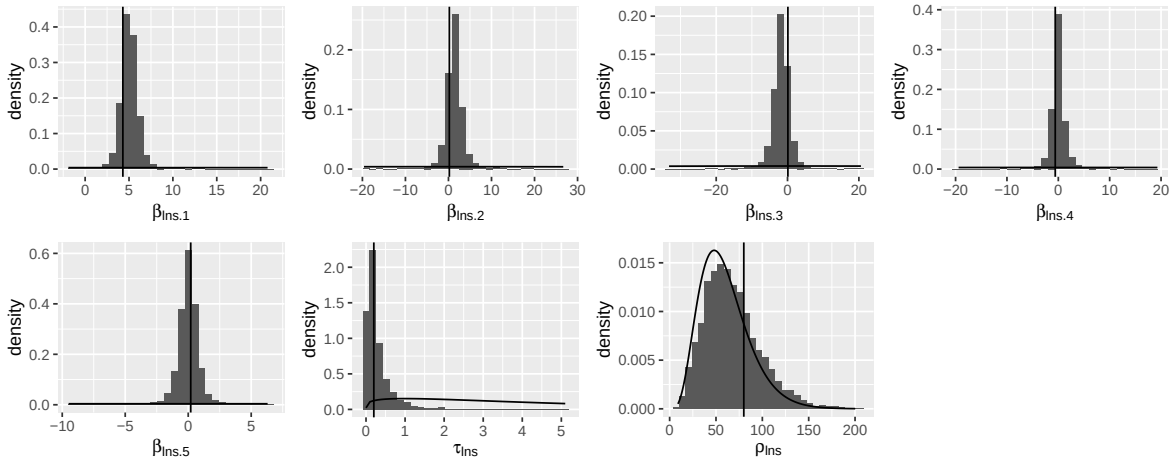


Figure 41: Histograms of the parameter estimates that describe the parameter $\ln s$. The estimates are based on simulated data from 8 locations, where 4 locations have quite short data series. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

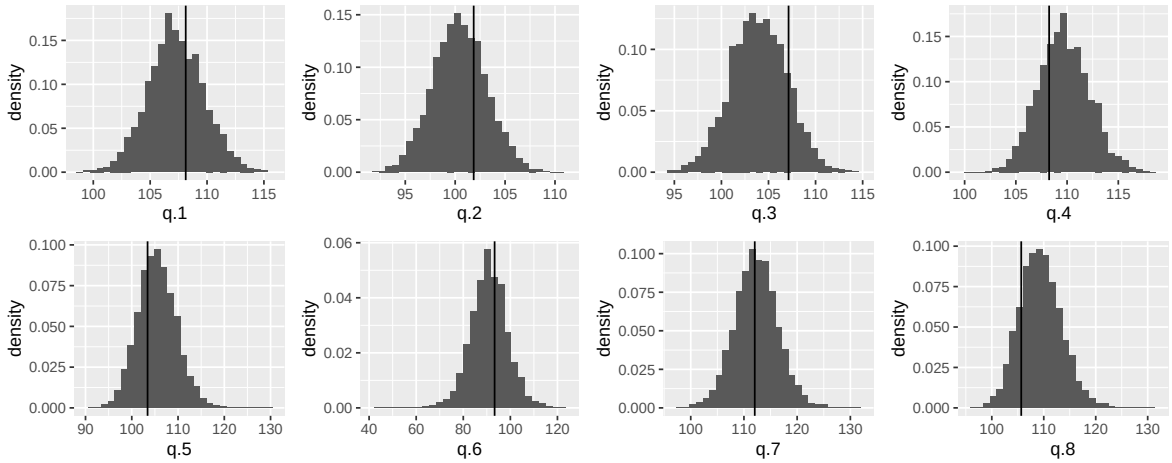


Figure 42: Histograms of the parameter estimates for **q** based on simulated data from 8 locations, where 4 of the locations have quite short data series. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

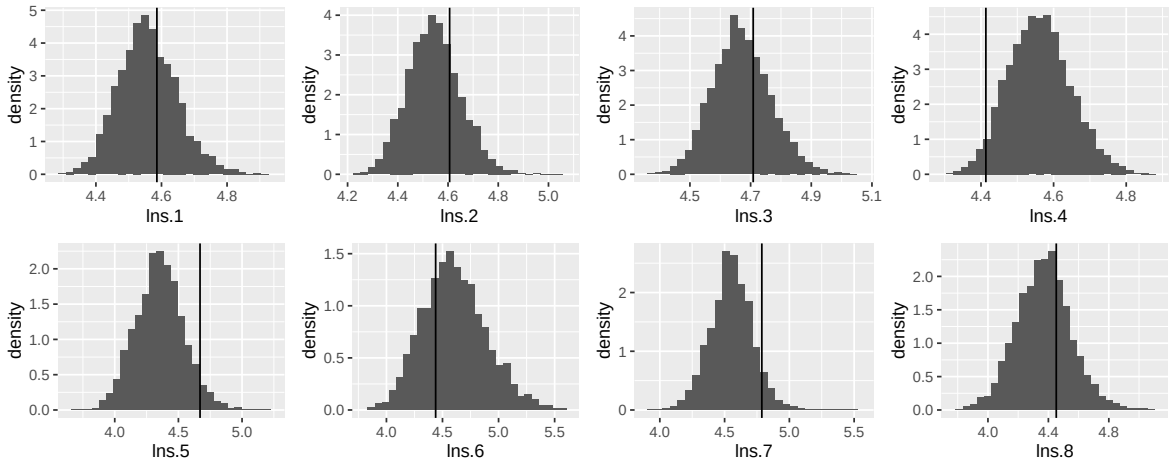


Figure 43: Histograms of the parameter estimates for **q** based on simulated data from 8 locations, where 4 of the locations have quite short data series. The chosen parameter values are depicted as vertical lines and the priors of the parameters are also added to the plots.

The return levels are plotted against the return periods for the 8 locations in Figure 44. One can see that the uncertainty bounds of the additional four locations all are contained within the interval $[100, 300]$. The locations 5, 7 and 8 have quite narrow uncertainty bounds considering the amount of data used in the estimation. The location with the fewest data points, that is location 6 with 9 data points, has the widest uncertainty bounds. This location does however only have two data points less than locations 5 and 8, which indicates that those two data points can make quite a large difference. This analysis indicates that if one can obtain more than 10 data points, then the resulting return level estimates based on a spatial model can be quite informative.

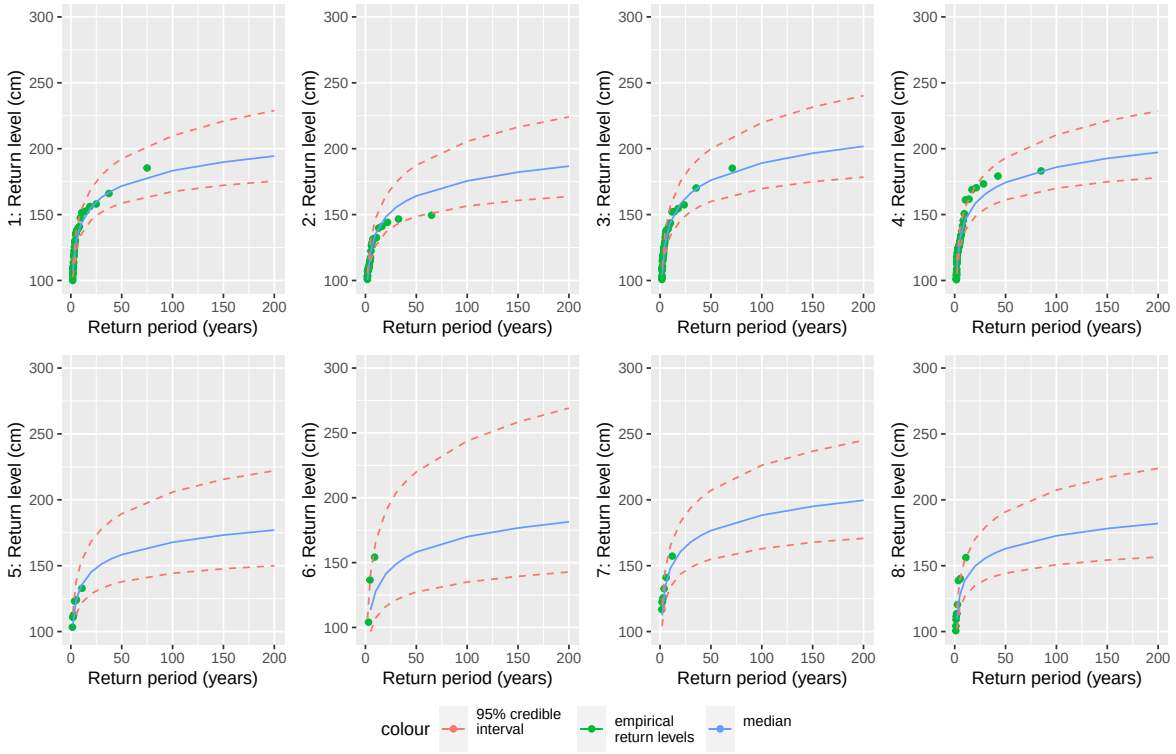


Figure 44: The return level plots of 8 locations based on simulated data. The plots contain the medians and the 95% credible intervals of the return levels plotted against the return periods. The empirical return levels based on the data are also shown.

We also try to do a spatial analysis based on the analysis of 8 locations. That is we would like to try to obtain return levels in areas where there is not data available. The spatial map of the 95% credible interval and the median of the return levels associated with the 200 year return period are shown in Figure 45. One can see that the largest difference between the lower and upper uncertainty bounds are about 300 cm, this is a huge improvement compared to the analysis of four locations shown in Figure 37. Furthermore, the largest difference only exist in a very narrow area where we have no data available. Most of the area has uncertainty bounds in the interval between 140 cm and 250 cm, which means that the difference between the upper and lower bounds is approximately 1 meter in most of the Oslo fjord.

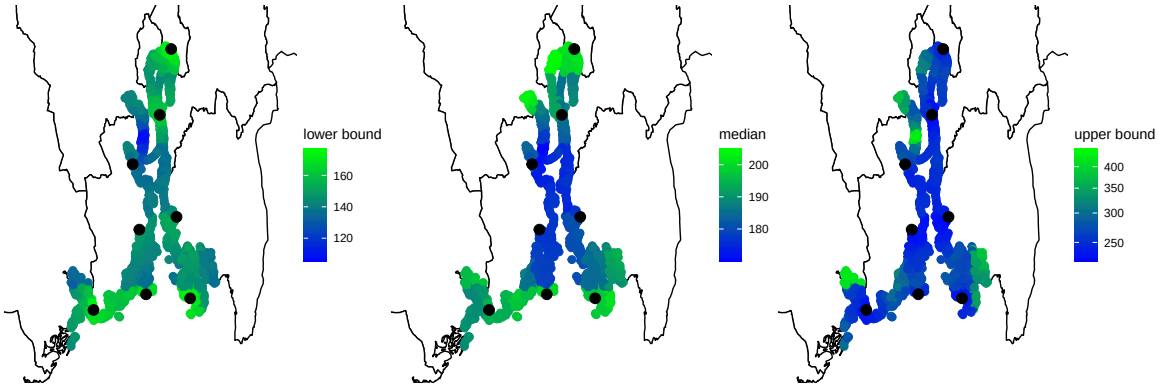


Figure 45: Maps of return levels in the entire Oslo fjord based on parameter estimates from 8 locations with simulated data, where 4 locations have quite short data series. In the plots the lower bound of the 95% credible interval on the left, the median in the middle and the upper bound of the 95% credible interval on the right.

5 Data Analysis

There are long series of sea level data available at four locations in the Oslo fjord. The four locations, Helgeroa, Oscarsborg, Oslo and Viker, are shown in Figure 46. We want to study the data so that we can obtain estimates that can give us information about the frequency of extreme sea levels. The estimates that we have the most interest in are return levels and return periods. We are interested in finding return level estimates for the four locations for which we have data. We will also try to obtain return level estimates for every other location in the Oslo fjord, however we do not expect this approach to give usable results based on the results obtained in the simulation study of 4 locations.

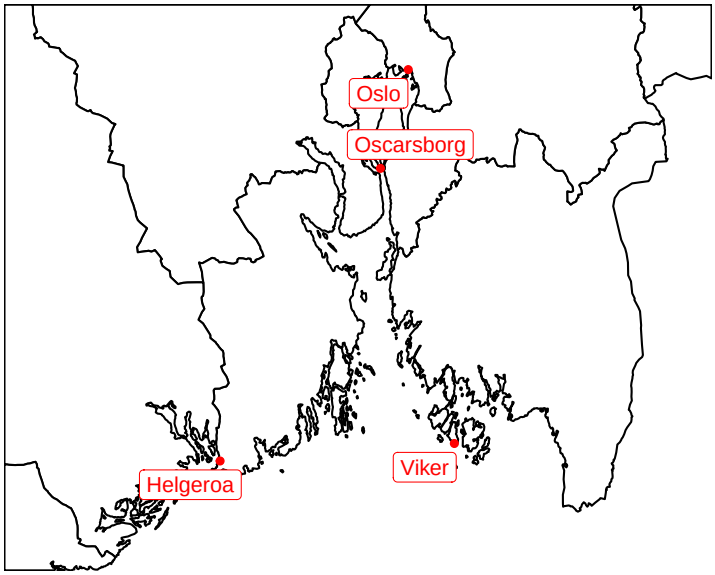


Figure 46: The four locations in the Oslo fjord for which there are long series of sea level data available. The locations are Helgeroa, Oscarsborg, Oslo and Viker.

We will study three different models in this section. We will first look at a univariate model, this analysis is performed so that we can compare the results to those of the multivariate model. In the next analysis we apply a model where the parameter ξ is assumed to be common across the locations, while the parameters q and $\ln s$ are assumed to be independent in space. We are studying this model due to the parameter ξ being quite difficult to estimate and also because we are interested in looking at the difference between this model and the multivariate model. Lastly, we study the multivariate model, which is the model that we are the most interested in.

5.1 Univariate Analysis

The univariate model is studied so that we can compare the results to the multivariate model and discuss whether the multivariate model leads to improved estimates.

We are studying a simplified model in the univariate analysis. We let z_j be the j-th observed maximum sea level at the location of interest. The model likelihood is

$$\begin{aligned} p(\mathbf{z}|q, \ln s, \tilde{\xi}) &= \prod_{j=1}^{k} \text{GEV}(z_j|q, s, \xi) \\ \text{GEV}(z_j|q, s, \xi) &= \exp\Big\{ -\Big[\Big(\frac{z_j - q}{s(l_{1-\beta/2,\xi} - l_{\beta/2,\xi})^{-1}} + l_{\alpha,\xi}\Big)\Big]_+^{-1/\xi}\Big\} \\ l_{\gamma,\xi} &= (-\log(\gamma))^{-\xi} \\ s &= \exp(\ln s) \\ \xi &= a + (b-a)\cdot\frac{\exp\{\tilde{\xi}\}}{1+\exp\{\tilde{\xi}\}} \end{aligned} \tag{44}$$

where k is the number of data available at the location. q_α is the $\alpha-$quantile and s_β is the difference between the $1-\beta/2$ and $\beta/2$ quantiles. a and b are the lower and upper bounds of the shape parameter ξ.

We use STAN to find samples of the parameters based on the posterior model, this model can be stated as follows

$$\begin{aligned} p(q, \ln s, \tilde{\xi}|\mathbf{z}) \propto\ & p(\mathbf{z}|q, \ln s, \tilde{\xi}) \\ & \cdot p(\tilde{\xi})p(q)p(\ln s). \end{aligned} \tag{45}$$

The probability distributions representing the prior knowledge can be explained with the following equations

$$\begin{aligned} \tilde{\xi} &\sim \text{normal}(0, 10) \\ q &\sim \text{normal}(100, 50) \\ \ln s &\sim \text{normal}(0, 10) \end{aligned} \tag{46}$$

We fit the model using the yearly maximum values from the sea level data. The sea level data has had the trend and the missing data removed before the maximum values were found. We have 77, 56, 101 and 30 data points for Helgeroa, Oscarsborg, Oslo and Viker respectively. The values of the parameters $\alpha = 0.5$ and $\beta = 0.05$ were chosen so that the parameters q and s represent the median and the 95% credible interval of the GEV distribution respectively. We choose a and b to be equal -0.5 and 0.5 for Helgeroa, Oscarsborg and Oslo, this is the same choice we made in the simulation study. We cannot use the same interval for Viker due to the parameter estimates for ξ being outside those bounds. The bounds chosen for Viker are $-0.95, 0.95$. The difference in bounds for Viker are probably caused by there being fewer data points available and it is therefore not possible to estimate the parameter ξ with more accuracy.

The parameter estimates obtained from fitting the univariate model are shown in Figures 47, 48 and 49. One can see that the priors are not noticeable, which means that these should not have had much influence on the parameter samples. One can also see that the histograms are generally quite wide. The parameter q, shown in Figure 48, is estimated to be larger in the north, that is Oslo and Oscarsborg, and smaller in the south. The parameter $\ln s$ is shown in Figure 49, a spatial pattern for this parameter is not immediately obvious. The shape parameter, shown in Figure 47, always covers the interval $[-0.25, 0]$, so it is reasonable to expect the common ξ in the next section to be estimated in that interval. One can see that the parameter estimates for Viker seem to have larger uncertainty than the other three locations, this is reasonable since Viker is the location with the least data, that is only 30 data points.

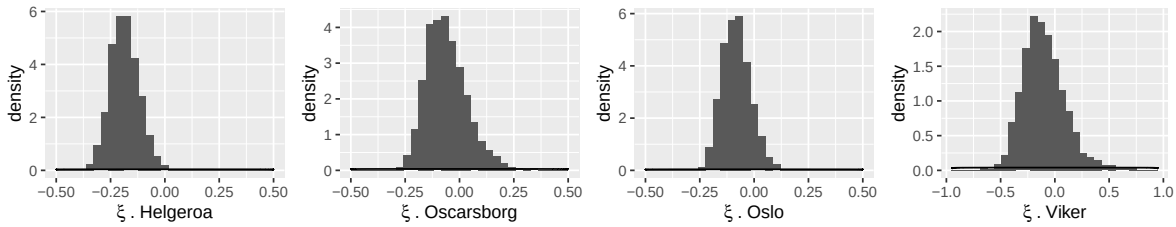


Figure 47: Histograms of the parameter estimates of ξ from a univariate analysis on the sea level data.

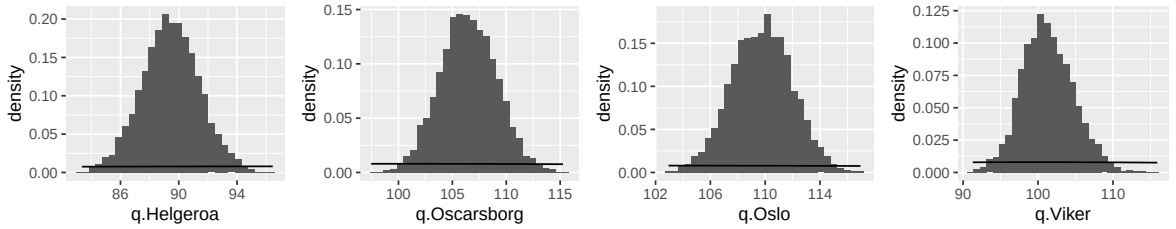


Figure 48: Histograms of the parameter estimates of q from a univariate analysis on the sea level data.

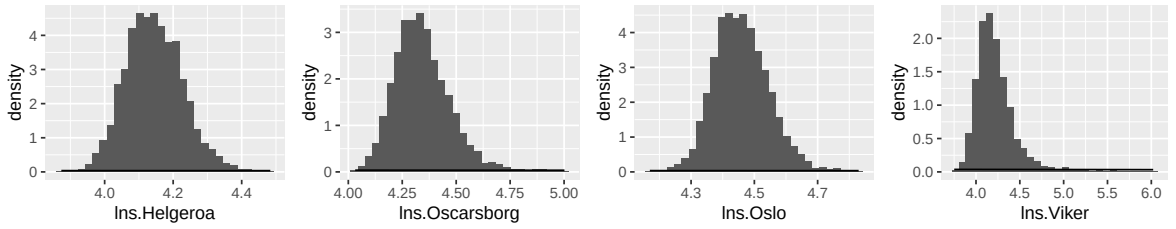


Figure 49: Histograms of the parameter estimates of $\ln s$ from a univariate analysis on the sea level data.

Return level plots show the median of the sea level that is expected to be exceeded every m years with uncertainty bounds plotted against the return period m. The return level plots for the univariate analysis are shown in Figure 50. One can see that the empirical return level values seem to be contained within the uncertainty bounds for all the locations. The width of the uncertainty bounds differ drastically from station to station. The bounds for Viker are extremely wide, whereas the bounds for Helgeroa are very narrow. It is somewhat surprising that Helgeroa has narrower bounds than Oslo, since Oslo has 20 more data points. The reason for this can potentially be the complicated geography of the Oslo area, whereas Helgeroa is located in a very open area. It is difficult to specify the risk of extreme sea levels when the bounds are very wide, as the return levels could vary quite drastically.

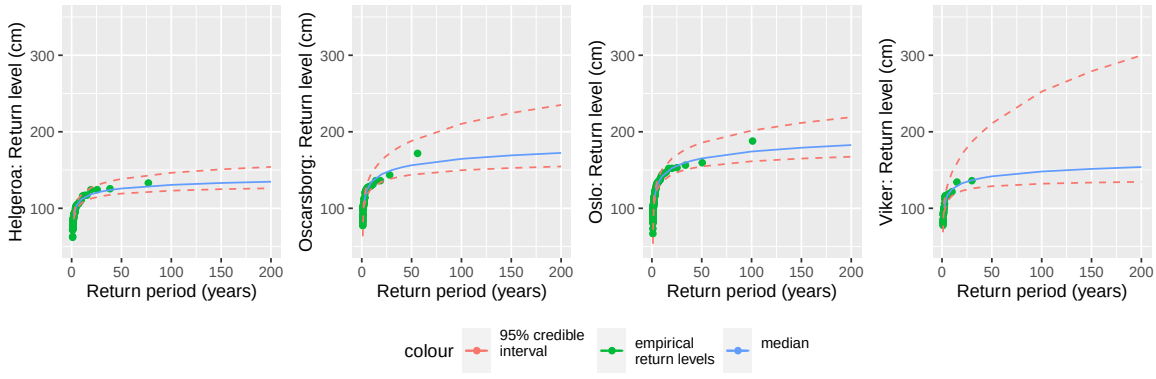


Figure 50: Return level plots for the univariate analysis of the four stations: Helgeroa, Oscarsborg, Oslo and Viker.

5.2 Analysis with Common Parameter ξ Across Locations

We perform an analysis where the parameter ξ is assumed to be common for the different locations while the parameters q and $\ln s$ are assumed to be independent in space. This analysis is performed due to the parameter ξ being quite difficult to estimate and we are also interested in comparing the results to the multivariate analysis. We let z_{ij} be the j-th observed maximum sea level at location i. The model likelihood is

$$
\begin{aligned}
p(\mathbf{z}|\mathbf{q}, \ln \mathbf{s}, \tilde{\xi}) &= \prod_{i=1}^{l} \prod_{j=1}^{k_i} \mathrm{GEV}(z_{ij}|q_i, s_i, \xi) \\
\mathrm{GEV}(z_{ij}|q_i, s_i, \xi) &= \exp\Big\{ -\Big[\Big(\frac{z_{ij} - q_i}{s_i(l_{1-\beta/2,\xi} - l_{\beta/2,\xi})^{-1}} + l_{\alpha,\xi}\Big)\Big]_+^{-1/\xi}\Big\} \\
l_{\gamma,\xi} &= (-\log(\gamma))^{-\xi} \\
s_i &= \exp(\ln s_i) \\
\xi &= a + (b-a) \cdot \frac{\exp\{\tilde{\xi}\}}{1 + \exp\{\tilde{\xi}\}}
\end{aligned}
\tag{47}
$$

where l is the number of locations and k_i is the number of data z available in location i. q is the $\alpha-$quantile and s is the difference between the $1-\beta/2$ and the $\beta/2$ quantiles. a and b are the lower and upper bounds of the shape parameter ξ. The limits of the parameter ξ are chosen to be $a = -0.5$ and $b = 0.5$ and the quantiles α and β are chosen to be 0.5 and 0.05 respectively.

We use the same data as was used in the last section and STAN to find samples of the parameters based on the posterior model. The posterior distribution is

$$\begin{aligned} p(\mathbf{q}, \ln \mathbf{s}, \tilde{\xi}|\mathbf{z}) \propto p(\mathbf{z}|\mathbf{q}, \ln \mathbf{s}, \tilde{\xi}) \\ \cdot p(\tilde{\xi})p(\mathbf{q})p(\ln \mathbf{s}). \end{aligned} \tag{48}$$

The probability distributions representing the prior knowledge can be explained with the following equations

$$\begin{aligned} \tilde{\xi} &\sim \text{normal}(0, 10) \\ \mathbf{q} &\sim \text{normal}(100, 50) \\ \ln \mathbf{s} &\sim \text{normal}(0, 10) \end{aligned} \tag{49}$$

The resulting parameter estimates are presented in Figures 51, 52 and 53. One can see that the histogram for the parameter ξ, shown in Figure 51, covers the exact interval that was expected from the univariate analysis. We compare the estimates for q in Figure 52 with the ones obtained in the univariate analysis in Figure 48. One can see that the estimates are close to identical. We also compare the estimates for $\ln s$ in Figure 53 with the ones obtained in the univariate analysis in Figure 49. One can see that the estimates are very similar for Helgeroa, Oscarsborg and Oslo, while the estimates for Viker differ. The upper bound for the parameter $\ln s$ was much larger in the univariate analysis.

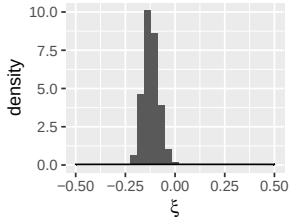


Figure 51: Histogram of the parameter estimates for ξ based on a model with common ξ across the locations, while the parameters q and $\ln s$ are not represented by spatial fields. The four locations are: Helgeroa, Oscarsborg, Oslo and Viker.

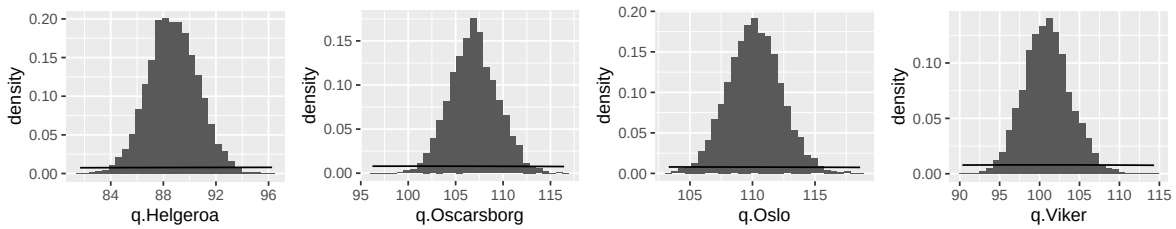


Figure 52: Histograms of the parameter estimates for q based on a model with common ξ across the locations, while the parameters q and $\ln s$ are not represented by spatial fields. The four locations are: Helgeroa, Oscarsborg, Oslo and Viker.

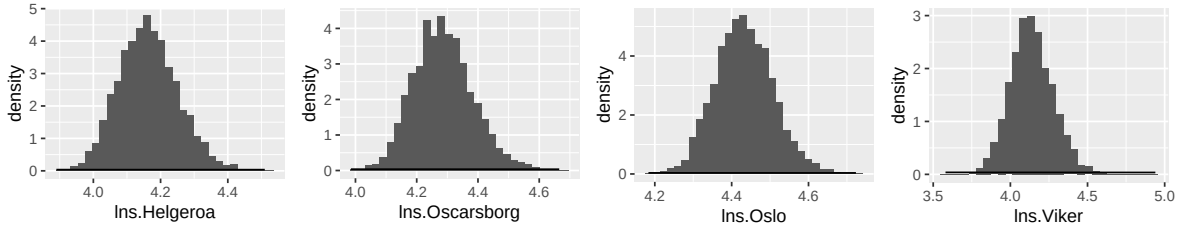


Figure 53: Histogram of the parameter estimates for $\ln s$ based on a model with common ξ across the locations, while the parameters q and $\ln s$ are not represented by spatial fields. The four locations are: Helgeroa, Oscarsborg, Oslo and Viker.

The return level plots are shown in Figure 54. One can see that most of the empirical return level values are contained within the 95% credible bounds, however the largest values for Oslo and Oscarsborg are just barely outside of the bounds. We compare the return level plots to the ones obtained in the univariate analysis in Figure 50. One can see that the uncertainty bounds are narrower for all locations except for Helgeroa, however Helgeroa had quite narrow uncertainty bounds also in the univariate analysis. This is a large improvement since the bounds were way to wide in the univariate analysis, which made it impossible to pinpoint the return levels.

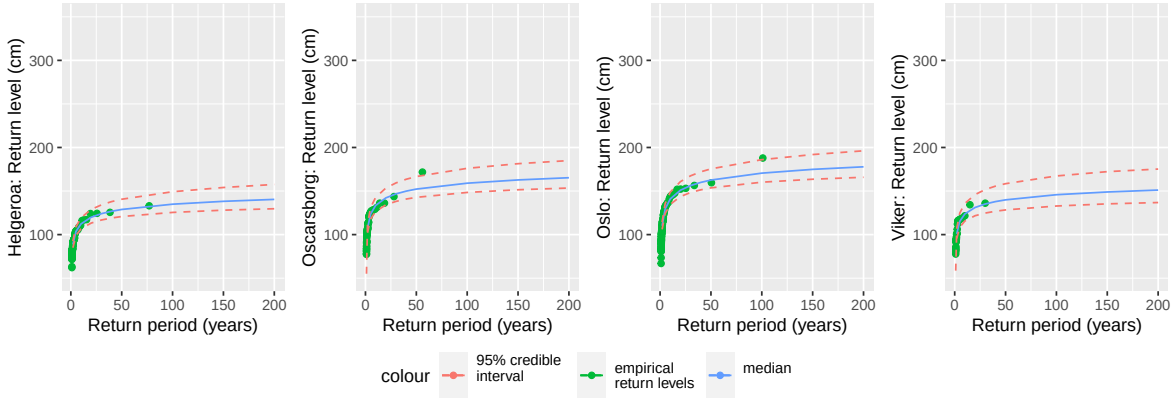


Figure 54: Return level plots for the univariate analysis of the four stations: Helgeroa, Oscarsborg, Oslo and Viker.

5.3 Multivariate Analysis

The multivariate model is applied to the data in this section, this is the main analysis we will perform on the sea level data. We restate the model defined in Section 3.1.2 and 4.1. We let z_{ij} be the j-th observed maximum sea level at location i. The model likelihood is

$$
\begin{aligned}
p(\mathbf{z}|\mathbf{q_\alpha}, \ln \mathbf{s_\beta}, \tilde{\xi}) &= \prod_{i=1}^{l} \prod_{j=1}^{n_i} \text{GEV}(z_{ij}|q_{\alpha,i}, s_{\beta,i}, \xi) \\
\text{GEV}(z_{ij}|q_{\alpha,i}, s_{\beta,i}, \xi) &= \exp\Big\{ -\Big[\Big(\frac{z_{ij} - q_{\alpha,i}}{s_{\beta,i}(l_{1-\beta/2,\xi} - l_{\beta/2,\xi})^{-1}} + l_{\alpha,\xi}\Big)\Big]_+^{-1/\xi}\Big\} \\
l_{\gamma,\xi} &= (-\log(\gamma))^{-\xi} \\
s_{\beta,i} &= \exp(\ln s_{\beta,i}) \\
\xi &= a + (b-a) \cdot \frac{\exp\{\tilde{\xi}\}}{1+\exp\{\tilde{\xi}\}}
\end{aligned}
\tag{50}
$$

where l is the number of locations and n_i is the number of data available in location i. q_α is the $\alpha-$quantile and s_β is the difference between the $1-\beta/2$ and $\beta/2$ quantiles. a and b are the lower and upper bounds of the shape parameter ξ. We set $a = 0$ and $b = 0.5$. We choose to look at the quantiles $\alpha = 0.5$ and $\beta = 0.05$ because the parameter q_α then represents the median while the parameter s_β represents the 95% credible interval of the GEV distribution.

Moreover we let

$$\begin{aligned} q_{\alpha,i} &= \mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{q}} + u_i \\ \mathbf{u} &\sim \text{normal}(0, \boldsymbol{\Sigma}_q) \\ \Sigma_{q,ik} &= \tau_q^2 \cdot \Sigma(d_{ik}|\nu, \rho_q) \end{aligned} \tag{51}$$

$$\begin{aligned} \ln s_{\beta,i} &= \mathbf{x}_i^T \boldsymbol{\beta}_{\mathbf{s}} + w_i \\ \mathbf{w} &\sim \text{normal}(0, \boldsymbol{\Sigma}_s) \\ \boldsymbol{\Sigma}_{s,ik} &= \tau_s^2 \cdot \Sigma(d_{ik}|\nu, \rho_s) \end{aligned}$$

where $\mathbf{x}_i$ is the covariate vector at location i consisting of an intercept and the covariates presented in Section 2.2. $\boldsymbol{\beta}_q$ and $\boldsymbol{\beta}_s$ are parameter vectors that together with the covariates decides the mean of the parameters q_α and $\ln s_\beta$ in every point i. u_i and w_i are the values of the Gaussian random fields at location i. $d_{ik} = |r_i - r_k|$ is the distance between two points in the Gaussian random field. The parameters τ_q, τ_s are the standard deviation of the Gaussian random fields and $\Sigma(d|\nu, \rho)$ is the Matern correlation function with ranges ρ_q and ρ_s.

We use STAN to find samples of the parameters based on the posterior model, this model can be stated as follows

$$\begin{aligned} p(\mathbf{q}, \mathbf{ln\, s}, \tilde{\xi}, \beta_q, \beta_s, \tau_q, \tau_s, \rho_q, \rho_s|\mathbf{z}) \propto\ & p(\mathbf{z}|\mathbf{q}, \mathbf{ln\, s}, \tilde{\xi}) \\ & \cdot p(\mathbf{q}|\beta_q, \tau_q, \rho_q) \\ & \cdot p(\mathbf{ln\, s}|\beta_s, \tau_s, \rho_s) \\ & \cdot p(\tilde{\xi})p(\beta_q)p(\beta_s)p(\tau_q)p(\tau_s)p(\rho_q)p(\rho_s). \end{aligned} \tag{52}$$

The probability distributions representing the prior knowledge can be explained with the following equations

$$\begin{aligned} \tilde{\xi} &\sim \text{normal}(0, 10) \\ \beta_q &\sim \text{normal}(0, 100) \\ \beta_s &\sim \text{normal}(0, 100) \\ \tau_q &\sim \text{gamma}(5/4, 1/4) \\ \tau_s &\sim \text{gamma}(5/4, 1/4) \\ \rho_q &\sim \text{gamma}(5, 1/12) \\ \rho_s &\sim \text{gamma}(5, 1/12) \end{aligned} \tag{53}$$

The parameter estimates obtained are displayed in Figures 55, 56, 57, 58 and 59. We study the results in Figures 56 and 57 and observe that the histograms of the parameter estimates of the β's are very wide and that neither of the parameters of the Gaussian random field, that is the standard deviations and the ranges, have changed from the priors. These results are

not unexpected due to small amount of data available and the same results were also observed in the simulation study.

One can see that the estimates of the parameters q and $\ln s$ seem to be approximately the same. The exception is the parameter $\ln s$ for Viker, where the upper bound is much larger in the univariate case. The histograms of the parameters in this section seem to be close to identical to the histograms obtained for the analysis with a common ξ, this indicates that most of the improvements seen in the multivariate analysis are caused by having a common ξ across the locations.

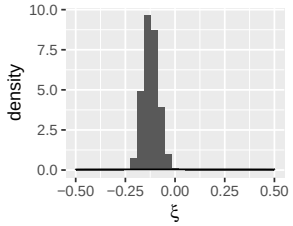


Figure 55: Histogram of the parameter estimate of ξ from the multivariate analysis of the sea level data.

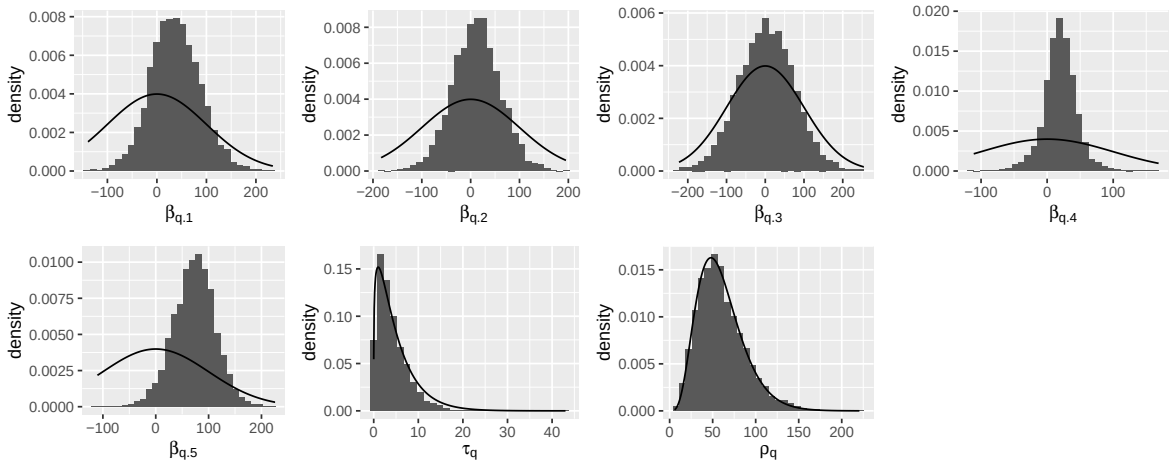


Figure 56: Histogram of the parameter estimates used to specify q from the multivariate analysis of the sea level data.

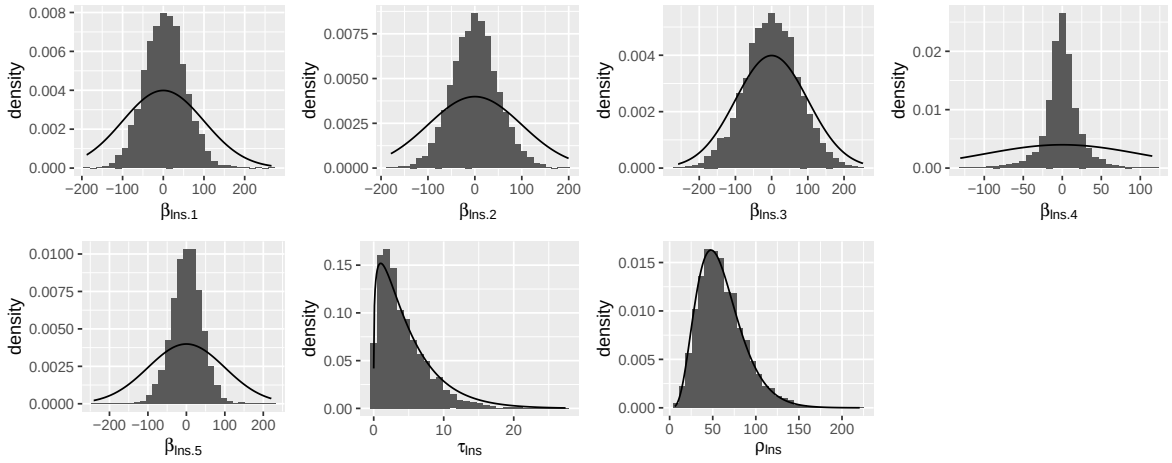


Figure 57: Histogram of the parameter estimates used to specify $\ln s$ from the multivariate analysis of the sea level data.

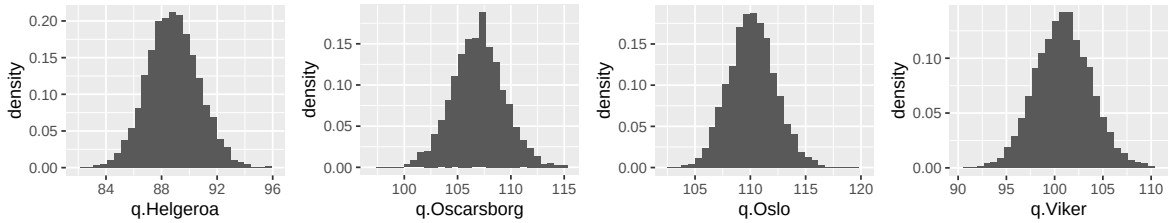


Figure 58: Histogram of the parameter estimates of q from the multivariate analysis of the sea level data.

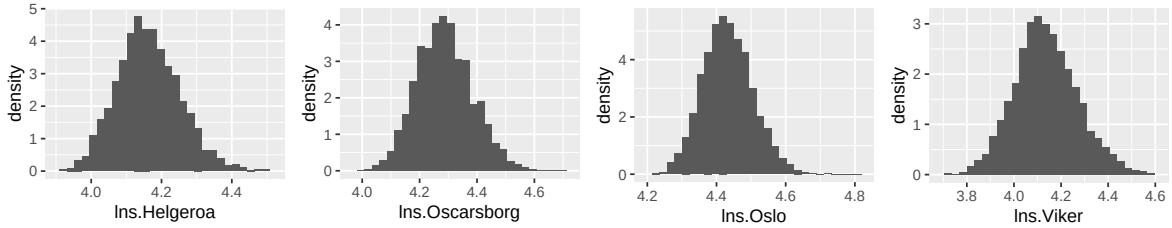


Figure 59: Histogram of the parameter estimates of $\ln s$ from the multivariate analysis of the sea level data.

The return level plots from the multivariate analysis are shown in Figure 60. One can see that most of the empirical return level estimates coincide very well with the median of the estimated return levels. The exceptions are the last points from Oscarsborg and Oslo, however these points do not seem to follow the same path that has been mapped out by the other points. We also compare the result to the equivalent plot from the analysis using a common

ξ, that is Figure 54. One can see that the plots are close to identical in the two cases, which again indicates that most of the improvements we see in the multivariate model are caused by having a common parameter ξ.

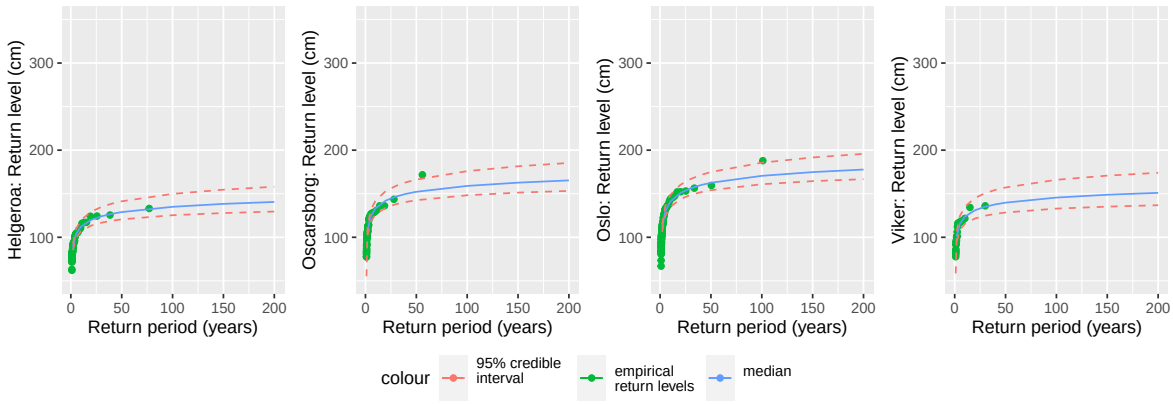


Figure 60: The return level plots based on sea level data from Helgeroa, Oscarsborg, Oslo and Viker. The plots contain the medians and the 95% credible intervals of the return levels plotted against the return periods. The empirical return levels based on the data are also shown.

We extend the return level estimates to the entire Oslo fjord by using conditional normality. The resulting medians and 95% credible bounds are shown in Figure 61. One can see that the upper bounds are extremely large in the areas where no data is available, meaning that we cannot make any conclusions about the return levels. This result is expected since the distances are quite large, meaning that the estimates are based on very little information. One can see that the lower bounds are slightly larger close to the locations were we have sea level data and that the upper bounds are significantly lower close to the same locations.

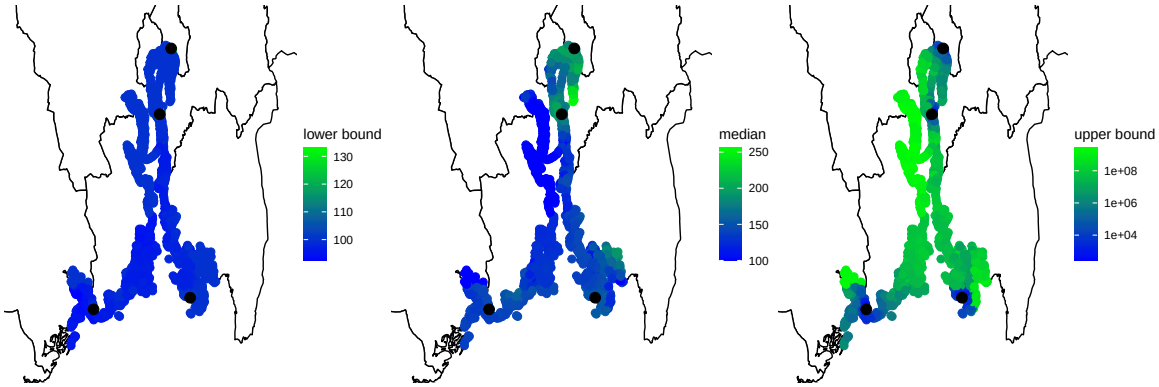


Figure 61: Maps of return levels in the entire Oslo fjord based on parameter estimates from the multivariate analysis of the sea level data from Helgeroa, Oscarsborg, Oslo and Viker. In the plots the lower bound of the 95% credible interval on the left, the median in the middle and the upper bound of the 95% credible interval on the right.

6 Discussion

The initial objective of this thesis was to fit a spatial extreme value model to the sea level data in the Oslo fjord. We performed a simulation study, this was done to see how the model behaves with optimal data and to see what results could have been obtained if more data was available. Next, we studied the real data, for this data we performed a univariate analysis, an analysis where only the shape parameter ξ of the GEV distribution had a spatial representation and a multivariate analysis. The exact same analyses were performed for the simulated data based on 4 locations. In the simulation study we also applied the multivariate model to simulated data for 15 and 8 locations. The simulation study based on 15 locations was performed on long data series, while the study based on 8 locations had 4 longer series and 4 shorter series to study the effect of obtaining additional data.

The spatial model seems to have lead to improvements in the estimation of return levels at the monitored locations. The improvements seem to be caused mainly by the common ξ parameter, and not so much by the spatial representation of q and $\ln s$. This result is not unexpected as the parameter ξ being quite difficult to estimate as well as the model being based on only 4 locations, which means that there is not much information available to estimate spatially varying parameters. The estimation of return levels at unmonitored locations was not very informative. There is however hope that the uncertainty would decrease significantly if some additional stations with some years of sea level data were available, as seen in the simulation study.

A large advantage of the spatial model is that one can share information between locations. This is very beneficial since the largest problem when modelling extremes is the small amounts of data available to perform inference. Extremes are rare by definition, and long data series are therefore necessary to model extremes accurately. A disadvantage of the spatial model is that it is difficult to estimate the spatial parameters, especially the range. The range parameter barely deviated from the chosen prior even in the simulation study based on the largest number of locations, that is 15. We tried two different informative priors for the range parameters. The results were close to identical for all the parameters except the range, which indicated that choosing a reasonable informative prior for the range should not greatly influence the rest of the results.

There are difficulties related to estimating the spatial parameters, and this is especially prominent when we try to fit the model with data from only a few locations. We compare the analysis of a model with a common shape parameter ξ and independent parameters q and $\ln s$ and the multivariate model. This comparison shows that the models have very similar parameter estimates and return levels. This indicates that the simpler model might be just as good of a choice when only very few sparse data are available. The downside of choosing the simpler model is that the spatial interpolation is not possible, the results from this analysis were however very poor when there was not much data. One can argue for keeping the spatial model since there is not much difference between the two models and as there is always the possibility of obtaining more data, adding more locations and data to the model is not difficult.

The biggest challenge we face in this project is the lack of data. An argument against using the GEV model is that the model is very wasteful with data. One year of hourly measurements is equivalent to one data point. This problem could potentially have been solved by using a threshold exceedance model, that is a model that uses all the data exceeding a threshold. Hand tuning is however necessary when one uses threshold exceedance models, as one needs to find a reasonable threshold at every location. The process of finding the threshold is in addition a potential source of error as the thresholds are chosen based on subjective inspection of plots. The threshold should probably also be space dependent, which would make it even more difficult to estimate. The GEV model is thus a more wasteful, but also an easier and a safer option.

We stated that one can obtain more data, the statement is true, however the process of obtaining new data is very tedious as one needs one year of hourly measurements to obtain one data point that can be used in the model. This means that the simulation study based on 15 locations, which obtained very good results, will not be possible in practice until somewhere between 50 and 100 years with data collection have occurred. The simulation study based on 8 locations, where the data series from the additional 4 stations were short, is an analysis for which it is much easier to obtain the necessary data. The results show that quite informative estimates for the return levels can be obtained in large portions of the Oslo fjord if 4 additional locations with between 5 and 15 years of data existed. This analysis is however based on simulated data, so an equivalent analysis based on real data will not necessarily be as accurate.

Checking the model is quite challenging in this project, this is due to the sparsity of the data as well as the interest in making estimates in areas where there is no data. The model checking was limited to comparing the results to simulation studies, univariate results and empirical results where those were available. Checking the spatial prediction by performing cross validation was not feasible as there was not enough data to get informative uncertainty bounds.

7 Conclusion

The aim of this project was to fit a spatial extreme value model to the sea level data in the Oslo fjord for the purpose of obtaining information about the expected frequency of flooding. This is a point of interest because flooding charts are needed by for example insurance and construction companies to make informed decisions. Return levels with associated return periods are often used to portray the information. The model used to describe the probability of extreme occurrences was a GEV model where the parameter ξ was assumed to constant in space, while the parameters q and $\ln s$ were assumed to vary in space.

The results of this project showed that the spatial model caused improvements of the estimates of return levels compared to a univariate model. It was also shown that the improvements seemed to be mostly caused by the shape parameter ξ being common across the locations. The spatial representation of the parameters q and $\ln s$ not affecting the results as much was most likely caused by sparse data, as this makes it quite difficult to estimate spatial dependencies. Another result that was observed was that a spatial interpolation to unmonitored locations was not informative due to very wide uncertainty bounds.

The sparse data was the main limitation of this study. The results from the simulation study indicates that additional data could greatly improve the results, especially the spatial interpolation. The problem with obtaining additional data is that it is quite time consuming. However, it was also shown that numerous shorter series of additional data can lead to quite large improvements.

The results from this thesis will hopefully be helpful for future projects that want to model the probability of occurrences of extreme sea levels in the Oslo fjord. A big challenge to overcome for future research is the sparse data and associated obstacles. An interesting extension of this project would be to study the entire Norwegian sea line. This is however challenging due to the different behaviour of the sea level that can be observed along the shore.

References

N. Beck, C. Genest, J. Jalbert, and M. Mailhot. Predicting extreme surges from sparse data using a copula-based hierarchical bayesian spatial model. *Environmetrics*, 31(5):e2616, 2020. doi: https://doi.org/10.1002/env.2616.

B. Bell, H. Hersbach, P. Berrisford, P. Dahlgren, A. Horányi, J. Muñoz Sabater, J. Nicolas, R. Radu, D. Schepers, A. Simmons, C. Soci, and J-N. Thépaut. (2020): Era5 hourly data on single levels from 1950 to 1978 (preliminary version). copernicus climate change service (c3s) climate data store (cds). `https://cds.climate.copernicus-climate.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-preliminary-back-extension?tab=overview`, 2021. (Accessed on 06-04-2021).

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo, 2018. URL `https://arxiv.org/abs/1701.02434`.

C. Bracken, B. Rajagopalan, L. Cheng, W. Kleiber, and S. Gangopadhyay. Spatial bayesian hierarchical modeling of precipitation extremes over a large domain. *Water Resources Research*, 52(8):6643–6655, 2016. doi: https://doi.org/10.1002/2016WR018768.

Daniela Castro-Camilo, Raphaël Huser, and Håvard Rue. Practical strategies for gev-based regression models for extremes, 2021. URL `https://arxiv.org/abs/2106.13110`.

Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001. doi: https://doi.org/10.1007/978-1-4471-3675-0.

Stuart Coles and Jonathan Tawn. Bayesian modelling of extreme surges on the uk east coast. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 363(1831): 1387–1406, 2005. ISSN 1364503X. URL `http://www.jstor.org/stable/30039659`.

Daniel Cooley, Douglas Nychka, and Philippe Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479):824–840, 2007. doi: https://doi.org/10.1198/016214506000000780.

Anita Verpe Dyrrdal, Alex Lenkoski, Thordis L. Thorarinsdottir, and Frode Stordal. Bayesian hierarchical modeling of extreme hourly precipitation in norway. *Environmetrics*, 26(2):89–106, 2015. doi: https://doi.org/10.1002/env.2301.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 3rd ed. edition, 2018. URL `http://www.stat.columbia.edu/~gelman/book/BDA3.pdf`.

Geonorge. Tilnærma maksimal tidevannsamplitude, avstand frå middelvatn til lågaste astronomiske tidvatn (lat). `https://kartkatalog.geonorge.no/metadata/dd1c9967-86c4-479c-b7b4-fc326c5d8783`, 2021. (Accessed on 23-04-2021).

G. H. Givens and J. A. Hoeting. *Computional Statistics*. John Wiley & Sons, Inc., 2013. doi: https://doi.org/10.1002/9781118555552.

H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. (2018): Era5 hourly data on single levels from 1979 to present. copernicus climate change service (c3s) climate data store (cds). 10.24381/cds.adbb2d47, 2021. (Accessed on 06-04-2021).

Magnus Hieronymus, Christian Dieterich, Helén Andersson, and Robinson Hordoir. The effects of mean sea level rise and strengthened winds on extreme sea levels in the baltic sea. *Theoretical and Applied Mechanics Letters*, 8(6):366–371, 2018. ISSN 2095-0349. doi: https://doi.org/10.1016/j.taml.2018.06.008.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. URL `http://jmlr.org/papers/v15/hoffman14a.html`.

Kartverket. Havnivå data fra permanente stasjoner i norge, 1990-2021. `http://api.sehavniva.no/tideapi_no.html`, 2021. (Accessed on 01-2021).

Finn Lindgren and Håvard Rue. Bayesian spatial modelling with r-inla. *Journal of Statistical Software, Articles*, 63(19):1–25, 2015. ISSN 1548-7660. doi: https://doi.org/10.18637/jss.v063.i19.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011. doi: https://doi.org/10.1111/j.1467-9868.2011.00777.x.

Marion Helen Røed. Extreme value analysis of sea levels on the norwegian coast, 2021. Unpublished Project Report.

Morten Skjong, Arvid Naess, and Ole Erik Brandrud Næss. Statistics of Extreme Sea Levels for Locations along the Norwegian Coast. *Journal of Coastal Research*, 29(5):1029–1048, 2013. doi: https://doi.org/10.2112/JCOASTRES-D-12-00208.1.

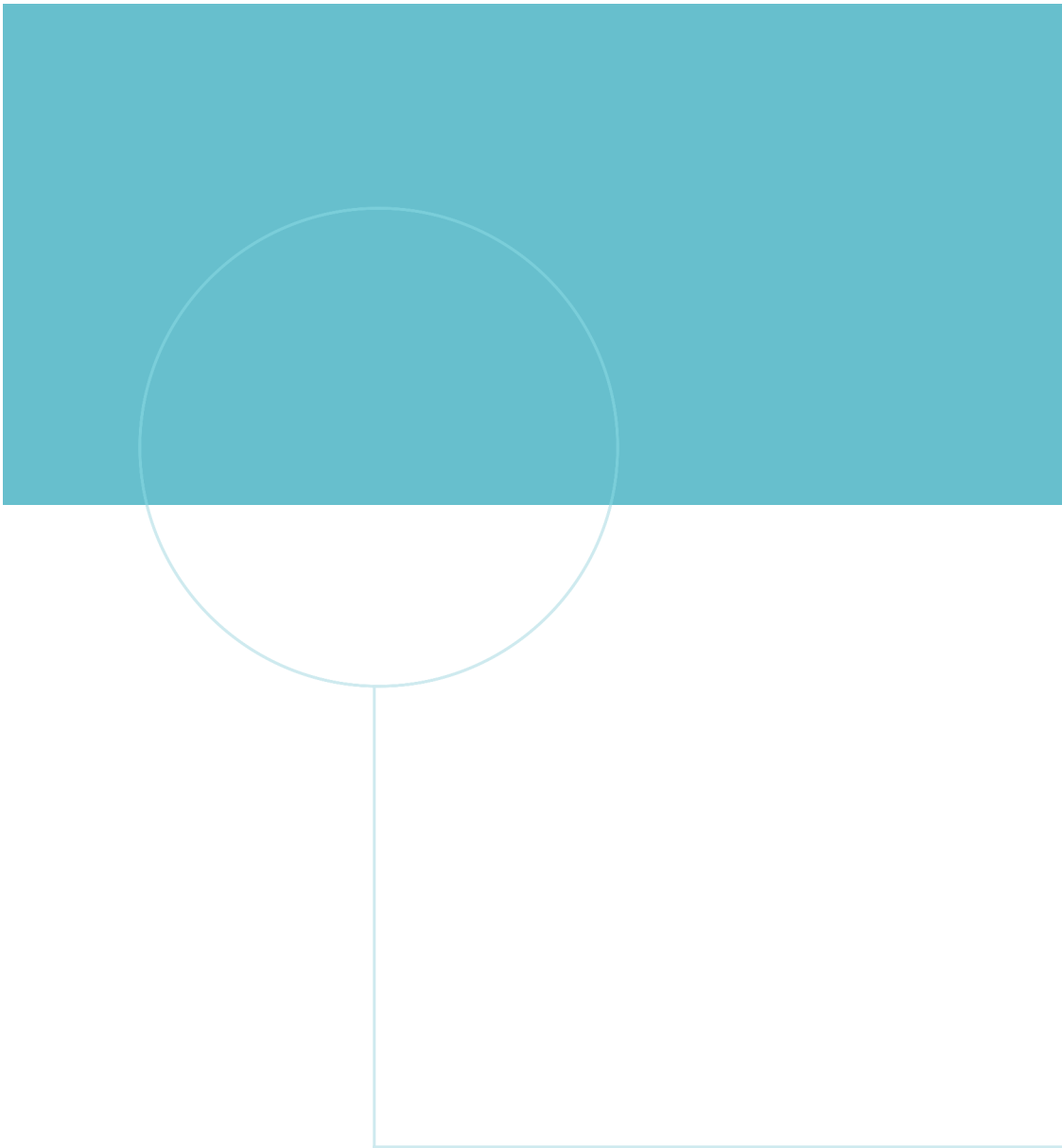
Stan Development Team. Stan modeling language users guide and reference manual, version 2.26, 2019. URL `https://mc-stan.org`.

Michael L. Stein. *Interpolation of Spatial Data*. Springer, New York, NY, 1999. doi: https://doi.org/10.1007/978-1-4612-1494-6.

A Code

The code used in this thesis is available at github:

https://github.com/marionhr/Spatial-Extreme-Value-Modelling-of-Sea-Level-Data-in-the-Oslo-Fjord



NTNU

Norwegian University of
Science and Technology