LUND UNIVERSITY
DEPARTMENT OF STATISTICS
Master thesis

On Risk Analysis Of Extreme Sea Levels In Falsterbo Peninsula

Author:
Kevin Persson

Supervisor:
Nader Tajvidi

June 13, 2017
Abstract

Coastal protection is vital for protecting infrastructure, coastal environments and human lives against flooding. Building efficient coastal protection requires a good understanding of maximum sea levels which might occur in different time periods in the future. Extreme value theory provides a mathematical framework for such analyses.

On November 13, 1872 the biggest recorded sea level surge devastated the Danish, German and Swedish Baltic Sea coast. This master thesis focuses on estimating the return period of 1872 storm using one-dimensional extreme value analysis based on historical data from the measure stations near Falsterbo Peninsula. A multivariate extreme value approach is applied to include covariates such as wind speed and wave height to further improve the understanding of which factors affect extreme sea levels.

Fit diagnostics show that the block maxima model based on observations from Klagshamn measure station provides the best fit for the data; hence it has been used to estimate the return period of the 1872 storm. Wind speed and wave height from nearby station were used to improve the accuracy of the analysis in a multivariate framework.
Acknowledgments

I would like to thank my supervisor Nader Tajvidi at the Center for Mathematical Science, Lund University, for his support and valuable statistics expertise. I would also like to thank Caroline Fredriksson and Magnus Larsson at the Water Resources Engineering, Lund University, for their help in introducing the subject and data management. Lastly I would like to thank all my friends and family who have supported me throughout my studies.
# Contents

1 Introduction .................................................. 1  
1.1 Background .................................................. 1  
1.2 Aim of thesis ................................................. 2  

2 Theory ......................................................... 3  
2.1 Stationary process .......................................... 3  
2.2 Extreme value theory ....................................... 4  
  2.2.1 Model formulation ...................................... 5  
  2.2.2 Peaks over threshold ................................... 7  
  2.2.3 Parameter estimation ................................... 8  
  2.2.4 Stationary sequence .................................... 9  
  2.2.5 Return period estimate ................................ 11  
2.3 Multivariate extremes ..................................... 12  
  2.3.1 Characterization of bivariate extreme value distribution 13  
  2.3.2 Bivariate extreme value distributions ................ 14  
2.4 Dependence measure ......................................... 18  
2.5 The auto-correlation function (ACF) ....................... 19  
2.6 Model validation ............................................ 21  
  2.6.1 Fit diagnostic ........................................... 21  
  2.6.2 Akaike information criteria ............................ 21  
  2.6.3 Goodness-of-fit test for BGPD ......................... 22  

3 Analysis ....................................................... 23  
3.1 The data ..................................................... 23  
3.2 Correlation analysis ....................................... 24  
3.3 Missing values .............................................. 27  

4 One dimensional analysis of the data ....................... 28  
4.1 Block-maxima ............................................... 28  
4.2 Generalized Pareto Distribution ............................ 30  
4.3 Recommendation ............................................ 31
5 Multivariate analysis, wave and wind
  5.1 Skanör and Klagshamn ................................................. 33
    5.1.1 Type 1 BGPD ..................................................... 34
    5.1.2 Type 2 BGPD ..................................................... 38
  5.2 Viken and wind speed ................................................ 42
    5.2.1 Type 1 BGPD ..................................................... 45
    5.2.2 Type 2 BGPD ..................................................... 49
  5.3 Angelholm, waves and sea level .................................... 53
    5.3.1 Type 1 BGPD ..................................................... 54
    5.3.2 Type 2 BGPD ..................................................... 58

6 Discussion ................................................................. 62
  6.1 Summary ............................................................... 62
  6.2 Discussion ............................................................ 63
Chapter 1

Introduction

1.1 Background

Extreme high sea levels cause coastal flooding which damages infrastructure, coastal environments and in some cases even takes human lives. Protection against flooding is based on the probability of flooding (return period). The return period is based on risk. Thus if the probability of a flood is large but the consequence are small a lower return period can be chosen. One must find a balance between cost, environmental and other effects on the society in order to build an efficient coastal protection.

On the 13 November 1872 the biggest recorded storm surge devastated the Danish, German and Swedish Baltic Sea coast, which took the lives of 271 people, destroyed 2800 buildings and left 15000 persons homeless. The estimated storm surge was 2.4 meters above normal sea level. The Falsterbo Peninsula, in Vellinge municipality in south Sweden, is prone to flooding since it’s low-lying. The municipality in Vellinge follows a protection plan that is stepwise implemented with purpose to protect the urban areas from flooding until year 2065. Falsterbo Peninsula is vulnerable since the outer part can be cut off from the mainland, leaving the rescue service unable to get access. Therefore it is especially important that return periods longer than 100 years are considered, possible varying return periods in different areas based on vulnerability.
1.2 Aim of thesis

The question many ask today is can this happen again? A published paper by [5] asked this question and answered it using extreme value analysis in one dimension. Variables such as wind speed was not used in their modeling but considered. The authors suggested that a more advanced study and models should be implemented in order do get a more precise return period.

This thesis will focus on estimating the return period of the 1872 storm using one dimensional extreme value theory. Analyze if covariates such as wind speed has an significant effect on extreme water levels in Falsterbo, if not do other measure stations have significant covariates? If so a multivariate extreme value approach is applied.
Chapter 2

Theory

2.1 Stationary process

A sequence of random variables $X_1, X_2, \ldots$ is said to be a random process. Example of an random process is sequence of independent identically distributed random variables. Often real-life phenomenon don’t exhibits this property, usually their is often dependence through time which depends on the history of the process. It’s not uncommon for the data to have seasonal behaviors as well, such as temperature data. One has to address both seasonal behaviors and time dependency separately. It’s often convenient to study a process which exhibits time dependency but its behavior is homogeneous though time. Now we are ready for the definition of a stationary process.

Definition 3.1 A random process $X_1, X_2, \ldots$ is said to be stationary if, given integers $\{i_1, \ldots, i_k\}$ and any integers $m$, the joint distribution of $\{X_{i_1}, \ldots, X_{i_k}\}$ and $\{X_{i_1+m}, \ldots, X_{i_k+m}\}$ are identical for any choice of $m$ [1, p. 25].

Stationarity implies that given any subset of variables the joint distribution of the same subset viewed $m$ time points later remains unchanged. The difference between an independent series and stationarity is that a stationary series does not preclude $X_i$ being dependent on its previous values, but $X_{i+m}$ must have the same dependence on its previously values. One exclude trends, seasonality and other deterministic cycles from the assumption of stationarity.
2.2 Extreme value theory

Extreme value theory has grown to be one of the more important statistical disciplines in the recent years. The techniques used in extreme value theory are widely used in other disciplines such as portfolio adjustment, hurricane damage, traffic prediction in telecommunication and sea level. The features of an extreme value analysis allows us to quantify a stochastic process where its unusually low or high. Often the required task is to estimate the probability that an event so extreme that it has not been observed.

An example of this is for coastal protection. In coastal protection one usually builds a sea-wall that should protect against all sea levels that can occur during its life span of 100-years. Data has been collected but only for instance during a 10-year period. Thus the challenge lies in design the sea wall that should protect during 100-years with only 10-years of historical data are available. Extreme value theory provides a framework of how one can solve this problem and similar of this nature. Since one often lacks the knowledge of the physical and empirical guidelines to extrapolate, one uses the asymptotic argument. Suppose we denote the hourly sea level with $X_1, X_2, \ldots$. Let

\[ M_n = \max\{X_1, X_2, \ldots, X_n\} \]

denote the maximum of sea level over a period of $n$. If we know the behavior of $X_i$ we would also know the behavior of $M_n$. Often the distribution of $X_i$ is unknown thus making it impossible for one to know the exact distribution of $M_n$. 
2.2.1 Model formulation

A natural approach is to study the distribution function of $M_n$, this will yield

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

However, this is not particularly useful since the distribution $F$ is unknown. Another drawback is also that the limiting distribution $G$ is a degenerate function in $x^F = \sup\{x \in \mathbb{R} : F(x) < 1\}$, i.e

$$\lim_{n \to \infty} F^n(x) = G(x) = \begin{cases} 
0, & x < x^F \\
1, & x \geq x^F.
\end{cases}$$

With this result one concludes that $M_n \xrightarrow{P} x^F$ and since the maximum is non-decreasing $M_n \xrightarrow{a.s} x^F$ even holds. This can be used when estimating $x^F$ but for statistical purpose this does not contribute.

To avoid getting a non-degenerate limiting distribution it is necessary to standardize the quantity of interest, similar to the Central Limit Theorem. The first important convergence result is in The Extremal Types Theorem, also more know as Fisher-Tippett-Gnedenko theorem [4]. Here we will use the version of [1, p. 46].
Theorem 2.2.1 Let \( X_1, \ldots, X_n \) be a sequence of iid random variables and let \( M_n = \max\{X_1, \ldots, X_n\} \). If there exist a sequence of constants \( \{a_n > 0\} \) and \( \{b_n\} \) such that

\[
P\left( \frac{M_n - b_n}{a_n} \leq x \right) \to G(x) \quad \text{as} \quad n \to \infty,
\]

where \( G \) is a non-degenerate distribution function, then \( G \) belongs to one of the following families:

\[I(Gumbel) : G(x) = \exp \left\{ - \exp \left[ - \left( \frac{x - b}{a} \right) \right] \right\}, \quad -\infty < x < \infty\]

\[II(Fr\'echet) : G(x) = \begin{cases} 0, & x \leq b \\ \exp \left\{ - \left( \frac{x - b}{a} \right)^{-\xi} \right\}, & x > b \end{cases}\]

\[III(Reversed Weibull) : G(x) = \begin{cases} \exp \left\{ - \left[ - \left( \frac{x - b}{a} \right) \right]^\xi \right\}, & x < b \\ 1, & x \geq b \end{cases}\]

for parameters \( a > 0 \), \( b \) and, in the case of families II and III, \( \xi > 0 \).

These three families of extreme value distributions are summarized to the Generalized Extreme Value distribution (GEV)

\[
G(x; \xi, \mu, \sigma) = \exp \left\{ - \left[ 1 + \xi \left( \frac{x - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} \right\}, \quad (2.1)
\]

where \(-\infty < \mu < \infty\), \( \sigma > 0 \) and \(-\infty < \xi < \infty \). If \( \xi = 0 \) one has Type I, if \( \xi > 0 \) one has Type II and if \( \xi < 0 \) one has Type III.
2.2.2 Peaks over threshold

A quite popular approach to model extreme value is Peaks Over Threshold (POT). This method uses all observations exceeding some threshold $u$. In contrast, the so-called block-maxima approach considers only the maxima within a given time interval. Fixing a threshold gives the advantage of not wasting information since the block maxima approach does not use other large observations which were observed within the same time interval. The main issues with POT is to determine the threshold level, as with block maxima fixing the size of the time interval.

The idea behind POT is based on the following result. Let $X_1, ..., X_n$ be a sequence of iid having the distribution function $F$ and let $X$ denote an arbitrary term in this sequence. If the conditions for the Extremal Types theorem are fulfilled one has $P(M_n \leq x) \approx G(x)$ for large $n$, where $G$ is GEV (2.1).

Fixing a threshold $u$ the conditional distribution of the exceedances $X - u$ given that $X > u$ can be calculated as

$$P(X - u > x|X > u) = \frac{1 - F(u + x)}{1 - F(u)}. \quad (2.2)$$

If the distribution of $F$ is known, the distribution of the threshold exceedances (2.2) would also be known. Since this is not the case one has to use approximations that are broadly applicable for high values of the threshold. By using the first order Taylor approximation $\log F(z) \approx -(1 - F(x))$, which only hold if $F(x) \approx 1$, one can conclude that

$$\frac{1 - F(u + x)}{1 - F(u)} \approx \left[ 1 + \frac{\xi(x/\sigma)}{1 + \xi(u - \mu)/\sigma} \right]^{-1/\xi} = \left[ 1 + \xi \frac{x}{\tilde{\sigma}} \right]^{-1/\xi} \quad (2.3)$$

where

$$\tilde{\sigma} = \sigma + \xi(u - \mu), \quad (2.4)$$

more details can be found in [1, p. 75-77]. The formula (2.3) implicates that the approximation is only valid if $F(u) \approx 1$, or equivalent if the threshold $u$ is large enough with the respect to the support of $F$. The distribution in (2.3) is known as the Generalized Pareto Distribution (GPD).
2.2.3 Parameter estimation

Below we will discuss estimation of parameters in GEV and GPD.

**GEV:** Under the assumption that \(X_1, \ldots, X_n\) are independent variables having a GEV distribution, the log-likelihood for the GEV parameters when \(\xi \neq 0\) is

\[
\log L(\mu, \sigma, \xi) = -n \log \sigma - (1+1/\xi) \sum_{i=1}^{n} \log \left[ 1 + \xi \left( \frac{x_i - \mu}{\sigma} \right) \right] - \sum_{i=1}^{n} \left[ 1 + \xi \left( \frac{x_i - \mu}{\sigma} \right) \right]^{-1/\xi},
\]

provided that

\[
1 + \xi \left( \frac{x_i - \mu}{\sigma} \right) > 0, \quad \text{for } i = 1, \ldots, n.
\]

In the case when \(\xi = 0\) one uses the Gumbel limit of the GEV distribution, this will result in the log-likelihood

\[
\log L(\mu, \sigma) = -n \log \sigma - \sum_{i=1}^{n} \left( \frac{x_i - \mu}{\sigma} \right) - \sum_{i=1}^{n} \exp \left\{ - \left( \frac{x_i - \mu}{\sigma} \right) \right\}.
\]

These log-likelihoods can be maximized numerically, there are no analytical solutions, [1, p. 55-56].

**GPD:** The cumulative distribution function of the GPD is given by

\[
G_{\xi, \tilde{\sigma}}(x) = \begin{cases} 
1 - \left(1 + \xi \frac{x}{\tilde{\sigma}} \right)^{-1/\xi}, & \xi \neq 0 \\
1 - \exp \left( - \frac{x}{\tilde{\sigma}} \right), & \xi = 0
\end{cases}
\]

(2.5)

where \(\tilde{\sigma}\) is given by (2.4), \(\tilde{\sigma} > 0\), and \(x \geq 0\) when \(\xi \geq 0\) and \(0 \leq x \leq -\tilde{\sigma}/\xi\), when \(\xi < 0\). \(\xi\) and \(\tilde{\sigma}\) are called the shape and scale parameter respectively.
The log-likelihood function of (2.5) this becomes

\[
\log L(\xi, \tilde{\sigma}) = -n \log \tilde{\sigma} - (1 + 1/\xi) \sum_{i=1}^{n} \log(1 + \xi x_i/\tilde{\sigma}), \quad \xi \neq 0.
\]

This also has to be optimized numerically, no analytical solution exists. In the exponential case \( \xi = 0 \), the log-likelihood becomes

\[
\log L(\tilde{\sigma}) = -n \log \tilde{\sigma} - \sum_{i=1}^{n} x_i/\tilde{\sigma},
\]
resulting in the ML estimate \( \tilde{\sigma}_{ML} = n^{-1} \sum_{i=1}^{n} x_i \), [1, p. 80].

### 2.2.4 Stationary sequence

So far the models we have derived have been under the assumption that the underlying process consists of a sequence of independent and identically distributed random variables. It is not uncommon that real life data exhibits dependence. The thesis will mainly focus on dealing with stationary sequences when applying threshold models but the main theorems for GEV will also be presented for consistency.

**GEV: Theorem 2.6** Let \( X_1, X_2, \ldots \) be a stationary process and \( X_1^*, X_2^*, \ldots \) be a sequence of independent variables with the same marginal distribution. Let \( M_n = \max(X_1, \ldots, X_n) \) and \( M_n^* = \max(X_1^*, \ldots, X_n^*) \). Under suitable regularity conditions,

\[
P \left( \frac{M_n^* - \mu_n}{\sigma_n} \leq x \right) \to G_1(x)
\]

as \( n \to \infty \) for normalizing sequences \( \{\sigma_n > 0\} \) and \( \{\mu_n\} \), where \( G_1 \) is a non-degenerate distribution function, if and only if

\[
P \left( \frac{M_n - \mu_n}{\sigma_n} \leq x \right) \to G_2(x)
\]

where

\[
G_2(x) = G_1^{\theta}(x)
\]

for a constant \( \theta \) such that \( 0 < \theta \leq 1 \).
Since $G_1$ is a GEV distribution so is $G^\theta_1$. If $G_1$ corresponds to the GEV distribution with parameters $(\mu, \sigma, \xi)$ then

\[
\begin{align*}
\mu^* &= \mu - \frac{\sigma}{\xi} \quad \text{and} \quad \sigma^* = \sigma \theta^\xi, \quad \xi \neq 0 \\
\mu^* &= \mu + \sigma \log \theta \quad \text{and} \quad \sigma^* = \sigma, \quad \xi = 0
\end{align*}
\]

corresponds to the parameters for $G^\theta_1(x)$. The quantity $\theta$ is called the extremal index. One usually interpreting the extremal index of a stationary series in terms of cluster at extreme levels [1, p. 92-97],

$$\theta = (\text{limiting mean cluster size})^{-1}.$$  

**GPD:** If one has a stationary sequence the extremes may have a tendency to cluster, hence applying GPD directly is not recommended. Many methods exist that deals with dependent exceedances in threshold exceedances models. The most common model is declustering, which corresponds to a filtering of the dependent observations and obtain exceedances over a threshold that are approximately independent. The declustering method works by:

1. Use an empirical rule to define clusters of exceedances.
2. Identifying the maximum excess within each cluster.
3. Assume cluster maxima to be independent. Assume that the excess is given by GPD.
4. Fit GPD to the cluster maxima.
### 2.2.5 Return period estimate

Once the estimated parameters of GEV or GPD have been obtained the return period can be estimated. The return period is calculated differently based in which approach one took, below both cases are explained.

**GEV:** The estimates of the extreme quantiles are obtained by inverting the GEV distribution (2.1)

\[
x_p = \begin{cases} 
\mu - \frac{\sigma}{\xi} \left[ 1 - \left\{ - \log(1 - p) \right\}^{-\xi} \right], & \text{for } \xi \neq 0, \\
\mu - \sigma \log \left( - \log(1 - p) \right), & \text{for } \xi = 0,
\end{cases}
\]

where \( G(x_p) = 1 - p \). \( x_p \) is referred to the *return level* associated with the return period \( 1/p \), [1, p. 49].

**GPD:** Similarly suppose GPD is a suitable model for exceedances of a threshold \( u \) by a variable \( X \). That is, for \( x > u \),

\[
P(X > x | X > x) = \left[ 1 + \xi \left( \frac{x - u}{\tilde{\sigma}} \right) \right]^{-1/\xi}.
\]

It follows that

\[
P(X > x) = \zeta_u \left[ 1 + \xi \left( \frac{x - u}{\tilde{\sigma}} \right) \right]^{-1/\xi},
\]

where \( \zeta_u = P(X > u) \). The level \( x_m \) that is exceeded on average once every \( m \) observations is the solution of

\[
\zeta_u \left[ 1 + \xi \left( \frac{x_m - u}{\tilde{\sigma}} \right) \right]^{-1/\xi} = \frac{1}{m}.
\]

Hence the return level can be obtained by rearranging the above expression, similarly for \( \xi = 0 \) one can follow the same procedure with (2.5), hence by [1, p. 81],

\[
x_m = \begin{cases} 
u + \tilde{\sigma} \left[ (m\zeta_u)^{\xi} - 1 \right], & \xi \neq 0 \\
u + \tilde{\sigma} \log(m\zeta_u), & \xi = 0
\end{cases}
\]

When dealing with a stationary sequence the return level is given by

\[
x_m = \begin{cases} 
u + \tilde{\sigma} \left[ (m\zeta_u\theta)^{\xi} - 1 \right], & \xi \neq 0 \\
u + \tilde{\sigma} \log(m\zeta_u\theta), & \xi = 0
\end{cases}
\]

where \( \theta \) is given by the extremal index. [1, p. 103].
2.3 Multivariate extremes

Let \( X_n, n \geq 1 \) be identically independent distributed random vector in \( \mathbb{R}^d \), \( X_k = (X_k^{(1)}, \ldots, X_k^{(d)}) \), \( k = 1, \ldots, n \). Define maxima component-wise,

\[
M_n = (M_n^{(1)}, \ldots, M_n^{(d)}) = (\max_{k=1,\ldots,n} X_k^{(1)}, \ldots, \max_{k=1,\ldots,n} X_k^{(d)}). \tag{2.6}
\]

Where one is interested in the asymptotic distribution of (2.6).

Suppose \( X_1, \ldots, X_n \) are independent identical distributed with distribution function \( F(X^{(1)}, \ldots, X^{(d)}) \),

\[
P(M_n \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = F^n(x), \ x \in \mathbb{R}^d.
\]

Thus the distribution of (2.6) even in this case \( M_n \) has a degenerate distribution, \( M_n \to x^F \) where \( x^F = \sup \{ x | F(x) < 1 \} \).

Assume there exist normalizing constant \( a_n^{(i)} > 0, \ b_n^{(i)} > 0 \in \mathbb{R} \) for \( i = 1, \ldots, d, n \geq 1 \) such that

\[
P \left( \frac{M_n^{(i)} - b_n^{(i)}}{a_n^{(i)}} \leq x^i, 1 \leq i \leq d \right) = F^n(a_n^{(1)}x^{(1)} + b_n^{(1)}, \ldots, a_n^{(d)}x^{(d)} + b_n^{(d)}) \to G(x_1, \ldots, x_d).
\]

Where the limiting distribution \( G \) has each margin distribution \( G_i, i = 1, \ldots, d \) is non-degenerate.

The \( i \)th margin distribution is

\[
F_i^n(a_n^{(i)}x^{(i)} + b_n^{(i)}) \to G_i(x).
\]

From the results for the univariate extreme value theory each \( G_i \) is a member of the GEV family (2.1),

\[
G(x; \gamma, \mu, \sigma) = \exp \left\{ -\left( 1 + \gamma \frac{x - \mu}{\sigma} \right)^{-\frac{1}{\gamma}} \right\}.
\]
2.3.1 Characterization of bivariate extreme value distribution

**Definition:** $G(x)$ is max-stable if for $i = 1, \ldots, d$ and every $t > 0$ there exist functions $\alpha(i)(t) > 0, \beta(i)(t) > 0$ such that

$$G^t(x) = G(\alpha(1)(t)x^{(1)} + \beta(1)(t), \ldots, \alpha(d)(t)x^{(d)} + \beta(d)(t)).$$

It can been shown that $G(x)$ is max-stable if and only if it is a multivariate extreme value distribution. Thus one needs to find all possible multivariate max-stable distribution assuming one of the three possible univariate marginal extreme value distribution.

It can be shown that any bivariate extreme value distribution with unit Fréchet margins can be written as

$$G^*(x, y) = \exp\left(-\frac{1}{x} + \frac{1}{y}A\left(\frac{x}{x+y}\right)\right) \quad (2.7)$$

where $A(w)$ is called the dependence function. Since (2.7) has unit Fréchet margins we have

$$\lim_{x \to \infty} G^*(x, y) = e^{-1/y}, \quad \lim_{y \to \infty} G^*(x, y) = e^{-1/x}.$$ 

Since

$$G^*(x, y) = G^*(x/n, y/n),$$ 

$G^*(x, y)$ is max-stable. It can be shown that for all $t > 0$

$$G^t(xt, yt) = G^*(x, y).$$

The dependence function $A(w)$ has the following properties

- $A(0) = A(1) = 1$.
- $\max(w, 1 - w) \leq A(w) \leq 1, \quad 0 \leq w \leq 1$.
- $A(w)$ is convex for $w \in [0, 1]$.

The lower bound for $A(w)$ is:

$$A(w) \begin{cases} 
1 - w, & w < 1/2 \\
1, & w > 1/2
\end{cases}$$

with upper bound $A(w) = 1$. There is no parametric family which gives all possible bivariate extreme value distributions.
2.3.2 Bivariate extreme value distributions

In the package *evd* [19] developed the software program R [13] nine parametric bivariate extreme value models are included. Each one of them will be presented but first some technicalities will be discussed.

Define

\[ y_i = y_i(x_i) = \left\{ 1 + \xi_i \left( \frac{x_i - \mu_i}{\sigma_i} \right) \right\}^{-\frac{1}{\xi_i}} \]

for \( i = 1, 2 \), where the marginal parameters are given by \((\mu_i, \sigma_i, \xi_i)\), \(\sigma_i > 0\).

If \( \xi_i = 0 \) then \( y_i \) is defined by continuity.

In each of the bivariate distributions functions \( G(x_1, x_2) \) given below, the univariate margins are GEV. That means that \( G(x_i) = \exp(-y_i) \) for \( i = 1, 2 \).

If \( 1 + \xi_i \left( \frac{x_i - \mu_i}{\sigma_i} \right) \leq 0 \) for some \( i = 1, 2 \), the value \( \xi_i \) is either greater than the upper end point (if \( \xi_i < 0 \)), or less than the lower end point (if \( \xi_i > 0 \)) of the \( i \)-th univariate marginal distribution.

The bivariate distribution function \( G(x_1, x_2) \) can be constructed by choosing an appropriate threshold for both margins or block maxima for both margins, in this paper the Bivariate Generalized Pareto Distribution (BGPD) will be used.

**Type 2 BGPD** The Type 2 BGPD aims to fit

\( \{(x, y) \mid (x, y) \not< (u_x, u_y)\} \)

where \( u_x \) and \( u_y \) are defined as above.

The models used for fitting these BGPD will be included in the package *mgpd* [14] in R. For further understanding about models based on Type 2 BGPD the reader is refereed to [15] and [16] for further reading.
**Type 1 BGPD**  In type 1 peaks over threshold a distribution is fitted to the observations

\[ \{(x, y) | (x, y) > (u_x, u_y)\} \]

where \(u_x\) and \(u_y\) are suitable threshold for each margin. The following models can be used for fitting BGPD of type 1, included in the package *evd* [19] in R.

**Model "log":**  The bivariate logistic distribution function is defined as

\[
G(x_1, x_2) = \exp \left\{ -\left[ \frac{1}{y_1} + \frac{1}{y_2} \right]^r \right\},
\]

where \(0 < r \leq 1\). Independence is obtained when \(r = 1\) and complete dependence when \(r\) approaches zero. The bivariate logistic distribution function is a special case of the bivariate asymmetric logistic model, [7].

**Model "alog":**  The bivariate asymmetric logistic distribution function is defined as

\[
G(x_1, x_2) = \exp \left\{ -(1 - t_1)y_1 - (1 - t_2)y_2 - \left[ (t_1y_1)^{\frac{1}{r}} + (t_2y_2)^{\frac{1}{r}} \right]^r \right\},
\]

where \(0 < r \leq 1\) and \(0 \leq t_1, t_2 \leq 1\). Independence is obtained when either \(r = 1, t_1 = 0\) or \(t_2 = 0\). Complete dependence is obtained in the limit when \(t_1 = t_2 = 1\) and \(r\) approaches zero. Note that if \(t_1\) and \(t_2\) are fixed and \(r\) approaches zero different limits occur. The asymmetric logistic model is equivalent to the logistic model when \(t_1 = t_2 = 1\), [20].

**Model "hr":**  The Hustler-Reiss distribution function is defined as

\[
G(x_1, x_2) = \exp \left\{ -y_1\phi \left\{ \frac{1}{r} + r \left[ \log \left( \frac{y_1}{y_2} \right) \right]/2 \right\} - y_2\phi \left\{ \frac{1}{r} + r \left[ \log \left( \frac{y_2}{y_1} \right) \right]/2 \right\} \right\},
\]

where \(\phi\) is the standard normal distribution function and \(r > 0\). Independence is obtained in the limit as \(r\) approaches zero. Complete dependence is obtained as \(r\) tends to infinity, [8].
Model "neglog": The bivariate asymmetric logistic distribution function is defined as

\[ G(x_1, x_2) = \exp \left\{ -y_1 - y_2 - \left[ y_1^{-r} + y_2^{-r} \right] \right\}, \]

where \( r > 0 \). Independence is obtained in the limit as \( r \) approaches zero. Complete dependence is obtained as \( r \) tends to infinity. The bivariate asymmetric logistic distribution function is a special case of the bivariate asymmetric negative logistic model, [6].

Model "aneglog": The bivariate asymmetric negative logistic distribution function is defined as

\[ G(x_1, x_2) = \exp \left\{ -y_1 - y_2 - \left[ \left( t_1 y_1 \right)^{-r} + \left( t_2 y_2 \right)^{-r} \right] \right\}, \]

where \( r > 0 \) and \( 0 < t_1, t_2 \leq 1 \). Independence is obtained in the limit as either \( r, t_1 \) or \( t_2 \) approaches zero. Complete dependence is obtained in the limit when \( t_1 = t_2 = 1 \) and \( r \) tends to infinity. Note that if \( t_1 \) and \( t_2 \) are fixed and \( r \) tends to infinity different limits occur. When \( t_1 = t_2 = 1 \) the bivariate asymmetric negative logistic distribution function is equivalent to the negative logistic model, [10].

Model "bilog": The bilogistic distribution function is defined as

\[ G(x_1, x_2) = \exp \left\{ -y_1 q^{1-\alpha} - y_2 (1 - q)^{1-\beta} \right\}, \]

where \( q \) is the root of the equation

\[ (1 - \alpha) y_1 (1 - q)^{\beta} - (1 - \beta) y_2 q^\alpha = 0, \quad 0 < \alpha, \beta < 1. \]

Independence is obtained as \( \alpha = \beta \) approaches one, and when one of \( \alpha, \beta \) is fixed and the other approaches one. Complete dependence is obtained in the limit as \( \alpha = \beta \) approaches zero. Different limits occur when one of \( \alpha, \beta \) is fixed and the other approaches zero. The bilogistic distribution is equivalent to the logistic distribution when \( \alpha = \beta \), [18].
Model "negbilog": The negative bilogistic distribution function is defined as
\[ G(x_1, x_2) = \exp \left\{ -y_1 - y_2 + y_1q^{1+\alpha} + y_2(1-q)^{1+\beta} \right\}, \]
where \( q \) is the root of the equation
\[ (1 + \alpha)y_1q^\alpha - (1 + \beta)y_2(1-q)^\beta = 0, \quad \alpha, \beta > 0. \]
Independence is obtained as \( \alpha = \beta \) tends to infinity, and when one of \( \alpha, \beta \) is fixed and the other tends to infinity. Complete dependence is obtained in the limit as \( \alpha = \beta \) approaches zero. Different limits occur when one of \( \alpha, \beta \) is fixed and the other approaches zero. The negative bilogistic distribution is equivalent to the negative logistic distribution when \( \alpha = \beta \) with reformulation \( \frac{1}{\alpha}, \frac{1}{\beta} \), [3].

Model "ct": The Coles-Tawn distribution function is defined as
\[ G(x_1, x_2) = \exp \left\{ -y_1 \left[ 1 - Be(q; \alpha + 1, \beta) \right] - y_2 Be(q; \alpha, \beta + 1) \right\}, \quad \alpha, \beta > 0 \]
where
\[ q = \frac{\alpha y_2}{\alpha y_2 + \beta y_1} \]
and \( Be(q; \alpha, \beta) \) is the beta distribution function evaluated at \( q \). Independence is obtained as \( \alpha = \beta \) approaches zero, and when one of \( \alpha, \beta \) is fixed and the other approaches zero. Complete dependence is obtained in the limit as \( \alpha = \beta \) tends to infinity. Different limits occur when one of \( \alpha, \beta \) is fixed and the other tends to infinity, [2].

Model "amix": The asymmetric mixed distribution function has a dependence function with the following cubic polynomial form
\[ A(t) = 1 - (\alpha + \beta)t + \alpha t^2 + \beta t^3. \]
Where \( \alpha \geq 0 \) and \( \alpha + 3\beta \geq 0 \), and where \( \alpha + \beta \leq 1 \) and \( \alpha + 2\beta \leq 1 \). Thus \( \alpha \in [0, 1.5] \) and \( \beta \in [-0.5, 0.5] \), though if \( \alpha > 1 \to \beta < 0 \). Independence is obtained when \( \alpha = \beta = 0 \). Complete dependence can not be obtained but the dependence increases for increasing \( \alpha \) when \( \beta \) is fixed [20].
2.4 Dependence measure

There are several different dependence measure, mostly common is Pearson’s $\rho$. Since Pearson’s correlation only measure linear dependence other measurements, namely Kendall’s $\tau$ and Spearman’s $\rho_s$, which are based on the concept of discordance and concordance are discussed to.

**Linear dependence:** Suppose $(X, Y) \sim F(X, Y)$, then the Pearson’s correlation $\rho_{X,Y}$ is defined as

$$
\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}, \quad -1 \leq \rho \leq 1.
$$

(2.8)

From (2.8) it is clear that $\rho$ is not defined for distributions which have non finite variance. Another drawback of this dependence measure is that it only measures linear dependence. If $T_1$ and $T_2$ are increasing transforms than $\rho_{T_1(X),T_2(Y)} \neq \rho_{X,Y}$.

**Concordance and discordance:** A pair of observation $(x_i, y_i)$ and $(x_j, y_j)$ is said to be concordant if $(x_i, y_i)(x_j, y_j) > 0$ and discordant if $(x_i, y_i)(x_j, y_j) < 0$. Drawing a line through the two points will thus have a positive slope if concordant, and negative if discordant.

Two distribution free dependence measures that is based on the concept of concordance and discordance is Kendall’s $\tau$ and Spearman’s $\rho_s$. These two measures does not have the issues that $\rho_{X,Y}$ has. Kendall’s $\tau$ and Spearman’s $\rho_s$ can also be used to find the dependence function of a bivariate max-stable distribution function.

**Kendall’s $\tau$:** Let $(X_1, Y_1) \sim F(X, Y)$ and $(X_2, Y_2) \sim F(X, Y)$ be two independent random vectors with the same bivariate distribution function $F$, [11]. Kendall’s $\tau$ is defined by

$$
\tau = P((X_1 - X_2)(Y_1 - Y_2) > 0) - P((X_1 - X_2)(Y_1 - Y_2) < 0), \quad -1 \leq \tau \leq 1.
$$

(2.9)

Equation (2.9) is the difference between the probability of concordance and discordance.

The sample version of Kendall’s $\tau$ can be estimated from observations $(x_1, y_1), ..., (x_n, y_n)$ of the variables. Let $c$ and $d$ denote the number of concordant and discordant pairs respectively, then the sample estimate is given by

$$
\hat{\tau} = \frac{c - d}{c + d}
$$
Spearman’s $\rho_s$: Let $(X_1, Y_1)$, $(X_2, Y_2)$ and $(X_3, Y_3)$ be three independent random vectors with the same bivariate distribution function $F(X,Y)$, Spearman’s $\rho_s$: is defined by

$$\tau = 3 \left[ P((X_1 - X_2)(Y_1 - Y_3) > 0) - P((X_1 - X_2)(Y_1 - Y_3) < 0) \right], \quad -1 \leq \tau \leq 1,$$

The sample version of Spearman’s rho is defined as the Pearson’s correlation of the ranked variables,

$$\rho_s = \frac{\text{cor}(rg_x, rg_y)}{\sqrt{\text{var}(rg_x)\text{var}(rg_y)}}, \quad -1 \leq \rho_s \leq 1.$$

Where $rg_x \sim \text{rank of } x$ and $rg_y \sim \text{rank of } y$.

2.5 The auto-correlation function (ACF)

The auto-correlation function is valuable tool of checking if the observations in a time series are uncorrelated. Usually ACF is used for model validation on the residuals on a fitted time series model, instead we will use it in this context for checking if our observations them self are independent.

To better understand the ACF the auto covariance will be introduced, being a measure of correlation within a time series. The auto-covariance at lag $k$ is defined as $r_y(k) \equiv C[y_t, y_{t-k}]$ and is estimated using either an unbiased estimation method

$$\hat{r}_y^u(k) = \frac{1}{N-k} \sum_{t=k+1}^{N} (y_t - \hat{m}_y)(y_{t-k} - \hat{m}_y),$$

or an asymptotically unbiased estimation method

$$\hat{r}_y^b(k) = \frac{1}{N} \sum_{t=k+1}^{N} (y_t - \hat{m}_y)(y_{t-k} - \hat{m}_y),$$

where $N$ is the sample size, $k$ is the lag and $\hat{m}_y$ the estimated mean of the sample.
It should be noted that $\hat{r}_g^u(k)$ is only unbiased, $E\{\hat{r}_g^u(k)\} = r_g(k)$ for $N \rightarrow \infty$. From these observations one may conclude that $\hat{r}_g^u(k)$ is a more appropriate estimate than $\hat{r}_g^b(k)$, however this is not the case. This becomes more clear if one looks at the variance of the resulting estimated auto-covariance. For high lags, the unbiased estimate will be formed by only averaging a few terms, thus giving estimate that has an increasing variance with the lag number.

The auto-correlation function of $y_t$ is defined as

$$\rho_y(k) = \frac{r_y(k)}{r_y(0)}, \quad k = 0, 1, 2, ...$$

Since $r_y(k)$ is largest for $k = 0$, the auto-correlation is bounded such that $|\rho_y(k)| \leq 1$.

When having observation that one would like to conduct extreme value theory on, the observation should be independent and identically distributed. Let $e_t$ be a sequence of zero-mean white Gaussian process with variance $\hat{\sigma}_e$, the expected value and variance of the auto-correlation estimates are given by:

$$E\{\hat{\rho}_e(k)\} = 0,$$

$$V\{\hat{\rho}_e(k)\} = \frac{1}{N},$$

for $k \neq 0$. It can be shown that $\hat{\rho}_e(k)$ is asymptotically normal distributed and the confidence interval can thus be created as

$$\hat{\rho}_e(k) \approx \pm \frac{\lambda_{\alpha}}{\sqrt{N}}, \quad k \neq 0.$$

Visually, observations might be independent if a proportion less than $\alpha$ exceeds the confidence interval [9].
2.6 Model validation

2.6.1 Fit diagnostic

When fitting a distribution to a data set inference should be made regarding about the fit. Thus using visual tools can be used.

Let $x_1, \ldots, x_n$ denote an sample of independent observations from $F$. An estimate of $F$, say $\hat{F}$, has been obtained. The empirical distribution function is defined by

$$\tilde{F}(x) = \frac{i}{n+1}, \quad \text{for } x_{(i)} \leq x \leq x_{(i+1)}.$$

Since $\tilde{F}$ is an estimate of the true probability distribution function $F$, it should be fairly similar to the estimated model $\hat{F}$. Hence comparing $\hat{F}$ with $\tilde{F}$ leads to various goodness of fit procedures, probability plot and quantile plot are two graphically techniques commonly used.

A probability plot consists of the points

$$\left\{ \left( \hat{F}(x_{(i)}), \frac{i}{n+1} \right) : i = 1, \ldots, n \right\}.$$

If $\hat{F}$ is a reasonable model the points of the probability plot should lie close to the unit diagonal.

A quantile plot consist of the points

$$\left\{ \left( \hat{F}^{-1} \left( \frac{i}{n+1} \right), x_{(i)} \right) : i = 1, \ldots, n \right\}.$$

Here the quantiles $x_{(i)}$ and $\hat{F}^{-1}(i/(n+1))$ each gives estimate of the $i/(n+1)$ quantile of the distribution $F$. If $\hat{F}$ is a good estimate of $F$, the points in the quantile plot should lie close to the unit diagonal [1].

2.6.2 Akaike information criteria

The Akaike information criteria (AIC) is commonly used for model order selection. The value of AIC is based on information theory, rewarding a high likelihood of the model but penalized an high model order. One prefers a model with a low AIC.
AIC comes with the package stats in R which is presented below. Here, $LL$ is the maximum of the log likelihood function and $m$ the number of parameter in the model [17].

$$AIC = -2LL + 2m$$

2.6.3 Goodness-of-fit test for BGPD

The goodness-of-fit test used test the null hypothesis $H_0 : G \in G_{\hat{\theta}}$, where $G$ is a bivariate extreme value distribution and $G_{\hat{\theta}}$ is a specific bivariate extreme value distribution. The test is based on comparing with the empirical bivariate cumulative distribution function, defined as

$$G_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x, Y_i \leq y),$$

where $X$ and $Y$ are pseudo-observations. This is used to compare the "distance" between the empirical distribution $G_n$ and the parametric estimation $G_{\hat{\theta}}$. For numerical purpose the Riemann sum can bu used as test statics

$$\tilde{S}_n = \sum_{i=1}^{n} (G_n(x_i, y_i) - G_{\hat{\theta}}(x_i, y_i))^2,$$

where $((x_1, y_1), ..., (x_n, y_n))$ is the obtained sample.

Following the procedure below lead to an approximation of the p-value

1. Calculate the empirical distribution $G_n$ from the sample and estimate $\theta$ with one of the available bivariate extreme value distribution.

2. Compute the test-statistic $s_0 := \tilde{S}_n$.

3. Simulate a sample $\tilde{Z}$ from the chosen bivariate extreme value distribution $G_{\theta}$ of the size as the original sample.

4. Estimate $\hat{\theta}$ from $\tilde{Z}$ and calculate its empirical distribution $\hat{G}_n$.

5. Repeat steps 3 to step 4 for a large $N$ and compute its test-statistic $s_i := \tilde{S}_n$ for any $1 \leq i \leq N$.

6. The approximate p-value is $\#\{1 \leq i \leq N | s_i > s_0\} / N$. 22
Chapter 3

Analysis

3.1 The data

The data comes from Sweden’s Meteorological and Hydrological Institute (SMHI) open access data with hourly observations. The stations that will be used for estimating the return period for the storm surge that occurred 1872 are Skanör (operated since 1992), Klagshamn (operated since 1929) and Ystad (operated since 1886-1987). Other measure stations that will be used for multivariate extreme value analyses on wave height, wind direction and wind speed and their impact on the sea level are Angelholm (observations from 1976) and Viken (observations from 1976). This data set originates from Water Resources Engineering, Lund university. A map overview of all stations is presented in Figure 3.1.

Figure 3.1: Map overview; the Falsterbo Peninsula is marked with a red rectangel.
The sea level from SMHI open access data is given by the elevation system RH 2000, while the data from Viken and Angelholm has been highly positive corrected since onshore winds will effect the water level positive. Thus the water level at these study areas are higher than the SMHI station water levels. Water level data from Viken and Angelholm was adjusted for wind setup accordingly by:

$$\delta h = \frac{\rho_{\text{air}}}{\rho_{\text{water}}} \frac{C_D W^2 L_B}{gd}.$$  

The wave heights used in the multivariate model of Angelholm is also adjusted for oblique wave angles. Since these calculations are not the main focus in this thesis the reader is refereed to [12] for more information on this topic.

SMHI asses the quality of the data points as green or yellow, where green stands for controlled and approved values and yellow for suspicious or aggregated values. For the measure station in Ystad, there is no information about the data quality. For the measure stations Skanör and Klagshamn there is no data quality until 2010 and thereafter all measurements are green.

### 3.2 Correlation analysis

To compare the different time series of Skanör, Klagshamn and Ystad without influence from sea level rise and difference in mean still water level between the stations, the water level will be presented in the local mean sea level (MSL). In this thesis we assume the sea level rise to be a constant linear process which is fitted to a liner equation and is subtracted from each data set. Even though sea level rise is not linear the associated error is negligible is these short time scales. The fitted linear equation to each data set is presented in Table 3.1.

<table>
<thead>
<tr>
<th>Stations</th>
<th>Intercept</th>
<th>time</th>
<th>100-year increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skanör, year 1992-2017</td>
<td>9.854 (0.09)</td>
<td>4.544e-05 (7.144e-07)</td>
<td>39.83 cm</td>
</tr>
<tr>
<td>Klagshamn, year 1929-2017</td>
<td>6.858 (0.04)</td>
<td>9.293e-06 (9.613e-08)</td>
<td>8.1 cm</td>
</tr>
<tr>
<td>Ystad, year 1886-1987</td>
<td>9.796 (0.06)</td>
<td>4.468e-06 (1.991e-07)</td>
<td>3.92 cm</td>
</tr>
<tr>
<td>Viken, year 1976-2017</td>
<td>4.172 (0.07)</td>
<td>1.383e-06 (3.461e-07)</td>
<td>12.12 cm</td>
</tr>
</tbody>
</table>

Table 3.1: Result of the fitted linear equation to the hourly sea level observations for each of the measure station. The brackets represent the standard error of the estimates.
Since the observation stations Skanör, Klagshamn and Ystad are relatively close to each other we hope that there are positive correlations between the simultaneously hourly observations. The correlation between different data sets are estimated by comparing the hourly values with Pearson’s correlation and Spearman’s correlation shown in Table 3.2.

Table 3.2: Result of correlation analysis between the stations. The values within the brackets is the simultaneously years used.

<table>
<thead>
<tr>
<th>Stations</th>
<th>$\rho$</th>
<th>$\rho_s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skanör-Klagshamn (1992-2017)</td>
<td>0.951</td>
<td>0.953</td>
</tr>
<tr>
<td>Ystad-Klagshamn (1964-1987)</td>
<td>0.951</td>
<td>0.946</td>
</tr>
<tr>
<td>Skanör-Viken (1992-2017)</td>
<td>0.036</td>
<td>0.118</td>
</tr>
</tbody>
</table>

The five first years of the sea level of Skanör are shown in Figure 3.2.

Figure 3.2: The first 5 years of hourly data in Skanör, one can clearly see a seasonal cycle.
Since the observations are hourly the observations themselves are correlated with each other, illustrated by Figure 3.3.

Figure 3.3: The ACF for the first 3 weeks of the sea level in Skanör. The dotted line represent the confidence bounds indicating that there is dependence between the observations.

To avoid dependence in the observation one can either use block-maxima or threshold models. Choosing a large enough block length will negate the dependence as when it comes to threshold models one has to decluster the observations to avoid the dependence. It is said that storm surges that occurs 48-hours apart can be considered independent of each other, this will be considered when declustering is performed on the data.

Investigating simultaneous extreme values of Skanör and Klagshamn there seems to be an one hour time shift between the extreme sea level occurrence, hence the extreme sea level in Skanör takes one hour before the same magnitude occurs in Klagshamn. To avoid this time difference the extreme sea values will be shifted one hourly accordingly.
3.3 Missing values

The data was analyzed for missing values. The Ystad, Viken and Ängelholm series is complete while a visual interpretation of the missing values in Skanör and Klagshamn is shown in Figure 3.4. Most extreme water levels occurs from October to March, hence if there is missing values during this period, there is a risk that the most extreme values are absent in the data. Thus for one-dimensional analyze values before 1960 in Klagshamn will be excluded. For the multivariate case all simultaneously values of Skanör and Klagshamn where no missing values occurred will be used.

(a) Missing values in the time series from Klagshamn, the black circles represent a missing values.

(b) Missing values in the time series from Skanör, the black circles represent a missing values.

As one can see in Table 3.2 the hourly sea levels of Skanör and Klagshamn is highly correlated with each other. Looking at the corresponding values for the missing data in Klagshamn one will obtain that 50 MSL is the highest value. Thus directly replacing the missing values in Skanör with Klagshamn will not effect our extreme value analysis. Note that if the analysis would have shown a higher MSL one is not recommended to simply replace the missing values since one does not even know if the data belongs to the same distribution. The only reason in this case we simply replaced the missing values with Klagshamn observations is that there was no extremes and we wanted to make the data set complete for simplicity.

Figure 3.4: Missing values of Klagshamn and Skanör.
Chapter 4

One dimensional analysis of the data

4.1 Block-maxima

Figure 3.2 shows a seasonal trend for the Skanör data set. This seasonal cycle appears for all sea level measure stations. Thus one has to choose a block maxima which negates this seasonal cycle. Since the cycle peaks around the winter month a block length of one year from summer to summer seems to be the best choice of the block length.

Now that the block length is chosen GEV and Gumbel models can be fitted to the Block maxima for Skanör, Klagshamn and Ystad. The models were tested with a likelihood ratio test with a significance level of 5 % to see if Gumbel or GEV fitted the data best. If GEV was significant better than Gumbel, Gumbel was rejected. If Gumbel proved to be significant better than GEV, Gumbel was kept for its advantage for being the simpler model, the parameters were estimated using maximum likelihood method.
Table 4.1: Maximum likelihood estimation of parameters for the GEV and Gumbel ($\xi = 0$) model, 95% confidence intervals estimated with normal approximation are presented within brackets.

<table>
<thead>
<tr>
<th>Station</th>
<th>location $\mu$</th>
<th>scale $\sigma$</th>
<th>shape $\xi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skanör 1992-2017</td>
<td>96.3 (88.2-104.3)</td>
<td>19.9 (13.8-25.9)</td>
<td>0</td>
</tr>
<tr>
<td>Klagshamn 1960-2017</td>
<td>89.7 (83.2-96.2)</td>
<td>22.9 (18.2-27.6)</td>
<td>-0.3 (-0.46-0.14)</td>
</tr>
<tr>
<td>Ystad 1886-1987</td>
<td>84.9 (81.3-88.5)</td>
<td>17.5 (14.9-19.9)</td>
<td>0</td>
</tr>
</tbody>
</table>

The estimated parameters are presented in Table 4.1.

Table 4.2: The estimated 100,250-year return level relative mean sea level with 95% confidence interval within the brackets, and the estimated return period for the 1872 storm

<table>
<thead>
<tr>
<th>Station</th>
<th>100-year return level</th>
<th>250-year return level</th>
<th>1872-storm return period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Klagshamn 1960-2017</td>
<td>146 (135-158)</td>
<td>150 (137-165)</td>
<td>Exceeds the upper limit of the distribution (165 cm)</td>
</tr>
<tr>
<td>Ystad 1886-1987</td>
<td>165 (152-178)</td>
<td>181 (166-196)</td>
<td>9664 (2552-57164) years</td>
</tr>
</tbody>
</table>

The point estimate for the 100, 250 -year return level and the return period for the 1872 storm is presented in Table 4.2.

Since the Skanör series is rather small the estimate of the 100, 250-year return level is rather difficult and thus not very precious. Looking at Table 4.2 one can see some differences in the estimates of the 100, 250-year return levels between Skanör and Klagshamn even though these measurement stations are close to each other. This is probably due to that the resent years have had relatively high sea levels which will capture more in the Skanör series than Klagshamn, due to that Klagshamn has more data than Skanör. Hence the Klagshamn model is to be trusted more over Skanör.

The Klagshamn model follows a Weibull-type distribution with upper limit of 165 cm MSL. This upper limit is not the nail in the coffin since the parameters can vary between the confidence bounds, hence this estimate can be bigger, the most extreme case is a upper limit of 293 cm, thus the 1872 storm and the storm occurring year 1904 which reached 185 cm above normal in Klagshamn is realistic. Comparing these result with [5, p. 139] these result seems more accurate. This is probably due to more extreme sea levels have occurred in the past two years, the 4 January 2017 is the most recent extreme event. Still one has to consider that mean of the upper limit is 165 cm.
4.2 Generalized Pareto Distribution

Since the measurements are hourly the observations are still correlated between each other, an example of this was shown in Figure 3.3. The method of declustering was discussed and will be used for fitting our GPD. The threshold was chosen based on fit diagnostic and then declustering was used on all the observation above that threshold, resulting in independent observations. The result is shown in Table 4.3.

Table 4.3: Maximum likelihood estimation of parameters for the GPD, 95 % confidence intervals estimated with normal approximation are presented within brackets.

| Station          | threshold | exceed | scale σ     | shape ξ 
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Skanör 1992-2017</td>
<td>80</td>
<td>74</td>
<td>17.7 (13.7-21.7)</td>
<td>0</td>
</tr>
<tr>
<td>Klagshamn 1960-2017</td>
<td>90</td>
<td>70</td>
<td>22.4 (16.1-28.7)</td>
<td>-0.31 (-0.48-0.14)</td>
</tr>
<tr>
<td>Ystad 1886-1987</td>
<td>80</td>
<td>183</td>
<td>15.7 (13.4-17.9)</td>
<td>0</td>
</tr>
</tbody>
</table>

The point estimate for the 100, 250-year return level and the return period for the 1872 storm is presented in Table 4.4.

Table 4.4: The estimated 100,250-year return level relative mean sea level with 95% confidence interval within the brackets, and the estimated return period for the 1872 storm

<table>
<thead>
<tr>
<th>Station</th>
<th>100-year return level</th>
<th>250-year return level</th>
<th>1872-storm return period</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skanör 1992-2017</td>
<td>181 (158-204)</td>
<td>197 (170-223)</td>
<td>3750 (665-58628) years</td>
</tr>
<tr>
<td>Klagshamn 1960-2017</td>
<td>146 (121-171)</td>
<td>149 (116-182)</td>
<td>Exceeds the upper limit of the distribution (162 cm)</td>
</tr>
<tr>
<td>Ystad 1886-1987</td>
<td>162 (149-173)</td>
<td>175 (162-190)</td>
<td>20491 (5405-122023) years</td>
</tr>
</tbody>
</table>

The estimated 100, 250-years return level for Skanör, Klagshamn and Ystad are similar to estimation based on the GEV-distributions, which indicates stability in the models. The main difference between the models is that GPD for Klagshamn has bigger confidence bounds when estimating the return level and the upper limit is lower with now 162 cm, all return periods for the 1872 storm is higher comparing to Table 4.2.
4.3 Recommendation

Comparing all density and QQ plots for the all models, both GEV and GPD will result in that the models fitted for Klagshamn is the best fit for the data. In Figure 4.1 one sees that the GEV model for Klagshamn is almost 1:1 in the QQ plot, indicating a perfect fit. Since the GEV model is fitted better in the extreme than GPD the GEV model for Klagshamn seems to be the best model of them all. Even thought the upper bound for this model is bounded it has been shown that this bound has some variety.

(a) Density plot for GEV-model based on Klagshamn data.
(b) QQ-plot for the GEV-model based on the Klagshamn data.

(c) Density plot for GPD-model based on Klagshamn data with a threshold of 90 cm.
(d) QQ-plot for the GPD-model based on the Klagshamn data with a threshold of 90 cm.

Figure 4.1: QQ-plot and density plot for the Klagshamn data based on GEV and GPD with a threshold of 90 cm.
In Table 4.5 return level for 100, 200, 300 and 500 years return period are calculated for the fitted GEV model. The results are presented with relative to the normal sea level (MSL) and the reference RH 2000. As these values are based on Klagshamn data, 6.6 cm have been added to the result to adjust the levels to conditions at Skanör, according to the GEV fit.

Table 4.5: Estimated return level for 100, 200, 300 and 500 years return level adjusted to Skanör. 95 % confidence interval are given within brackets.

<table>
<thead>
<tr>
<th>return period</th>
<th>cm rel MSL</th>
<th>cm rel RH 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>153 (142-165)</td>
<td>169 (158-181)</td>
</tr>
<tr>
<td>200</td>
<td>157 (143-170)</td>
<td>173 (159-186)</td>
</tr>
<tr>
<td>300</td>
<td>158 (144-173)</td>
<td>174 (160-189)</td>
</tr>
<tr>
<td>500</td>
<td>160 (144-176)</td>
<td>176 (160-192)</td>
</tr>
</tbody>
</table>

Since the GEV model of Klagshamn is very consistence with the result from GPD model of same time period this is most likely the best fit for estimating return levels for sea levels in Falsterbo Peninsula.
Chapter 5

Multivariate analysis, wave and wind

It is important to study the relationship of multiple measure stations at once, thus we now direct our attention to the multivariate extreme value theory. Here we will quickly go through multivariate extreme value of the Klagshamn and Skanör data. We will also discuss a models which takes into account wind speed and wind direction for the Viken measure station and wave height for a dataset in Ängelholm. The models which are presented for the Viken data and Ängelholm are not meant for Falsterbo but for stations which wind speed and wave height effects the sea level dramatically, such as cities close to large-size ocean since wind conditions have an increasing effect if the ocean is larger.

5.1 Skanör and Klagshamn

In the future analysis we want to find a suitable multivariate model for the Skanör and Klagshamn. The reason for this is that it can be used to calculate that given that Skanör has a sea level at 145 cm or higher what is the probability that Klagshamn has a sea level at 100 cm? These kind of questions can be answered with multivariate extreme value theory and hence it can be of great importance to us.

Models will be fitted to the Type 1 BGPD and Type 2 BGPD. Model selection will be based on goodness-of-fit test and AIC.
5.1.1 Type 1 BGPD

Since we have hourly observations once again we will have dependence in the data. To overcome this we will have to declustered the data, as we did before in the one dimensional case. A suitable threshold seemed to be around 80 for both margins thus

\[ \{(x, y) | (x, y) > (80, 80)\} \]

Having decided a threshold all values above 80 for both margins will be declustered, this will result in Figure 5.1.

![Declustered values of Klagshamn and Skanör above 80 cm](image)

**Figure 5.1:** Simultaneous values of Klagshamn and Skanör, missing values of Skanör have been replaced with Klagshamn values. Values above 80 have been declustered, the dotted line represent the threshold for each margin.
All nine bivariate models where fitted to the values above. If the shape parameter was not significant different from 0 for a model it was set to 0. The estimated parameters for the models where

Table 5.1: The parameter estimation for all nine parametric bivariate extreme value models in the package evd [19] in R. The AIC ahas been calculated for each model, the goodness-of-fit test based on 1000 simulations has been preformed on all models. The model "alog" has the lowest AIC.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_1$</th>
<th>$\xi_1$</th>
<th>$\sigma_2$</th>
<th>$\xi_2$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$r$</th>
<th>AIC</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>log</td>
<td>19.12</td>
<td>0</td>
<td>17.75</td>
<td>0</td>
<td>0.57</td>
<td>3365</td>
<td>0.664</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alog</td>
<td>17.47</td>
<td>0</td>
<td>27.2</td>
<td>-0.35</td>
<td>0.99</td>
<td>0.49</td>
<td>0.299</td>
<td>3290</td>
<td>0.799</td>
<td></td>
<td></td>
</tr>
<tr>
<td>amix</td>
<td>17.47</td>
<td>0.004</td>
<td>27.15</td>
<td>-0.37</td>
<td>1.059</td>
<td>-0.059</td>
<td></td>
<td></td>
<td></td>
<td>3366</td>
<td>0.906</td>
</tr>
<tr>
<td>hr</td>
<td>18.16</td>
<td>0</td>
<td>20.05</td>
<td>0</td>
<td>1.504</td>
<td>3371</td>
<td>0.875</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aneglog</td>
<td>17.49</td>
<td>-0.15</td>
<td>27.05</td>
<td>-0.39</td>
<td>1</td>
<td>0.356</td>
<td>1.28</td>
<td>3344</td>
<td>0.961</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neglog</td>
<td>18.82</td>
<td>0</td>
<td>18.78</td>
<td>0</td>
<td>1.05</td>
<td>3368</td>
<td>0.762</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bilog</td>
<td>17.5</td>
<td>0</td>
<td>26.54</td>
<td>-0.26</td>
<td>0.1</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
<td>3306</td>
<td>0.861</td>
</tr>
<tr>
<td>negbilog</td>
<td>17.99</td>
<td>0</td>
<td>26.47</td>
<td>-0.32</td>
<td>1.49</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td>3325</td>
<td>0.632</td>
</tr>
<tr>
<td>ct</td>
<td>15.1</td>
<td>0</td>
<td>21.98</td>
<td>-0.23</td>
<td>0.42</td>
<td>29.99</td>
<td></td>
<td></td>
<td></td>
<td>3330</td>
<td>0.635</td>
</tr>
</tbody>
</table>

From Table 5.1 we can see all parameter estimation for all nine bivariate extreme value distributions available in the package evd in R. The $p$-value tells us that all models seems reasonable but for model selection we will choose the model with the lowest AIC, ”alog” has the lowest AIC.

It should be noted that the ”alog” model has zero in shape for the Skanör margin and negative shape for the Klagshamn margin, similar to the GPD models for in the one-dimensional analysis. Negative shape parameter for the Klagshamn margin indicates that there is a upper bound for Klagshamn, unbounded for Skanör. The upper limit for Klagshamn is 157 cm in MSL. This does not seem unreasonable since in the one-dimensional case the upper limit was 162 cm for Klagshamn and unbounded for Skanör, thus the results are somewhat accurate.
Now that we have our model it can be in our interest to calculate the conditional density that

\[ g_Y(Y = y|X = 120) = \frac{g_{X,Y}(120, y)}{g_X(120)}. \]

To see that given that Skanør has a sea level at 120 cm what is the most likely value that Klagshamn will have. The conditional density for this is shown in Figure 5.2.

![Conditional density for alog](image.png)

Figure 5.2: Conditional density for the ”alog” model given that Skanør has a sea level at 120 cm. Most weight is around 120 cm.

So far this model seems to reasonable for the data. The main issue with this model is easiest explained by computing some probabilities with the model.

A: \( P(Y > 100|X > 130) = \frac{1 - G_X(130) - G_Y(100) + G_{X,Y}(130, 100)}{1 - G_X(130)} \)

B: \( P(X > 100|Y > 130) = \frac{1 - G_X(100) - G_Y(130) + G_{X,Y}(100, 130)}{1 - G_Y(130)} \)

C: \( P(Y > 160|X > 180) = \frac{1 - G_X(180) - G_Y(160) + G_{X,Y}(180, 160)}{1 - G_X(180)} \)

D: \( P((X,Y) < (188, 150)|(X,Y) \notin (80,80)) = \frac{G_{X,Y}(188, 150) - G_{X,Y}(80, 80)}{1 - G_{X,Y}(80, 80)}. \)
Table 5.2: Calculated probabilities with the "alog" model.

<table>
<thead>
<tr>
<th>Probability</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.9903</td>
</tr>
<tr>
<td>B</td>
<td>0.6258</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0.9962</td>
</tr>
</tbody>
</table>

The probability C says that since the upper limit is 157 cm for Klagshamn the probability that given Skanör has a sea level of 180 cm or above what is the probability that Klagshamn has a sea level of 160 cm or above is zero. Probability D tells us that the model does expect points above (188,150) cm, this is due to that the sea level for Skanör is unbounded apart from Klagshamn.

Since the model can not use all available data from Klagshamn it may be estimating the upper point in the distribution wrong. If more data was available the model would most likely perform better. Thus this model is not well suited for doing any analysis regarding coastal protection, more data is needed.

Thus the conclusion is that this model is not better than performing models on the margins separately.
5.1.2 Type 2 BGPD

We will now focus on Type 2 BGPD, which means that we will try to fit

\[
\{(x, y) | (x, y) \not\in (u_x, u_y)\}
\]

where \((u_x, u_y) = (80, 80)\).

Transforming the data so that \((80, 80) = (0, 0)\) will result in Figure 5.3.

![Image](imageurl)

**Figure 5.3:** Data transformed to fit Type 2 BGPD.

We will fit nine models to the Type 2 BGPD, note that the estimation is based on a likelihood function and therefore strongly dependent on good starting values. Finding suitable good starting values for all nine models we got eight of them to converge.
Table 5.3: All estimated parameters for all nine models. The estimated parameters in the "negbilog" model did not converge.

<table>
<thead>
<tr>
<th>Models</th>
<th>loc 1</th>
<th>scale 1</th>
<th>shape 1</th>
<th>loc 2</th>
<th>scale 2</th>
<th>shape 2</th>
<th>dep</th>
<th>asy</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;log&quot;</td>
<td>3.3392</td>
<td>9.8434</td>
<td>1.0984</td>
<td>2.8344</td>
<td>18.2410</td>
<td>0.3209</td>
<td>2.5046</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>&quot;bilog&quot;</td>
<td>2.6688</td>
<td>7.6621</td>
<td>0.9551</td>
<td>2.6387</td>
<td>18.1990</td>
<td>0.1226</td>
<td>(0.1711, 0.5852)</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>&quot;negbilog&quot;</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>&quot;taj&quot;</td>
<td>1.1100</td>
<td>6.7944</td>
<td>0.9193</td>
<td>-1.111</td>
<td>20.4342</td>
<td>0.4023</td>
<td>(4.1658, 5.6053)</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>&quot;psilog&quot;</td>
<td>3.1702</td>
<td>7.5202</td>
<td>0.5699</td>
<td>3.1702</td>
<td>18.1702</td>
<td>0.0702</td>
<td>2.1202</td>
<td>0.0683</td>
<td>3.1702</td>
</tr>
<tr>
<td>&quot;philog&quot;</td>
<td>3.2966</td>
<td>9.6029</td>
<td>0.9812</td>
<td>-0.9235</td>
<td>17.1664</td>
<td>0.3216</td>
<td>2.6137</td>
<td>-0.0113</td>
<td>1.6587</td>
</tr>
<tr>
<td>&quot;phineglog&quot;</td>
<td>3.0967</td>
<td>7.5914</td>
<td>0.6665</td>
<td>3.068</td>
<td>17.9941</td>
<td>-0.0195</td>
<td>1.817</td>
<td>0.0003</td>
<td>1.96505</td>
</tr>
<tr>
<td>&quot;neglog&quot;</td>
<td>2.033</td>
<td>8.3063</td>
<td>1.064</td>
<td>0.7381</td>
<td>17.6398</td>
<td>0.3119</td>
<td>1.7612</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>
Since the estimation is based on a likelihood function and therefore strongly
dependent on good stating values it is not possible to preform goodness-of-fit
tests for these models. Thus the model selection will be based on AIC and
some evaluated probabilities for these models.

Now that we have fitted our eight models to our data we would like once
again to calculate the probability

\[ E : P(X, Y) < (180, 150)|(X, Y) \not< (80, 80)) = G_{X,Y}(180 - 80, 150 - 80). \]

Doing this for all models and calculate the AIC for all of them will result in
the following table:

Table 5.4: The AIC calculated for all eight models and the probability of E.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>Prob E</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;log&quot;</td>
<td>-1813.726</td>
<td>0.9287246</td>
</tr>
<tr>
<td>&quot;bilog&quot;</td>
<td>-1779.032</td>
<td>0.9568159</td>
</tr>
<tr>
<td>&quot;taj&quot;</td>
<td>-1771.349</td>
<td>0.9188793</td>
</tr>
<tr>
<td>&quot;psilog&quot;</td>
<td>-1843.840</td>
<td>0.9775748</td>
</tr>
<tr>
<td>&quot;psineglog&quot;</td>
<td>-1789.548</td>
<td>0.9299905</td>
</tr>
<tr>
<td>&quot;philog&quot;</td>
<td>-1801.550</td>
<td>0.9376223</td>
</tr>
<tr>
<td>&quot;phineglog&quot;</td>
<td>-1839.478</td>
<td>0.9798314</td>
</tr>
<tr>
<td>&quot;neglog&quot;</td>
<td>-1812.327</td>
<td>0.9315942</td>
</tr>
</tbody>
</table>

From Table 5.4 we can see that all probabilities are around 0.92 which is
lower than the result in Table 5.2. The AIC varies quite much between all
models.
We know proceed to plot the estimate of predictions region for $p = 0.75, 0.9, 0.95, 0.99$. The four models we will do this to is the most interesting once.

- psilog(asymmetric): The "psilog" model has the smallest AIC and the second highest probability between all models.
- phineglog(asymmetric): The "phineglog" model has the second lowest AIC and highest probability of all models.
- log(symmetric): The "log" model has the lowest AIC of the symmetric models.
- bilog(symmetric): The "bilog" model has the highest probability between the symmetric models.

And the 0.75, 0.9, 0.95 and 0.99 quantile for these models is represented below in the following Figure 5.4:

![Quantile plot for Type 2 BGPD.](image)

Figure 5.4: Quantile plot for Type 2 BGPD.
From Figure 5.4 we see that the "psilog" model follows the shape best of all the models, its quantiles curve is more broader than the "phineglog" model. Comparing the symmetric models one sees that the "log" model shape is not as good as the asymmetric "psilog" model. The symmetric "bilog" model is not recommended to use since its quantile curves does not follow the shape at all.

Thus it seems that the "psilog" model performs the best between all eight models. Since the estimated shape parameters are both positive for both margins the problem which occurred in Type 1 BGPD will not occur here, thus this model is preferred in further analysis is to be done.

5.2 Viken and wind speed

Viken is a measure station which is not close to Falsterbo Peninsula and has near independent sea level values to Skånör. The reason we will still study this measure station is that there is said to be strong dependence between wind speed and rising sea level, a factor which was not significant for the other measure stations which we have been studying. Before any analysis takes place we will remove the linear trend as seen in Table 3.1

To verify the hypothesis about wind speed and rising sea level the simultaneously hourly values of them is shown in Figure 5.5.

![Simultaneously values of sea level and wind speed](image)

Figure 5.5: Simultaneously values of wind speed and Vikens sea level. The more wind speed the more likely it is to have higher sea level.
As we can see from Figure 5.5 the higher wind speed the more likely it is for higher sea level to occur. Looking at Spearman’s correlation and Pearson’s correlation will give the following values.

$$\rho_s = 0.3550957, \quad \rho = 0.4278931$$

Since the wind direction also effect the sea level positive, the right wind direction allows the sea to build up more water, it can be of our interest to study how the wind direction effect the sea level at Viken.

Looking at simultaneously values of wind direction one also find a nice correlation here, illustrated by Figure 5.6.

Figure 5.6: Simultaneously values of wind direction and Vikens sea level. When the wind direction is around 200-360 to have higher sea level.

Thus having a model which both takes into account wind speed and wind direction seems to be natural.
To tackle this we construct a variable that both takes into account the wind direction and the wind speed. To get a suitable variable the following steps were taken:

1. Transform wind direction values from degrees to Radians.
2. Sinus transform the wind direction (values between -1 and 1) and multiply with the simultaneously wind speed values.
3. Multiply with -1 to shift the variable.

The main idea of the procedure above is to reward wind direction values that are most likely to lead to high sea level. The procedure above can be modified for other measure stations, for example if a measure station has the most likely high sea level occurrence around 90 degrees step 3 is not needed. Following above steps results in Figure 5.7:

![Figure 5.7: Simultaneously values of transformed wind data and sea level of Viken.](image)

Looking at Spearman’s correlation and Pearson’s correlation will give the following values:

$$\rho_s = 0.5686749, \quad \rho = 0.6155174$$
Comparing with Figure 5.5 following the above procedure results in a variable that has more correlation than just using wind speed, hence we have stronger correlation combining wind speed and wind direction. Figure 5.7 shows that if the wind direction is around 270 degrees with a strong wind it is more likely to have a high sea level, which seems very natural.

Further extreme analysis will now be preformed on the transformed wind data resulting in Figure 5.7.

### 5.2.1 Type 1 BGPD

Once again we will have to declustered the data since their are still dependence left in the hourly observations of the sea level. A suitable threshold seems to be around 90 for the sea level and around 16 for the transformed wind data.

\[ \{(x, y)| (x, y) > (16, 90)\} \]

This is shown in Figure 5.8:

![Declustered values above the threshold](image)

**Figure 5.8:** Simultaneously values of transformed wind data and sea level of Viken. All values above 16, 90 has been declustered.

All nine bivariate models where fitted to the values above. If the shape parameter was not significant different from 0 for a model it was set to 0. The estimated parameters for the models is presented in Table 5.5.
Table 5.5: The parameter estimation for all nine parametric bivariate extreme value models in the package `evd` in R. The AIC has been calculated for each model, the goodness-of-fit test has been performed on all models, based on 1000 simulations. The models "bilog" and "ct" have the lowest AIC.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_1$</th>
<th>$\xi_1$</th>
<th>$\sigma_2$</th>
<th>$\xi_2$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$r$</th>
<th>AIC</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>log</td>
<td>2.0719</td>
<td>0</td>
<td>17.351</td>
<td>0</td>
<td>0.8163</td>
<td>10260</td>
<td>0.366</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>alog</td>
<td>2.1466</td>
<td>-0.028</td>
<td>13.896</td>
<td>0.163</td>
<td>0.745</td>
<td>0.999</td>
<td>0.7794</td>
<td>10262</td>
<td>0.229</td>
<td></td>
<td></td>
</tr>
<tr>
<td>amix</td>
<td>2.1046</td>
<td>0</td>
<td>18.036</td>
<td>0</td>
<td>0.194</td>
<td>0.195</td>
<td>0.842</td>
<td>10264</td>
<td>0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hr</td>
<td>2.102</td>
<td>0</td>
<td>17.374</td>
<td>0</td>
<td>0.842</td>
<td>10263</td>
<td>0.404</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>aneglog</td>
<td>2.0766</td>
<td>0</td>
<td>17.216</td>
<td>0</td>
<td>0.462</td>
<td>0.787</td>
<td>0.7323</td>
<td>10260</td>
<td>0.384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>neglog</td>
<td>2.0914</td>
<td>0</td>
<td>17.286</td>
<td>0</td>
<td>0.4826</td>
<td>10261</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bilog</td>
<td>2.0734</td>
<td>0</td>
<td>17.275</td>
<td>0</td>
<td>0.871</td>
<td>0.749</td>
<td>10259</td>
<td>0.402</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>negbilog</td>
<td>2.085</td>
<td>0</td>
<td>17.352</td>
<td>0</td>
<td>1.515</td>
<td>2.886</td>
<td>10262</td>
<td>0.392</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ct</td>
<td>2.0733</td>
<td>0</td>
<td>17.247</td>
<td>0</td>
<td>0.382</td>
<td>0.164</td>
<td>10259</td>
<td>0.373</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From Table 5.5 we can see all parameter estimation for all nine bivariate extreme value distribution. The $p$-value shows that all models seem to fit the data good. For model selection the models with the lowest AIC will be investigated. The models "bilog" and "ct" has similar AIC and further analysis will be done to choose the best of them. Note that the both shape parameters are zero for both models, indicating that the margins has no upper bound. Now that we have all our models we can calculate the conditional density that

- Case 1: $g_y(Y = y|X = 30) = \frac{g_{X,Y}(30, y)}{g_X(30)}$

- Case 2: $g_y(Y = y|X = 50) = \frac{g_{X,Y}(50, y)}{g_X(50)}$.

To see that given that we have a wind speed around 30, 50 m/s in the right wind direction what the density looks like for different MSL. The densities will be calculated for both the "bilog" and "ct" model. The reason the wind speed was set so high is that to check that the models behave correctly under strong wind assumptions, this will result in Figure 5.9.
Figure 5.9: Conditional density calculated for both "bilog" and "ct" model for both cases. One can clearly see the difference between the models.

From Figure 5.9 one can clearly see that the "bilog" model does not perform well under strong wind assumptions. Since Figure 5.7 clearly indicates an up-going trend the conditional density calculated for Case 2 for "bilog" shows us that this model is not the best to choice. Instead the "ct" model seems to preforming better. Looking at Figure 5.7 one can see that one has a large variance during wind assumption, thus the "ct" model does catch this behavior accordingly. The "ct" model seems to give reasonable result during very extreme wind condition.
Now that the chosen model "ct" has been selected we can calculate some interesting probabilities for this model.

\[ A : P((X,Y) \neq (20,150)) = 1 - G_{X,Y}(20,150) \]

\[ B : P((X,Y) \neq (30,150)) = 1 - G_{X,Y}(30,150) \]

\[ C : P((X,Y) \neq (20,240)) = 1 - G_{X,Y}(20,240) \]

\[ D : P((X,Y) \neq (30,240)) = 1 - G_{X,Y}(30,240). \]

Table 5.6: Calculated probabilities with the "ct" model.

<table>
<thead>
<tr>
<th>Probability</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.1493892</td>
</tr>
<tr>
<td>B</td>
<td>0.03090318</td>
</tr>
<tr>
<td>C</td>
<td>0.1352068</td>
</tr>
<tr>
<td>D</td>
<td>0.001246608</td>
</tr>
</tbody>
</table>

From Table 5.6 one can see that the calculated probability A and C is high, indicating that the model does not expect a sea level with corresponding wind speed and wind direction. Since the probability B and D is lower the model expect that its more likely that wind conditions of 30 is more likely to give a sea level of 150 MSL, respectively 240 MSL for probability D.

The "ct" model seems to work properly given all test that has been carried out. The model presented would be practically used to calculate sea level given wind conditions. Since SMHI can predict storm of large magnitudes and which wind direction they will strike this model could be used to predict what sea level would be most likely to occur.
5.2.2 Type 2 BGPD

We will now focus on Type 2 BGPD, which means that we will try to fit

\[ \{(x, y) | (x, y) < (u_x, u_y)\} \]

where \((u_x, u_y) = (16, 90)\).

Transforming the data so that \((16, 90) = (0, 0)\) will result in Figure 5.10.

![Figure 5.10: Data transformed to fit Type 2 BGPD.](image)

We will try to fit nine models of Type 2 BGPD, note that the estimation is based on a likelihood function and therefore strongly depends on good starting values. Finding suitable good starting values for all nine models eight of them converge.
Table 5.7: All estimated parameter for all nine models. The estimated parameters in the “negbilog” model did not converge.

<table>
<thead>
<tr>
<th>Models</th>
<th>loc 1</th>
<th>scale 1</th>
<th>shape 1</th>
<th>loc 2</th>
<th>scale 2</th>
<th>shape 2</th>
<th>dep</th>
<th>asy</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>“log”</td>
<td>-0.4184</td>
<td>2.0199</td>
<td>0.0640</td>
<td>-8.8360</td>
<td>12.3603</td>
<td>0.1474</td>
<td>1.4090</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>“bilog”</td>
<td>-0.9912</td>
<td>2.8750</td>
<td>0.08005</td>
<td>-20.1505</td>
<td>29.2647</td>
<td>-0.0939</td>
<td>(0.3661, 0.6397)</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>“negbilog”</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>“taj”</td>
<td>-0.3193</td>
<td>2.0502</td>
<td>0.0522</td>
<td>-7.9505</td>
<td>12.6568</td>
<td>0.1422</td>
<td>(1.4666, 0.1947)</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>“psilog”</td>
<td>-1.006</td>
<td>3.9289</td>
<td>-0.0086</td>
<td>-16.2709</td>
<td>27.8324</td>
<td>0.1493</td>
<td>2.1162</td>
<td>-0.9508</td>
<td>3.4278</td>
</tr>
<tr>
<td>“psineglog”</td>
<td>-0.8328</td>
<td>4.1671</td>
<td>-0.2208</td>
<td>-17.8328</td>
<td>27.8421</td>
<td>-0.1162</td>
<td>1.3421</td>
<td>0.6671</td>
<td>3.5171</td>
</tr>
<tr>
<td>“philog”</td>
<td>1.7250</td>
<td>2.0859</td>
<td>0.0654</td>
<td>4.8268</td>
<td>14.4886</td>
<td>0.1485</td>
<td>1.4053</td>
<td>0.0035</td>
<td>1.3197</td>
</tr>
<tr>
<td>“phineglog”</td>
<td>-1</td>
<td>5.350</td>
<td>-0.2787</td>
<td>-18</td>
<td>28.012</td>
<td>0.1162</td>
<td>3</td>
<td>-8.237e-07</td>
<td>3</td>
</tr>
<tr>
<td>“neglog”</td>
<td>4.5066</td>
<td>2.5131</td>
<td>0.06008</td>
<td>24.8592</td>
<td>19.5951</td>
<td>0.1494</td>
<td>0.70183</td>
<td>None</td>
<td>None</td>
</tr>
</tbody>
</table>
Since the estimation is based on a likelihood function and therefore strongly depends on the starting parameters it is not possible to perform the goodness-of-fit test as we did in Type 1 BGPD. Thus the model selection will be based on AIC and some evaluated probabilities.

Now that eight of the nine models is fitted to the data we can calculate the probability

\[ E : P((X,Y) < (30,240) | (X,Y) \not< (16,90)) = G_{X,Y}(30 - 16, 240 - 90) \]

for all eight models. Calculating this probability and AIC for the eight models following Table 5.8:

Table 5.8: The AIC calculated for all eight models and the probability of E.

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>Prob E</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;log&quot;</td>
<td>-6182.928</td>
<td>0.9971644</td>
</tr>
<tr>
<td>&quot;bilog&quot;</td>
<td>-6300.156</td>
<td>0.9856383</td>
</tr>
<tr>
<td>&quot;taj&quot;</td>
<td>-6176.264</td>
<td>0.9975019</td>
</tr>
<tr>
<td>&quot;psilog&quot;</td>
<td>-6367.331</td>
<td>0.9741044</td>
</tr>
<tr>
<td>&quot;psineglog&quot;</td>
<td>-6403.426</td>
<td>0.9990279</td>
</tr>
<tr>
<td>&quot;philog&quot;</td>
<td>-6182.237</td>
<td>0.9974704</td>
</tr>
<tr>
<td>&quot;phineglog&quot;</td>
<td>-6999.834</td>
<td>0.9877179</td>
</tr>
<tr>
<td>&quot;neglog&quot;</td>
<td>-6186.827</td>
<td>0.9957060</td>
</tr>
</tbody>
</table>

From Table 5.8 we can see that all probabilities are around 0.99, the AIC varies a bit for all models.

We now proceed to plot the estimate of predictions region for \( p = 0.75, 0.9, 0.95, 0.99 \). The four models we will do this to is the most interesting one.

- **phinlog** (asymmetric): The "phinlog" model has the smallest AIC and the second lowest probability between all asymmetric models.
- **psineglog** (asymmetric): The "psineglog" model has the second lowest AIC of all models and highest probability of all models.
- **bilog** (symmetric): The "bilog" model has the lowest AIC of the symmetric models.
- **taj** (symmetric): The "taj" model has the highest probability between the symmetric models but the highest AIC.
And the 0.75, 0.9, 0.95 and 0.99 quantile for these models is represented below in Figure 5.11.

Figure 5.11: Quantile plot for Type 2 BGPD.

From Figure 5.11 we can see that the model that had the lowest AIC, ”phineglog”, has the most extreme shape between the four selected models. The ”phineglog” model suggest very extreme sea level when the wind speed is around 30 in the right wind direction. Comparing this to the ”bilog” or ”taj” which seems to give lower results. Since the quantile curves of the asymmetric models seem to extreme the symmetric model ”bilog” seems to fit the process the best.

Since there is no available goodness-of-fit test for models of Type 2 BGPD one should trust the models of Type 1 BGPD more than these.
5.3 Ängelholm, waves and sea level

In previous section wind speed and wind direction we used to help us modeling extreme sea levels, now we will try using wave height instead. The data is from Ängelholm and sea level has been patched for window clamping, thus no linear sea level trend will be removed from this data set. Though wave height and wind speed are somewhat correlated its always preferred to represent different kinds of approaches so that one can use the one that suits their measure station the best.

![Simultaneously values of wave height and sea level](image)

Figure 5.12: Simultaneously values of wave height and sea level in Ängelholm.

As one can see from Figure 5.12 there is dependence between high sea levels and the wave height. Even though there is a huge variance in the low wave heights the most extreme wave heights results in extreme sea levels, for this measure station. Figure 5.12 suggest a up going trend, the suspicious are confirmed by calculating Pearson’s correlation $\rho$ and Spearman’s correlation $\rho_s$ one will get:

$$\rho_s = 0.5429468 \quad \rho = 0.605578$$

indicating a positive correlation between wave heights and sea level.
5.3.1 Type 1 BGPD

Once again we will have to declustered the data since there are still dependence left in the hourly observations of the sea level. A suitable threshold seems to be around 80 for the sea level and 3 for the wave height [m] for the margins

\[ \{(x, y) | (x, y) > (2.9, 78)\}. \]

The data with corresponding threshold are shown in Figure 5.13:

![Declustered values of wave height and sea level above (2.9,78)](image)

Figure 5.13: Simultaneously values of wave height and sea level in Ängelholm. All Values above (2.9,78) has been declustered.

All nine bivariate models where fitted to the values above. If the shape parameter was not significant different from 0 for a model it was set to 0. The estimated parameters and their corresponding p-value is shown in Table 5.9.
Table 5.9: The parameter estimation for all nine parametric bivariate extreme value models in the package evd in R. The AIC has been calculated for each model, the goodness-of-fit test has been performed on all models, based on 1000 simulations. The model “amix” has the lowest AIC.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\sigma_1$</th>
<th>$\xi_1$</th>
<th>$\sigma_2$</th>
<th>$\xi_2$</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$t_1$</th>
<th>$t_2$</th>
<th>$r$</th>
<th>AIC</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>log</td>
<td>0.3717</td>
<td>0</td>
<td>14.1706</td>
<td>0.1998</td>
<td></td>
<td></td>
<td>0.8104</td>
<td>8523</td>
<td>0.724</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alog</td>
<td>0.3515</td>
<td>0</td>
<td>13.8669</td>
<td>0.2188</td>
<td></td>
<td>0.6700</td>
<td>0.9990</td>
<td>0.7367</td>
<td>8522</td>
<td>0.592</td>
<td></td>
</tr>
<tr>
<td>amix</td>
<td>0.3717</td>
<td>0</td>
<td>13.9201</td>
<td>0.2348</td>
<td>0.2118</td>
<td>0.2017</td>
<td></td>
<td></td>
<td></td>
<td>8516</td>
<td>0.691</td>
</tr>
<tr>
<td>hr</td>
<td>0.3892</td>
<td>0</td>
<td>14.0678</td>
<td>0.1973</td>
<td></td>
<td></td>
<td>0.8575</td>
<td>8532</td>
<td>0.807</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aneglog</td>
<td>0.3696</td>
<td>0</td>
<td>13.9313</td>
<td>0.2266</td>
<td></td>
<td>0.4342</td>
<td>0.6581</td>
<td>0.9602</td>
<td>8518</td>
<td>0.688</td>
<td></td>
</tr>
<tr>
<td>neglog</td>
<td>0.3815</td>
<td>0</td>
<td>14.1471</td>
<td>0.1967</td>
<td></td>
<td></td>
<td>0.4961</td>
<td>8528</td>
<td>0.795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bilog</td>
<td>0.3732</td>
<td>0</td>
<td>13.8678</td>
<td>0.2234</td>
<td>0.8588</td>
<td>0.7188</td>
<td></td>
<td></td>
<td></td>
<td>8522</td>
<td>0.715</td>
</tr>
<tr>
<td>negbilog</td>
<td>0.3786</td>
<td>0</td>
<td>13.8268</td>
<td>0.2293</td>
<td>1.0410</td>
<td>3.5267</td>
<td></td>
<td></td>
<td></td>
<td>8525</td>
<td>0.754</td>
</tr>
<tr>
<td>ct</td>
<td>0.3738</td>
<td>0</td>
<td>13.8671</td>
<td>0.2206</td>
<td>0.5004</td>
<td>0.1740</td>
<td></td>
<td></td>
<td></td>
<td>8522</td>
<td>0.72</td>
</tr>
</tbody>
</table>

From Table 5.9 we can all parameter estimation for all nine bivariate extreme value models. The $p$-value indicates that all models are suitable candidates but for model selection we will choose the model with the lowest AIC, ”amix” has the lowest AIC. We will compare the ”amix” model with the ”hr”, which had the highest $p$-value. Further analysis on ”amix” and ”hr” will be carried out. First we will calculate the conditional density that

- Case 1: $g_Y(Y = y | X = 3.5) = \frac{g_{X,Y}(3.5, y)}{g_X(3.5)}$

- Case 2: $g_Y(Y = y | X = 6) = \frac{g_{X,Y}(6, y)}{g_X(6)}$.

To see that given that we have a wave height around 3.5, 6 meters what the density looks like for different sea levels. The densities will be calculated for both ”amix” and the ”hr” model, this will result in Figure 5.14.
Figure 5.14: Conditional density calculated for both "hr" and "amix" model, for both cases. One can clearly see the difference.

From Figure 5.14 we can see that the "hr" model and "amix" model preform very differently between the calculated conditional densities between the two different cases. The main difference is seen in the second case. The "hr" model has more weight around sea level of 200 cm than the "amix" model, the "amix" model has more weight around 100 and some around 400. Thus even though the "amix" model had the lowest AIC between all the models actually the "hr" model seems to preform better during the tests, this really shows the importance of doing proper analysis.

Now that the chosen model "hr" has been selected we can calculate some interesting probabilities for this model.

\[ A : P(X,Y \not< (3,150)) = 1 - G_{X,Y}(3,150) \]
\[ B : P(X,Y \not< (4,150)) = 1 - G_{X,Y}(4,150) \]
\[ C : P(X,Y \not< (3,240)) = 1 - G_{X,Y}(3,240) \]
\[ D : P(X,Y \not< (4,240)) = 1 - G_{X,Y}(4,240) \].
Table 5.10: Calculated probabilities with the “hr” model.

<table>
<thead>
<tr>
<th>Probability</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.5421904</td>
</tr>
<tr>
<td>B</td>
<td>0.07547131</td>
</tr>
<tr>
<td>C</td>
<td>0.5386328</td>
</tr>
<tr>
<td>D</td>
<td>0.05818069</td>
</tr>
</tbody>
</table>

From Table 5.10 one can see that the calculated probability A and C is high, indicating that the model does not expect a that given sea level with that given wave height. Since the probability B and D is lower the model expect that its more likely that a wave height if 4 meters is more likely to give a sea level of 150 cm, respectively 240 cm for probability D.

The ” hr” model seems to work accordingly to all test that has been carried out. The model presented could be practically used to calculate sea level given wave heights. This would provide more information about the wave heights during storm surges and thus allowing to build sea protection accordingly to what wave height one expects.
5.3.2 Type 2 BGPD

We will now focus on Type 2 BGPD, which means that we will try to fit
\[ \{(x, y) | (x, y) \not< (u_x, u_y)\} \]
where \((u_x, u_y) = (2.9, 78)\).

Transforming the data so that \((2.9, 78) = (0, 0)\) which results in Figure 5.15:

![Exceedances](image)

Figure 5.15: Data transformed to fit Type 2 BGPD.

We will fit nine models of Type 2 BGPD, note that the estimation is based on a likelihood function and therefore strongly depends on good starting values. Finding suitable good starting values for all nine models eight of them converge.
Table 5.11: All estimated parameter for all nine models. The estimated parameters in the "negbilog" model did not converge.

<table>
<thead>
<tr>
<th>Models</th>
<th>loc 1</th>
<th>scale 1</th>
<th>shape 1</th>
<th>loc 2</th>
<th>scale 2</th>
<th>shape 2</th>
<th>dep</th>
<th>dep</th>
<th>dep</th>
<th>dep</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;log&quot;</td>
<td>-0.3538</td>
<td>0.9461</td>
<td>-0.1803</td>
<td>-4.8538</td>
<td>26.7961</td>
<td>0.2247</td>
<td>2.1461</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>&quot;bilog&quot;</td>
<td>-0.2073</td>
<td>0.4368</td>
<td>0.1266</td>
<td>-3.6713</td>
<td>25.2046</td>
<td>0.1332</td>
<td>(0.4671, 0.6885)</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>&quot;negbilog&quot;</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>No conver</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>&quot;taj&quot;</td>
<td>-0.1949</td>
<td>0.3524</td>
<td>0.0799</td>
<td>-1.9969</td>
<td>14.1389</td>
<td>0.1483</td>
<td>(1.4713, 0.1749)</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>&quot;psilog&quot;</td>
<td>-0.4618</td>
<td>0.8381</td>
<td>-0.1819</td>
<td>-4.9618</td>
<td>26.0381</td>
<td>0.1270</td>
<td>3.0381</td>
<td>1.0381</td>
<td>3.9881</td>
<td></td>
</tr>
<tr>
<td>&quot;psineglog&quot;</td>
<td>-0.4618</td>
<td>0.8381</td>
<td>-0.1819</td>
<td>-4.9618</td>
<td>26.6881</td>
<td>0.1270</td>
<td>2.0381</td>
<td>1.0381</td>
<td>4.3381</td>
<td></td>
</tr>
<tr>
<td>&quot;philog&quot;</td>
<td>-0.2760</td>
<td>0.4984</td>
<td>0.0842</td>
<td>-4.895</td>
<td>27.8963</td>
<td>0.0253</td>
<td>1.6969</td>
<td>0.0145</td>
<td>1.7203</td>
<td></td>
</tr>
<tr>
<td>&quot;phaseglog&quot;</td>
<td>-0.4923</td>
<td>0.9498</td>
<td>-0.1923</td>
<td>-4.9923</td>
<td>28.2826</td>
<td>-0.0298</td>
<td>1.9602</td>
<td>-0.00084</td>
<td>2.0076</td>
<td></td>
</tr>
<tr>
<td>&quot;neglog&quot;</td>
<td>-0.2380</td>
<td>0.3662</td>
<td>0.0477</td>
<td>-3.7820</td>
<td>14.4285</td>
<td>0.1268</td>
<td>0.6692</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
</tbody>
</table>
Since the estimation is based on a likelihood function and therefore strongly depends on the starting parameters it is not possible to preform the goodness-of-fit test as we did in Type 1 BGPD. Thus the model selection will be based on AIC and some evaluated probabilities.

Now that eight of the nine models is fitted to the data we can calculate the probability

\[ E : P((X, Y) < (5, 240) | (X, Y) \geq (2.9, 78)) = G_{X,Y}(5 - 2.9, 240 - 78) \]

for all eight models. Calculating this probability and AIC for the eight models following Table 5.12:

<table>
<thead>
<tr>
<th>Model</th>
<th>AIC</th>
<th>Prob E</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;log&quot;</td>
<td>-4137.078</td>
<td>0.9660805</td>
</tr>
<tr>
<td>&quot;bilog&quot;</td>
<td>-4017.449</td>
<td>0.9814784</td>
</tr>
<tr>
<td>&quot;taj&quot;</td>
<td>-3941.804</td>
<td>0.9952398</td>
</tr>
<tr>
<td>&quot;psilog&quot;</td>
<td>-4371.371</td>
<td>0.9854985</td>
</tr>
<tr>
<td>&quot;psineglog&quot;</td>
<td>-4268.003</td>
<td>0.9849739</td>
</tr>
<tr>
<td>&quot;philog&quot;</td>
<td>-4046.303</td>
<td>0.9825979</td>
</tr>
<tr>
<td>&quot;phineglog&quot;</td>
<td>-4222.457</td>
<td>0.9779043</td>
</tr>
<tr>
<td>&quot;neglog&quot;</td>
<td>-3946.954</td>
<td>0.9961808</td>
</tr>
</tbody>
</table>

From Table 5.12 we can see that all probabilities are around 0.99, the AIC varies a bit for all models.

We know proceed to plot the estimate of predictions region for \( p = 0.75, 0.9, 0.95, 0.99 \).

The four models we will do this to is the most interesting once.

- psilog(asymmetric): The "psilog" model has the smallest AIC and the highest probability between all asymmetric models.

- phineglog(asymmetric): The "phineglog" model has the lowest probability between all asymmetric models.

- taj(symmetric): The "taj" model has the highest AIC and the highest probability of all models.

- log(symmetric): The "log" model has the lowest AIC between the symmetric models and the highest probability of all models.
And the 0.75, 0.9, 0.95 and 0.99 quantile for these models is represented below in the Figure 5.16:

![Quantile plot for Type 2 BGPD](image)

Figure 5.16: Quantile plot for Type 2 BGPD.

From Figure 5.16 one sees that the "psilog", "phineglog" and the "log" model have very similar shapes in the quantiles, the main difference is that the "log" model has wider quantiles curves. Since these quantile curves are so extreme for the 0.99 % quantile it suggest that the symmetric "taj" model is the preferred model for this.

Once more a reminder that the models based on Type 2 BGPD does not have a developed goodness-of-fit test and thus should not be trusted wholeheartedly.
Chapter 6

Discussion

6.1 Summary

The purpose of this master’s thesis was to estimate the return period of the 1872 incident in Falsterbo Peninsula using the available SMHI data. The selected stations where stations close to Falsterbo or had strong correlation with this measure station.

The data was model both with block-maxima and Generalized Pareto Distribution (GPD) to compare, and see similarities between the estimated parameters in the models. If the models were similar it would indicate that the model are consistent. The parameters in the models was estimated using maximum likelihood, fit diagnostic was used to choose the model fitting best. The best model for estimating the return period for the 1872 incident proved to be the GEV fit of Klagshamn. Once the chosen model was selected we converted its values to match Falsterbo sea level accordingly to estimate different return periods for this model.

Further analysis was made in Falsterbo to see if other covariates affect extreme sea level such as wind speed and wave heights. Since no significant correlation was shown for this measure station it was not included but nearby measure stations that showed a strong correlation was investigated. Viken showed a strong correlation between wind speed and wind direction and Angelholm with wave height. Thus a multivariate extreme value approach was done for these to measure station. Model selection was based on goodness-of-fit, the AIC and some evaluated probabilities for the models.
6.2 Discussion

All in all, the results seems quite reasonable. The chosen model GEV Klagshamn model for calculating the return period of the 1872 incident has almost a perfect 1:1 in the QQ plot. Though one has to keep in mind with this model is that the shape parameter is negative, as seen in Table 4.1, the upper limit of the distribution is finite. Thus the model might not be reasonable if some drastic sea level occurrence happen in Klagshamn that would overshot the mean of the upper limit. Even thought the shape parameter negative the measure station has almost 30 more years of data than Skanör and has recent extreme events recorded, making this a more reliable model than Ystad.

Comparing GEV and GPD parameters in Table 4.1 and Table 4.3 one sees not a huge difference, indicating that the models are reliable. Non-stationary models was tested but proved not to be significant better than the chosen stationary model.

Handling the missing values of Skanör by replacing them with Klagshamn values seems like the appropriate approach since the largest replaced value was sea level of 50 cm, though simply replacing the values otherwise is not recommended if it had been above 50 cm. One can argue that simply removing the early years of Klagshamn is not efficient, waste of data. Since the missing data as seen in Figure 3.4 happen frequently in the most extreme months from October to Mars it becomes rather difficult to handle this data set in a different way, though it should be further investigated.

The multivariate approach, adding wind speed, wind direction and wave height seems to further increase the understanding what causes extreme sea levels, the models seems quite good. The goodness-of-fit test indicates that all models of Type 1 BGPD seems trustworthy. Different threshold levels for Viken were tested to further improve the goodness-of-fit result in Table 5.5, thought to no further success. The Type 2 BGPD seems to work rather well but since there are no goodness-of-fit test that works for these kind of models, since the heavily rely on good starting points one has to be extra careful. More test that fits the Type 2 BGPD has to be developed in order increase its certainty.
If wind speed and wave heights have no significant effect on extreme sea level one-dimensional extreme value analysis should be carried out for that measure stations, if needed looking at nearby stations also. If the measure station is located near large-size ocean chances are that wave height and wind speed are heavily correlated with extreme sea levels, thus an multivariate approach should be carried out. The advantages with a multivariate approach that takes into account wind speed or wave height is that one can predict sea levels based on the other covariate. If we know that from weather observations that a hurricane will take place in a couple of days ahead one can calculate the expected sea level based on the wind speed and wind-direction and thus take accordingly preparations.
Bibliography


