Analysis of Extreme Events in the Coastal Engineering Environment

Cornel Stander

Thesis presented in partial fulfilment of the requirements for the degree of Master of Science in Applied Mathematics in the Faculty of Science at Stellenbosch University.

Supervisor: Dr GPJ Diedericks
Co-Supervisor: Dr S Fidder-Woudberg

December 2015

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 2 October 2015

Abstract

Coastal zones are subject to storm events and extreme waves with certain return periods. The return period of such events is defined as the average time interceding two independent, consecutive events, similar in nature, i.e., with the same return level. Coastal structures have to be designed to provide sufficient protection against flooding or erosion to a desired return level associated with a particular return period, for example 100 years. Statistical analyses of measured wave data over a time series are used for these estimations.

In this study, wave data, measured by a Datawell Waverider buoy, is analysed by means of extreme value analyses. This dataset covers only approximately 18 years. Extreme value theory provides a framework that enables extrapolation in order to estimate the probability of events that are more extreme than any that have already been observed. It can, for example, be used to estimate wave return levels over the next 100 years given only an 18 year history. Different methods for making these estimations are implemented and evaluated.

Datasets containing periods where data values are absent (i.e., gaps in a dataset), as well as the effects these missing values have on the estimation of extreme values, are also investigated. Methods for the treatment of gaps are evaluated by using NCEP (National Centre for Environmental Prediction) hindcast data, containing no missing values, and creating incomplete datasets from this data. Estimations are then made based on these incomplete sets. The resulting estimations are compared to the estimations made based on the complete NCEP dataset.

Finally, recommendations are made for conducting optimal extreme value analyses, based on this study.

Opsomming

Kusgebiede is onderhewig aan storms en ekstreme golwe met sekere terugkeer-periodes. Die terugkeer-periode van 'n gebeurtenis word gedefinieer as die gemiddelde tyd tussen twee onafhanklike, opeenvolgende gebeurtenisse, gelyksoortig van aard, met ander woorde, met dieselfde terugkeer-vlak. Kusstrukture moet ontwerp word om genoegsame beskerming teen oorstromings of erosie tot 'n spesifieke terugkeer-vlak, geassosieer met 'n spesifieke terugkeer-periode, byvoorbeeld 'n 100 jaar, te bied. Statistiese analise van gemete golfdata oor 'n tydsreeks word gebruik vir hierdie benaderings.

In hierdie studie word golfdata, gemeet deur 'n *Datawell Waverider* boei, geanaliseer deur middel van ekstreemwaarde analise. Hierdie datastel dek slegs ongeveer 18 jaar. Ekstreemwaarde-teorie bied 'n raamwerk wat ekstrapolasie moontlik maak om die waarskynlikheid van gebeurtenisse te bepaal wat meer ekstreem is as enige gebeurtenisse wat reeds waargeneem is. Dit kan byvoorbeeld gebruik word om golf-terugkeer-vlakke oor die volgende 100 jaar te voorspel, gegewe slegs 'n 18 jaar geskiedenis. Verskillende metodes om hierdie beramings te maak, word geïmplementeer en geëvalueer.

Datastelle met periodes waar datawaardes afwesig is (ook genoem gapings in 'n datastel), asook die uitwerking van hierdie afwesige datawaardes op die beraming van ekstreemwaardes, word ondersoek. Metodes vir die hantering van gapings word geëvalueer deur onvolledige datastelle te skep uit volledige (met ander woorde, sonder enige afwesige waardes) NCEP (*National Centre for Environmental Prediction*) data. Beramings word dan gemaak gebasseer op die onvolledige datastelle. Hierdie resulterende beramings word vergelyk met die beramings gemaak gebasseer op die volledige NCEP datastel.

Uiteindelik word aanbevelings gemaak vir die optimale uitvoering van ekstreemwaarde analise, gebasseer op hierdie studie.

Acknowledgements

I would like to thank the following people sincerely for playing a part in my journey towards the completion of my thesis:

- The Council for Scientific and Industrial Research (CSIR) for providing wave data.
- My supervisors, dr Diedericks and dr Fidder-Woudberg, for their guidance and time.
- Luther Terblanche, that formed part of the supervisor team, for his inputs.
- My parents, H.P. and Suné van der Merwe, for their wisdom, guidance, and financial support.
- My brother, H.P., for providing humoristic inputs in gloomy times.
- My husband, Liam Stander, for his unconditional support and encouragement.
- God, for blessing me with all of the above mentioned people as well as the privilege and ability to complete my studies.

Contents

Nomenclature

Abbreviations

CDF	Cumulative Distribution Function
pdf	probability density function
MLE	Maximum Likelihood Estimate
POT	Points Over Threshold
GEV	Generalized Extreme Value
GPD	Generalized Pareto Distribution
NCEP	National Centre for Environmental Prediction

Standard characters

H_{mo}	significant wave height
u	threshold level
F	cumulative distribution function
f	probability density function
F_{GEV}	generalized extreme value cumulative distribution function
f_{GEV}	generalized extreme value probability density function
F_{GPD}	generalized Pareto cumulative distribution function
f_{GPD}	generalized Pareto probability density function
F_W	Weibull cumulative distribution function
f_W	Weibull probability density function
p	non-exceedance probability
x_p	return value corresponding to non-exceedance probability p

Greek symbols

θ	location parameter
α	scale parameter
ξ	shape parameter

Miscellaneous

$E(\cdot)$	expected value

Chapter 1

Introduction

1.1 Problem Statement and Data

Coastal zones are subject to storm events with associated extreme waves. Various parameters are of interest, but in this study emphasis is put on the return periods of the extreme wave conditions. The return period (also called the average recurrence interval) of such events is defined as the average time interceding two independent, consecutive events, similar in nature, i.e., with the same return level (Corbella & Stretch (2012)). For example, the return period of a 5 m high (return level) wave refers to the average time interceding the occurence of two consecutive 5 m high waves over a certain time period.

The estimation of return levels and their uncertainty are of utmost importance in the design of coastal defence structures such as seawalls and breakwaters (Figure 1.1.1). Figure 1.1.2 shows a picture of a large wave striking the breakwater at Kalk Bay Harbour, Cape Town, South Africa. Typically, coastal structures are designed to provide sufficient protection against flooding or erosion to a desired return level associated with a particular return period, for example 100 years (Thompson et al. (2009)). Under- or overdesign can lead to very costly consequences and even fatalities. Proper statistical analyses of measured wave data of a time series are therefore required for these estimations.

In this study, wave data, measured by a Datawell Waverider buoy (Figures 1.1.3 and 1.1.4), is analysed. The spherical buoy floats at the sea surface and undergoes upward and downward movement along with the waves. The vertical acceleration of the buoy is measured with an onboard accelerometer (supplemented with an artificial horizon to define the vertical) and integrated twice in order to obtain the vertical displacement as a function of time. The small horizontal movements of the buoy are ignored (Holthuijsen (2007)).

The Datawell Waverider buoy is situated at Slangkop, off the Cape Peninsula (Figure 1.1.5). Three-hourly measurements of ten wave parameters were provided by the CSIR. The only parameter of interest for this study is the significant wave height (H_{mo}), in metres.

The height of a wave is defined as the vertical distance between the crest and the

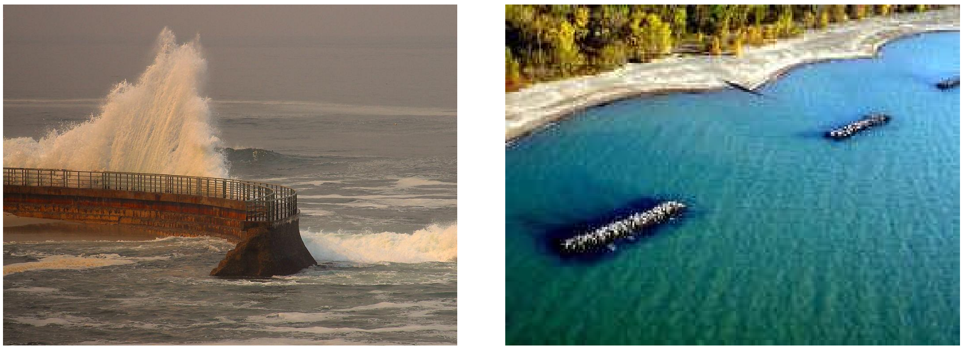


Figure 1.1.1: Examples of coastal defence structures – a seawall on the left and breakwaters on the right. [Source: http://www.studyblue.com/notes/note/n/coastal-protection-measures-hard-engineering/deck/269414]

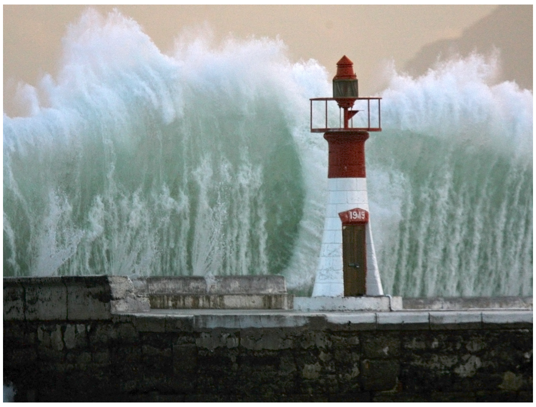


Figure 1.1.2: Photo of a large wave slamming against the breakwater at Kalk Bay Harbour. [Source: http://hqworld.net/gallery/details.php?image_id=23677&sessionid=f0feb15c9f8403838e9ff02ae3c3d692]

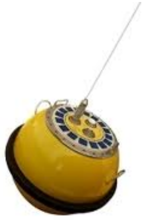


Figure 1.1.3: Datawell Waverider Mk3.

preceding trough (Snodgrass (1951)), as shown in Figure 1.1.6. The significant wave height is the average (mean) height of the highest third of all the waves (Holthuijsen (2007), Sverdrup & Munk (1946)). The significant wave height has been found by experiments to be a quantity that closely resembles the visually estimated wave height (Holthuijsen (2007)).

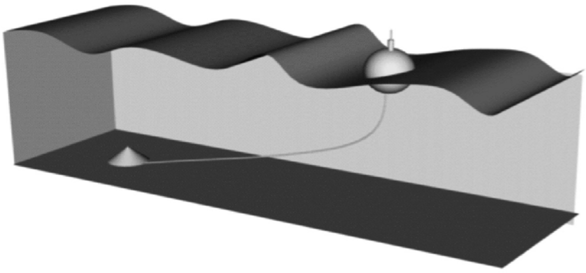


Figure 1.1.4: The Waverider buoy at sea (Holthuijsen (2007)).

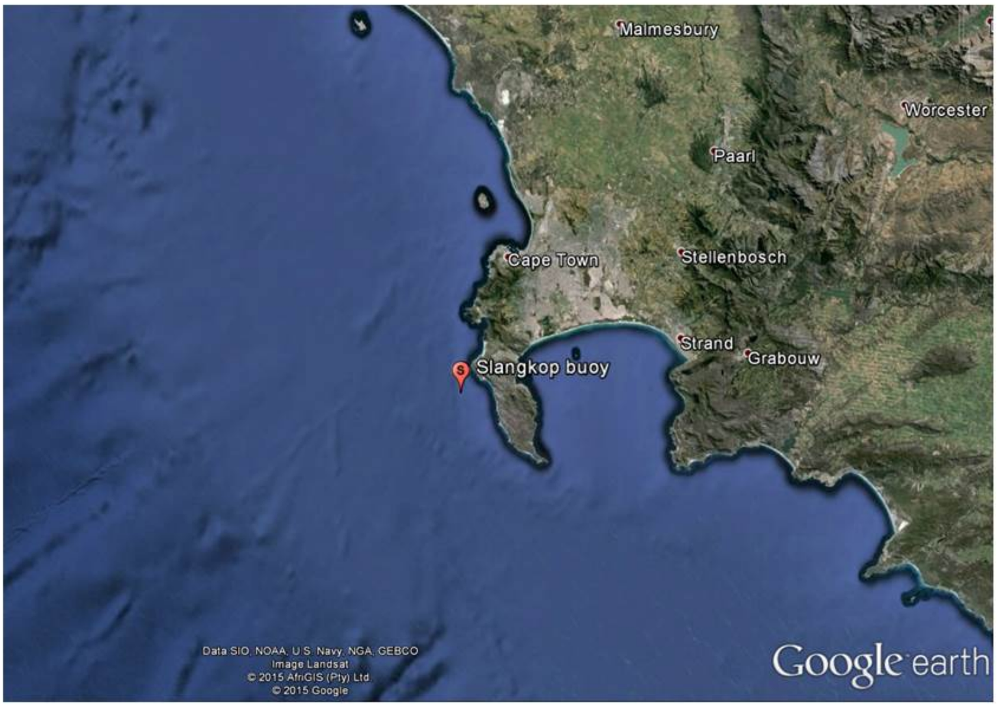


Figure 1.1.5: The position of the Datawell Waverider buoy at Slangkop, off the Cape Peninsula. [Source: Google earth]

Along with the estimation of return levels and their uncertainty, another problem, often encountered in data analyses, is that of missing data values. In the case of this study, for example, the dataset (provided by the CSIR, measured by a Datawell bouy) has gaps (i.e., no measurements) at certain points in time. The effect of such gaps on estimations have to be analysed and was done qualitatively in this study.

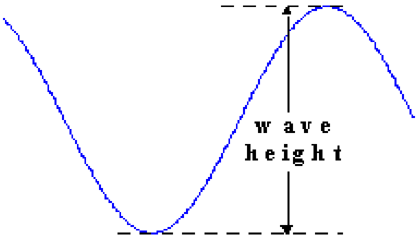


Figure 1.1.6: Wave height is measured from trough to crest. [Source: `http://web.vims.edu/physical/research/TCTutorial/longwaves.htm`]

1.2 Extreme Value Theory

Extreme value theory provides a framework that enables extrapolation in order to estimate the probability of events that are more extreme than any that have already been observed (Coles (2001)). In other words, it is used to determine limiting distributions (Thompson et al. (2009)) by estimating statistical models that best fit the extreme values of the observed data.

In statistics there are two main methods for defining extremes. The first is the block maxima approach. For this approach, the time period covered by the data set is divided into blocks, with the most extreme value in each block being used for future analyses (e.g. daily or monthly maxima). The second method is based on exceedances over a chosen threshold (u), and is called the Points Over Threshold (POT) method. These methods are illustrated in Figure 1.2.1. It is customary to use the abbreviation POT for *Peaks* Over Threshold, but in this study it refers to *Points* Over Threshold.

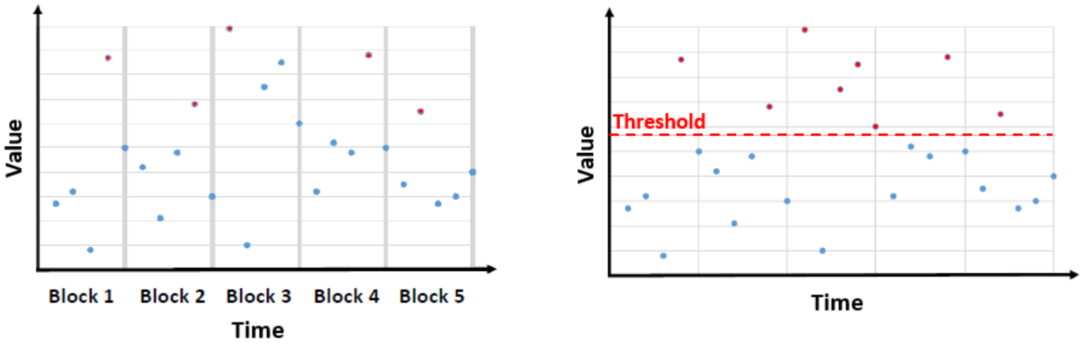


Figure 1.2.1: The block maxima approach is illustrated in the graph on the left and the POT approach is illustrated in the graph on the right. The red dots indicate the values to be used for analysis.

A disadvantage of the block maxima approach is that it only considers a single maximum within each block. This creates the riks of omitting significant data. For example, if two large values occur in one block, only the larger of the two is taken into account. The POT method, on the other hand, eliminates this risk by taking all values greater than the selected threshold into consideration. Only the POT method will be implemented in this study.

Extreme value analysis can be applied in a variety of fields of which the coastal engineering environment is but one example. Other fields include finance, traffic prediction, geological engineering and earth sciences. For instance, an example of its application in the field of earth sciences is in the study of ozone in Houston, Texas, by Smith (1989). Aspects such as the frequency with which specified high levels of ozone are exceeded are estimated. Gilli & Këllezi (2006) illustrates a possible use of extreme value analysis in finance by dealing with the behaviour of the tails of financial series for the assessment of tail related risk.

The method(s) determined to be optimal for the extreme value analysis of the wave data in this study may therefore possibly also be applied in other fields. The emphasis of this study is therefore not on the specific values of the determined estimations (for example the return levels or return periods), but rather on the *efficiency of the methods* used in order to obtain those estimations. It is important to note that this study focusses mainly on the statistical and mathematical aspects of extreme values, rather than on the physical coastal engineering aspects.

1.3 Objectives

The first main objective of this study is to explore different methods for predicting the highest wave (return level) one can expect to observe in a certain time period (return period). Waves with a 100-year return period are of particular interest, since the one-in-one-hundred-year wave is the one often used by engineers in the design of coastal structures.

In this study, focus is placed on waves with return periods of 5, 10, 18 (since the Slangkop dataset covers a time period of approximately 18 years), 50, and 100 years, although the available dataset only covers a period of approximately 18 years. Extreme value theory is therefore used to estimate wave return levels over the next 100 years given only this 18-year history.

The second objective of this study is to investigate the effect of gaps (i.e., missing data values) in a dataset on estimations. Different spreads as well as different gap sizes, along with their effects on statistical estimations, are considered. Conclusions are drawn regarding the reliability of estimations made based on incomplete datasets. A measure of an ‘acceptable’ gap size and spread is also determined.

1.4 Layout of Thesis

Firstly, in Chapter 2, some statistical theory to be used in later chapters of the study is introduced. A number of statistical distributions and parameter inference methods are discussed. Chapter 3 covers the points over threshold method. Section 3.1.1 considers techniques for threshold selection and in Section 3.1.2 the statsitical distributions and parameter estimation methods (introduced in Chapter 2) are implemented in order to estimate return levels by the points over threshold method.

Chapter 3.3 discusses return level estimation from a different angle, by using a least squares approximation for the estimation of the probability distribution parameters. A basic linear least squares problem is firstly considered in Section 3.3.1, whereafter the least squares method is applied to the Slangkop wave data in order to estimate parameters of the probability distributions presented in Section 3.3.2.

In Chapter 4, the fits of the generalized Pareto distribution (GPD) and Weibull distribution to the H_{mo}-data are evaluated by means of a goodness-of-fit statistics, namely the χ^2-test. Chapter 4.2 discusses techniques for determining the uncertainty of quantile estimates. These techniques include the Bootstrap technique (Section 4.2.1), the Monte Carlo simulation technique (Section 4.2.4), and the Jack-knife resampling technique (Section 4.2.7). These techniques are applied to the GPD and Weibull distribution along with three different methods for parameter estimation, namely the method of maximum likelihood, the method of moments, and the method of L-moments. In Chapter 5, the fits of the probability distributions, used along with different parameter estimation methods, to the empirical data are evaluated visually by means of frequency and probability plots.

Chapter 6 explores the presence of 'gaps' (i.e., missing data values) in a dataset. It considers the effects that gaps of different sizes and spreads have on estimations made based on an incomplete dataset.

In Chapter 7, findings from earlier chapters are implemented in order to analyse the Slangkop data optimally. Analyses on the dataset are done based on three different methods, namely the points over threshold, peaks over threshold, and block maxima methods, respectively. Estimations made based on these methods are compared.

Chapter 8 draws some overall conclusions on the study and presents some recommendations based on certain findings resulting from this study.

Chapter 2

Statistical Theory

This chapter introduces some of the statistical theory to be used in later chapters.

2.1 Return Level and Return Period

The return level and return period are two very important concepts in this study and it is therefore necessary to define them clearly. If the **return period**, **T**, of a wave is measured in years, the **return level**, **X**, is the threshold that is exceeded in one year with probability $\frac{1}{T}$. This is equivalent to saying that the return level X is exceeded on average once in T years. For example, a wave of height 8 m has a return period of 5 years if and only if the probability of observing a wave higher than 8 m in a year is $\frac{1}{5}$ (Coles (2001)).

For calculation purposes, when the points over threshold (POT) method is used, the quantity T is not expressed in years, but is calculated as:

$$T = \underbrace{\underbrace{\frac{\text{number of observed exceedances over the threshold}}{\text{length of the dataset in years}}}_{\text{average number of POTs per year}} \times y}_{\text{expected number of POTs per } y\text{-years}} . \tag{2.1.1}$$

It follows that T is then the number of POTs between the occurence of two, consecutive waves, both with a return period of y-years. Hence,

$$\frac{1}{T} = \text{the probabilty of observing a wave with a return period of } y\text{-years in one year.}$$

If $F(x)$ is a cumulative distribution function (CDF) of a specified probability distribution and $F(x) = 1 - \frac{1}{T}$, then $F(x)$ is the probability of observing any wave with a height less than or equal to x in a year. More generally, a CDF of a random variable X, denoted by $F(x)$, is the probability that X takes on a value less than or equal to x (Wackerly et al. (2008)). In other words, it is the cumulative sum (or integral in the case where the values that X can take on are given by a continuous interval) of the probabilities associated with all possible values of X which are less than or equal to x. In the next section, probability distributions to be used for predictions and estimations regarding wave return levels and return periods are considered.

2.2 Probability Distributions

In order to extrapolate the available data and to make estimations regarding events, more extreme than any already observed, the use of suitable probability distributions are necessary. The block maxima and POT approaches have different applicable distributions associated with each of them. Even though the block maxima approach is not implemented in this study, the probability distribution to be used in conjunction with this approach is still considered, since the probability distribution used for the POT method is derived from it.

The Fisher-Tippett theorem (Fisher & Tippett (1928)) (also known as the Extremal Types Theorem (Coles (2001))) provides an answer to the question of which probability distributions can be used in conjunction with the block maxima approach. This theorem states that the maxima of sequences of independent and identically distributed random variables (i.e., they have a common distribution function) can, after normalization, only converge to one of the three members of the Generalized Extreme Value (GEV) family. More detail on this theorem is given in Appendix A.1. The probability distributions to be used in conjunction with the POT approach are the Generalized Pareto Distribution (GPD) (derived from the GEV distributions) as well as the Weibull distribution. All of these probability distributions will be elaborated on in the next subsections.

2.2.1 Generalized Extreme Value Distributions

The three members of the generalized extreme value (GEV) family are the Gumbel, Fréchet and Weibull distributions (Coles (2001)). The CDF of the GEV family is given by

$$F_{\text{GEV}}(x;\phi,\alpha,\xi)=\begin{cases} e^{-\left[1+\xi\left(\frac{x-\phi}{\alpha}\right)\right]^{-\frac{1}{\xi}}}, & \xi\neq 0,\\ e^{-e^{\left(-\frac{x-\phi}{\alpha}\right)}}, & \xi=0,\end{cases}\tag{2.2.1}$$

where $\phi\in\mathbb{R}$ is the location parameter, $\alpha>0$ the scale parameter, and $\xi\in\mathbb{R}$ the shape parameter. When $\xi\neq 0$, the restriction $1+\frac{\xi(x-\phi)}{\alpha}>0$ holds. When $\xi=0$ there is no restriction on x. The three members of the generalized extreme value family are distinguished by the tail behaviour of the distributions. This tail behaviour is determined by the shape parameter, ξ. When $\xi=0$ it yields the Gumbel (type I) distribution, when $\xi>0$ it yields the Fréchet (type II) distribution, and when $\xi<0$ it yields the Weibull (type III) distribution.

The probability density function (pdf) of the GEV family is given by

$$f_{\text{GEV}}(x;\phi,\alpha,\xi)=\begin{cases} \frac{1}{\alpha}\left[1+\frac{\xi(x-\phi)}{\alpha}\right]^{-\frac{1}{\xi}-1}e^{-\left[1+\frac{\xi(x-\phi)}{\alpha}\right]^{-\frac{1}{\xi}}}, & \xi\neq 0,\\ \frac{1}{\alpha}e^{-\left(\frac{x-\phi}{\alpha}\right)}e^{-e^{\left(-\frac{x-\phi}{\alpha}\right)}}, & \xi=0.\end{cases}\tag{2.2.2}$$

Again the restriction $1 + \frac{\xi(x-\phi)}{\alpha} > 0$ holds for $\xi \neq 0$.

Plots of the probability density functions of the three individual members of the GEV family are shown in Figure 2.2.1 to illustrate the influence of the different parameters on the shape of the distributions. Changing ϕ from 0 to 1 results in the distributions undergoing a shift of 1 unit to the right. In general, an increase in the value of ϕ causes the distributions to shift horizontally to the right, whereas decreasing values of ϕ cause them to shift horizontally to the left. Increasing the value of α from 1 to 2 results in the probability densities spreading out. In general, an increase in the value of α results in the spreading out of the probability densities, whereas a decrease in the value of α leads to a narrowing in the spread of the probability densities.

For the Fréchet case the restriction $1 + \frac{\xi(x-\phi)}{\alpha} > 0$ results in a lower bound, $x > \phi - \frac{\alpha}{\xi}$, while for the Weibull case it results in a upper bound, $x < \phi - \frac{\alpha}{\xi}$. This simply means that the probability density outside of these bounds are zero. Note that the Weibull distribution as a member of the GEV family differs from the ordinary Weibull distribution in that the ordinary Weibull distribution has a lower bound. The Weibull distribution as a GEV family member can therefore also be called the "Reversed Weibull" and is used to model minima (in contrast to the ordinary Weibull distribution which is used to model maxima and will be defined later).

Next, probability distributions which can be used for the POT method will be considered.

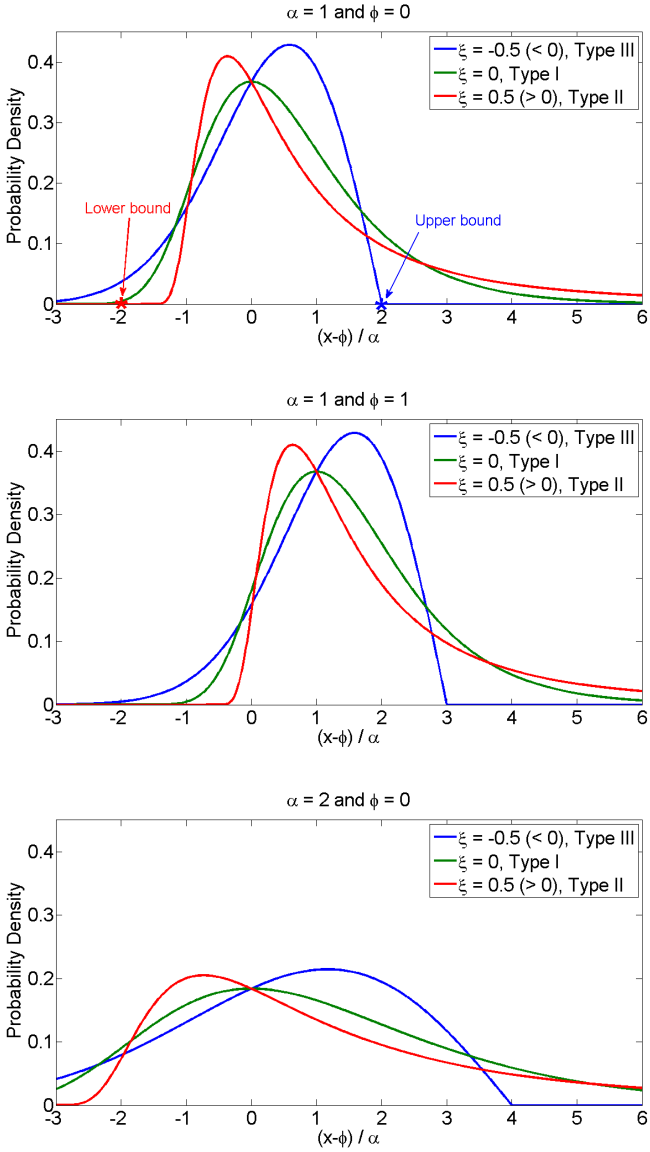
2.2.2 Generalized Pareto Distribution

The generalized Pareto distribution (GPD) is often used for modelling excesses above a sufficiently high threshold (this is based on theoretical arguments which will not be discussed in this study, however, more detailed information on this is given in Appendix A.2) (Scarrott & MacDonald (2012), Pickands III (1975), Balkema & De Haan (1974)). In other words, if the variable X represents the observed H_{mo}s, then the exceedances above a specified threshold u (i.e., $X - u$, where $X > u$) can be modelled by the GPD (e.g., Coles (2001)). Certain conditions do however have to hold for this to be true, such as that the observed values have to be independent and identically distributed (i.i.d.) (Holthuijsen (2007)). Refer to Appendix B.1 for more information on i.i.d. distributed values.

The CDF of the GPD is given by

$$F_{\text{GPD}}(x; u, \alpha_u, \xi) = \begin{cases} 1 - \left[1 + \frac{\xi(x-u)}{\alpha_u}\right]^{-\frac{1}{\xi}}, & \xi \neq 0, \\ 1 - e^{-\left(\frac{x-u}{\alpha_u}\right)}, & \xi = 0, \end{cases} \tag{2.2.3}$$

where $x > u$. When $\xi \neq 0$ the restriction $1 + \frac{\xi(x-u)}{\alpha_u} > 0$ holds and when $\xi = 0$ there is no restriction on x. The location parameter is $u \in \mathbb{R}$, $\alpha_u > 0$ is again, similarly as in the case of the GEV distribution, the scale parameter and $\xi \in \mathbb{R}$ is also again the shape parameter. The scale parameter, α_u, changes with the threshold (or location parameter), u, whereas the shape parameter, ξ, does not (Thompson et al. (2009)).


α = 1 and ϕ = 0

α = 1 and ϕ = 1

α = 2 and ϕ = 0

Figure 2.2.1: Probability density functions of the three members of the GEV family.

The pdf of the GPD is given by

$$f_{\text{GPD}}(x;u,\alpha_u,\xi)=\begin{cases}\frac{1}{\alpha_u}\left[1+\frac{\xi(x-u)}{\alpha_u}\right]^{-\frac{1}{\xi}-1}, & \xi\neq 0,\\ e^{-\left(\frac{x-u}{\alpha_u}\right)}, & \xi=0.\end{cases}\tag{2.2.4}$$

When $\xi \neq 0$, the restriction $1+\frac{\xi(x-u)}{\alpha_u}>0$ holds again and when $\xi=0$ there is again no restriction on x.

For illustration some plots of the pdf of the GPD are shown in Figure 2.2.2 for the cases where $\xi = -0.25(< 0)$, $\xi = 0$, and $\xi = 1(> 0)$. Chages is the location parameter, u, and in the shape parameter, α_u, have the same effects on pdf as in the case of the GEV distribution. In the figures, α represents α_u.

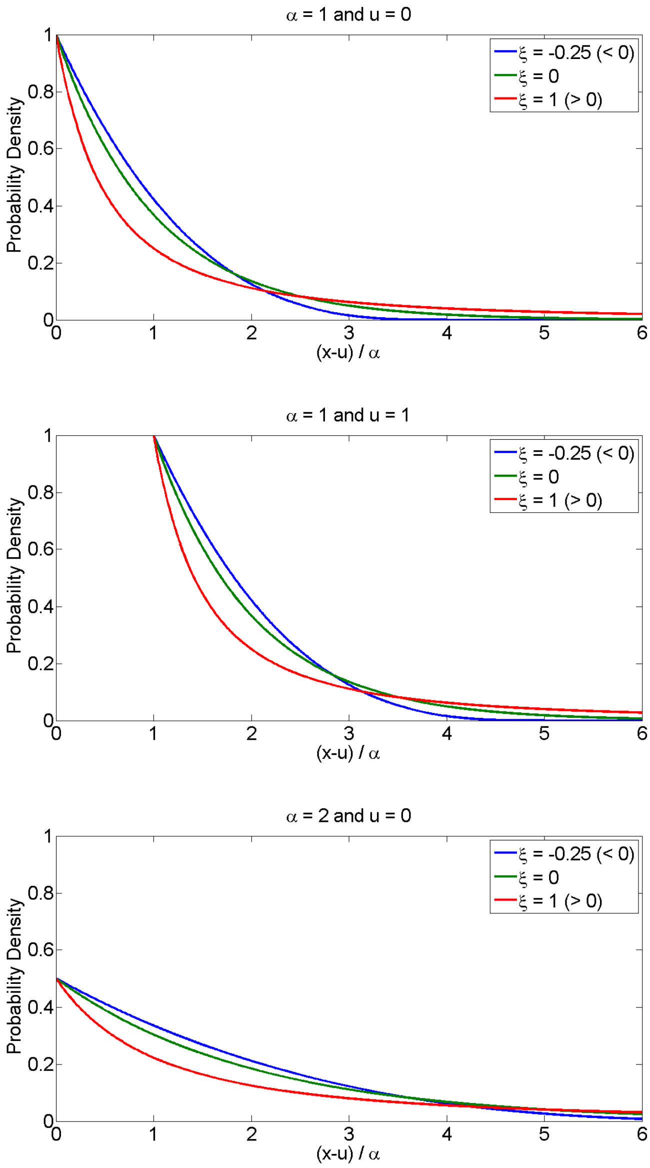


Figure 2.2.2: Probability density functions of the GPD.

In the case where $\xi = 0$ and $u = 0$, the GPD is equivalent to the exponential distribution. The case where $u = 0$ is not of interest and therefore the exponential distribution is not considered in this study. The next distribution that can be used with the POT method is the Weibull distribution.

2.2.3 Weibull Distribution

The CDF of the Weibull distribution is given by

$$F_{\mathrm{W}}(x; u, \alpha, \xi) = \begin{cases} 1 - e^{-\left(\frac{x-u}{\alpha}\right)^{\xi}}, & x \geq u, \\ 0, & x < u, \end{cases} \tag{2.2.5}$$

where $u > 0$ is the location parameter, $\alpha > 0$ is the scale parameter, and $\xi > 0$ is the shape parameter.

The Weibull pdf is given by

$$f_{\mathrm{W}}(x; u, \alpha, \xi) = \begin{cases} \frac{\xi}{\alpha}\left(\frac{x-u}{\alpha}\right)^{\xi-1} e^{-\left(\frac{x-u}{\alpha}\right)^{\xi}}, & x \geq u, \\ 0, & x < u. \end{cases} \tag{2.2.6}$$

The pdf is very sensitive to changes in the value of the shape parameter, ξ. Figure 2.2.3 shows that the forms of the probability density functions differ quite significantly for different values of ξ.

It is also once again clear that an increase in the value of u (location parameter) results in the probability density functions undergoing a shift to the right and an increase in the value of α (shape parameter) results in a flattening of the shape of the probability density functions which is equivalent to a spreading out of the probabilities.

2.3 Inference on Parameters

All of the above probability distributions contain two or three unknown parameters. These parameters have to be determined by means of statistical inference (Coles (2001)). This means that the values of the parameters have to be estimated by making conclusions from the sample, i.e., the observed H_{mo}-values. Methods for doing so include the Method of Maximum Likelihood, the Method of Moments, and the Method of L-Moments. These methods are described in the following three subsections.

2.3.1 Method of Maximum Likelihood

According to the literature, the method of maximum likelihood is a very popular and widely used technique for deriving estimators (Casella & Berger (2002)). It maximizes the probability of obtaining the observed sample by selecting as estimates the values of the parameters that maximize the likelihood (the joint density function) of the observed sample (Wackerly et al. (2008)). Its characteristics can be examined mathematically (Corbella & Stretch (2012)).

The likelihood (or joint density) of a sample is the product of the values of the specified pdf at each of the sample observations. In other words, if $y_1, y_2, \ldots, y_n$ are sample observations and $f(y; \theta)$ is the specified pdf (with parameter(s) θ), then the likelihood of the sample is given by (Wackerly et al. (2008))

$$L(y_1, y_2, \ldots, y_n | \theta) = \prod_{i=1}^{n} f(y_i; \theta). \tag{2.3.1}$$

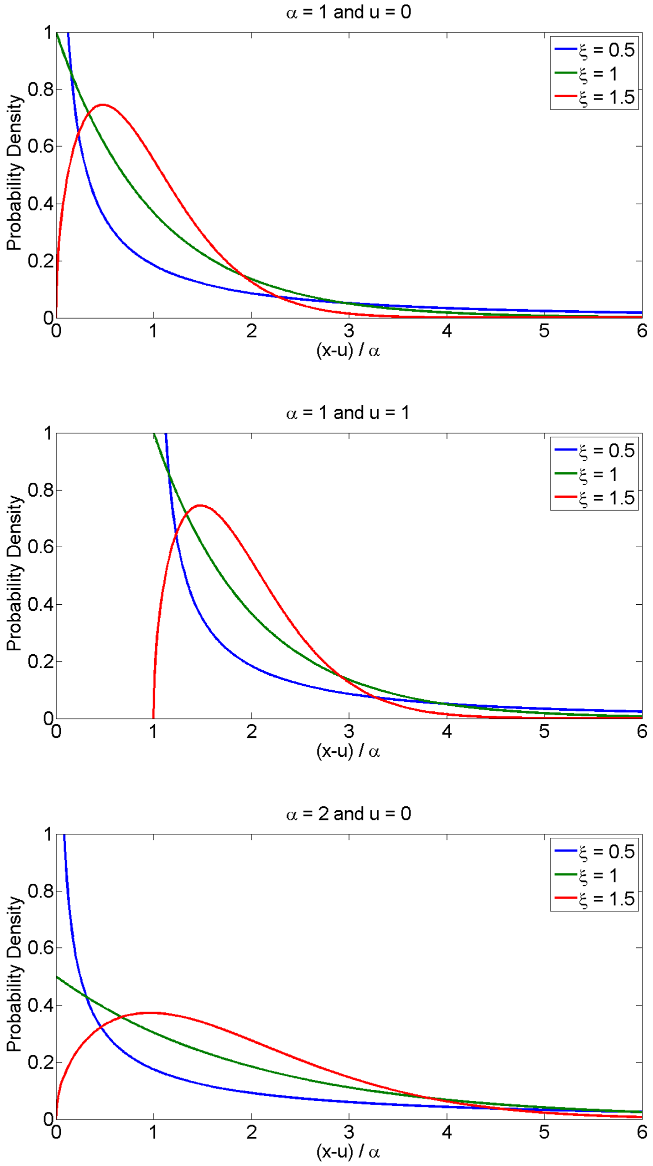


Figure 2.2.3: Probability density functions of the Weibull distribution.

For the sake of simplicity, the log-likelihood,

$$\ell(y_1, y_2, \ldots, y_n|\theta) = \log L(y_1, y_2, \ldots, y_n|\theta), \tag{2.3.2}$$

is maximized instead of the likelihood function itself. This can be done since the logarithmic function is monotonic (i.e., non-decreasing) and therefore takes its maximum at the same point as the likelihood function (Coles (2001)).

If $\theta = [\theta_1, \theta_2, \ldots, \theta_k]$, the maximum likelihood estimates, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, are determined by solving the following *score (or likelihood) equations* (Martins & Stedinger

(2000))

$$\frac{\partial \ell(y_1, y_2, \ldots, y_n|\theta)}{\partial \theta_1} = 0, \quad \frac{\partial \ell(y_1, y_2, \ldots, y_n|\theta)}{\partial \theta_2} = 0, \quad \ldots, \quad \frac{\partial \ell(y_1, y_2, \ldots, y_n|\theta)}{\partial \theta_k} = 0. \tag{2.3.3}$$

According to Casella & Berger (2002) there are two drawbacks when using the method of maximum likelihood which are caused by two possible problems that can arise when trying to determine the maximum of a function. The first is verifying that the obtained maximum is indeed a global maximum and the second is that the determined estimates may be very sensitive to small changes in the sample, which makes them unreliable. Next, the Method of Moments is considered.

2.3.2 Method of Moments

The method of moments determines estimates for parameters by solving equations which relate the population moments to the parameters by means of substituting the (unknown) population moments by the (known) sample moments. This is one of the oldest methods for finding estimators, is relatively simple to use and can almost always be used (Casella & Berger (2002), Hansen (2007)).

If $\mu'_k = E[X^k]$ is the kth population moment about the origin of a random variable X, and $m'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ is the corresponding kth sample moment, μ'_k can be estimated (and therefore substituted) by m'_k. The population moments are functions of the population parameters. Therefore, if there are n population parameters (say, $\theta_1, \theta_2, \ldots, \theta_n$), n equations are required to determine estimates for each one. The following system of equations therefore has to be solved ($\mu'_k = \mu'_k(\theta_1, \theta_2, \ldots, \theta_n)$, since μ'_k is a function of the n population parameters)

$$\begin{aligned} m'_1 &= \mu'_1(\theta_1, \theta_2, \ldots, \theta_n) \\ m'_2 &= \mu'_2(\theta_1, \theta_2, \ldots, \theta_n) \\ &\vdots \\ m'_n &= \mu'_n(\theta_1, \theta_2, \ldots, \theta_n). \end{aligned} \tag{2.3.4}$$

A short example follows to illustrate this method: A Bernoulli random variable, X, can take on one of two possible values, 0 (representing a failure) or 1 (representing a success), with p the probability of a success and $1-p$ the probability of a failure. In other words, X has the probability mass function

$$\begin{aligned} p(0) &= P[X = 0] = 1 - p \\ p(1) &= P[X = 1] = p, \end{aligned} \tag{2.3.5}$$

where $0 \leq p \leq 1$ (Ross (2010)). Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli population with parameter p. For the Bernoulli random variable, $\mu'_1 = E[X] = p$, and

m_1' is used to estimate p. That is,

$$m_1' = \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{2.3.6}$$

Consider, for example, the random sample $0, 1, 1, 0, 1$ from a Bernoulli population with parameter $p = \frac{1}{2}$. From equation (2.3.6) it follows that

$$m_1' = \hat{p} = \frac{1}{5}\left(0 + 1 + 1 + 0 + 1\right) = \frac{3}{5}.$$

The value $\hat{p} = \frac{3}{5}$ will therefore be used as an estimation for p based on the sample values. The last method to be considered for the esimation of parameters is the Method of L-Moments.

2.3.3 Method of L-Moments

The "L" in L-moments stands for "linear", since they are expected values of certain linear combinations of order statistics, called L-statistics (Hosking (1990)). According to Hosking (1990), L-moments are less subject to bias in estimation than conventional moments and they are sometimes even more accurate in small samples than maximum likelihood estimates.

If X is a random variable and $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ (i.e., $X_{k:n}$ denotes the k^{th} smallest value in a sample of size n) are the order statistics of a random sample of size n, the r^{th} population L-moment is

$$\lambda_r = r^{-1}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} E\left[X_{r-k:r}\right], \quad r = 1, 2, \ldots, \tag{2.3.7}$$

where E is the expected value, and

$$E\left[X_{j:r}\right] = \frac{r!}{(j-1)!(r-j)!}\int X\{F(X)\}^{j-1}\{1 - F(X)\}^{r-j}\mathrm{d}F(X), \tag{2.3.8}$$

where F is the cumulative distribution function of the variable X and $X(F)$ is the quantile function (David (1981) as cited by Landwehr & Matatlas (1989)).

Equation (2.3.7) shows that λ_r is a linear function of the expected order statistics. It follows from equations (2.3.7) and (2.3.8) that the first four L-moments are (Hosking

(1990))

$$\begin{aligned}\lambda_1 &= E[X] = \int_0^1 X(F)\mathrm{d}F,\\ \lambda_2 &= \frac{1}{2}E[X_{2:2} - X_{1:2}] = \int_0^1 X(F)(2F-1)\mathrm{d}F,\\ \lambda_3 &= \frac{1}{3}E[X_{3:3} - 2X_{2:3} + X_{1:3}] = \int_0^1 X(F)(6F^2 - 6F + 1)\mathrm{d}F, \quad \text{and}\\ \lambda_4 &= \frac{1}{4}E[X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}] = \int_0^1 X(F)(20F^3 - 30F^2 + 12F - 1)\mathrm{d}F.\end{aligned} \tag{2.3.9}$$

Let

$$\beta_r = E\left\{X[F(X)]^r\right\} = \int_0^1 X(F)F^r\mathrm{d}F, \quad r = 1, 2, \ldots, \tag{2.3.10}$$

which is the probability weighted moment (PWM) of order r. Then equations (2.3.9) can be simplified to (DHI (2003))

$$\begin{aligned}\lambda_1 &= \beta_0,\\ \lambda_2 &= 2\beta_1 - \beta_0,\\ \lambda_3 &= 6\beta_2 - 6\beta_1 + \beta_0, \quad \text{and}\\ \lambda_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0.\end{aligned} \tag{2.3.11}$$

The following unbiased PWM estimators are used for the estimation of L-moments (Landwehr & Matatlas (1989) as cited by DHI (2003)):

$$\begin{aligned}\hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n} x_i,\\ \hat{\beta}_1 &= \frac{1}{n}\sum_{i=1}^{n} \frac{n-i}{n-1} x_{(i)},\\ \hat{\beta}_2 &= \frac{1}{n}\sum_{i=1}^{n} \frac{(n-i)(n-i-1)}{(n-1)(n-2)} x_{(i)}, \quad \text{and}\\ \hat{\beta}_3 &= \frac{1}{n}\sum_{i=1}^{n} \frac{(n-i)(n-i-1)(n-i-2)}{(n-1)(n-2)(n-3)} x_{(i)},\end{aligned} \tag{2.3.12}$$

where $x_{(n)} \leq x_{(n-1)} \leq \ldots \leq x_{(1)}$ is the descending ordered sample of observations. The PWM's in equation (2.3.11) ($\beta_0, \beta_1, \beta_2$ and β_3) are then replaced by the PWM estimators in equation (2.3.12) ($\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ and $\hat{\beta}_3$) in order to obtain the L-moment estimates ($\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3$ and $\hat{\lambda}_4$), i.e.,

$$\begin{aligned}
\hat{\lambda}_1 &= \hat{\beta}_0, \\
\hat{\lambda}_2 &= 2\hat{\beta}_1 - \hat{\beta}_0, \\
\hat{\lambda}_3 &= 6\hat{\beta}_2 - 6\hat{\beta}_1 + \hat{\beta}_0, \quad \text{and} \\
\hat{\lambda}_4 &= 20\hat{\beta}_3 - 30\hat{\beta}_2 + 12\hat{\beta}_1 - \hat{\beta}_0.
\end{aligned} \tag{2.3.13}$$

The three methods described above for estimating parameters will all be implemented later in this study and the fits of the probability distributions will be compared using the different methods. A less frequently used method for estimating the parameters of a probability distribution, is the Least Squares Method. This method is discussed at the end of Chapter 3.

Chapter 2 introduced statistical theory, such as the concepts of return level and return period, probability distributions to be used in later chapters, as well as methods for estimating the parameters of probability distributions. In the next chapter, the points over threshold (POT) method is discussed and implemented in order to fit different probability distributions to the H_{mo}s.

Chapter 3

Points Over Threshold Method

The POT method only considers H_{mo}s higher than a specified threshold, u. Since it takes all values greater than u into account, the risk of omitting significant data is eliminated. For independence assumptions made in this study, refer to Appendix B.1.1.

This method can be implemented in two different ways. The first is by the use of a predetermined, *fixed threshold*, u. Here, all events (values) above u are taken into account (i.e., all $x_i > u, i = 1, 2, \ldots, n$), which implies that the number of exceedances, n, becomes a random variable. The second way of implementing this method is by the use of an *estimated threshold*. In this case the n largest events $x_{(n)} \leq x_{(n-1)} \leq \ldots \leq x_{(1)}$ are extracted, which implies that the threshold level, u, becomes a random variable (DHI (2003)). Only the fixed threshold method is considered in this study.

The provided Slangkop dataset has periods where data is absent. All the analyses done in this study up to the end of Chapter 6, is done with the replacement of the absent data by zeros. In other words, where there were no H_{mo} measurements in the dataset, zeros were inserted. This is done for the sake of simplicity and since the emphasis of this study is on the methods used to make estimations, rather than the specific values of the methods themselves. More sophisticated methods for the treatment of missing data is covered in Chapter 6.

3.1 Fixed Threshold

The first aspect to consider is that of threshold selection. The methods considered for threshold selection in this study are purely statistictical. In practice, however, the statistics of wave height are concerned with their physical consequences. These physical consequences can be used to define the threshold. Possible threshold selection procedures which take these physical consequences into consideration, include: the assesment of the stability of the estimated parameters for the specified probability distribution (Northrop et al. (2015)), using a standardized least squares criterion or goodness-of-fit statistics (Tanaka & Takara (2000), Northrop et al. (2015)), minimizing the asymptotic mean-squared error of estimators of the shape parameter (Northrop et al. (2015)), etc. These procedures are not considered in this study. In the next section, threshold selection on a purely statistical basis is considered.

3.1.1 Threshold Selection

The selection of an optimal threshold requires a bias-variance trade-off. If the threshold is too low, the results are biased because of the model asymptotics being invalid. In other words, a threshold that is too low will result in the exceedances over the threshold not coverging to the chosen probability distribution, since the probability distribution is chosen based on its capability of fitting *extreme values*. On the other hand, if the threshold is too high, the variance is large due to few data points. In the specific case where the GPD is the chosen probability distribution, Teena et al. (2012) state that the threshold must be high enough for the exceedances over threshold to converge to the GPD, while the sample size should be large enough to ensure that there is enough data points left for satisfactory determination of the GPD parameters. According to Coles (2001), the standard practice when choosing a threshold, is to select the lowest threshold possible for which the limit model (i.e., the chosen probability distribution) provides a reasonable approximation for the exceedances. Teena et al. (2012), Coles (2001) and Thompson et al. (2009) all use the GPD as their probability distribution of choice for fitting exceedances over a threshold.

According to Coles (2001), there are two methods for selecting a threshold. The first is to inspect graphical aspects of the data and to make a threshold choice based on its graphical features. A disadvantage of this method is that it is very much subjective and based on the opinion of the interpreter. It also requires substantial expertise and can be very time consuming (Scarrott & MacDonald (2012)).

An example of a plot that can be used to evaluate the graphical features of the data is the mean residual life (MRL) plot. A MRL plot involves plotting the threshold, u, against the mean exceedance over u, i.e., $\frac{1}{n_u}\sum_{i=1}^{n_u}\left(x_{(i)}-u\right)$, where $x_{(1)},\ldots,x_{(n_u)}$ are the n_u observations that exceed u. Figure 3.1.1 shows a MRL plot for the H_{mo}s of the dataset used in this study.

The mean exceedances are empirical estimates of the expected values of exceedances over a threshold u, $E(X-u|X>u)$. If the GPD is a valid model for exceedances over a threshold of u_0, the expected value of the exceedance is $E(X-u_0|X>u_0)=\frac{\alpha_{u_0}}{1-\xi}$, defined for $\xi<1$ to ensure the mean exists (Scarrott & MacDonald (2012)). For a threshold $u>u_0$ the expected value becomes a linear function of u (Coles (2001)) and is given by (Scarrott & MacDonald (2012)) as

$$E(X-u|X>u)=\frac{\alpha_{u_0}+\xi u}{1-\xi}=\frac{\xi}{1-\xi}u+\frac{\alpha_{u_0}}{1-\xi}. \tag{3.1.1}$$

Since the sample means are empirical estimates of the expected threshold exceedances, the MRL plot is evaluated to determine for which threshold levels it is approximately continuously linear in u (Coles (2001)). Figure 3.1.1 shows that this is definitely not a straightforward task. For example, for threshold levels lower than approximately 5 m, the plot looks piece-wise linear, but it is unclear which straight line section should be used. In other words, no no clear indication is given as to which threshold <5 m to

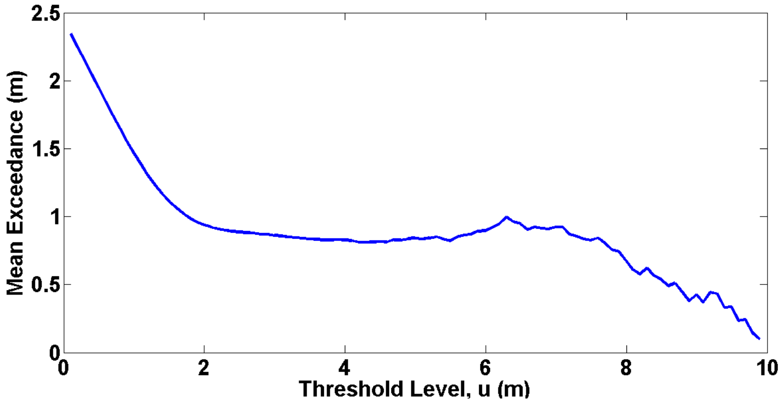


Figure 3.1.1: MRL plot for the H_{mo}s of the dataset used in this study.

select. The MRL plot will not be used for threshold selection in this study since no clear interpretation can be made from it.

The second method for selecting a threshold mentioned by Coles (2001) is to fit the GPD to a range of thresholds and to look for stability of the parameter estimates. This is based on the threshold stability property of the GPD (Scarrott & MacDonald (2012), Coles (2001)): if the GPD is a valid model for exceedances over a threshold of u_0, then for any threshold $u > u_0$, the exceedances over u also follow a GPD with the same shape (i.e., the shape parameter ξ remains unchanged) but with a shifted scale $\alpha_u = \alpha_{u_0} + \xi(u - u_0)$. In order to ensure that the scale parameter does not change with u it can be reparameterized as

$$\alpha^* = \alpha_u - \xi u. \tag{3.1.2}$$

The new scale parameter α^* is then constant with respect to u, since

$$\begin{aligned}\alpha^* &= \alpha_u - \xi u\\ &= [\alpha_{u_0} + \xi(u - u_0)] - \xi u\\ &= \alpha_{u_0} + \xi u - \xi u_0 - \xi u\\ &= \alpha_{u_0} - \xi u_0.\end{aligned}$$

The argument that forms the basis of the following technique for threshold selection is therefore that estimates of the parameters α^* and ξ should both be constant for thresholds above u_0.

Automated threshold selection technique

Thompson et al. (2009) describe a technique for the automated selection of a threshold. It is based on the above stated argument, that estimates of both α^* and ξ should be constant for thresholds $u > u_0$, if u_0 is a valid threshold for exceedances to follow the GPD. This technique plots estimates of the GPD parameters, determined by using a range of different thresholds, against the thresholds. The goal is therefore to determine

u_0, i.e., the lowest threshold for which exceedances over the threshold follows the GPD. An outline and implementation of this technique, on the data used in this study, follows.

Equally spaced, increasing candidate thresholds, $u_1 < u_2 < \ldots < u_n$, are selected, and $\hat{\alpha}_{u_j}$ and $\hat{\xi}_{u_j}$ represent the maximum likelihood estimators of the scale and shape parameters, respectively, associated with threshold $u_j, j = 1, 2, \ldots, n$. If u_0 is a valid threshold for exceedances to follow the GPD, it follows that

$$\alpha_{u_{j-1}} = \alpha_{u_0} + \xi(u_{j-1} - u_0) \text{ and } \alpha_{u_j} = \alpha_{u_0} + \xi(u_j - u_0), \tag{3.1.3}$$

and hence,

$$\alpha_{u_j} - \alpha_{u_{j-1}} = \xi(u_j - u_{j-1}). \tag{3.1.4}$$

Standard maximum likelihood theory states that the expected value of a maximum likelihood estimate, $\hat{\theta}$, is approximately equal to the actual population parameter, θ, i.e., $E[\hat{\theta}] \approx \theta$ (Coles (2001)). therefore, $E[\hat{\alpha}_{u_j}] \approx \alpha_{u_j}$ and $E[\hat{\xi}_{u_j}] \approx \xi_{u_j}$. Now, τ_{u_j} is defined as

$$\tau_{u_j} = \hat{\alpha}_{u_j} - \hat{\xi}_{u_j} u_j, \quad j = 1, \ldots, n, \tag{3.1.5}$$

and the expected values of the differences $\tau_{u_j} - \tau_{u_{j-1}}, j = 2, \ldots, n$, are considered:

$$\begin{aligned} E[\tau_{u_j} - \tau_{u_{j-1}}] &= E[\hat{\alpha}_{u_j} - \hat{\xi}_{u_j} u_j - \hat{\alpha}_{u_{j-1}} + \hat{\xi}_{u_{j-1}} u_{j-1}] \\ &\approx (\alpha_{u_j} - \alpha_{u_{j-1}}) - \xi_{u_j} u_j + \xi_{u_{j-1}} u_{j-1} \\ &= \xi(u_j - u_{j-1}) - \xi(u_j - u_{j-1}) \\ &\quad \text{(from equation (3.1.4) and since } \xi_{u_j} = \xi_{u_{j-1}} = \xi) \\ &= 0. \end{aligned} \tag{3.1.6}$$

In other words, the expected value (or mean) of the differences $\tau_{u_j} - \tau_{u_{j-1}}$ is 0 if u_j and u_{j-1} are valid thresholds for exceedances over the threshold to follow the GPD. Hence, the objective is to find the first u_j (i.e., smallest u_j) for which this is true.

Standard maximum likelihood theory also states (e.g., Coles (2001)), that a maximum likelihood estimate, $\hat{\theta}_0$, follows an approximate normal distribution. This implies that $\tau_{u_j} - \tau_{u_{j-1}}$ will approximately follow a normal distribution as well. Thompson et al. (2009) use this distributional result to suggest the following procedure for finding a suitable threshold u:

1. Select suitable, equally spaced candidate thresholds $u_1 < u_2 < \ldots < u_n$, with u_1 the median of the data and u_n the 98% quantile. If, however, fewer than 100 values exceed the 98% quantile of the data, u_n is set to the 100th largest data value (i.e., the 100th data value if the data is in descending order). Thompson et al. (2009) state that setting $n = 100$ gives good results.

2. From the theory stated above, it is known that if u is a suitable threshold, then the differences $\tau_{u_j} - \tau_{u_{j-1}}$ (for $u \leq u_{j-1} \leq u_j$) have to follow an approximate normal distribution with mean 0. If u is not suitable these differences may not follow a normal distribution. therefore, Thompson et al. (2009) suggest that in

order to determine u a normality test has to be applied. Pearson's χ^2-test (chi-square test) is used for this purpose (refer to Chapter 4, Section 4.1.1 for more detail on the χ^2-test). It is a goodness-of-fit test which is used to establish whether or not the differences $\tau_{u_j} - \tau_{u_{j-1}}$ are consistent with a normal distribution with mean 0 (Greenwood & Nikulin (1996) as cited by Thompson et al. (2009)). First, $u = u_1$ is considered and Pearson's test is performed based on the differences $\tau_{u_2} - \tau_{u_1}, \tau_{u_3} - \tau_{u_2}, \ldots, \tau_{u_n} - \tau_{u_{n-1}}$. If these differences are normally distributed with mean 0, u is taken to be a suitable threshold. Otherwise, $u = u_2$ is considered, $\tau_{u_2} - \tau_{u_1}$ is removed from the set of differences, and the normality test is performed on the set of remaining differences, $\tau_{u_3} - \tau_{u_2}, \tau_{u_4} - \tau_{u_3}, \ldots, \tau_{u_n} - \tau_{u_{n-1}}$. Thompson et al. (2009) found that a size 0.2 significance level χ^2-test works optimally.

3. Step 2 is repeated until the set of differences is found to be consistent with a normal distribution with mean 0 by the χ^2-test.

Implementing the above procedure to the data used in this study results in a threshold choice of $u = u_{44} = 3.3718$ m. In other words, the first (i.e., lowest) threshold, u, which results in the differences $\tau_{u_j} - \tau_{u_{j-1}}$ (for $u \leq u_{j-1} \leq u_j$) having an approximate normal distribution with mean 0, is the 44th threshold. This means that the differences $\tau_{u_{45}} - \tau_{u_{44}}, \tau_{u_{46}} - \tau_{u_{45}}, \ldots, \tau_{u_{100}} - \tau_{u_{99}}$ are approximatly normally distributed with mean 0.

Figure 3.1.2 is a scatter plot of the three-hourly H_{mo} measurements against time. The automated threshold selection choice of 3.3718 m is indicated by the red line in the figure.

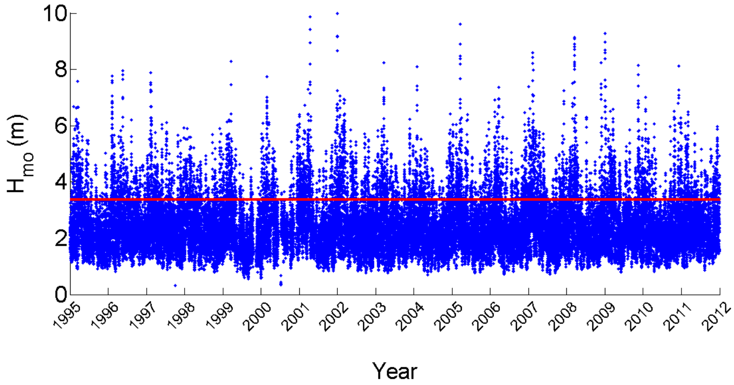


Figure 3.1.2: Scatter plot of the three-hourly H_{mo} data against time. The red line shows the automated threshold choice of $u = 3.3718$ m.

Figures 3.1.3 and 3.1.4 display histograms of $\tau_{u_j} - \tau_{u_{j-1}}$, where $j = k, \ldots, 100$ and $k = 2, \ldots, \mathbf{44}$. The hypothesis of normality was accepted when the Pearson normality test was performed on the differences $\tau_{45} - \tau_{\mathbf{44}}, \tau_{46} - \tau_{45}, \ldots, \tau_{100} - \tau_{99}$. The threshold $u = u_{44} = 3.3718$ m is displayed by the horizontal red line in Figure 3.1.5, which is a graph of the differences $\tau_{u_j} - \tau_{u_{j-1}}$ against the threshold u_{j-1} for the H_{mo}s of the wave data.

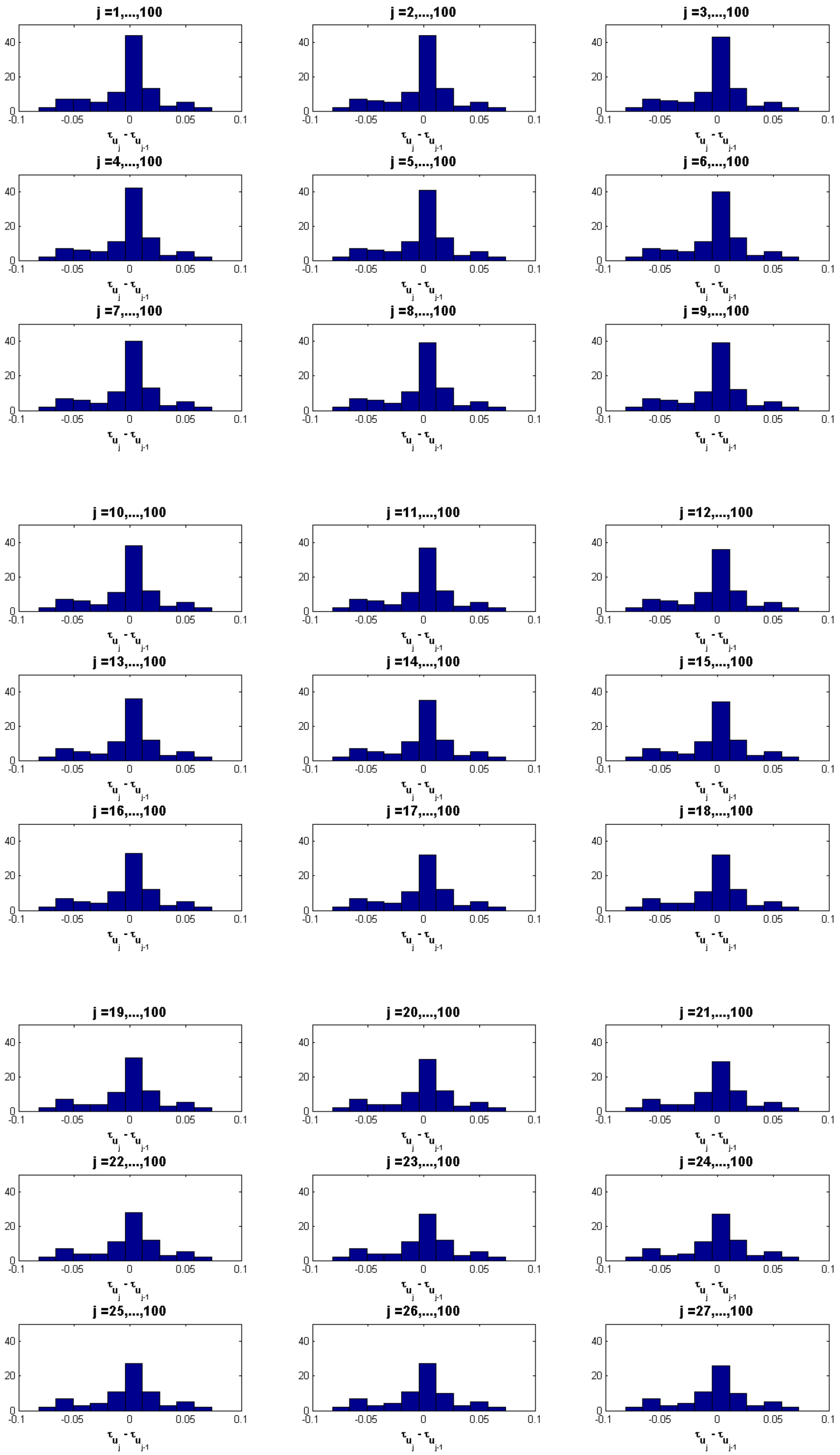


Figure 3.1.3: Histograms of $\tau_{u_j} - \tau_{u_{j-1}}$, where $j = k, \ldots, 100$ and $k = 2, \ldots, 27$. These differences should follow an approximate normal distribution with mean 0, in order for u_j to be an acceptable threshold. The vertical axes display frequencies.

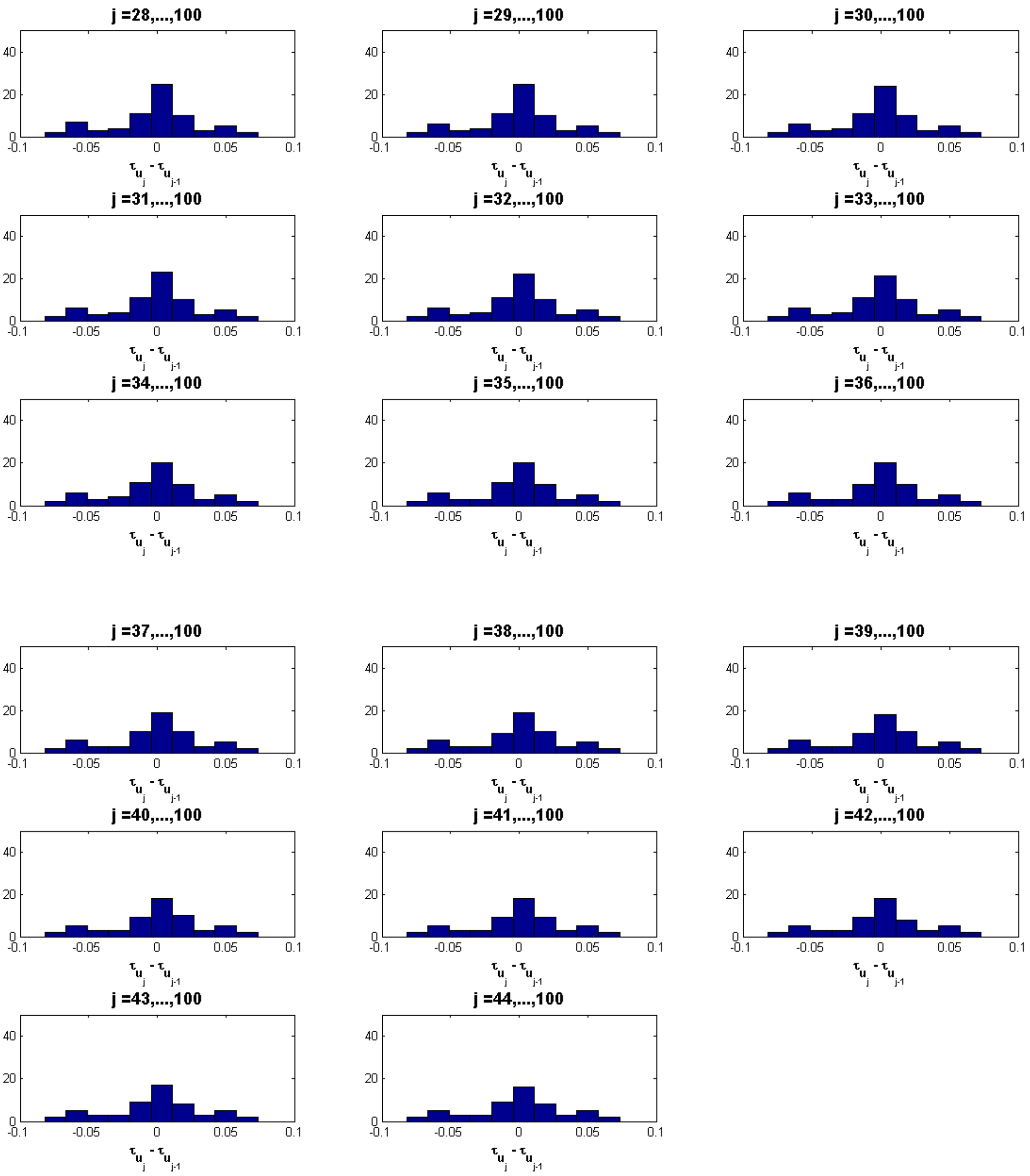


Figure 3.1.4: Histograms of $\tau_{u_j} - \tau_{u_{j-1}}$, where $j = k, \ldots, 100$ and $k = 28, \ldots, 44$. These differences should follow an approximate normal distribution with mean 0, in order for u_j to be an acceptable threshold. The vertical axes display frequencies.

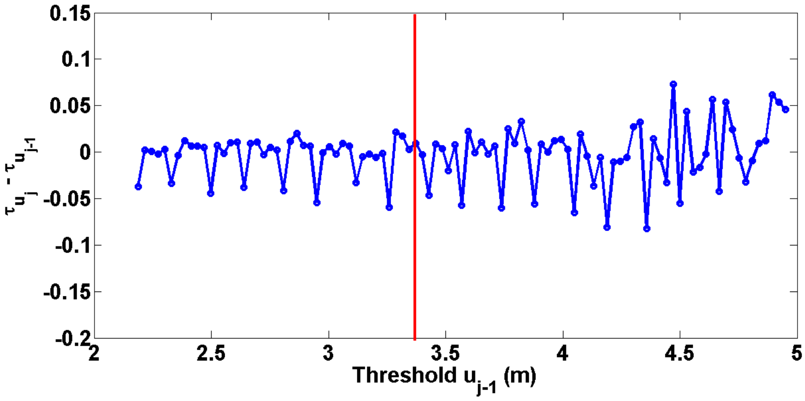


Figure 3.1.5: Graph of the differences $\tau_{u_j} - \tau_{u_{j-1}}$, against the threshold u_{j-1} for the H_{mo}s of the wave data used in this study. The red vertical line indicates the automated threshold selection choice of 3.3718 m.

Table 3.1 contains the means of $\tau_{u_j} - \tau_{u_{j-1}}$, for $i = k, \ldots, 100$. These means are plotted in Figure 3.1.6. The 10 means closest to 0 are in yellow and the numbers in brackets indicate their positions when sorted from closest to furthest from 0, with position (1) begin closest. Table 3.1 shows that the mean, when threshold u_{41} is used, is closest to 0. However, this threshold is not selected, since, by Pearson's normality test of size 0.2, the differences $\tau_{u_j} - \tau_{u_{j-1}}$ only become sufficiently consistent with a normal distribution with mean 0 when $k = 44$. Hence, $u = u_{44} = 3.3718$ m is the automated threshold selection choice.

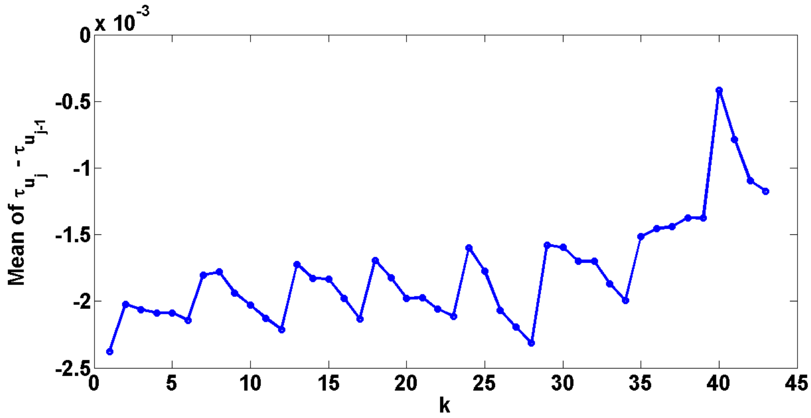


Figure 3.1.6: Plot of the means of $\tau_{u_j} - \tau_{u_{j-1}}$ (for $i = k, \ldots, 100$).

There is, however, a level of uncertainty associated with the selection of a threshold. This will be discussed next.

Table 3.1: Table of the means of $\tau_{u_j} - \tau_{u_{j-1}}$ (for $i = k, \ldots, 100$), with the 10 means closest to 0 in yellow and the numbers in brackets indicating their position when ordered from closest to furthest from 0, with position (1) being closest to 0. These means are plotted in Figure 3.1.6.

k	**Mean of $\tau_{\mathbf{u_j}} - \tau_{\mathbf{u_{j-1}}}$ (for $\mathbf{i = k, \ldots, 100}$)**
2	-0.002375485179321
3	-0.002021296855657
4	-0.002061184538372
5	-0.002087468533050
6	-0.002086703554676
7	-0.002142386567552
8	-0.001803433833953
9	-0.001781117462444
10	-0.001936730737749
11	-0.002030517202536
12	-0.002126736507383
13	-0.002211614318762
14	-0.001723686328898
15	-0.001827209373328
16	-0.001832892146729
17	-0.001977767084531
18	-0.002132608696762
19	-0.001693233601028
20	-0.001826317504252
21	-0.001980894467109
22	-0.001971906156508
23	-0.002058657059279
24	-0.002112787964101
25	-0.001598133411084
26	-0.001772575309779
27	-0.002065963537005
28	-0.002195123008166
29	-0.002310870174473
30	-0.001578575577248 **(10)**
31	-0.001596317512668
32	-0.001700875995531
33	-0.001699891507797
34	-0.001867936200428
35	-0.001994611582543
36	-0.001514875237410 **(9)**
37	-0.001455742436363 **(8)**
38	-0.001440142856207 **(7)**
39	-0.001373285966668 **(5)**
40	-0.001376626797391 **(6)**
41	-0.000415137311517 **(1)**
42	-0.000783755959896 **(2)**
43	-0.001096430832422 **(3)**
44	-0.001172914381230 **(4)**

Return Level Uncertainty

The uncertainty in the choice of a threshold has to be taken into account. Thompson et al. (2009) use the following procedure (of which Mooney & Duval (1993) and Efron & Tibshirani (1993) provide a basic summary), based on the Bootstrap procedure (which is discussed in further detail in Chapter 4, Section 4.2), to assess return level uncertainty:

1. Set $b = 1$.

2. A sample of size m (i.e., $y_1, y_2, \ldots, y_m$) is drawn randomly, with replacement, from the original dataset. Such a sample is called a *Bootstrap sample.*

3. The quantity of interest, in this case a specific return level, is calculated for the bootstrap sample and denoted by $\hat{\theta}^*_b$. The return level is calculated as follows:

   (i) A threshold is selected by using the automated threshold selection technique described previously.

   (ii) The selected threshold is then used to estimate the GPD model (i.e., to determine the other two unknown parameters in the distribution, α and ξ, by means of the method of maximum likelihood).

   (iii) Finally, the GPD parameter estimates are then used to calculate the estimated return level.

4. Next, b is increased by 1 and steps (ii) and (iii) are repeated B times, where B is a large number. $B = 1000$ is used here.

5. Finally, a probability distribution is constructed by attaching a probability $\frac{1}{B}$ to each return value estimate $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$.

Thompson et al. (2009) uses a Bootstrap percentile to determine the level of uncertainty associated with a specific return level and this will be described briefly next. However, as stated previously, the Bootstrap technique is considered in greater detail in Section 4.2.

The values $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$ are sorted in ascending order. To obtain a 95% confidence level, the $(\frac{1-0.95}{2}B)^{\text{th}}$ and $(1 - \frac{1-0.95}{2}B)^{\text{th}}$ values are selected as the confidence interval bounds. If these values are not integers, the integer below and above are used, respectively.

Figure 3.1.7 shows a plot of the histogram of the 100-year return levels after a 1000 Bootstrap iterations. The bounds of the 95% Bootstrap percentile interval are shown, with the lower bound at 10.14 m and the upper bound at 15.17 m, as well as the return level based on the original data which is 11.69 m. The 100-year return level based on the actual data does therefore fall in the 95% Bootstrap interval.

Next, the different probability distributions will be used to estimate H_{mo} return levels for a fixed threshold by using all three different paramater estimation methods.

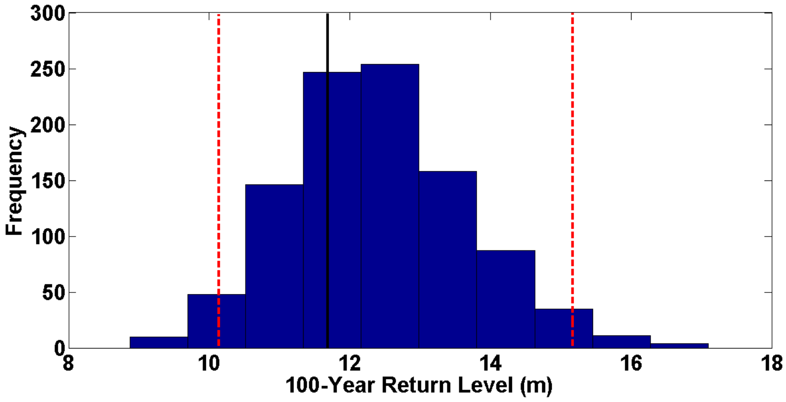


Figure 3.1.7: Histogram of the 100-year return levels and the associated 95% bootstrap percentile interval for 1000 Bootstrap iterations. The red dashed lines indicate the bounds of the percentile interval. The solid black line indicates the return level based on the original data.

3.1.2 Estimated Return Levels for a Fixed Threshold

In order to determine return levels associated with specific return periods for the various probability distributions, the following equation will be rearranged to make x_p the subject of the equation:

$$p = F(x_p), \tag{3.1.7}$$

where F represents the cumulative distribution function of the specified probability distribution and x_p denotes the return level corresponding to the non-exceedance probability p. In other words, the probability for an H_{mo} to be less than the level x_p, is p.

For consistency, the threshold $u = 3.3718$ m, as chosen by the automated threshold selection technique described earlier, will be used throughout. The GPD is considered first.

Return Levels Estimated by the GPD

For the GPD, equation (3.1.7) is as follows:

$$p = F_{GPD}(x_p) = 1 - \left[1 + \frac{\xi(x_p - u)}{\alpha}\right]^{-\frac{1}{\xi}}, \tag{3.1.8}$$

which is rearranged in order to make x_p the subject, yielding

$$x_p = u + \frac{\alpha}{\xi}\left[(1-p)^{-\xi} - 1\right]. \tag{3.1.9}$$

Table 3.2 displays the estimates of the GPD parameters, namely, ξ (the shape parameter) and α (the scale parameter), as determined by the method of maximum likelihood,

Table 3.2: Parameter estimates of ξ (shape parameter) and α (scale parameter) as well as the 100-year return level ($X(100)$) for each of the three different parameter estimation methods for the GPD.

Parameter Estimation Method	**Parameter Estimates**	**X(100)**
Maximum Likelihood	$\hat{\xi} = -0.0180$, and $\hat{\alpha} = 0.8628$	11.6881 m
Moments	$\hat{\xi} = -0.0194$, and $\hat{\alpha} = 0.8312$	11.3263 m
L-Moments	$\hat{\xi} = -0.0392$, and $\hat{\alpha} = 0.8143$	10.4273 m

the method of moments, and the method of L-moments, respectively. It also displays the 100-year return level as predicted by the GPD for each of these estimation methods.

The return levels and associated non-exceedance probabilities as well as return periods, as determined by the use of equation (3.1.9), are plotted in Figures 3.1.8, 3.1.10, and 3.1.12. The method of maximum likelihood was used for parameter estimation in Figure 3.1.8, the method of moments was used in Figure 3.1.10, and the method of L-moments was used in Figure 3.1.12. The empirical data values are also plotted together with their associated non-exceedance probabilities (red dots).

Figures 3.1.9, 3.1.11, and 3.1.13 again show plots of the return levels and their associated return periods as determined by the GPD for the method of maximum likelihood, the method of moments, and the method of L-moments, respectively. These figures also show the empirical data values plotted together with their associated return periods (black dots). Note that the green lines in Figures 3.1.8 and 3.1.9 are identical, as well as the green lines in Figures 3.1.10 and 3.1.11, and lastly, also the green lines in Figures 3.1.12 and 3.1.13. The calculation of the empirical non-exceedance probabilities is described in Section 3.2.

Discussion of Return Levels Estimated by the GPD

Based on Figures 3.1.8 to 3.1.13 certain conclusions can be made on the efficiency of each of the three parameter estimation methods (method of maximum likelihood, method of moments, and method of L-moments) when used in conjunction with the GPD for making estimations regarding return levels. Overall, the estimation of the GPD parameters by the method of maximum likelihood leads to the highest estimation of return levels. The 5-, 10-, 18-, 50-, and 100-year return levels are indicated by red dotted lines in Figures 3.1.9, 3.1.11, and 3.1.13. For each of these return levels, the estimations made when the method of maximum likelihood is used, are the highest compared to when the method of moments or the method of L-moments are used.

From Figure 3.1.9 it is evident that the GPD based on parameters as estimated by the method of maximum likelihood seem to fit the empirical data (black dots) for greater return periods the best when comparing it to estimations made by the GPD

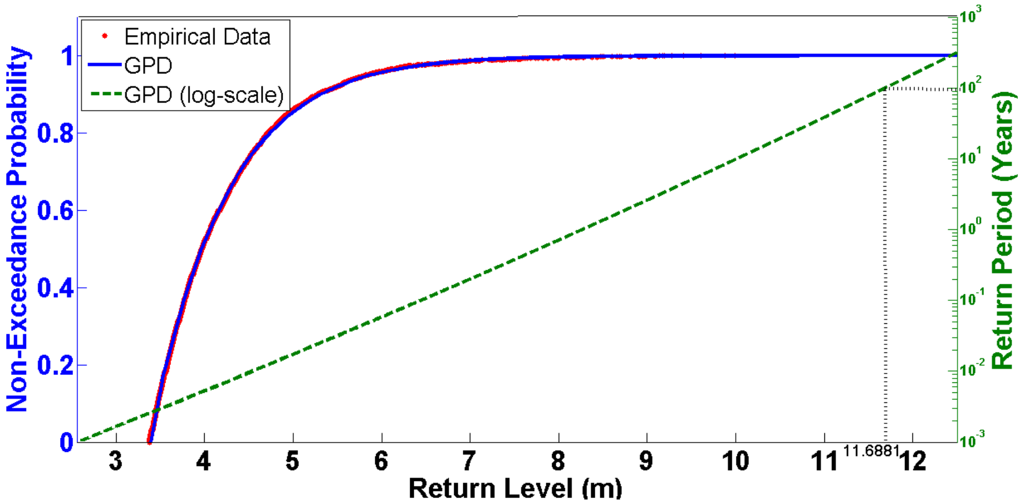


Figure 3.1.8: The return levels and associated non-exceedance probabilities as well as return periods determined by the GPD, when a threshold of $u = 3.3718$ m is used. The GPD parameters were estimated by the *method of maximum likelihood*. The 100-year return level is indicated by the black dotted line and is 11.6881 m. The green line is identical to the green line in Figure 3.1.9.

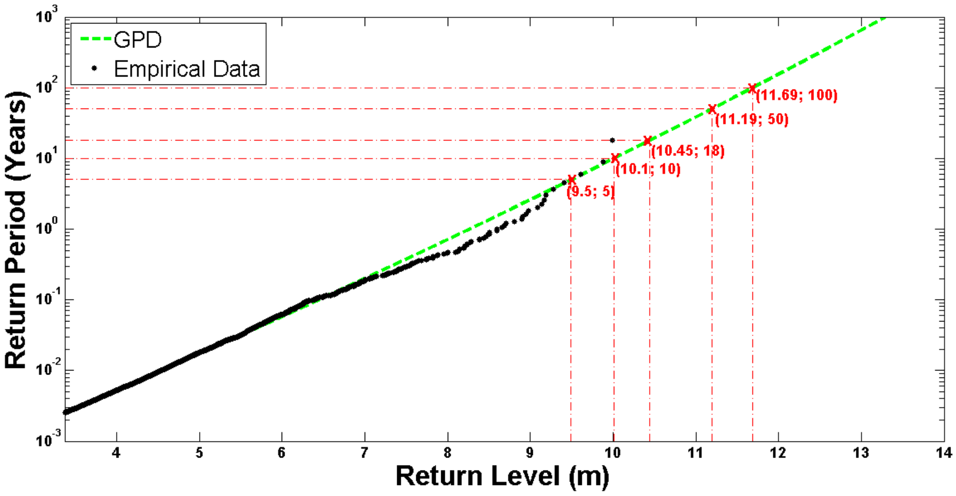


Figure 3.1.9: The return levels and associated return periods as determined by the GPD, when a threshold of $u = 3.3718$ m is used. The GPD parameters were estimated by the *method of maximum likelihood*.

when one of the other two methods for parameter estimation is used (Figures 3.1.11 and 3.1.13). This can be concluded since the empirical data seems to deviate least from the GPD estimation in Figure 3.1.9. For the return levels with greater return periods (which are the more important return periods) the method of L-moments seems to be

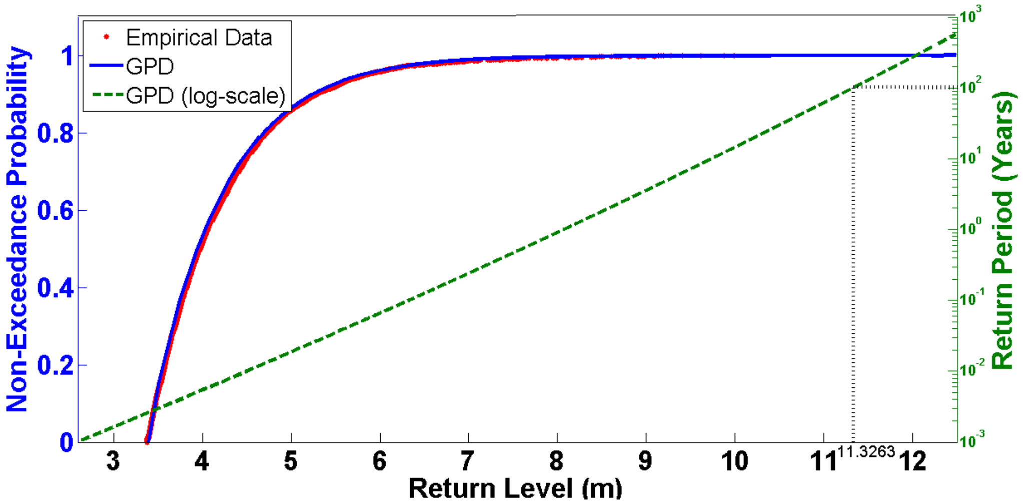


Figure 3.1.10: The return levels and associated non-exceedance probabilities as well as return periods determined by the GPD, when a threshold of $u = 3.3718$ m is used. The GPD parameters were estimated by the *method of moments*. The 100-year return level is indicated by the black dotted line and is 11.3263 m. The green line is identical to the green line in Figure 3.1.11.

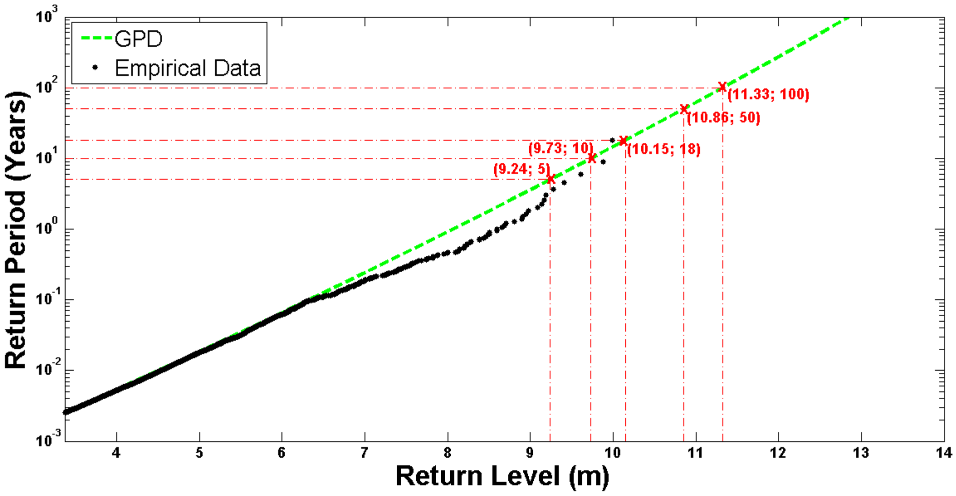


Figure 3.1.11: The return levels and associated return periods as determined by the GPD, when a threshold of $u = 3.3718$ m is used. The GPD parameters were estimated by the *method of moments*.

the least reliable method for parameter estimation, since the deviation of the GPD from the empirical data is the largest. When the method of moments or the method of L-moments are used, it seems like the probability of the under-estimation of return levels appears to be higher than when the method of maximum likelihood is used.

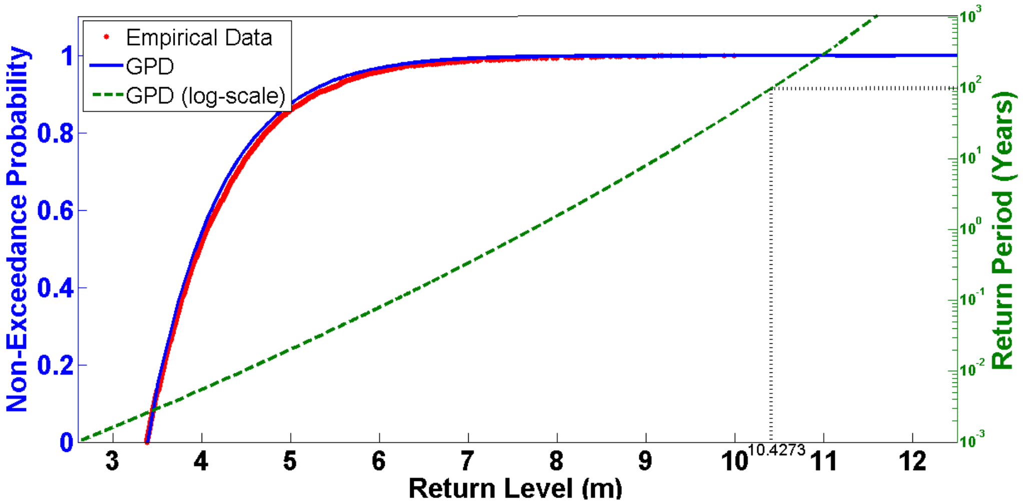


Figure 3.1.12: The return levels and associated non-exceedance probabilities as well as return periods determined by the GPD, when a threshold of $u = 3.3718$ m is used. The GPD parameters were estimated by the *method of L-moments*. The 100-year return level is indicated by the black dotted line and is 10.4273 m. The green line is identical to the green line in Figure 3.1.13.

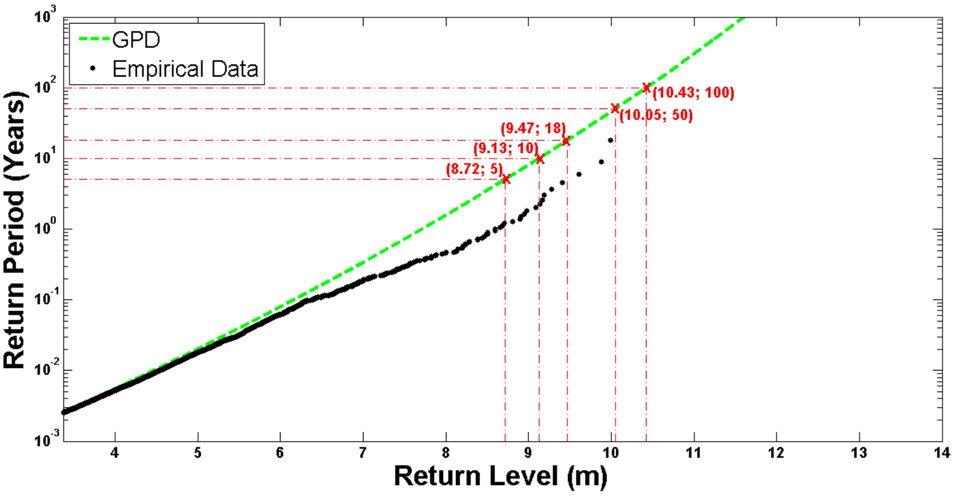


Figure 3.1.13: The return levels and associated return periods as determined by the GPD, when a threshold of $u = 3.3718$ m is used. The GPD parameters were estimated by the *method of L-moments*.

The conclusion on which parameter estimation method is the most reliable to be used in conjunction with the GPD in order to estimate return levels is the method of maximum likelihood. Next, the return levels estimated by the "ordinary" Weibull distribution will be considered.

Return Levels Estimated by the Weibull Distribution

In the case of the Weibull distribution, equation (3.1.7) yields

$$p = F_W(x_p) = 1 - e^{-\left(\frac{x_p - u}{\alpha}\right)^{\xi}}, \tag{3.1.10}$$

from which x_p is again made the subject, resulting in

$$x_p = u + \alpha \left[-\ln(1-p)\right]^{\frac{1}{\xi}}. \tag{3.1.11}$$

Table 3.3 displays the estimates of the Weibull parameters, namely, ξ (the shape parameter) and α (the scale parameter), as determined by the method of maximum likelihood, the method of moments, and the method of L-moments, respectively. It also displays the 100-year return level as predicted by the Weibull distribution for each of these estimation methods.

Table 3.3: Parameter estimates of ξ (shape parameter) and α (scale parameter) as well as the 100-year return level for each of the three different parameter estimation methods for the Weibull distribution.

Parameter Estimation Method	**Parameter Estimates**	**X(100)**
Maximum Likelihood	$\hat{\xi} = 1.0361$, and $\hat{\alpha} = 0.8600$	11.7574 m
Moments	$\hat{\xi} = 1.0192$, and $\hat{\alpha} = 0.8543$	12.0224 m
L-Moments	$\hat{\xi} = 1.0283$, and $\hat{\alpha} = 0.8573$	11.8768 m

The return levels and associated non-exceedance probabilities as well as return periods, as determined by the use of equation (3.1.11), are plotted in Figures 3.1.14, 3.1.16, and 3.1.18. The method of maximum likelihood was used for parameter estimation in Figure 3.1.14, the method of moments was used in Figure 3.1.16, and the method of L-moments was used in Figure 3.1.18. The empirical data values are also again plotted (red dots).

Figures 3.1.15, 3.1.17, and 3.1.19 again show plots of the return levels and their associated return periods as determined by the Weibull distribution for the method of maximum likelihood, the method of moments, and the method of L-moments, respectively. These figures also show the empirical data values plotted together with their associated return periods (black dots). Note that the green lines in Figures 3.1.14 and 3.1.15 are identical, as well as the green lines in Figures 3.1.16 and 3.1.17, and lastly, also the green lines in Figures 3.1.18 and 3.1.19.

Discussion of Return Levels Estimated by the Weibull Distribution

Based on Figures 3.1.14 to 3.1.19 certain conclusions can again be made regarding the efficiency of each of the three parameter estimation methods when used in conjunction

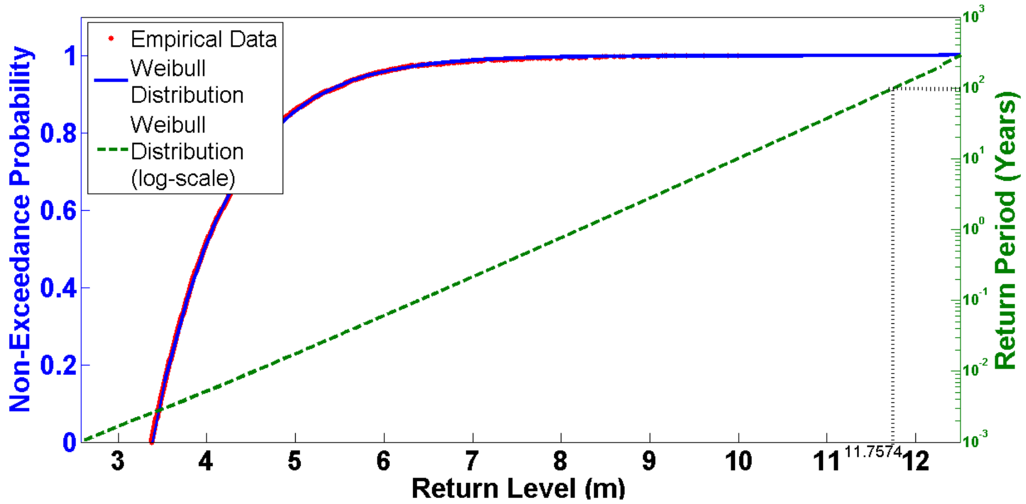


Figure 3.1.14: The return levels and associated non-exceedance probabilities as well as return periods determined by the Weibull distribution, when a threshold of $u = 3.3718$ m is used. The Weibull parameters were estimated by the *method of maximum likelihood*. The 100-year return level is indicated by the black dotted line and is 11.7574 m. The green line is identical to the green line in Figure 3.1.15.

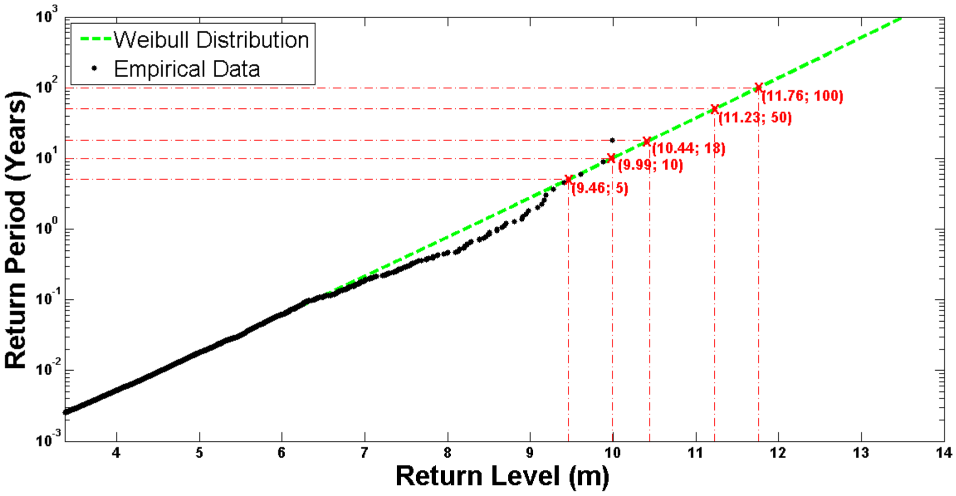


Figure 3.1.15: The return levels and associated return periods as determined by the Weibull distribution, when a threshold of $u = 3.3718$ m is used. The Weibull parameters were estimated by the *method of maximum likelihood*.

with the Weibull distribution for making estimations regarding return levels. In contrast to when the GPD is used, the estimation of the Weibull parameters by the method of maximum likelihood leads to the lowest estimation of return levels, overall, in comparison to when one of the other two parameter estimation methods are used. The 5-, 10-, 18-,

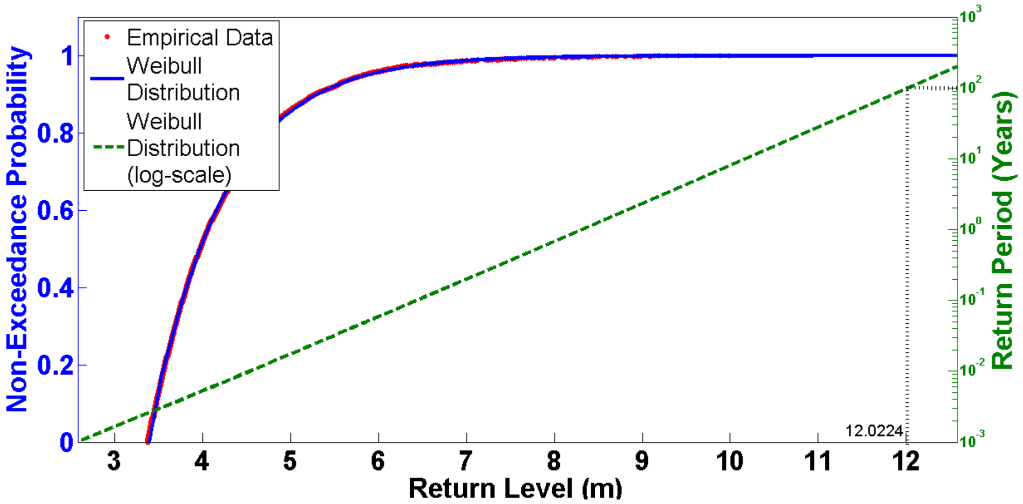


Figure 3.1.16: The return levels and associated non-exceedance probabilities as well as return periods determined by the Weibul distributionl, when a threshold of $u = 3.3718$ m is used. The Weibull parameters were estimated by the *method of moments*. The 100-year return level is indicated by the black dotted line and is 12.0224 m. The green line is identical to the green line in Figure 3.1.17.

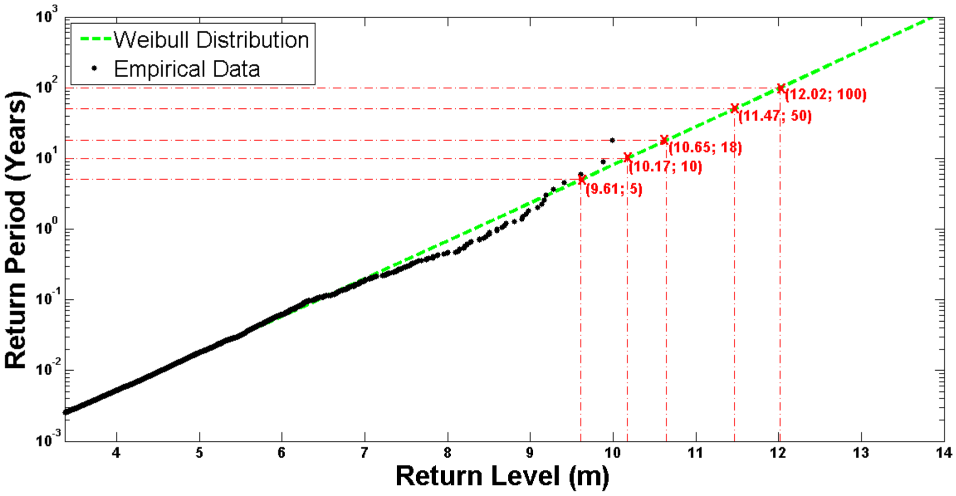


Figure 3.1.17: The return levels and associated return periods as determined by the Weibull distribution, when a threshold of $u = 3.3718$ m is used. The Weibull parameters were estimated by the *method of moments*.

50-, and 100-year return levels are indicated by the red dotted lines in Figures 3.1.15, 3.1.17, and 3.1.19. For each of these return levels, the estimations made when the method of maximum likelihood is used, are the lowest compared to when the method of moments or the method of L-moments are used.

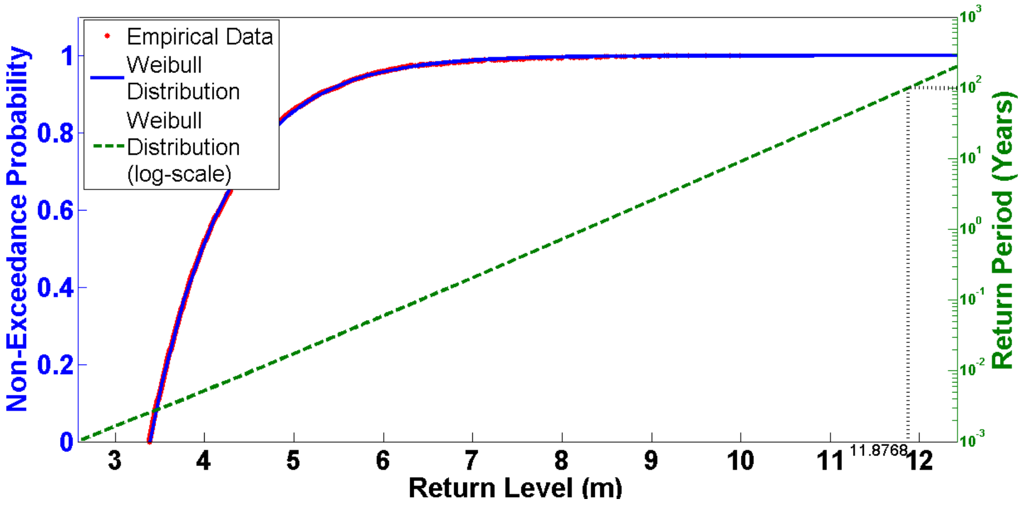


Figure 3.1.18: The return levels and associated non-exceedance probabilities as well as return periods determined by the Weibull distribution, when a threshold of $u = 3.3718$ m is used. The Weibull parameters were estimated by the *method of L-moments*. The 100-year return level is indicated by the black dotted line and is 11.8768 m. The green line is identical to the green line in Figure 3.1.19.

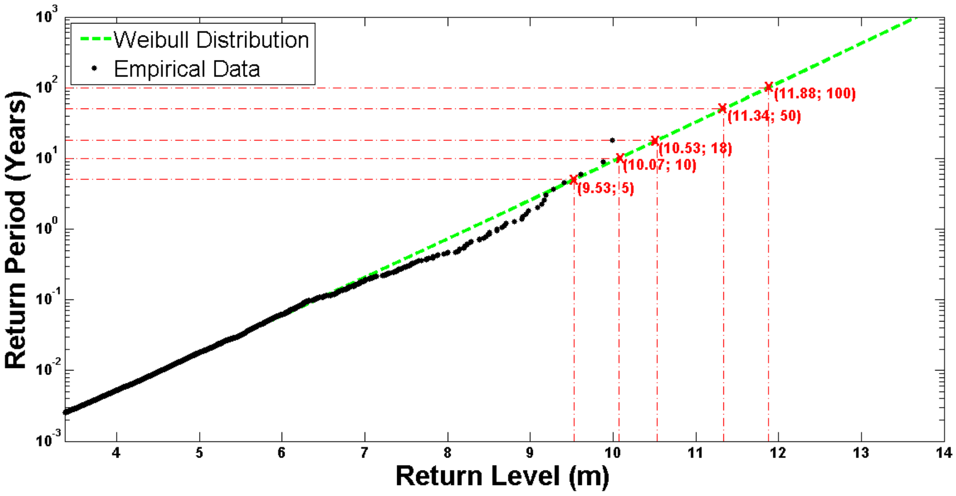


Figure 3.1.19: The return levels and associated return periods as determined by the Weibull distribution, when a threshold of $u = 3.3718$ m is used. The Weibull parameters were estimated by the *method of L-moments*.

It can be noted, however, that the return level estimations made by the use of the method of maximum likelihood in conjunction with the Weibull distribution are overall all higher than the estimations made when the method of maximum likelihood is used with the GPD. In other words, the same parameter estimation method that leads to the lowest estimation of return levels for the Weibull distribution, leads to the highest estimation of return levels for the GPD. Hence, the return levels estimated by the Weibull distribution are overall all higher than the return levels estimated by the GPD, irrespective of the parameter estimation method that is used.

Visually comparing the fit of the Weibull distribution to the empirical data in Figures 3.1.15, 3.1.17, and 3.1.19 does not lead to much insight, since little difference can be seen in the fit of the empirical data to the distribution. Conclusions that can be drawn, are that when the method of maximum likelihood is used for parameter estimation (Figure 3.1.15), the return level estimations are overall the lowest, and when the method of moments are used (Figure 3.1.19), the return level estimations are overall the highest. The estimations made by the use of the three methods do, however, only differ very slightly. This can be seen in Figure 3.1.20. Hence, the conclusion can be made that the use of all three of the parameter estimation methods in conjunction with the Weibull distribution yield very similar results.

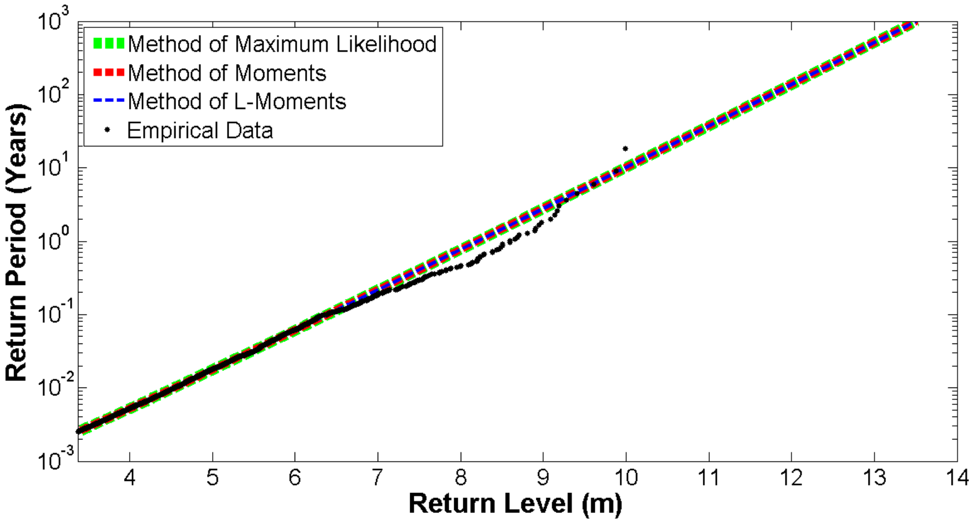


Figure 3.1.20: Comparison of the Weibull distributions when the three different methods for parameter esimation are used, namely, the method of maximum likelihood, the method of moments, and the method of L-moments.

Since return level estimations made by the Weibull distribution are all overall higher than the estimations made by the GPD distribution, and the Weibull distribution fits the empirical data well, the conclusion can be made that the use of the Weibull distribution is prefered when trying to minimize the risk of under-estimating return levels. As stated before, the method of maximum likelihood is a very popular and widely used technique for deriving parameters (Casella & Berger (2002)). It also has other advantages (refer to Section 2.3.1). Based on the fact of this method's popularity and advantages, as well as the the fact that estimations made by the three different parameter estimation methods used in conjunction with the Weibull distribution differ very slightly, it can be concluded that the use of the method of maximum likelihood together with the Weibull distribution is efficient and reliable for making estimations regarding return levels. This is the combination of parameter estimation method and probability distribution that will be used in chapters that follow. In the following section, the calculation of the empirical non-exceedance probability will be explained.

3.2 Empirical Non-Exceedance Probability

In order to determine the empirical non-exceedance probability, p (refer to equation (3.1.8)), associated with each of the observed H_{mo}s exceeding a specified threshold, the following equation is used:

$$p_{x_k} = 1 - \frac{n + 1 - \text{rank}(x_k)}{n} = \frac{\text{rank}(x_k) - 1}{n}. \tag{3.2.1}$$

Here, n is the number of observed exceedances over the chosen threshold, x_k is the observed H_{mo}, and $\text{rank}(x_k)$ is the rank of the kth value if the n exceedances are arranged in ascending order.

For example, if the ten H_{mo}s, $x_1 = 4$ m, $x_2 = 5$ m, $x_3 = 11$ m, $x_4 = 11$ m, $x_5 = 5$ m, $x_6 = 12$ m, $x_7 = 10$ m, $x_8 = 8$ m, $x_9 = 10$ m, and $x_{10} = 5$ m, were observed (exceeding a specified threshold) they can be arranged in ascending order as in the first column of Table 3.4, with their associated ranks in the third column. The empirical non-exceedance probability of $x_7 = 10$ m (i.e., $P(X < x_7)$) is then $p_{x_7} = \frac{\text{rank}(x_7)-1}{10} = \frac{6-1}{10} = 0.5$.

Table 3.4: Table containing H_{mo}s in ascending order along with their associated ranks.

$\mathbf{x_k}$	**k**	**rank($\mathbf{x_k}$)**
4	1	1
5	2	2
5	5	3
5	10	4
8	8	5
10	7	6
10	9	7
11	3	8
11	4	9
12	6	10

So far in this chapter, the implementation of the POT method was considered. A threshold selection technique was described and estimations regarding return levels of waves were made based on the Slangkop data by using both the GPD and the Weibull distribution. These distributions were used along with different parameter estimation methods and the results were compared. Next, an alternative method for estimating the parameters of probability distributions will be considered, namely, *least squares approximation*.

3.3 Least Squares Approximation

In this section the least squares approximation method will be used to determine the parameters of the different statistical distributions. The resulting estimations regarding return levels will then be compared to the estimations determined in the previous sections by the POT method.

3.3.1 The Basic Linear Least Squares Problem

Firstly, a basic discrete linear least squares problem is considered as an illustrative example. Given the data in Table 3.5, the objective is to find a straight line that will best approximate the data. The values in Table 3.5 are plotted in Figure 3.3.1.

Table 3.5: Data for discrete linear least squares example.

x_i	y_i
1	1.3
2	3.5
3	4.2
4	5.0
5	7.0
6	8.8
7	10.1
8	12.5
9	13.0
10	15.6

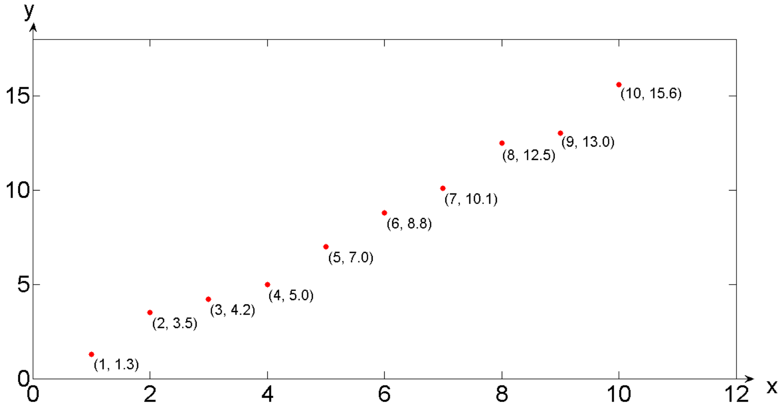


Figure 3.3.1: Plotted values of Table 3.5.

In order to fit a straight line, with equation $y = mx + c$, through the data points in Figure 3.3.1 by means of a least squares approach, the *least squares error* has to be minimized (Lawson & Hanson (1974)). The least squares error is the sum of the squared differences between the actual data values (i.e., the y_is, $i = 1, 2, \ldots, 10$) and the values predicted by the chosen model (i.e., $f(x_i)$, $i = 1, 2, \ldots, 10$). When fitting a straight line $f(x_i) = mx_i + c$, the least squares error to be minimized is given by (Burden & Faires (2001))

$$S = \sum_{i=1}^{10} \left[y_i - (mx_i + c)\right]^2. \tag{3.3.1}$$

In other words, the values of m and c that minimize equation (3.3.1) have to be determined. In order to do this, the derivative of S with respect to both m and c are taken, respectively, and set equal to zero. This yields the following two equations, called the *normal equations*:

$$10c + m\sum_{i=1}^{10} x_i = \sum_{i=1}^{10} y_i, \tag{3.3.2}$$

and

$$c\sum_{i=1}^{10} x_i + m\sum_{i=1}^{10} x_i^2 = \sum_{i=1}^{10} x_i y_i. \tag{3.3.3}$$

Solving the system of normal equations results in $m = 1.538$ and $c = -0.360$, and hence, the best approximating line is given by the equation $y = 1.538x - 0.360$. The code for determining m and c is given in Appendix D.1. The line is plotted in Figure 3.3.2.

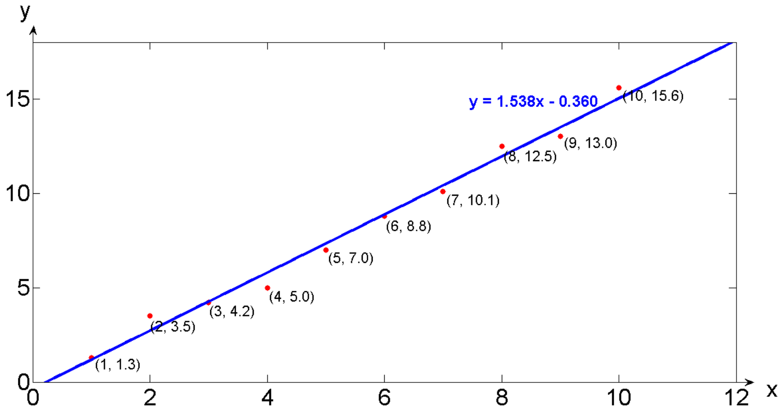


Figure 3.3.2: Graph of the best approximating line, $y = 1.538x - 0.360$, to the values in Table 3.5.

In general, for a set of n data points and an approximating linear function of the form $y = mx + c$, the normal equations are given by

$$nc + m\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \quad \text{and} \quad c\sum_{i=1}^{n} x_i + m\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{3.3.4}$$

The solution to this system of equations, determined by using the same method as in the illustrative example, is (Burden & Faires (2001))

$$c = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{3.3.5}$$

and

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}. \tag{3.3.6}$$

In the next section the application of the least squares method to the wave data used in this study will be considered.

3.3.2 Application of Least Squares Method to Wave Data

In practice, data is very rarely approximately linearly related, as is the case with the wave data in this study. When the POT method is used, the GPD and the Weibull distribution are models used to represent the approximate relation between the data points. In a previous chapter, the unknown parameters of these distributions were determined by means of the Method of Maximum Likelihood, the Method of Moments, and the Method of L-Moments, respectively. The next objective is to determine these parameters by the use of a least squares method. The GPD is considered first.

Least Squares Fit of GPD

The GPD has three parameters, namely u (the selected threshold level), α, and ξ. Here, as in the previous chapter, a fixed threshold of $u = 3.3718$ m is used. Only the other two parameters, α and ξ, have to be determined by means of the least squares method.

The normal equations used for the linear least squares problem (equations (3.3.5) and (3.3.6)) are used to solve α and ξ by defining new variables. Equation (3.1.9) (i.e., $x_p = u + \frac{\alpha}{\xi}\left[(1-p)^{-\xi} - 1\right]$) is non-linear and relates the return level, x_p, to the non-exceedance probability, p, for the GPD. This equation is rearranged and then the logarithm is taken, yielding

$$\ln\left(x_p - u + \frac{\alpha}{\xi}\right) = -\xi \ln(1-p) + \ln\left(\frac{\alpha}{\xi}\right). \tag{3.3.7}$$

By comparing equation (3.3.7) to the linear equation $y = mx + c$, used in the case of linearly related data, the following substitutions are made

$$\underbrace{\ln\left(x_p - u + \frac{\alpha}{\xi}\right)}_{y} = \underbrace{-\xi}_{m}\, \underbrace{\ln(1-p)}_{x} + \underbrace{\ln\left(\frac{\alpha}{\xi}\right)}_{c}.$$

Hence, the x_i and y_i in equations (3.3.5) and (3.3.6) are replaced by $\ln(1-p_i)$ and $\ln\left(x_{p_i} - u + \frac{\alpha}{\xi}\right)$, respectively. In this case, however, equation (3.3.5) cannot be used directly to solve $c = \ln\left(\frac{\alpha}{\xi}\right)$, since the right hand side of the equation contains sums over the y_i, where $y_i = \ln\left(x_{p_i} - u + \frac{\alpha}{\xi}\right)$ and $\frac{\alpha}{\xi}$ are unknown. It is therefore necessary to use a non-linear solver to find $\frac{\alpha}{\xi}$. Newton-Rhapson iteration is used for this purpose.

The Newton-Rhapson method starts with an initial approximation p_0 and generates the sequence $\{p_n\}_{n=0}^{\infty}$ by (Burden & Faires (2001)):

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \quad \text{for } n \geq 1. \tag{3.3.8}$$

The iteration continues until convergence to a specified tolerance level is achieved. The Matlab code for implementing the Newton-Rhapson method is given in Appendix D.2. It is implemented here by using equation (3.3.5) to define:

$$f\left(\frac{\alpha}{\xi}\right) = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2} - c, \tag{3.3.9}$$

and substituting the x_i by $\ln(1-p_i)$, the y_i by $\ln\left(x_{p_i} - u + \frac{\alpha}{\xi}\right)$, and c by $\ln\left(\frac{\alpha}{\xi}\right)$.

The value of $\frac{\alpha}{\xi}$, as determined by Newton-Rhapson iteration, is substituted into $y_i = \ln\left(x_{p_i} - u + \frac{\alpha}{\xi}\right)$ and equations (3.3.5) and (3.3.6) are then used to solve $c = \ln\left(\frac{\alpha}{\xi}\right)$ and $m = -\xi$, respectively. The estimated values of $\hat{\alpha}$ and $\hat{\xi}$ are shown in Table 3.6. The 100-year return level, according to the GPD, when these parameter estimates are used, is also shown in the table.

Table 3.6: Parameter estimates of ξ (shape parameter) and α (scale parameter) as well as the 100-year return level as determined when using the least squares method to fit the GPD to the empirical data.

Parameter Estimation Method	**Parameter Estimates**	**X(100)**
Least Squares	$\hat{\xi} = -0.0168$ $\hat{\alpha} = 0.8625$	11.7366 m

The return levels and associated non-exceedance probabilities as well as return periods, as determined by fitting the GPD to the empirical data in a least squares sense, is plotted in Figure 3.3.3. Figure 3.3.4 also shows a plot of the return levels and their associated return periods, as well as the empirical data values plotted together with their associated return periods (black dots). Note that the green lines in Figures 3.3.3 and 3.3.4 are identical.

Next, the least squares fit of the Weibull distribution to the empirical data is considered.

Least Squares Fit of Weibull Distribution

For the Weibull distribution a fixed threshold, $u = 3.3718$ m, is again used. The other two unknown parameters, α and ξ, will also be determined by means of the least squares method.

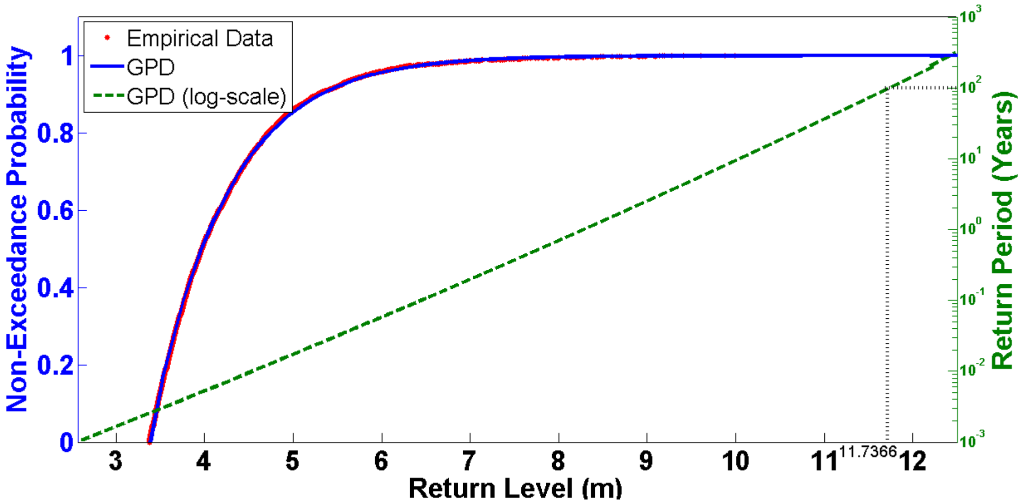


Figure 3.3.3: The return levels and associated non-exceedance probabilities as well as return periods as determined by the GPD, when a threshold of $u = 3.3718$ m is used. The GPD parameters were estimated by the *least squares method.* The 100-year return level is indicated by the black dotted line and is 11.7366 m. The green line is identical to the green line in Figure 3.3.4.

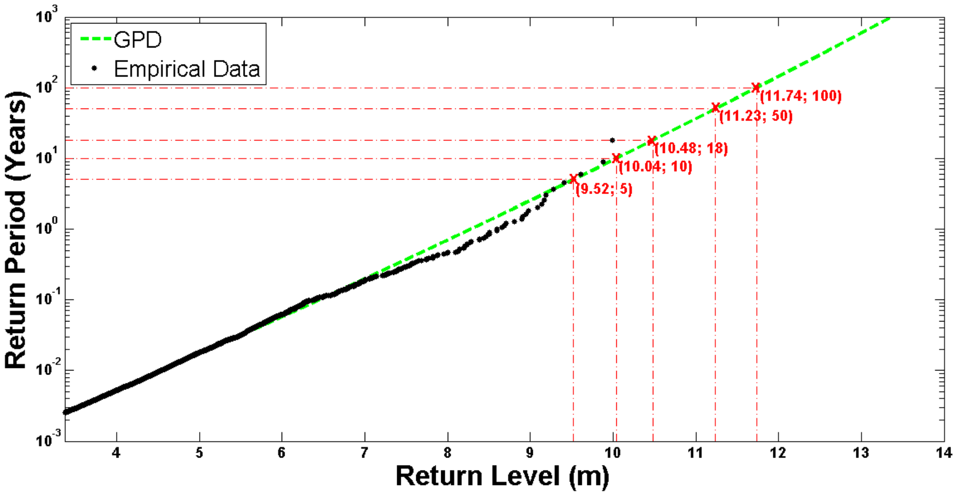


Figure 3.3.4: The return levels and associated return periods as determined by the GPD, when a threshold of $u = 3.3718$ m is used. The GPD parameters were estimated by the *least squares method.*

Equation (3.1.11) (i.e., $x_p = u + \alpha\,[-\ln(1-p)]^{\frac{1}{\xi}}$) is the non-linear equation for the Weibull distribution that relates the return level, x_p, to the non-exceedance probability, p. As in the case of the GPD, this equation is rearranged and the logarithm is taken, resulting in

$$\ln(x_p - u) = \frac{1}{\xi} \ln\left[-\ln(1-p)\right] + \ln\alpha. \tag{3.3.10}$$

By comparing equation (3.3.10) to the linear equation $y = mx + c$, used in the case of linearly related data, the following substitutions are made

$$\underbrace{\ln(x_p - u)}_{y} = \underbrace{\frac{1}{\xi}}_{m} \underbrace{\ln\left[-\ln(1-p)\right]}_{x} + \underbrace{\ln\alpha}_{c}.$$

In this case, the x_i and y_i in equations (3.3.5) and (3.3.6) are replaced by $\ln\left[-\ln(1-p_i)\right]$ and $\ln(x_{p_i} - u)$, respectively. Equation (3.3.5) is then used to solve $\ln\alpha$ and equation (3.3.6) is used to solve $\frac{1}{\xi}$. These values are determined to be -0.1632 and 0.9327, respectively, from which $\hat{\alpha}$ and $\hat{\xi}$ are estimated to be the values shown in Table 3.7. The 100-year return level, according to the Weibull distribution, when these parameter estimates are used, is also shown in the table.

Table 3.7: Parameter estimates of ξ (shape parameter) and α (scale parameter) as well as the 100-year return level as determined when the using the least squares method to fit the Weibull distribution to the empirical data.

Parameter Estimation Method	**Parameter Estimates**	**X(100)**
Least Squares	$\hat{\xi} = 1.0721$ $\hat{\alpha} = 0.8494$	11.0440 m

The return levels and associated non-exceedance probabilities as well as return periods, as determined by fitting the Weibull distribution to the empirical data in a least squares sense, is plotted in Figure 3.3.5. Figure 3.3.6 also shows a plot of the return levels and their associated return periods, as well as the empirical data values plotted together with their associated return periods (black dots). Note that the green lines in Figures 3.3.5 and 3.3.6 are identical.

3.3.3 Conclusions on Least Squares Fits of GPD and Weibull Distribution

The estimations of return levels when the GPD is fitted to the data in a least squares sense are overall higher than the estimations given by the least squares fitting of the Weibull distribution. This can be seen in Figure 3.3.7, which displays the GPD, Weibull distribution and empirical data on the same axes.

The specific 5-, 10-, 18-, 50-, and 100-year return levels for the GPD and the Weibull distribution, respectively, are indicated by the red dashed-dotted lines in Figures 3.3.4 and 3.3.6. When using the least squares method to fit a probability distribution to the data, the conclusion can be made that the use of the GPD is safer than the use of the Weibull distribution, since the risk of under-estimating return levels when using the GPD is less. However, for both the GPD and Weibull distribution the return level

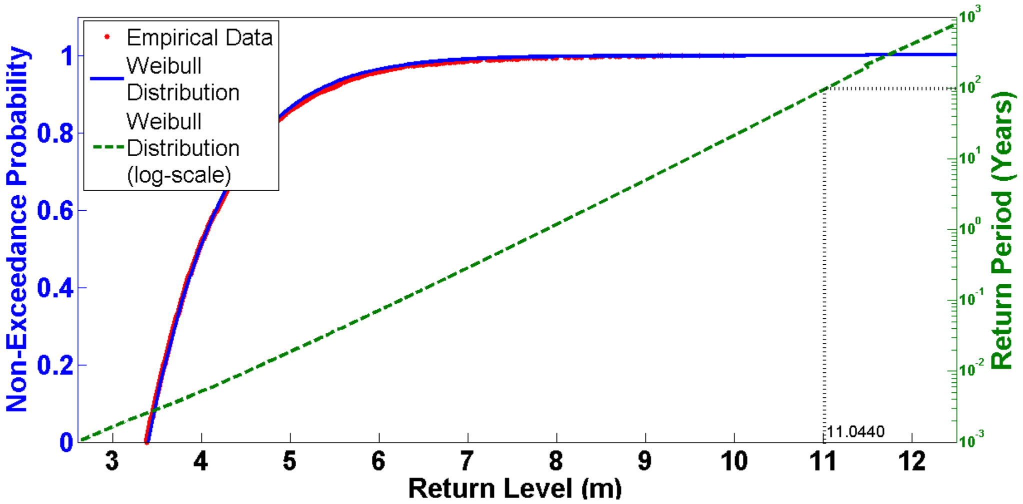


Figure 3.3.5: The return levels and associated non-exceedance probabilities as well as return periods as determined by the Weibull distribution, when a threshold of $u = 3.3718$ m is used. The Weibull parameters were estimated by the *least squares method*. The 100-year return level is indicated by the black dotted line and is 11.0440 m. The green line is identical to the green line in Figure 3.3.6.

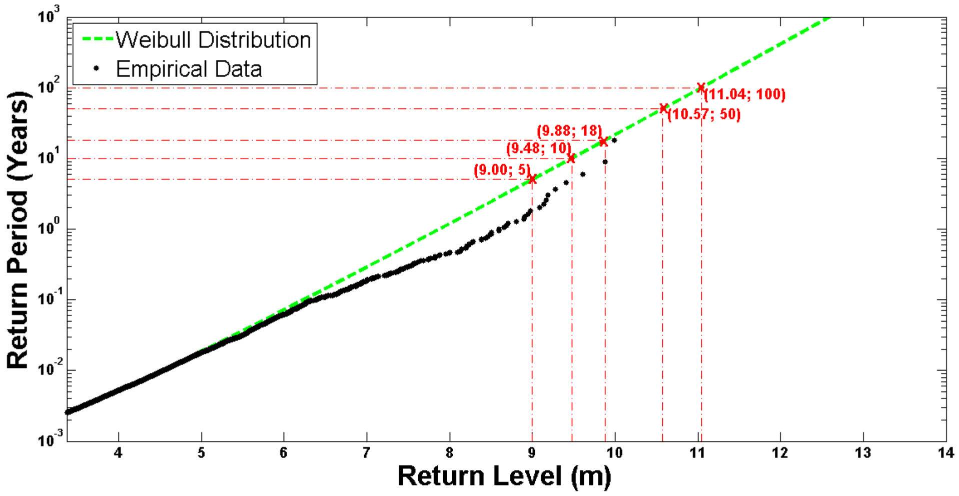


Figure 3.3.6: The return levels and associated return periods as determined by the Weibull Distribution, when a threshold of $u = 3.3718$ m is used. The Weibull parameters were estimated by the *least squares method.*

estimations are overall lower when the least squares approximation method is used for determining the unknown parameters of the distributions than when one of the other three parameter estimation methods are used. Hence, it can be concluded that the least squares approximation method is the least safest method for esimating unknown

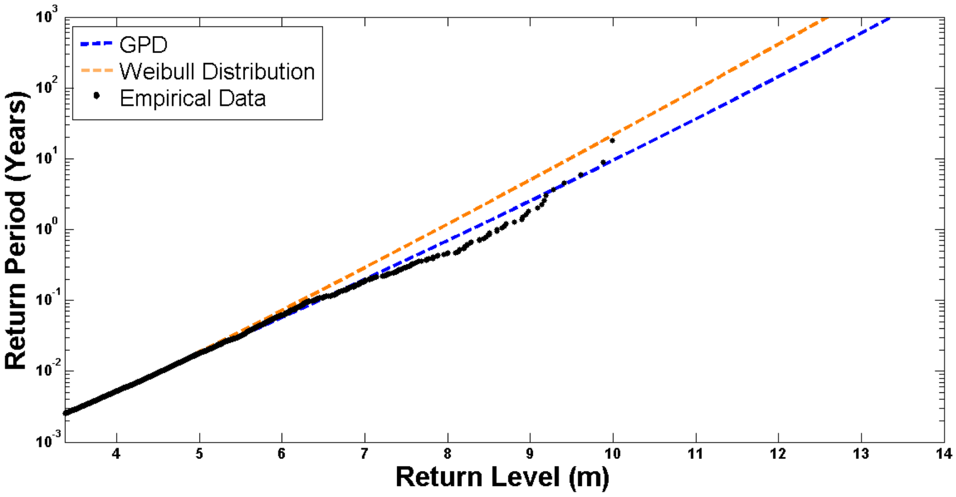


Figure 3.3.7: The return levels and associated return periods as determined by the GPD and Weibull distribution, respectively. The probability distributions' parameters were estimated by the *least squares method*.

parameters.

In this chapter, the POT method was used along with an automated threshold selection of 3.3718 m. Four different techniques were implemented for estimating the unknown parameters of the GPD and Weibull distribution, namely, the method of maximum likelihood, the method of moments, the method of L-moments, and the least squares approximation method. The overall conclusion is that the use of the method of maximum likelihood in conjunction with the Weibull distribution yields reliable estimations of wave return levels.

Chapter 4

Probability Distribution Fit Evaluation

The fits of the different probability distributions to the H_{mo}-data need to be evaluated in terms of how well they fit the data. Based on these evaluations conclusions can be drawn on whether or not (and to what extent) the sample data indicate that a specific statistical model (i.e., probability distribution) for a population distribution fits the data (Wackerly et al. (2008)). This is done by the use of goodness-of-fit statistics in Section 4.1. It has to be noted that there is a difference between the goodness-of-fit and accuracy of probability distributions. The accuracy of distributions are evaluated by determining the uncertainty of quantile estimates, which is done in Section 4.2.

4.1 Goodness-of-Fit Statistics

Goodness-of-fit statistics are used to describe how well a specified statistical model fits a dataset. In other words, it is used to evaluate how much the values predicted by the statistical model (i.e., a chosen probability distribution) differ from the actual, observed data values. Goodness-of-fit statistics can only be used to evaluate the fit of a probability distribution to the available data. In other words, if 18 years of H_{mo}-data is available, the fits of the probability distributions to the data can only be analysed for those 18 years and not, for instance, for a period of 20 years (or any period longer than 18 years).

The goodness-of-fit statistic that will be used in this study is the χ^2-test (Chi-Squared Test). Other goodness-of-fit statistics include the Kolmogrov-Smirnov Test, the Standardised Least Squares Criterion, the Probability Plot Correlation Coefficient, and the Log-Likelihood Measure. The χ^2-test is considered next.

4.1.1 χ^2-Test

The χ^2-test statistic evaluates the fit of the chosen probability distribution to the observed data by comparing the number of observed values and the number of expected values (as determined by the probability distribution) in class intervals covering the range of the variable. The test comprises of the following elements:

- firstly, the *test statistic*. This is a function of sample values on which a hypothesis is tested (Wackerly et al. (2008)). In the case of the χ^2-test, the test statistic is given by
$$z = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}, \tag{4.1.1}$$
where m is the number of classes, n_i is the actual number of observed values in class i, n is the sample size, and p_i is the probability corresponding to class i (as determined by the chosen probability distribution of which the fit is being evaluated), which implies that np_i is the number of expected values in class i (DHI (2003)). This test statistic was proposed by Karl Pearson in 1900 and can be shown to have an approximate χ^2-distribution when n is large.

- secondly, the *null hypothesis*, H_0. This is the hypothesis to be tested and this is done by evaluating the claim that the test statistic, z, is distributed according to an approximate χ^2-distribution (Wackerly et al. (2008)). The null hypothesis is assumed unless a specified level of statistical evidence is reached in order for it to be rejected (Rind (2014)).

- thirdly, the *alternative hypothesis*, H_a. If H_0 is rejected, it is done in favour of H_a.

- finally, the *rejection region*, RR. If the computed value of z falls within the specified values of the RR, H_0 is rejected in favour of H_a. If the computed value of z does not fall in the RR, H_0 cannot be rejected (Wackerly et al. (2008)).

Besides the elements specified above, there are additional specifications that have to be made in order to carry out a χ^2-test. The degrees of freedom associated with the test statistic, z, have to be taken into account. If the probability distribution, which is being evaluated, has q parameters, z has $m-1-q$ degrees of freedom. The appropriate number of degrees of freedom is determined as the number of classes (m) less one degree of freedom for each linear restriction placed on the class probabilities. Therefore, one degree of freedom is lost because the sum of the probabilities associated with each of the classes has to be equal to 1, i.e., $\sum_{i=1}^{m} p_i = 1$, and a further q degrees of freedom are lost for the estimation of the probability distribution's q parameters.

Another specification that needs to be made, is the significance level of the test, α. This is the probability of rejecting H_0 incorrectly, i.e., rejecting H_0 when, in actual fact, H_0 is true. It is also referred to as a type I error and is used to define the RR(Wackerly et al. (2008)). H_0 is rejected in favour of H_a at a significance level of α if (DHI (2003))
$$z > \chi^2_\alpha(m-1-q), \tag{4.1.2}$$
where $\chi^2_\alpha(m-1-q)$ is the $(1-\alpha)$-quantile in the χ^2-distribution with $m-1-q$ degrees of freedom, and can be determined from a χ^2-distribution table (refer to Figure C.1.1 in

Appendix C.1). The value $\chi^2_\alpha(m-1-q)$ is referred to as the *critical value* (Wackerly et al. (2008)).

Furthermore, according to DHI (2003), the test is more powerful if each of the classes has approximately the same probability, i.e., $p_i = p = \frac{1}{m}$, $i = 1, 2, \ldots, m$. The number of classes is also determined such that the expected number of events in a class is not smaller than 5. In the following sections, the χ^2-test is applied to test the hypotheses that the data values are distributed according to the GPD and the Weibull distribution, respectively. In both cases, the parameters of the distributions are estimated by the method of maximum likelhood.

Generalized Pareto Distribution

The null hypothesis in this case is stated as follows:
H_0: *The exceedances of the threshold of 3.3718 m are distributed according to a GPD, with parameters estimated by the method of maximum likelihood.*
In order to apply the χ^2-test, the data is first divided into eight classes in Table 4.1. These classes were chosen in such a way that the probability of obtaining a measurement in any of the classes is very similar (i.e., the p_i's are similar in size). The probability, p_i, for class i: $[a, b)$, is obtained by using the GPD CDF (refer to equation (2.2.3)), as follows

$$p_i = F_{GPD}(b) - F_{GPD}(a). \tag{4.1.3}$$

Table 4.1: Division of exceedances of the threshold into eight class intervals for use in the χ^2-test on the GPD.

i	**class**	$\mathbf{n_i}$	$\mathbf{p_i}$	$\mathbf{np_i}$
1	[3.37; 3.49]	896	0.1282	916.5263
2	(3.49; 3.62]	870	0.1224	875.0759
3	(3.62; 3.77]	884	0.1203	860.5200
4	(3.77; 3.96]	911	0.1255	897.4256
5	(3.96; 4.21]	886	0.1283	917.6826
6	(4.21; 4.53]	902	0.1183	846.1522
7	(4.53; 5.1]	908	0.1269	907.6857
8	(5.1; ∞)	894	0.1300	929.9304
Total		7151	1	7151

The data in Table 4.1 is substituted into equation (4.1.1), yielding

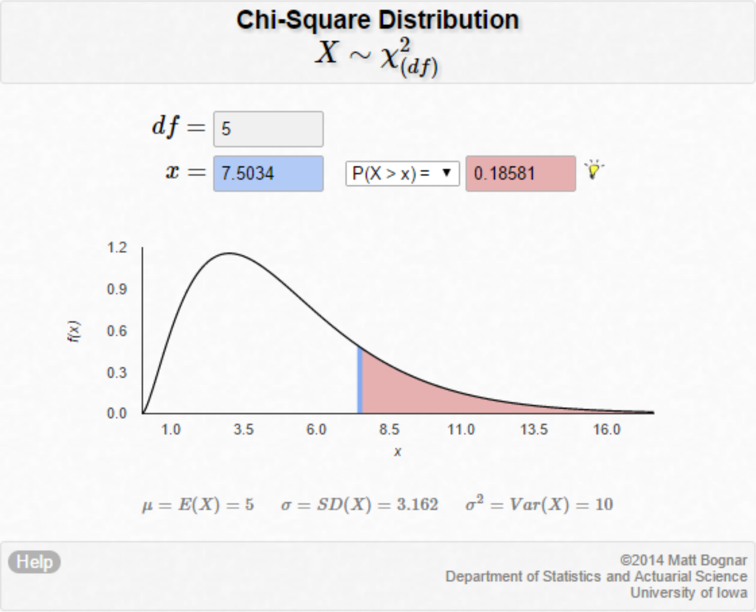
$$z_{GPD} = \sum_{i=1}^{8} \frac{(n_i - np_i)^2}{np_i} = 7.5034. \tag{4.1.4}$$

Since the exceedances are divided into eight classes and the GPD has two estimated parameters, α and ξ (the threshold level, $u = 3.3718$ m, is fixed), the test statistic has an approximate χ^2-distribution with $8 - 1 - 2 = 5$ degrees of freedom. If a significance level of $\alpha = 0.10$ is selected, the null hypothesis is rejected if

$$z_{GPD} > \chi^2_{0.10}(5) = 9.236, \tag{4.1.5}$$

(refer to Figure C.1.1 in Appendix C.1 for the obtained value of $\chi^2_{0.10}(5)$). Since the calculated value of z_{GPD} in equation (4.1.4) is less than the tabulated critical value of $\chi^2_{0.10}(5)$ in equation (4.1.5) (i.e., $z_{GPD} < \chi^2_{0.10}(5)$), H_0 cannot be rejected. It can therefore be concluded that the data do not present sufficient evidence to reject the hypothesis that the points over the threshold of 3.3718 m possess a GPD.

Another quantity which can be calculated is the p-value associated with the test. The meaning of this value is commonly misunderstood, but yet it is very important in interpreting the results of a test (Rind (2014)). The p-value is the smallest significance level for which the observed data indicate that H_0 should be rejected (Wackerly et al. (2008)). In other words, it is the probability of seeing a result as extreme or more extreme than the test statistic, if H_0 were true (Rind (2014)). Hence, the larger the p-value, the less the evidence to reject the claim that the points over the threshold possess a GPD distribution. In this case, the p-value is given by $P(z_{GPD} > 7.5034)$. From Figure C.1.1, Appendix C.1, it is known that p-value > 0.10. The exact p-value is established as 0.18581 (refer to Figure 4.1.1).



Chi-Square Distribution

$X \sim \chi^2_{(df)}$

$df =$ 5

$x =$ 7.5034   P(X > x) = ▾   0.18581

$\mu = E(X) = 5$ $\sigma = SD(X) = 3.162$ $\sigma^2 = Var(X) = 10$

Figure 4.1.1: The p-value such that $P(z_W > 7.5034)$ [source: `http://homepage.stat.uiowa.edu/~mbognar/applets/chisq.html`].

Next, the χ^2-test will be applied to test a hypothesis that the data values are distributed according to a Weibull distribution.

Weibull Distribution

The null hypothesis is stated as follows:
H_0: The exceedances of the threshold of 3.3718 m are distributed according to a Weibull distribution, with parameters estimated by the method of maximum likelihood.
Once again, the data is divided into eight classes as shown in Table 4.2. The classes were also again chosen such that the probability of the classes are similiar in size. The probability, p_i, for class i: $[a, b)$, is obtained by using the Weibull CDF (refer to equation (2.2.5)), as follows

$$p_i = F_W(b) - F_W(a). \tag{4.1.6}$$

Table 4.2: Division of exceedances of the threshold into eight class intervals for use in the χ^2-test on the Weibull distribution.

i	**class**	$\mathbf{n_i}$	$\mathbf{p_i}$	$\mathbf{np_i}$
1	[3.37; 3.49]	896	0.1201	858.7951
2	(3.49; 3.62]	870	0.1210	865.6209
3	(3.62; 3.77]	884	0.1214	868.3854
4	(3.77; 3.96]	911	0.1281	915.8582
5	(3.96; 4.21]	886	0.1317	941.5947
6	(4.21; 4.53]	902	0.1213	867.6654
7	(4.53; 5.1]	908	0.1290	922.3234
8	(5.1; ∞)	894	0.1274	910.7541
Total		7151	1	7151

The data in Table 4.2 is substituted into equation (4.1.1), yielding

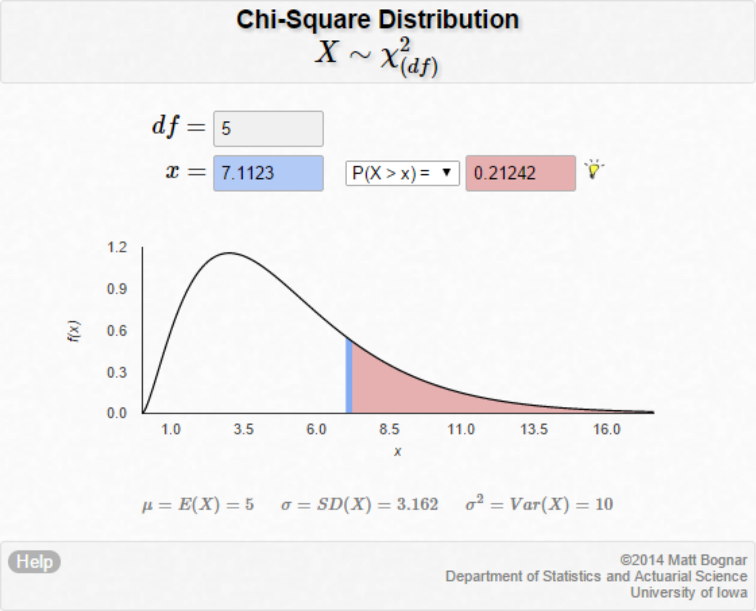
$$z_W = \sum_{i=1}^{8} \frac{(n_i - np_i)^2}{np_i} = 7.1123. \tag{4.1.7}$$

Here, the test statistic again has an approximate χ^2-distribution with $8 - 1 - 2 = 5$ degrees of freedom. If a significance level of $\alpha = 0.10$ is used, the null hypothesis is once again rejected if (the same as in equation (4.1.5))

$$z_W > \chi^2_{0.10}(5) = 9.236. \tag{4.1.8}$$

The calculated value of z_W in equation (4.1.7) is less than the tabulated critical value of $\chi^2_{0.10}(5)$ in equation (4.1.8) (i.e., $z_W < \chi^2_{0.10}(5)$), and, hence, H_0 is not rejected. It follows that the data do not present sufficient evidence to reject the hypothesis that the points over the threshold of 3.3718 m possess a Weibull distribution. In this case the p-value is established to be $P(z_W > 7.1123) = 0.21242$ (refer to Figure 4.1.2).

As stated previously, the larger the p-value, the less the evidence to reject the claim that the points over threshold are distributed according to the specified probability distribution. The p-value in the case of the Weibull distribution is larger than in the case of the GPD. Hence, based on the χ^2-test, the evidence in favour of rejecting the hypothesis



Chi-Square Distribution

$X \sim \chi^2_{(df)}$

$df =$ 5

$x =$ 7.1123 P(X > x) = ▾ 0.21242

$\mu = E(X) = 5$ $\sigma = SD(X) = 3.162$ $\sigma^2 = Var(X) = 10$

Help

©2014 Matt Bognar
Department of Statistics and Actuarial Science
University of Iowa

Figure 4.1.2: The p-value such that $P(z_W > 7.1123)$ [source: `http://homepage.stat.uiowa.edu/~mbognar/applets/chisq.html`].

that the points over threshold are distributed according to the Weibull distribution is less than the evidence in favour of rejecting the hypothesis that the points over threshold are distributed according to the GPD. In other words, even though neither of the null hypotheses that the data points are distributed according to the GPD or the Weibull distribution are rejected at a significance level of 10%, there is less evidence against the hypothesis that states the points are distributed according to the Weibull distribution than to the GPD. Stated simpler, analysing the results of the test leads to a conclusion more in favour of the points over threshold being distributed according to a Weibull distribution than according to a GPD.

Along with the fit of a specified probability distribution to a dataset, the degree uncertainty of estimations made by the distribution can also be evaluated. Techniques for doing so are considered in the following section.

4.2 Uncertainty of Quantile Estimates

A quantile estimate, x_p, is the smallest value such that $F(x_p) = P(X \leq x_p) = p$ (also refer to equation (3.1.7) (Wackerly et al. (2008)). Quantile estimates have a degree of uncertainty attached to it. This is mostly due to the restricted duration of most datasets, as is the case in this study, where only approximately 18 years of H_{mo}-data is available. Knowledge concerning the statistical confidence that can be attributed to the prediction of extreme wave heights are very important, since these estimations are used to determine the appropriate levels of coastal protection required (Li et al. (2008)).

Three techniques for determining degrees of uncertainty are considered, namely, the Bootstrap technique, the Monte Carlo similation, and Jack-knife resampling. The first technique discussed, is the Bootstrap technique.

4.2.1 The Bootstrap Technique

The bootstrap technique assigns measures of accuracy to statistical estimates (Efron & Tibshirani (1993)). Inferences about unknown populations (represented by statistical models) are made from sample data (Boos (2003)). The bootstrap is used to determine how accurately a statistic, calculated from a sample, estimates the corresponding quantity for the whole population (Efron & Tibshirani (1993)).

Drawing different samples from the same population yields different estimates – the set of estimates obtained represents the sampling distribution (i.e., the probability distribution based on a random sample) of the statistics. *Resampling* is the process of drawing a large number of samples from the dataset and then making numerical calculations to infer the sampling distribution of an estimate. An advantage of the bootstrap technique is that it does not require any assumption about the distribution of the data. It can be used to infer the sampling distribution of a statistic via repeated samples drawn from the sample itself, as opposed to the hypothetical resampling from the population (Chong & Choo (2011)).

Bootstrap samples are used in the boostrap technique. A bootstrap sample is a random sample of size n $(x_1^*, x_2^*, \ldots, x_n^*)$ drawn *with* replacement from the population of n objects $(x_1, x_2, \ldots, x_n)$. The Bootstrap data set $(x_1^*, x_2^*, \ldots, x_n^*)$ consists of members of the original data set $(x_1, x_2, \ldots, x_n)$, some appearing zero times, some appearing once, some appearing twice, etc. An example of a bootstrap sample might look as follows: $x_1^* = x_7, x_2^* = x_3, x_3^* = x_3, x_4^* = x_{22}, \ldots, x_n^* = x_7$ (Efron & Tibshirani (1993)). The complete technique is explained by the simple implementation below.

A Simple Implementation

The Bootstrap technique can be illustrated by the following simple implementation to the entire H_{mo}-dataset:
Let the mean H_{mo} be the statistic of interest. Only 52789 values of the H_{mo} values are available in the dataset. This is the sample data. From this single sample, only one value of the mean can be obtained. However, some information about the variability of the mean is required in order to reason about the population (which, in this case, will be H_{mo}-values for a much longer time period, say 100 years). The bootstrap technique is implemented on the sample data as follows:

1. A new sample (i.e., a "resample" or Bootstrap sample), that is also of size 52789, is formed from the original data set of size 52789. The Bootstrap sample is taken using *sampling with replacement* and therefore it is not identical to the original sample.

2. This process is repeated a large number of times, a 1000 times in this case.

3. For each Bootstrap sample, the mean is calculated. These means are called the Bootstrap estimates.

The 1000 determined Bootstrap estimates are then used to plot a histogram of bootstrap means in Figure 4.2.1, which provides an estimate of the shape of the distribution of the mean H_{mo} and can answer questions about how much the mean varies.

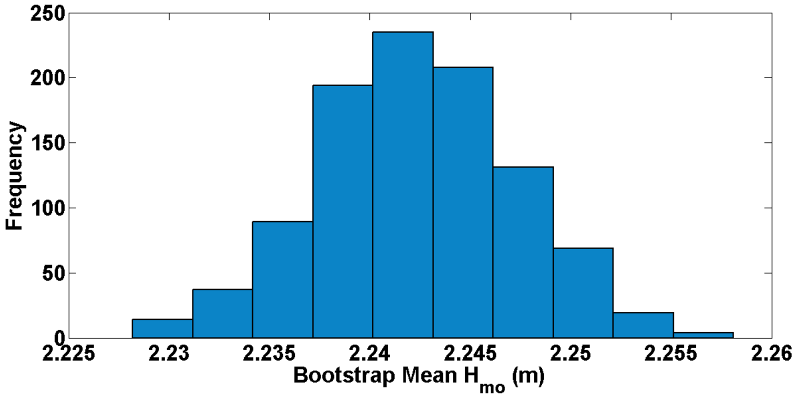


Figure 4.2.1: Histogram of 1000 Bootstrap mean H_{mo}-values.

The 1000 bootstrap samples are generated by means of Matlab's random number generator - each sample is generated by using the random number generator to generate 52789 integer values between the numbers 1 and 52789. The original H_{mo}-values are all stored in an array (of length 52789) and the Bootstrap sample is then made up of the H_{mo}-values at the generated indices. In order to prove that the generated indices are, in fact, random and also covers the range of integers from 1 to 52789 uniformly, a χ^2-test is performed.

Figure 4.2.2 is a bar chart, consiting of 52789 bars (one for each of the indices from 1 to 52789), displaying the number of times each index was generated during the generation of the 1000 samples.

From Figure 4.2.2, it is clear that the selected indices form a fairly uniform distribution, i.e., each index was selected relatively close to a 1000 times during the generation of 1000 Bootstrap samples of size 52789 each. In order to test whether these indices are, however, actually distributed uniformly, a χ^2-test can be applied.

If X is a uniform random variable on the interval (α, β), the probability density function of X is given by (Wackerly et al. (2008))

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if} \quad \alpha < x < \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2.1}$$

Therefore, a uniform random variable on the interval $(0, 52790)$ (or $[1, 52789]$) will have the following probability density function:

$$f(x) = \begin{cases} \frac{1}{52789} & \text{if} \quad 1 \leq x \leq 52789, \\ 0 & \text{otherwise.} \end{cases} \tag{4.2.2}$$

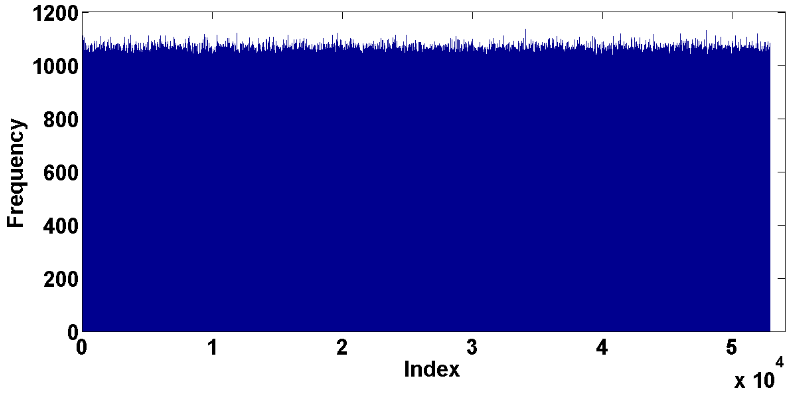


Figure 4.2.2: The number of times each of the indices (between 1 and 52789) was randomly generated in the generation of 1000 samples (each of size 52789).

The χ^2 test statistic is determined by making the following substitutions:

- the number of classes is $m = 52789$ (i.e., each index represents a class),
- the number of observed events in class i, n_i, is given by the number of times index i was randomly generated ($i = 1, 2, \ldots, 52789$),
- the sample size is $n = 52789 \times 1000$,
- the probability corresponding to index i, from equation (4.2.2), is $p_i = p = f(x) = \frac{1}{52789}$, and
- the expected number of times index i is generated is $np_i = np = 1000$.

This yields the following test statistic

$$z = \sum_{i=1}^{52789} \frac{(n_i - 1000)^2}{1000}, \tag{4.2.3}$$

which is approximately χ^2-distributed with $m - 1 - q = 52789 - 1 - 0 = 52788$ degrees of freedom ($q = 0$ since the uniform distribution has no parameters that have to be estimated). H_0 is the hypothesis that the indices are distributed according to a uniform distribution.

After running the χ^2-test, H_0 is accepted at a significance level of $\alpha = 0.05$. It can therefore be said that the generated indices are distributed according to a uniform distribution at a 95% confidence level.

Next, the bootstrap technique will be applied to determine standard deviations of estimations made by the GPD when the three differents methods for parameter estimation (method of maximum likelihood, method of moments, and method of L-moments) are used.

4.2.2 Bootstrap Technique Applied to the GPD

Table 4.3 displays the standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the GPD is used to make estimations based on a 1000 generated Bootstrap samples. The standard deviation is determined for each of the three different parameter estimation methods, namely, the Method of Maximum Likelihood, the Method of Moments, and the Method of L-Moments.

Table 4.3: Standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the GPD is used to make estimations based on 1000 generated Bootstrap samples (i.e., 1000 estimations of each return level are made). The standard deviation is determined for each of the three parameter estimation methods.

Parameter Estimation Method	**Return Level Standard Deviation**				
	5-year	**10-year**	**18-year**	**50-year**	**100-year**
Method of Maximum Likelihood	0.1973	0.2363	0.2720	0.3398	0.3897
Method of Moments	0.3472	0.4015	0.4502	0.5406	0.6059
Method of L-Moments	0.3554	0.4078	0.4543	0.5392	0.5996

Figures 4.2.3 to 4.2.5 the estimations made by the GPD based on the 1000 generated Bootstrap samples (blue lines) as well as the estimation made by the GPD based on the actual dataset (green, dotted line). The method of maximum likelihood, method of moments, and methods of L-moments, respectively, are used for GPD parameter estimation.

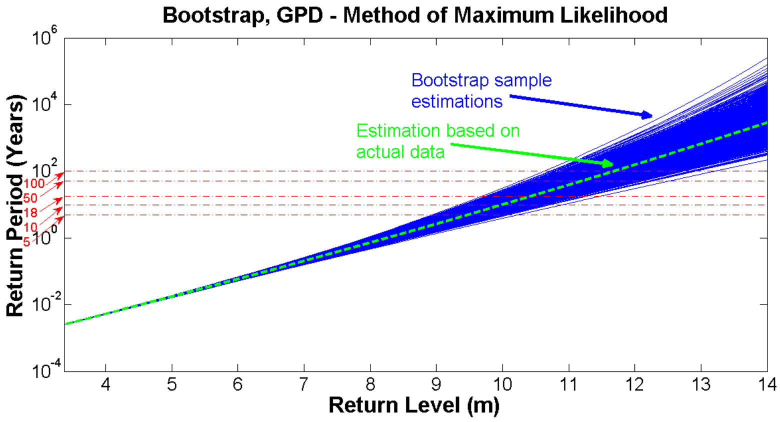


Figure 4.2.3: Estimations made by the GPD when the *method of maximum likelihood* is used for parameter estimation. The blue lines show the estimations made based on the 1000 generated Bootstrap samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

From Table 4.3 and Figures 4.2.3 to 4.2.5, it can be seen that the standard deviations when the method of maximum likelihood is used for parameter estimation, are overall

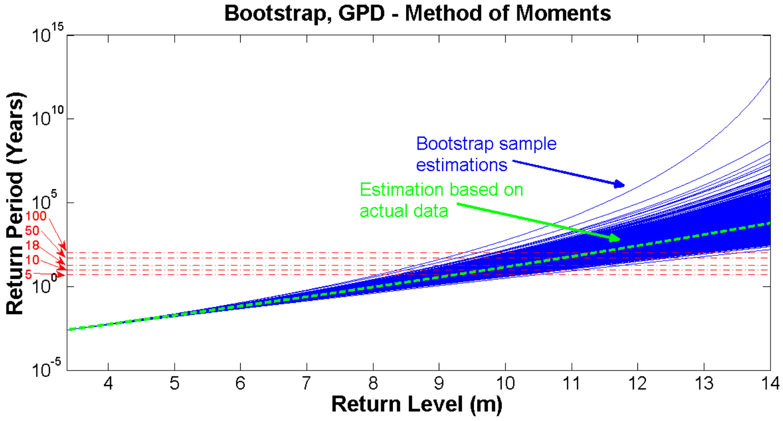


Figure 4.2.4: Estimations made by the GPD when the *method of moments* is used for parameter estimation. The blue lines show the estimations made based on the 1000 generated Bootstrap samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

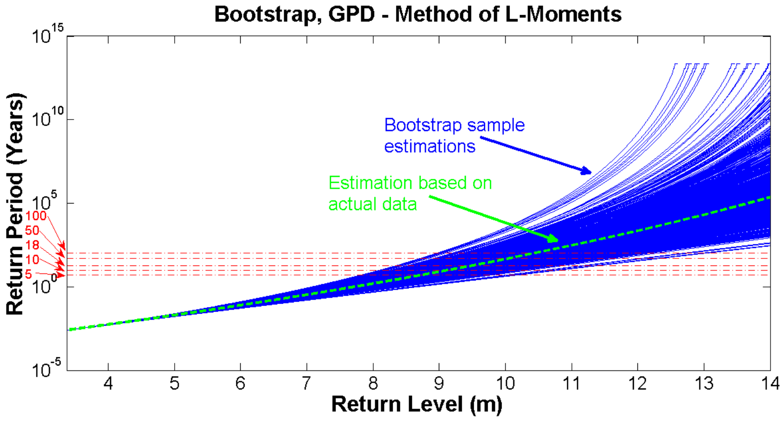


Figure 4.2.5: Estimations made by the GPD when the *method of L-moments* is used for parameter estimation. The blue lines show the estimations made based on the 1000 generated Bootstrap samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

significantly lower than when either the method of moments or the method of L-moments is used. A conclusion to be made from this is that when the GPD is used for making estimations regarding return levels, the method of maximum likelihood seems to be the parameter estimation method that yields the most reliable results.

The bootstrap technique will now be applied to determine standard deviations of estimations made by the Weibull distribution when the three different methods for parameter estimation are used.

4.2.3 Bootstrap Technique Applied to the Weibull Distribution

Table 4.4 displays the standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the Weibull distribution is used to make estimations based on a 1000 generated Bootstrap samples. The standard deviation is determined for each of the three different parameter estimation methods, namely, the Method of Maximum Likelihood, the Method of Moments, and the Method of L-Moments.

Table 4.4: Standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the Weibull distribution is used to make estimations based on 1000 generated Bootstrap samples (i.e., 1000 estimations of each return level are made). The standard deviation is determined for each of the three parameter estimation methods.

Parameter Estimation Method	**Return Level Standard Deviation**				
	5-year	**10-year**	**18-year**	**50-year**	**100-year**
Method of Maximum Likelihood	0.1136	0.1274	0.1394	0.1608	0.1756
Method of Moments	0.0716	0.0781	0.0835	0.0929	0.0993
Method of L-Moments	0.1188	0.1341	0.1474	0.1711	0.1876

Figures 4.2.6 to 4.2.8 show the estimations made by the Weibull distribution based on the 1000 generated Bootstrap samples (blue lines) as well as the estimation made by the Weibull distribution based on the actual dataset (green, dotted line). The method of maximum likelihood, method of moments, and methods of L-moments, respectively, are used for Weibull parameter estimation.

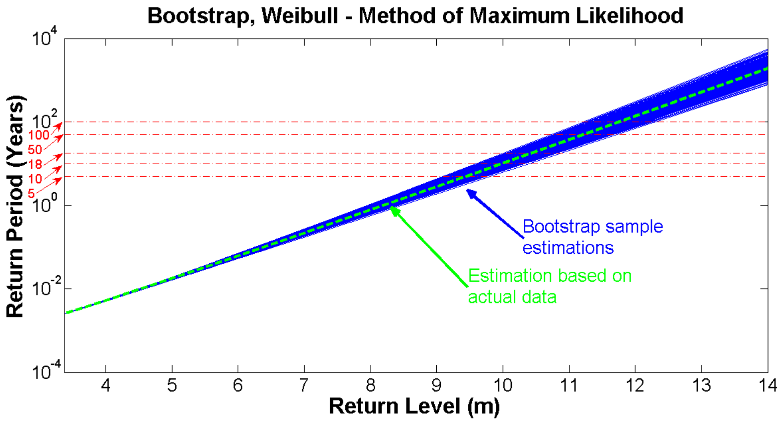


Figure 4.2.6: Estimations made by the Weibull distribution when the *method of maximum likelihood* is used for parameter estimation. The blue lines show the estimations made based on the 1000 generated Bootstrap samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

From Table 4.4 and Figures 4.2.6 to 4.2.8, it can be seen that the standard deviations when the method of moments is used for parameter estimation, are overall the lowest

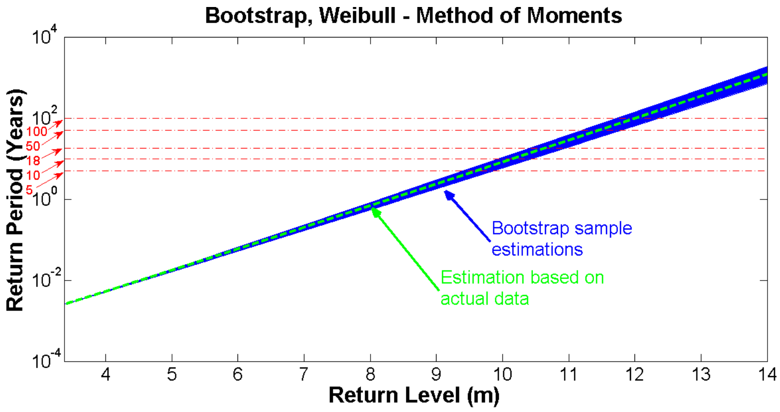


Figure 4.2.7: Estimations made by the Weibull distribution when the *method of moments* is used for parameter estimation. The blue lines show the estimations made based on the 1000 generated Bootstrap samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

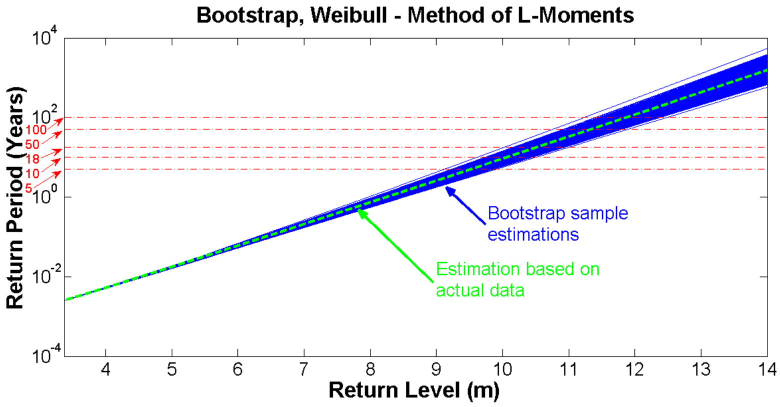


Figure 4.2.8: Estimations made by the Weibull distribution when the *method of L-moments* is used for parameter estimation. The blue lines show the estimations made based on the 1000 generated Bootstrap samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

when compared to the other two methods. However, the standard deviations do not differ significantly for the different methods, and are all relatively low. The standard deviations when the method of maximum likelihood is used is overall the second lowest.

When comparing the standard deviations when the GPD is used for estimations (in Table 4.3 and Figures 4.2.3 to 4.2.5) to the standard deviations when the Weibull distribution is used (in Table 4.4 and Figures 4.2.6 to 4.2.8), the GPD yields higher standard deviations overall (irrespective of the parameter estimation method used). It can therefore be concluded that the Weibull distribution is more reliable for making estimations, based on the data used in this study, than the GPD.

Once again, based on the fact that the method of maximum likelihood is a reliable and popular parameter estimation method (Casella & Berger (2002)), and the fact that the standard deviations when the Weibull distribution is used (irrespective of the method of parameter estimation) are low, reinforces the decision to consider the method of maximum likelihood in conjunction with the Weibull distribution as a reliable estimation method.

In the following section, another method for evaluating the uncertainty of quantile estimates will be considered, namely the Monte Carlo simulation.

4.2.4 Monte Carlo Simulation

The Monte Carlo simulation generates a large number of samples that have the same statistical characteristics as the observed sample and uses those samples to determine the bias and standard deviation of the quantile estimate (DHI (2003)).

A set of m random data points is generated from the chosen probability distribution by using the determined parameter estimates, $\hat{\theta}$, i.e.,

$$x_i = F^{-1}(r_i;\hat{\theta}) \quad , \quad i = 1, 2, \ldots, m, \tag{4.2.4}$$

where r_i is a randomly generated number between 0 and 1 and F^{-1} is the inverse of the cumulative distribution function. In the case of the POT method, where a fixed threshold is used, the number of events (values), i.e., m, exceeding the threshold is a random variable that is assumed to be Poisson distributed (DHI (2003)). This assumption can be made because of the *Poisson approximation to the binomial*, which will be briefly discussed next.

A binomial random variable, X, with parameters (n, p), represents the number of successes that occur in n trials, where each trial results in a success with probability p and in a failure with probability $1 - p$. The Poisson random variable may be used to approximate a binomial random variable with parameters (n, p) when n is large and p is small enough so that np is of moderate size. In other words, if n independent trials are performed, each resulting in a success with probability p, with n large and p small enough to make np moderate, the number of successes is approximately Poisson distributed with parameter $\lambda = np$, where λ equals the expected number of successes (Ross (2010)).

In the case of the 52789 independently observed H_{mo} values, each value can either exceed the specified threshold (success), or not (failure). The number of exceedances, m, will therefore be approximately Poisson distributed and can therefore be randomly generated from a Poisson distribution with parameter $\hat{\lambda}t$, where $\hat{\lambda}$ is the estimated annual number of events (i.e., exceedances over the threshold) for the observed sample, and t is the observation period. therefore,

$$\hat{\lambda}t = \frac{\text{total number of exceedances over threshold in sample}}{52789} \times 52789, \tag{4.2.5}$$

where the probability of a success (i.e., an exceedance) is

$$p = \frac{\text{total number of exceedances over threshold in sample}}{52789}, \tag{4.2.6}$$

and the sample size is $n = 52789$. It follows that the average annual number of events for the generated sample, say sample j, is estimated as

$$\hat{\lambda}^{(j)} = \frac{m}{t}. \tag{4.2.7}$$

For each newly generated sample, new parameters of the probability distribution, represented by $\hat{\theta}^{(j)}$, are estimated. The N-year event estimate is determined for each sample by using the newly determined parameters. After generating a large number of samples (for example, 10 000, as regarded as large by DHI (2003)) and repeating this process each time, the mean and standard deviation of the N-year event estimate can be determined (DHI (2003)). Here, only the standard deviations will be considered.

Next, the Monte Carlo simulation will be applied to determine standard deviations of estimations made by the GPD when the three differents methods for parameter estimation (method of maximum likelihood, method of moments, and method of L-moments) are used.

4.2.5 Monte Carlo Simulation Applied to the GPD

Table 4.5 displays the standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the GPD is used to make estimations based on 10 000 generated Monte Carlo samples. The standard deviation is determined for each of the three different parameter estimation methods, namely, the Method of Maximum Likelihood, the Method of Moments, and the Method of L-Moments.

Table 4.5: Standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the GPD is used to make estimations based on 10 000 generated Monte Carlo samples (i.e., 10 000 estimations of each return level are made). The standard deviation is determined for each of the three parameter estimation methods.

Parameter Estimation Method	**Return Level Standard Deviation**				
	5-year	**10-year**	**18-year**	**50-year**	**100-year**
Method of Maximum Likelihood	0.2052	0.2468	0.2849	0.3574	0.4108
Method of Moments	0.3139	0.3633	0.4076	0.4899	0.5495
Method of L-Moments	0.3196	0.3671	0.4093	0.4866	0.5415

As in the case of the Bootstrap method, it can be seen from Table 4.5 and Figures 4.2.9 to 4.2.11, that the standard deviations when the method of maximum likelihood is used is lower than when one of the other two methods of parameter estimation is used. This therefore leads to the same conclusion as before, which is that when the GPD is used for making estimations regarding return levels, the method of maximum likelihood as parameter estimation method leads to the most reliable, least deviating, results.

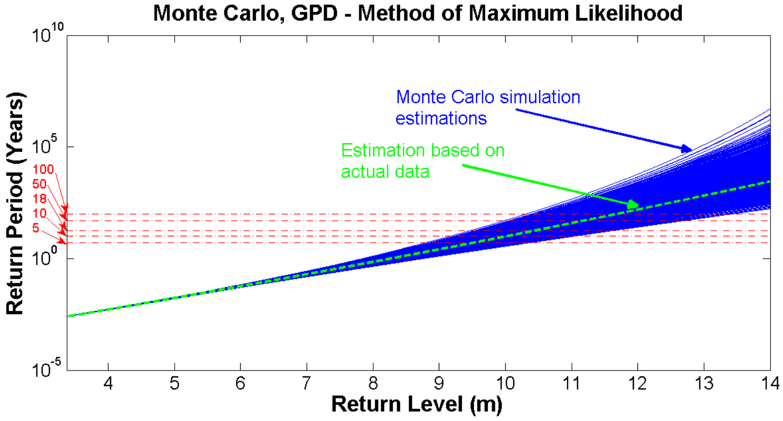


Figure 4.2.9: Estimations made by the GPD when the *method of maximum likelihood* is used for parameter estimation. The blue lines show the estimations made based on the 10 000 generated Monte Carlo samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

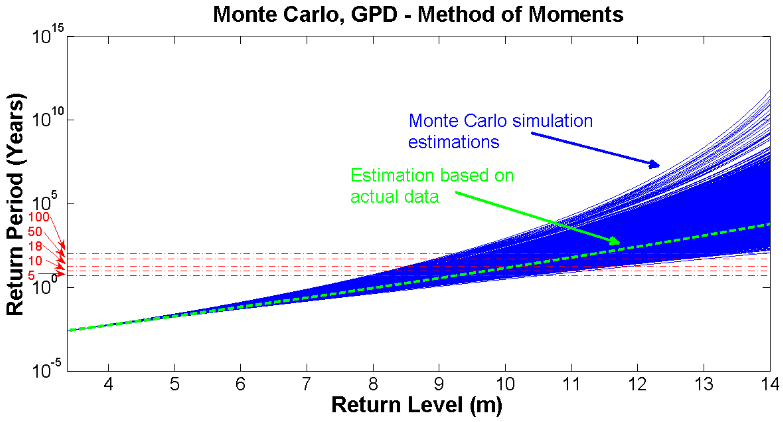


Figure 4.2.10: Estimations made by the GPD when the *method of moments* is used for parameter estimation. The blue lines show the estimations made based on the 10 000 generated Monte Carlo samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

In the following subsection, the Monte Carlo simulation will be applied to determine standard deviations of estimations made by the Weibull distribution when the method of maximum likelihood, the method of moments, and the method of L-moments are used, respectively, for parameter estimation.

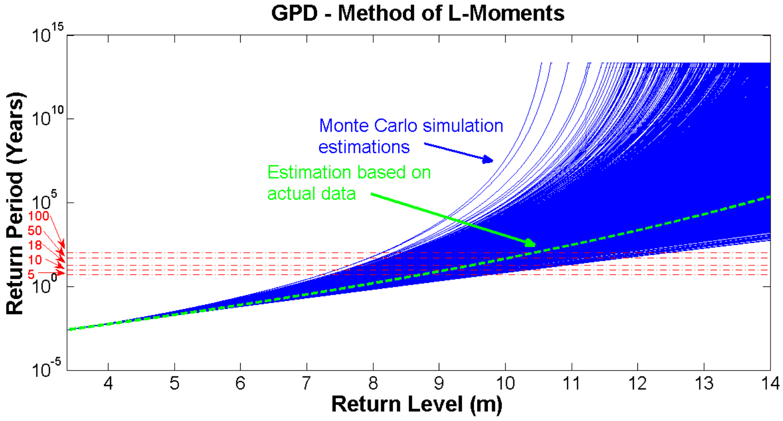


Figure 4.2.11: Estimations made by the GPD when the *method of L-moments* is used for parameter estimation. The blue lines show the estimations made based on the 10 000 generated Monte Carlo samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

4.2.6 Monte Carlo Simulation Applied to the Weibull Distribution

Table 4.6 displays the standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the Weibull distribution is used to make estimations based on 10 000 generated Monte Carlo samples. The standard deviation is determined for each of the three different parameter estimation methods, namely, the Method of Maximum Likelihood, the Method of Moments, and the Method of L-Moments.

Table 4.6: Standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the Weibull distribution is used to make estimations based on 10 000 generated Monte Carlo samples (i.e., 10 000 estimations of each return level are made). The standard deviation is determined for each of the three parameter estimation methods.

Parameter Estimation Method	**Return Level Standard Deviation**				
	5-year	**10-year**	**18-year**	**50-year**	**100-year**
Method of Maximum Likelihood	0.1103	0.1241	0.1360	0.1573	0.1721
Method of Moments	0.0729	0.0794	0.0850	0.0946	0.1010
Method of L-Moments	0.1178	0.1327	0.1457	0.1689	0.1851

Once again, as in the case of the Boostrap method for generating samples, it can be seen from Table 4.6 and Figures 4.2.12 to 4.2.14 that the standard deviations when the method of moments is used for parameter estimation, are overall the lowest when compared to the standard deviations when either the method of maximum likelihood or the method of L-moments is used. Also as in the case of the Bootstrap method, the standard deviations for the different methods are not extremely different, and are all relatively low. Once again, the standard deviations when the method of maximum

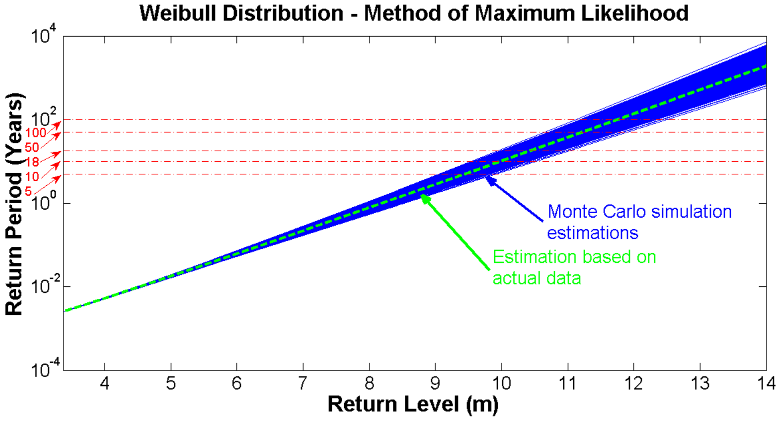


Figure 4.2.12: Estimations made by the Weibull distribution when the *method of maximum likelihood* is used for parameter estimation. The blue lines show the estimations made based on the 10 000 generated Monte Carlo samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

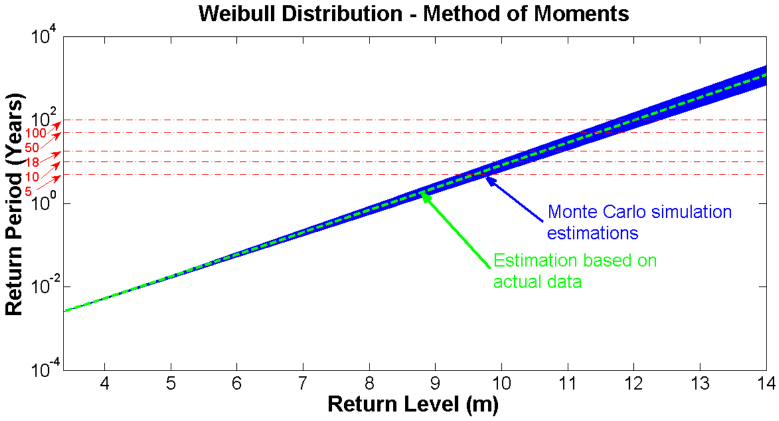


Figure 4.2.13: Estimations made by the Weibull distribution when the *method of moments* is used for parameter estimation. The blue lines show the estimations made based on the 10 000 generated Monte Carlo samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

likelihood is used, is the second lowest.

Similarly as in the previous section (the Bootstrap Technique), when comparing the standard deviations in the case of the GPD (in Table 4.5 and Figures 4.2.9 to 4.2.11) to the standard deviations in the case of the Weibull distribution (in Table 4.6 and Figures 4.2.12 and 4.2.13), the use of the GPD leads to higher overall standard deviations than when using the Weibull distribution. Hence, it can again be concluded that the Weibull distribution is the more reliable distribution of the two for making estimations.

In the last section of this chapter, a final method for evaluating the uncertainty of

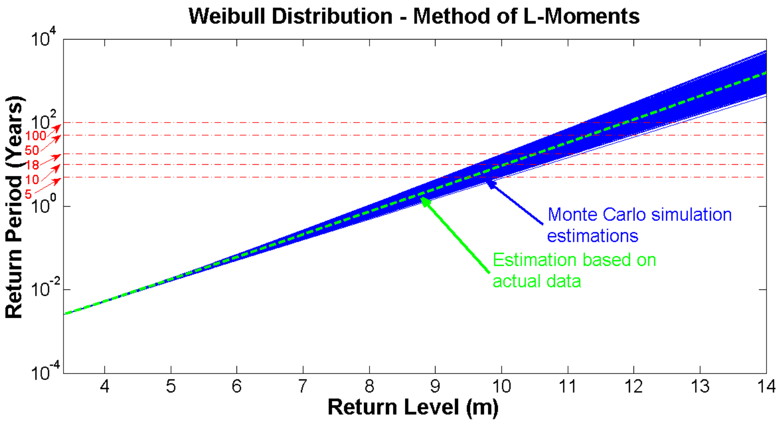


Figure 4.2.14: Estimations made by the Weibull distribution when the *method of L-moments* is used for parameter estimation. The blue lines show the estimations made based on the 10 000 generated Monte Carlo samples, and the green, dashed line shows the estimation made based on the actual dataset. The red, dashed-dotted lines indicate the 5-, 10-, 18-, 50-, and 100-year return periods, respectively.

quantile estimates is discussed, namely Jack-knife resampling.

4.2.7 Jack-knife Resampling

The Jack-knife resampling technique predates the bootstrap method and has similarities to it (Efron & Tibshirani (1993)). It is used to determine the bias and standard deviation of the quantile estimate by sampling n data sets of $(n-1)$ elements from the original data set (DHI (2003)). If $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the original sample, the ith *Jack-knife sample* is given by

$$\mathbf{x}_{(i)} = (x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n), \tag{4.2.8}$$

for $i = 1, 2, \ldots, n$. In other words, the ith Jack-knife sample is the original sample with the ith observation removed (Efron & Tibshirani (1993)).

If $\hat{\theta}$ represents the distribution parameter estimates, $\hat{\theta}^{(i)}$ are the distribution parameters estimated from the ith Jack-knife sample. These parameters are then used to obtain the N-year event estimate for each of the n samples. The Jack-knife estimates of the N-year event have to be corrected for bias as follows

$$\tilde{x}_N = n\hat{x}_N - (n-1)\bar{x}_N \quad , \quad \bar{x}_N = \frac{1}{n}\sum_{i=1}^{n} \hat{x}_N^{(i)}, \tag{4.2.9}$$

where $\hat{x}_N$ is the N-year return level as estimated from the original sample (DHI (2003)).

Hence, the standard deviation of the Jack-knife N-year return level is given by

$$
\begin{aligned}
s_N^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_N^{(i)} - \tilde{x}_N\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\hat{x}_N^{(i)} - \{n\hat{x}_N - (n-1)\bar{x}_N\}\right]^2 \\
&= \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{x}_N^{(i)} - \bar{x}_N\right)^2 . \qquad (4.2.10)
\end{aligned}
$$

Jack-knife resampling is now applied to determine standard deviations of estimations made by the GPD when the three different methods for parameter estimation (method of maximum likelihood, method of moments, and method of L-moments) are used.

4.2.8 Jack-knife Resampling Applied to the GPD

Table 4.7 displays the standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the GPD is used to make estimations based on 7151 Jack-knife samples. This number of samples is generated, since 7151 is the number of exceedances over the threshold in the original sample and Jack-knife sample i is generated by removing data value x_i ($i = 1, 2, \ldots, 7151$) from the original sample The standard deviation is determined for each of the three different parameter estimation methods, namely, the Method of Maximum Likelihood, the Method of Moments, and the Method of L-Moments.

Table 4.7: Standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the GPD is used to make estimations based on 7151 Jack-knife samples (i.e., 7151 estimations of each return level are made). The standard deviation is determined for each of the three parameter estimation methods.

Parameter Estimation Method	**Return Level Standard Deviation**				
	5-year	**10-year**	**18-year**	**50-year**	**100-year**
Method of Maximum Likelihood	1.03 $\times 10^{-15}$	2.04 $\times 10^{-15}$	3.31 $\times 10^{-15}$	6.73 $\times 10^{-15}$	10.14 $\times 10^{-15}$
Method of Moments	11.55 $\times 10^{-15}$	16.23 $\times 10^{-15}$	21.22 $\times 10^{-15}$	32.49 $\times 10^{-15}$	42.31 $\times 10^{-15}$
Method of L-Moments	1.962 $\times 10^{-15}$	2.847 $\times 10^{-15}$	3.805 $\times 10^{-15}$	6.004 $\times 10^{-15}$	7.939 $\times 10^{-15}$

When Jack-knife resampling is used, the standard deviations (shown in Table 4.7) is significantly lower than when the Bootstrap or the Monte Carlo simulation techniques is used. This makes sense since Jack-knife samples i and $i+1$ only differ from each other by one data value (sample i contains data value x_i and not x_{i+1}, whereas sample $i+1$, on the other hand, contains data value x_{i+1}, and not x_i). In the case of the Bootstrap and Monte Carlo simulation techniques, samples differ much more.

The standard deviations for the different parameter estimation methods do, however, differ relative to each other which is similar to the cases when the Bootstrap technique and Monte Carlo simulations are used. In other words, the standard deviations when the method of maximum likelihood is used in conjunction with the GPD are, again, overall the lowest, compared to the standard deviations when one of the other two parameter estimation methods are used.

Lastly, Jack-knife resampling is applied in order to determine standard deviations of estimations made by the Weibull distribution when each of the three parameter estimation methods are used.

4.2.9 Jack-knife Resampling Applied to the Weibull Distribution

Table 4.8 displays the standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the Weibull distribution is used to make estimations based on 7151 Jack-knife samples. The standard deviation is determined for each of the three different parameter estimation methods, namely, the Method of Maximum Likelihood, the Method of Moments, and the Method of L-Moments.

Table 4.8: Standard deviations of the estimated 5-year, 10-year, 18-year, 50-year, and 100-year return levels, respectively, when the Weibull distribution is used to make estimations based on 7151 Jack-knife samples (i.e., 7151 estimations of each return level are made). The standard deviation is determined for each of the three parameter estimation methods.

Parameter Estimation Method	**Return Level Standard Deviation**				
	5-year	**10-year**	**18-year**	**50-year**	**100-year**
Method of Maximum Likelihood	5.020 $\times 10^{-15}$	6.300 $\times 10^{-15}$	7.500 $\times 10^{-15}$	9.850 $\times 10^{-15}$	11.63 $\times 10^{-15}$
Method of Moments	3.316 $\times 10^{-15}$	4.118 $\times 10^{-15}$	4.866 $\times 10^{-15}$	6.310 $\times 10^{-15}$	7.395 $\times 10^{-15}$
Method of L-Moments	0.952 $\times 10^{-15}$	1.120 $\times 10^{-15}$	1.265 $\times 10^{-15}$	1.525 $\times 10^{-15}$	1.703 $\times 10^{-15}$

Once again, the standard deviations in Table 4.8 are much lower than in the cases of the Bootstap and Monte Carlo simulation techniques. However, also as before, the standard deviations for the different parameter estimation methods do differ relatively to each other similary as when one of the other techniques is used. The method of moments yields the lowest overall standard deviations, and the method of maximum likelihood the second lowest.

The standard deviations for all of the methods are very similar to when the Weibull distribution is used (also as in the cases of the Bootstrap technique and the Monte Carlo similation). Overall the standard deviations for the Weibull distribution are lower than for the GPD. Hence, using a popular, and reliable technique for parameter estimation, namely the method of maximum likelihood, in conjunction with the Weibull distribution for estimations regarding return levels and return periods seems like a good choice.

As previously mentoined, the fact that the standard deviations in both the case of the GPD and Weibull distribution are significantly lower when the Jack-knife resampling technique is used in comparison to when either the Bootstrap or the Monte Carlo similation techniques are used, can be attributed to the fact that the 7151 Jack-knife samples differ very slightly from one another. Each sample differs from each other sample by only one value. Since the *points* over threshold technique is used, the difference in only one point over the threshold from one sample to the next has a very small effect on the estimation of return levels. However, if the *peaks* over threshold method were used, the difference might have been more significant, since the difference in one peak-value from one sample to the next can be more influential. It can therefore be concluded that the use of the Jack-knife resampling technique in conjunction with the points over threshold method is not very effective or advisable.

Chapter 4 focussed on evaluating the fits of probability distributions to the Slangkop data as well as the accuracy of the estimations made by these distributions. In Section 4.1, the χ^2-test was employed as a goodness-of-fit statistic, in order to evaluate the fits of the GPD and Weibull distribution, respectively, to the data. Section 4.2 covered techniques for determining the uncertainty of quantile estimates in order to evaluate the accuracy of estimations made by the probability distributions. Three techniques for doing so were considered and implemented, namely, the Bootstrap technique, the Monte Carlo simulation technique, and the Jack-knife resampling technique. A conclusion made from these techniques, was that that the use of the method of maximum likelihood in conjunction with the Weibull distribution seems to be the most efficient combination of parameter estimation method and probability distribution for making estimations regarding return levels and return periods. Another conclusion was that the Jack-knife resampling technique is not well suited to be used together with the points over threshold method, since generated samples differ too little from one another and, hence, estimations based on these samples differ only extremely slightly. In Chapter 5 the visual fit of the probability distributions to the Slangkop data is evaluated by means of frequency and probability plots.

Chapter 5

Frequency and Probability Plots

In this chapter, the fit of probability distributions to the empirical data is evaluated visually by considering the fit of probability density functions to histograms.

5.1 Histograms and Probability Density Functions

Histograms are plots of empirical probability density functions. It displays the number of observations in classes covering the range of the variable (Coles (2001)). The appropriateness of a certain probability distribution for making estimations based on a certain dataset can therefore be evaluated by how well the pdf of the probability distribution fits the histogram of the data.

In the following subsection, the GPD pdf is plotted with the histogram of the dataset, with its parameters estimated by the method of maximum likelihood, the method of moments, and the method of L-moments, respectively.

5.1.1 Generalized Pareto Probability Density Function

Figures 5.1.1 to 5.1.3 all show plots of the histogram of the dataset, together with the GPD pdf, with the parameters determined by the method of maximum likelihood, the method of moments, and the method of L-moments, respectively. The right-hand panel shows close-ups of the tails of the figures (i.e., it zooms in on the higher return levels, ≥ 7 m). The fit of the pdf on the higher return levels are of greatest importance, since estimations regarding extreme values are what are of interest in this study.

Figure 5.1.4 is a combination of Figures 5.1.1 to 5.1.3. It shows the histogram of the dataset, together with the GPD pdf with parameters determined by each of the three parameter estimation methods.

Based on Figures 5.1.1 to 5.1.4 it seems as though the method of maximum likelihood is the safest method to be used for GPD parameter estimation. This conclusion can be made since the GPD pdf, with parameters estimated by the method of maximum likelihood, best estimates the frequency of the higher return levels. For these return levels, the pdf intersects the histogram bars approximately in the middle, which is indicative of

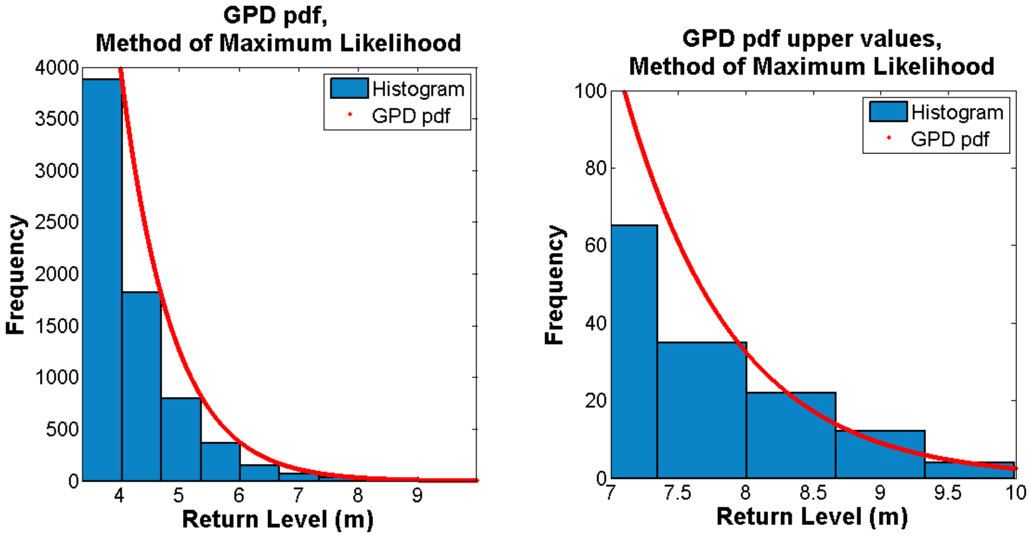


Figure 5.1.1: The GPD pdf, with parameters determined by the *method of maximum likelihood*, together with the histogram of the dataset. The figure on the right is a close-up of the higher return levels of the figure on the left.

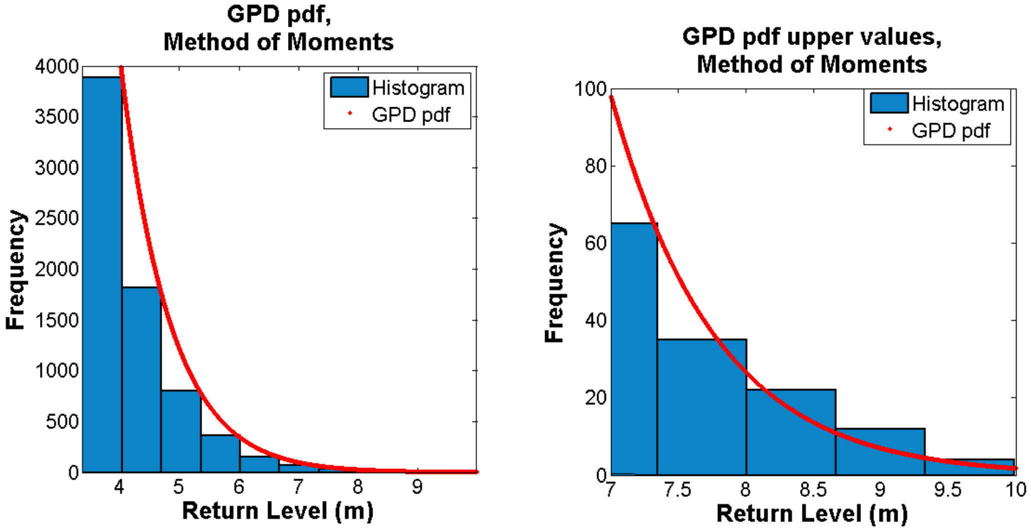


Figure 5.1.2: The GPD pdf, with parameters determined by the *method of moments*, together with the histogram of the dataset. The figure on the right is a close-up of the higher return levels of the figure on the left.

a good estimation. When one of the other two methods for parameter estimation (i.e., the method of moments or the method of L-moments) is used, the under-estimation of the frequency of the higher return levels are more likely, as can be seen in Figures 5.1.1 to 5.1.4.

The conclusion can therefore be made that when the GPD is used for return level

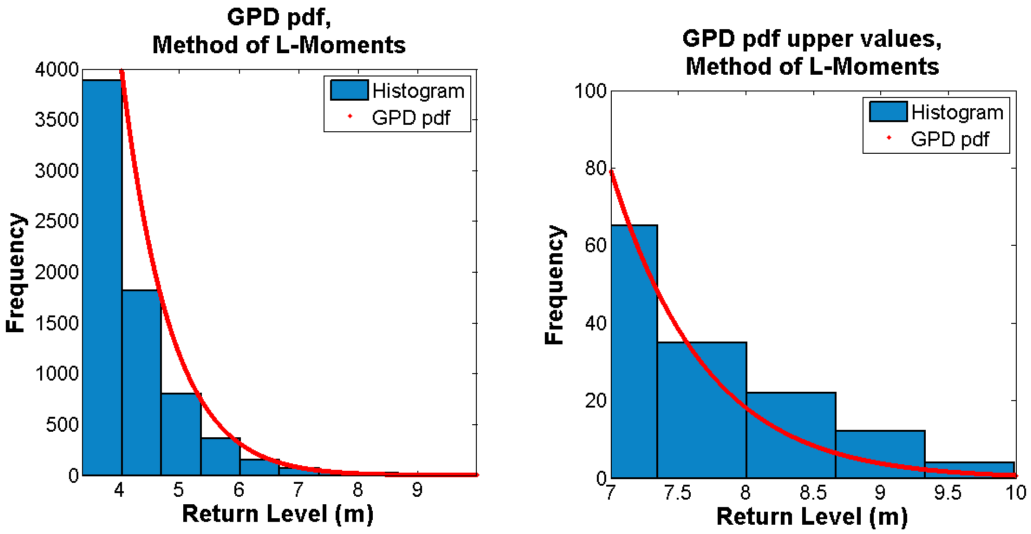


Figure 5.1.3: The GPD pdf, with parameters determined by the *method of L-moments*, together with the histogram of the dataset. The figure on the right is a close-up of the higher return levels of the figure on the left.

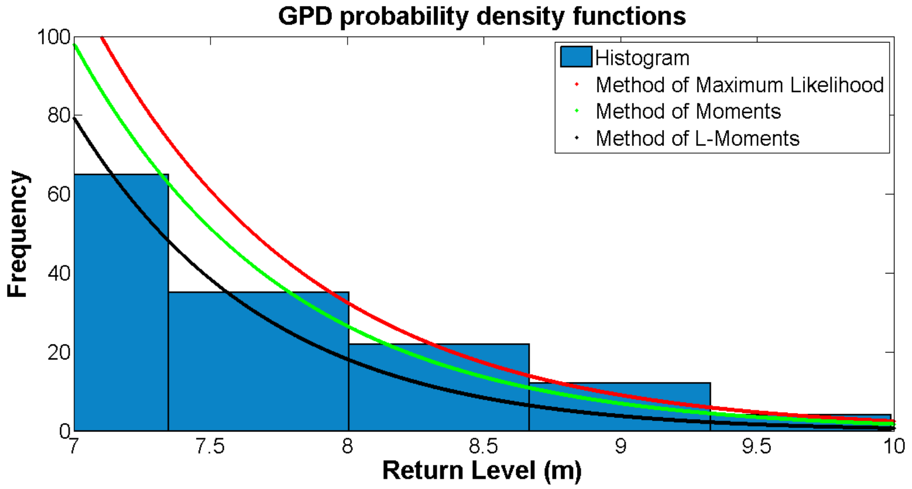


Figure 5.1.4: The GPD pdf, with parameters determined by the method of maximum likelihood, the method of moments, and the method of L-moments, respectively, together with the histogram of the dataset.

estimations, the method of maximum likelihood is the optimal method to be used for parameter estimation in order to avoid under-estimation.

In the next subsection, the Weibull pdf is plotted with the histogram of the dataset, with its parameters estimated by each of the three methods for parameter estimation.

5.1.2 Weibull Probability Density Function

Figures 5.1.5 to 5.1.7 all show plots of the histogram of the dataset, together with the Weibull pdf, with the parameters estimated by the method of maximum likelihood, the method of moments, and the method of L-moments, respectively. Each of these figures also shows a close-up of the higher return levels.

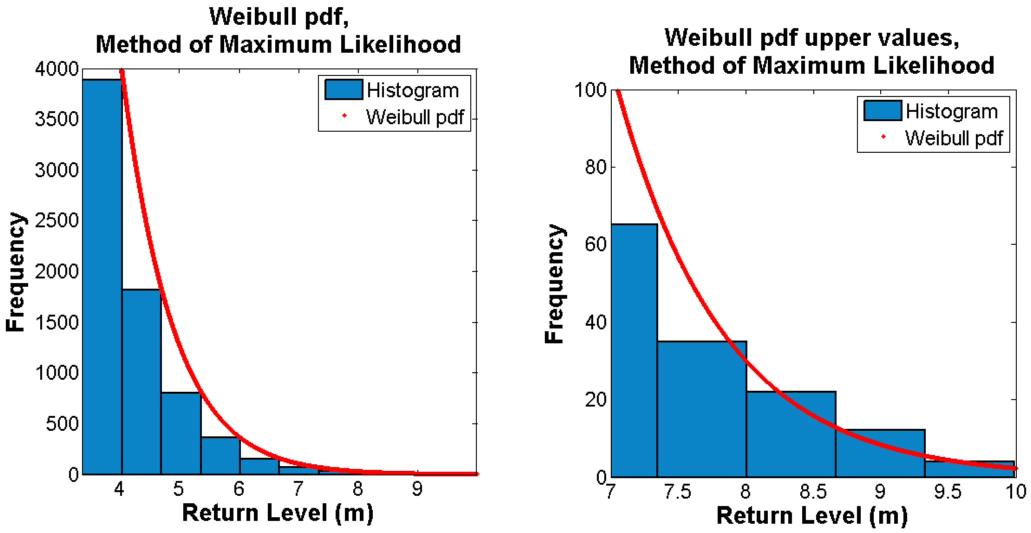


Figure 5.1.5: The Weibull pdf, with parameters determined by the *method of maximum likelihood*, together with the histogram of the dataset. The figure on the right is a close-up of the higher return levels of the figure on the left.

Figure 5.1.8 is a combination of Figures 5.1.5 to 5.1.7. It contains a comparison of the fits of the Weibull probability density functions for the three different methods of parameter estimation.

Based on Figures 5.1.5 to 5.1.8 it seems as though the method of moments is slightly safer for the estimation of Weibull parameters than the other two estimation methods when considering to avoid the under-estimation of the frequency of return levels. As can be seen in Figure 5.1.8, however, the three methods yield very similar fits of the Weibull pdf to the histogram. In each of the three cases the pdf intersects the histogram bar approximately in the middle for the higher return levels. Hence, the use of any of the three parameter estimation methods in conjunction with the Weibull distribution seems to yield reliable results.

In this chapter, the generalized Pareto and the Weibull probability density functions were plotted with histograms of the empirical data. The parameters of the probability density functions were estimated by each of three estimation methods, namely, the method of maximum likelihood, the method of moments, and the method of L-moments. A conclusion that was made, is that when the GPD distribution is used, the method of maximum likelihood is the optimal method for estimating the distribution's parameters to avoid under-estimation of return levels. In the case of the Weibull distribution, all

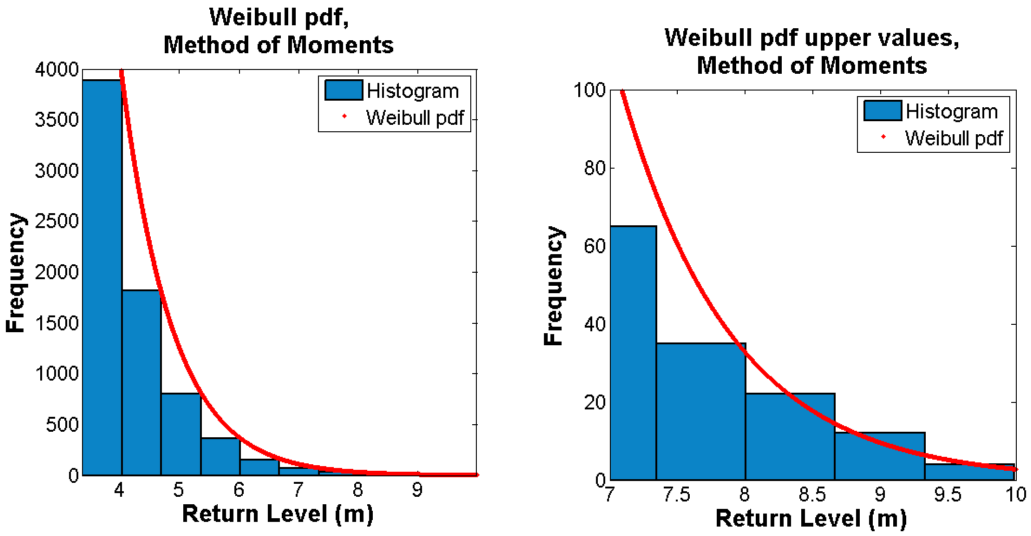


Figure 5.1.6: The Weibull pdf, with parameters determined by the *method of moments*, together with the histogram of the dataset. The figure on the right is a close-up of the higher return levels of the figure on the left.

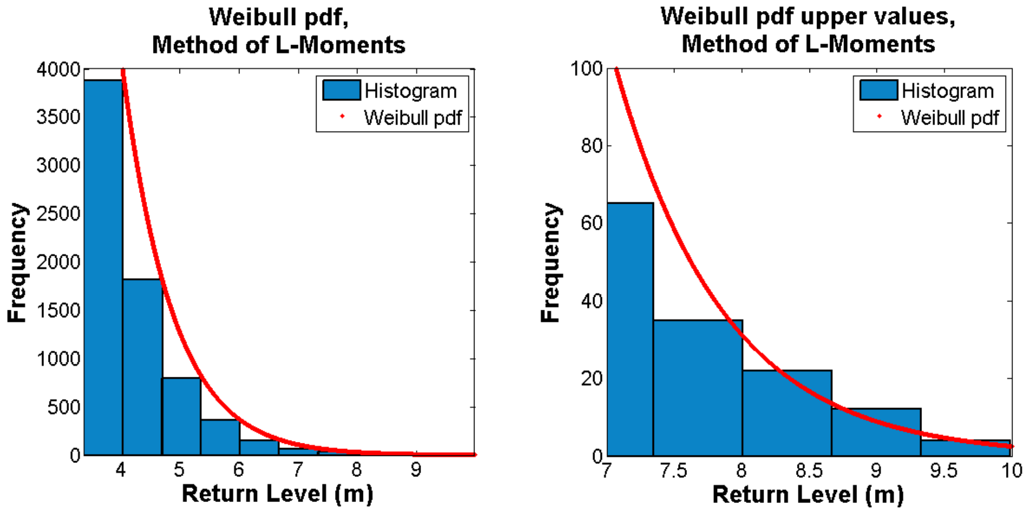


Figure 5.1.7: The Weibull pdf, with parameters determined by the *method of L-moments*, together with the histogram of the dataset. The figure on the right is a close-up of the higher return levels of the figure on the left.

three methods for parameter estimation yielded similar results, which all fitted the histograms well. Hence, the conclusion was made the the use of any of the three methods in conjunction with the Weibull distribution seems to yield reliable results.

Figure 5.1.9 contains a comparison of the GPD and Weibull probability distribution functions, when the parameters of both distributions are estimated by the method of

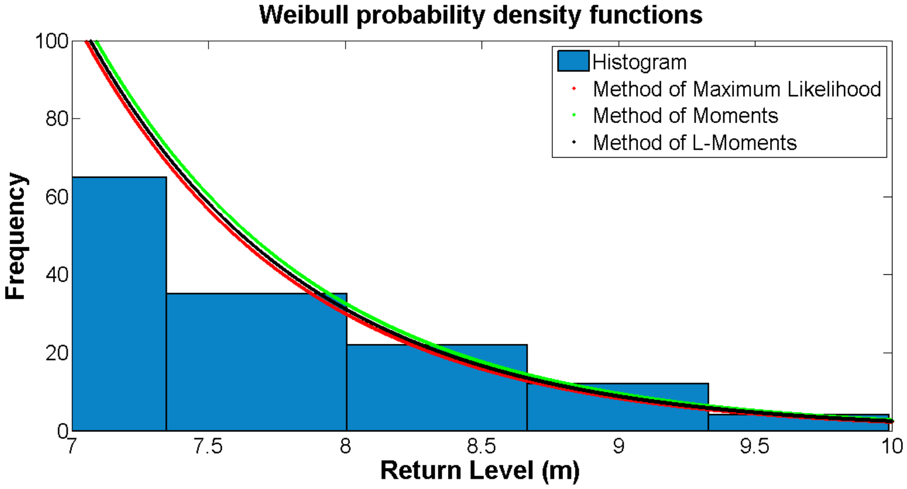


Figure 5.1.8: The Weibull pdf, with parameters determined by the method of maximum likelihood, the method of moments, and the method of L-moments, respectively, together with the histogram of the dataset.

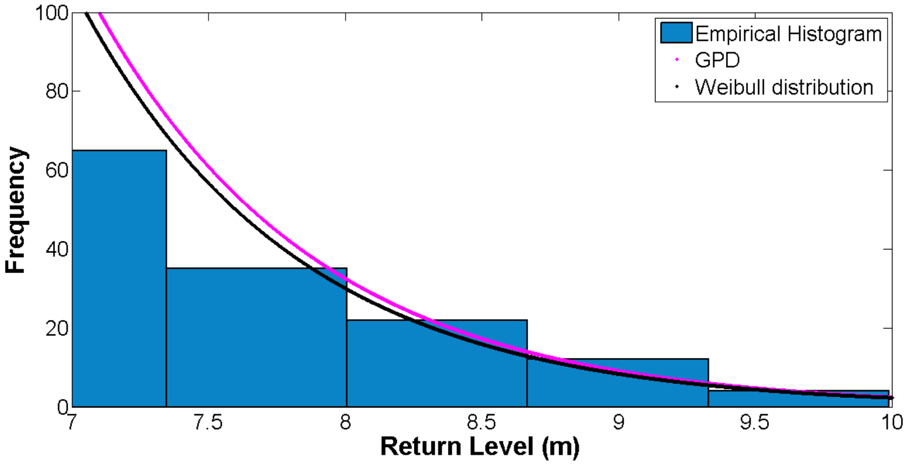


Figure 5.1.9: The GPD and Weibull probability distributions, with parameters determined by the method of maximum likelihood, together with the histogram of the dataset.

maximum likelihood. It can be seen that both distributions produce similar results, with the GPD curve being only slightly above the curve of the Weibull distribution. The higher the return levels become, the closer the curves (i.e., the probability distribution functions) move together. It can therefore be concluded that the use of both the GPD and Weibull distribution, in conjunction with the method of maximum likelihood as parameter estimation method, yield reliable return level estimations. The GPD pdf gave only slightly lower estimations than the Weibull pdf.

The next subject to consider, is the fact that the dataset used in this study contains missing values. In the next chapter, the handling of these missing values (i.e., “gaps”) will be discussed.

Chapter 6

Gaps in the Dataset

Many datasets have periods where data are absent which are referred to as gaps in the data (Burke (2001)). In particular, the dataset used in this study (i.e., the Slangkop dataset) has gaps due to the Datawell bouy not having been able to take measurements at certain points in time (for example when the bouy was removed for maintenance). Figure 6.0.1 shows plots of the H_{mo}s for the years 2001 to 2003. The red circles in the plot of the year 2001 indicate positions where absent data values are clearly visible. The other years also have absent data, but they do not have as many consecutive missing values as to make it visible when considering their H_{mo} plots.

All the analyses done on the dataset therefore far, was done with the replacement of the absent data (or gaps) by zeros. This was done in order to keep the duration of the dataset fixed. Different results would have been obtained if the zeros were removed and the length of the dataset shortened. However, the spesific values of estimations based on the dataset were not of particular interest therefore far, but the focus was rather on the *methods used* to make these estimations.

There are, however, much more sophisticated methods for the treatment of missing data. These methods includes casewise deletion, mean substitution and imputation and will be discussed briefly in the following sections.

6.1 Treatment of Gaps in the Dataset

6.1.1 Casewise Deletion

Casewise deletion excludes all cases that have missing data (Burke (2001)). In other words, it excludes the points in time where no measurements were made and as a result the length of the time series is also reduced. An example of casewise deletion is given in Table 6.1.

Casewise deletion may be a reasonable solution to a missing data problem if only a small fraction of the data is missing (Schafer (1997) gives five percent or less as an example). Otherwise casewise deletion may be ineffecient as it will tend to introduce bias (Schafer (1997)). The next possible method considered for dealing with missing

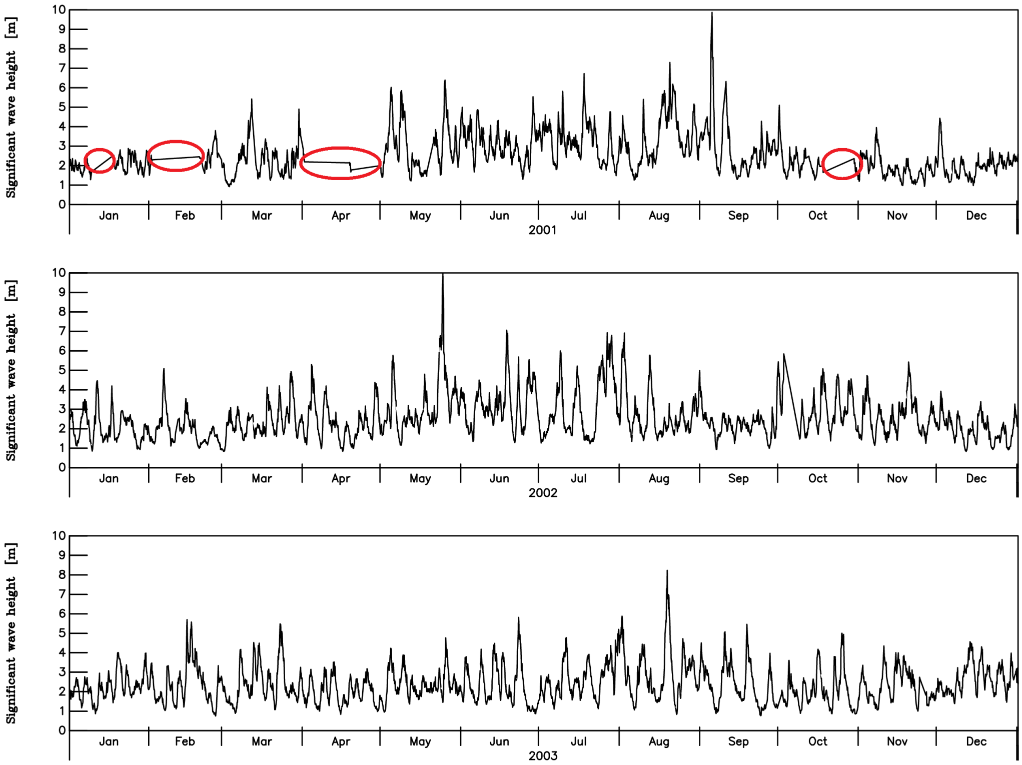


Figure 6.0.1: H_{mo}s from the Slangkop dataset for the years 2001 to 2003. The red circles in the year 2001 indicate clearly visible gaps in the dataset.

Table 6.1: A simple illustration of casewise deletion for a dataset containing one variable.

Date	**Time**	$\mathbf{H_{mo}}$
1994/06/23	03:00	4.72
1994/06/23	06:00	4.89
1994/06/23	09:00	5.79
1994/06/23	12:00	-
1994/06/23	15:00	4.93
1994/06/23	18:00	4.33
1994/06/23	21:00	3.90
1994/06/24	00:00	4.23
1994/06/24	03:00	4.90
1994/06/24	06:00	-
1994/06/24	09:00	4.84
1994/06/24	12:00	4.90
1994/06/24	15:00	5.14
1994/06/24	18:00	5.26

→

Date	**Time**	$\mathbf{H_{mo}}$
1994/06/23	03:00	4.72
1994/06/23	06:00	4.89
1994/06/23	09:00	5.79
1994/06/23	15:00	4.93
1994/06/23	18:00	4.33
1994/06/23	21:00	3.90
1994/06/24	00:00	4.23
1994/06/24	03:00	4.90
1994/06/24	09:00	4.84
1994/06/24	12:00	4.90
1994/06/24	15:00	5.14
1994/06/24	18:00	5.26

data is mean substitution.

6.1.2 Mean Substitution

Mean substitution replaces all missing data by the mean of the observed data. An example of this method is given in Table 6.2.

Table 6.2: Illustration of mean substitution.

Date	**Time**	$\mathbf{H_{mo}}$
1994/06/23	03:00	4.72
1994/06/23	06:00	4.89
1994/06/23	09:00	5.79
1994/06/23	12:00	-
1994/06/23	15:00	4.93
1994/06/23	18:00	4.33
1994/06/23	21:00	3.90
1994/06/24	00:00	4.23
1994/06/24	03:00	4.90
1994/06/24	06:00	-
1994/06/24	09:00	4.84
1994/06/24	12:00	4.90
1994/06/24	15:00	5.14
1994/06/24	18:00	5.26

→

Date	**Time**	$\mathbf{H_{mo}}$
1994/06/23	03:00	4.72
1994/06/23	06:00	4.89
1994/06/23	09:00	5.79
1994/06/23	12:00	**4.82**
1994/06/23	15:00	4.93
1994/06/23	18:00	4.33
1994/06/23	21:00	3.90
1994/06/24	00:00	4.23
1994/06/24	03:00	4.90
1994/06/24	06:00	**4.82**
1994/06/24	09:00	4.84
1994/06/24	12:00	4.90
1994/06/24	15:00	5.14
1994/06/24	18:00	5.26

Disadvantages of mean substitution include that it decreases the variability in the dataset in direct proportion to the number of missing data points. This leads to underestimation of dispersion (the spread of the data). Another disadvantage of this method is that it may also change the values of some other statistics considerably, such as linear regression statistics (Burke (2001), Burke (1998)). Mean substitution preserves the observed sample mean, but it biases estimated variances and covariances toward zero (Schafer (1997)).

6.1.3 Imputation

Another method that can be used to handle missing data is imputation. Mean substitution is a simple example of imputation. In a more general form, imputed missing values are predicted from patterns in the real (non-missing) data (Burke (2001)). For each missing value in the dataset, m possible imputed values are calculated (using a suitable statistical model derived from patterns in the data). This leads to obtaining m complete datasets (each set obtained by one of m statistical models). The m possible complete datasets are, in turn, analysed by the selected statistical method. The m intermediate results are then combined to yield the final result (statistic) and an estimate of its uncertainty.

Imputation, by using regression models to predict values, may inflate observed correlations and bias them away from zero. Devising an imputation scheme that preserves important aspects of the variables can also be very complicated. Even if the joint distribution of all variables could be adequetly preserved, standard errors, p-values and other measures of uncertainty could be misleading, because they do not reflect any uncertainty due to missing data (Schafer (1997)).

There are, however, methods for the analysis of incomplete multivariate data that account for the missing values, and the uncertainty they introduce, at each step of the analysis in a formal way. These methods, unlike the three methods already mentioned (case deletion, mean substitution and imputation), do not simply modify the data in an ad hoc manner to make them appear complete. An example of such a method is the Expectation-Maximization (EM) Algoritm (Schafer (1997)). However, these methods will not be elaborated on in this study.

When doing extreme value analysis on a dataset, the number of data values used is already not large, since only the extreme values (i.e., high values, defined by a specified method) are extracted from the dataset. If a dataset contains missing data, the available number of extreme values may be even smaller than it would have been in the case of a complete dataset. In the rest of this chapter the effect of gaps on the reliability of estimations made based on an incomplete dataset will be considered.

6.2 Effect of Gaps on Estimations of Extreme Values

The dataset used in this study contains various time periods where data are absent. An example of periods of absent data for 1994 is presented in Table 6.3. Previously, all missing data values, indicated by a dash, were replaced by zeros. This was done since the values of specific estimations made by different methods were not of as much interest as the methods themselves.

To evaluate the effects of gaps on the dataset of wave data, a complete dataset is firstly required. An NCEP (National Centre for Environmental Prediction) hindcast dataset is used for this purpose. This dataset covers a time period of approximately 17 years, which is close to the length of the Slankop dataset of approximately 18 years. The extraction point of the data is at (Latitude; Longitude) = (34°South; 17.5°East). This is the closest NCEP output point to the Slangkop Waverider buoy. Figure 6.2.1 shows the position of this output point as well as the position of the Slangkop buoy. This complete dataset is then used to create different incomplete datasets by removing data values from it in one of the following ways:

1. Remove $p\%$ of the data values from the dataset as one connected/single block (or 'chunk') from a certain position in the dataset.

2. Remove $p\%$ of the data values from the dataset in the form of randomly spread 1-week periods (i.e., the sum of all the 1-week periods makes up $p\%$ of the dataset). The choice of 1-week periods is based on the subjective assumption that 1 week is the longest period it would take to service or repair a buoy.

Table 6.3: Illustration of incomplete dataset. A missing data value is indicated by a dash.

Date	Time (hh:mm)	H_{mo}(m)
1994\06\08	00:00	3.05
1994\06\08	03:00	3.71
1994\06\08	06:00	–
1994\06\08	09:00	–
1994\06\08	12:00	–
1994\06\08	15:00	–
1994\06\08	18:00	–
1994\06\08	21:00	–
1994\06\09	00:00	–
1994\06\09	03:00	–
1994\06\09	06:00	–
1994\06\09	09:00	4.28
1994\06\09	12:00	3.82
1994\06\09	15:00	3.76
1994\06\09	18:00	3.21
1994\06\09	21:00	3.14
1994\06\10	00:00	2.60
1994\06\10	03:00	2.55
1994\06\10	06:00	–
1994\06\10	09:00	–
1994\06\10	12:00	2.50

An illustration of removing a 10% connected 'chunk' (method 1 above) from the NCEP dataset is given in Figure 6.2.2. Figure 6.2.3 is an illustration of creating gaps in the dataset by removing randomly spread 1-week periods (method 2 above) from the year 1998.

In order to evaluate how gaps of different sizes and spreads influence estimations of return levels associated with specified return periods, histograms (representing different cases) are plotted. These cases are summarised in Table 6.4.

6.2.1 Single Block Gaps

Return values were calculated based on a dataset with a particular gap size, in the form of one, connected block. The gap was then randomly moved to produce another set of return levels. This procedure was repeated 1000 times. Figures 6.2.4 to 6.2.8 show histograms that were generated as follows: The first histogram in each of the figures (in the top, left-hand corner, entitled '10% gap - NCEP data') displays 1000 return levels associated with a return period of 5 years, 10 years, 17 years (chosen since the length of the complete dataset is approximately 17 years), 50 years, and 100 years, respectively. These estimations are made by means of the POT method and the Weibull distribution

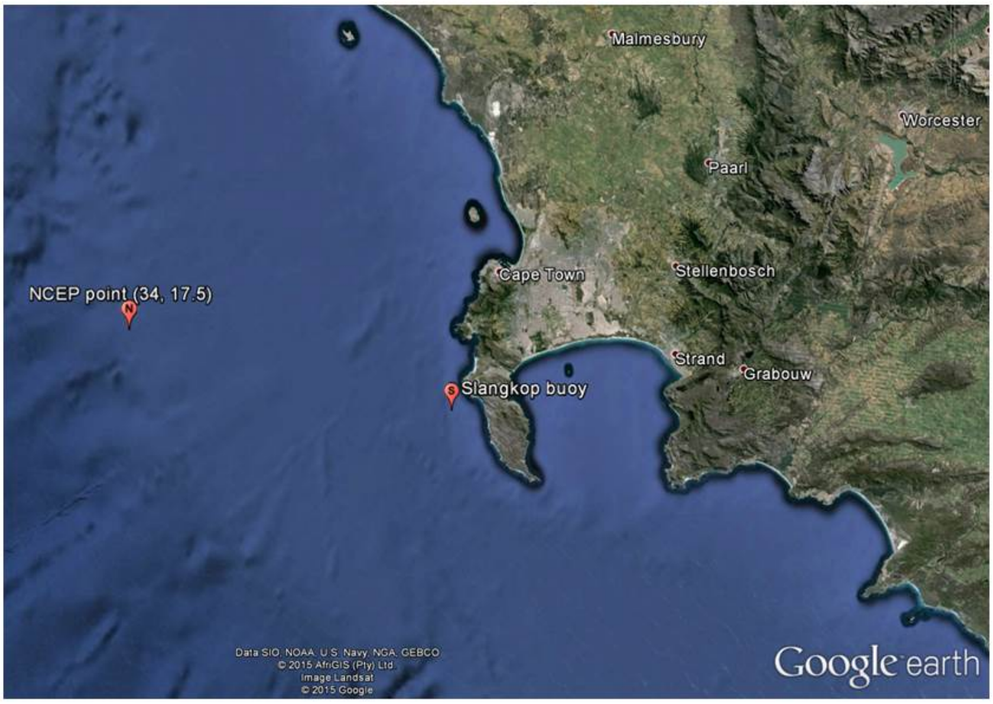


Figure 6.2.1: Position of the NCEP output point as well as the Slangkop buoy. [Source: Google earth]

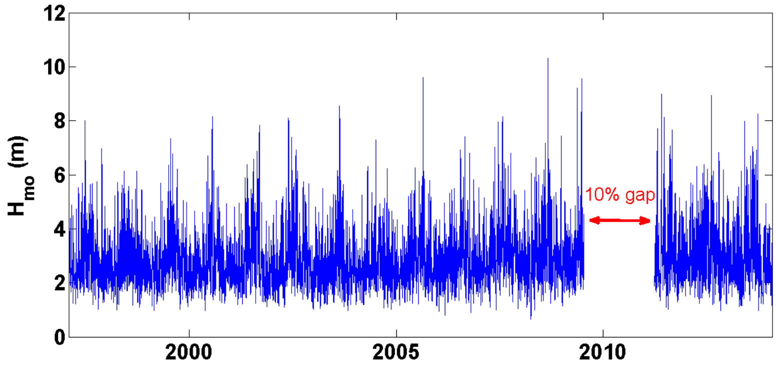


Figure 6.2.2: H_{mo}-values of the NCEP-dataset with a 10% gap created in the dataset, indicated by the red arrow.

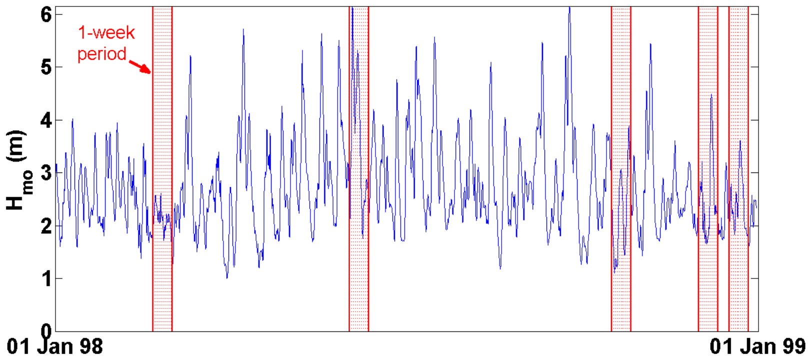


Figure 6.2.3: H_{mo}-values of the NCEP-dataset with gaps created in the dataset in the form of randomly spread 1-week periods during 1998. The 1-week periods are indicated by the red columns and add up to 10% of the dataset.

Table 6.4: Datasets with gaps.

Dataset	Type of Gap	Percentage Gap	Return Level	Figure
NCEP-dataset	Chunk \ Block	10%; 20%; 30%; 40%; 50%, 60%; 70%; 80%; 90%	5 year	6.2.4
			10 year	6.2.5
			17 year	6.2.6
			50 year	6.2.7
			100 year	6.2.8
	Random 1-week gaps	10%; 20%; 30%; 40%; 50%, 60%; 70%; 80%; 90%	5 year	6.2.18
			10 year	6.2.19
			17 year	6.2.20
			50 year	6.2.21
			100 year	6.2.22

along with the method of maximum likelihood (used for parameter estimation) based on the NCEP data with a 10% chunk of data values removed from the set. The 10% chunk removed from the first dataset makes up the first 10% of the set. For each new dataset, the 10% chunk is translated to the right in such a way that the last (1000th) 10% chunk removes the last 10% of the dataset (accuracy of this may vary slightly based on the size of the dataset and rounding). A more detailed explanation of this procedure, along with Matlab code for the removal of single, block gaps, is given in Appendix E.1. The red, dashed lines on the histograms indicate the estimated specified return level, based on the complete NCEP-dataset.

The rest of the histograms in Figures 6.2.4 to 6.2.8 are generated in the same manner as explained in the previous paragraph, except that the gap sizes increase from 20% up to 90%. When considering the first histogram in Figures 6.2.4 to 6.2.8 (i.e., the cases where 10% blocks of data are removed from the dataset) the estimations of return levels based on the incomplete datasets show relatively large deviations from the estimations of return levels based on the complete dataset. For example, in the case of the estimation of the 5-year return level (in Figure 6.2.4) the range of the estimations is [10.2; 10.7]. In line with one's intuition, the deviations of the return level estimations based on incomplete datasets from the estimated return levels based on the complete dataset increases as the size of the gap increases. The histograms also form no specific shape, such as, for example, a normal distribution. In a normal distribution, the return level frequencies closer to the return level based on the complete dataset will be higher and the frequencies of return levels will decrease as these levels become further form the return level based on the complete dataset. In other words, the spread of the return level frequencies does not resemble a normal distribution or any specific distribution, for that matter, in any of the cases.

Another observation that can be made, is that the histograms in Figures 6.2.4 to 6.2.8 all follow the same trend. As the gap sizes increase, the bars of the histograms spread out more. In other words, with an increase in gap-size, the range of the estimations increases and, hence, the deviations from the estimations based on the complete dataset increase.

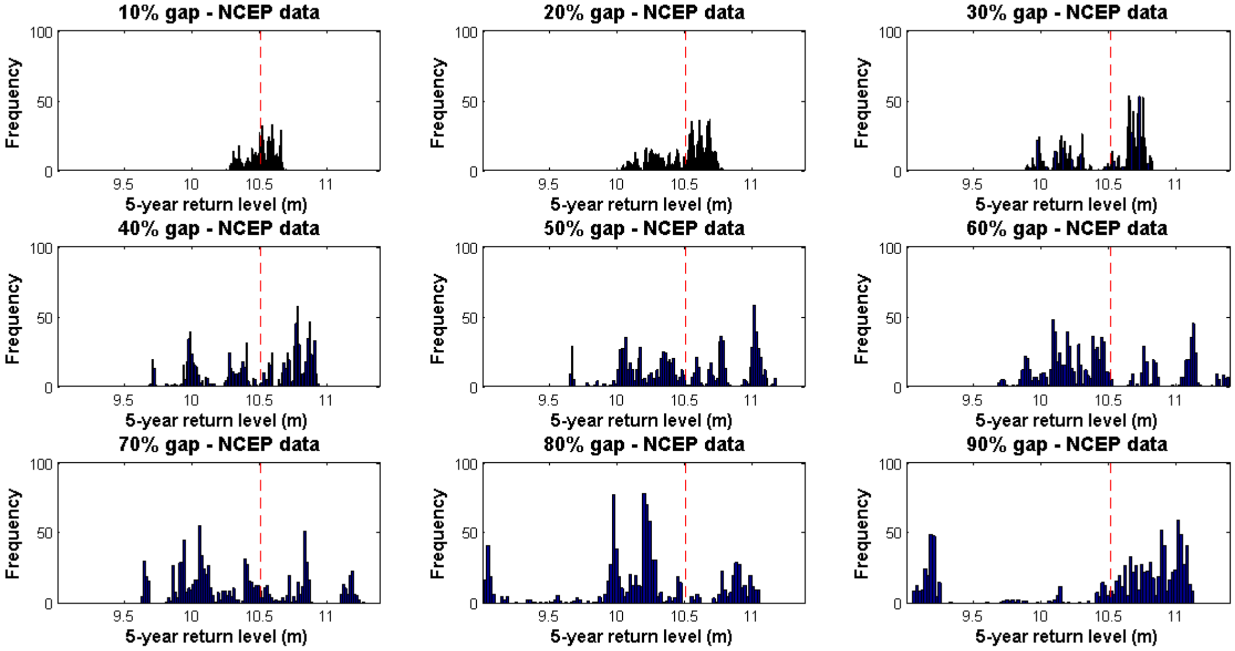


Figure 6.2.4: Histograms of 1000 estimated 5-year return levels when different percentage 'chunks' of data values are removed from the dataset. The red, dashed lines indicate the estimated 5-year return level based on the complete NCEP-dataset.

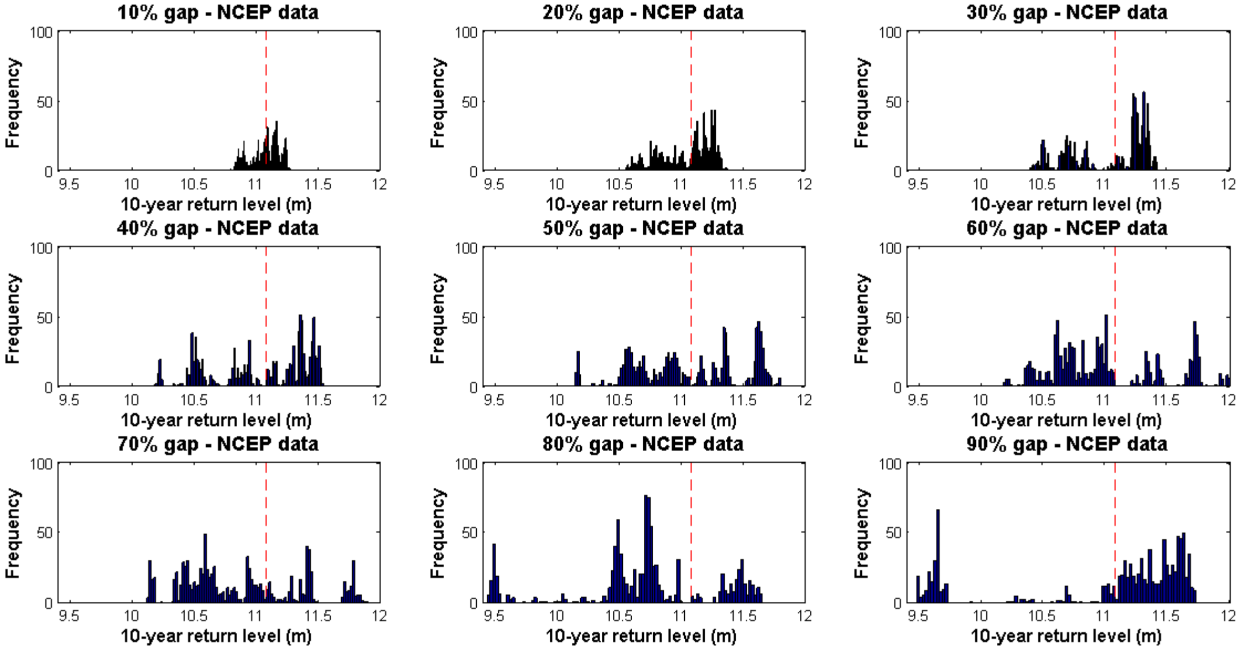


Figure 6.2.5: Histograms of 1000 estimated 10-year return levels when different percentage 'chunks' of data values are removed from the dataset. The red, dashed lines indicate the estimated 10-year return level based on the complete NCEP-dataset.

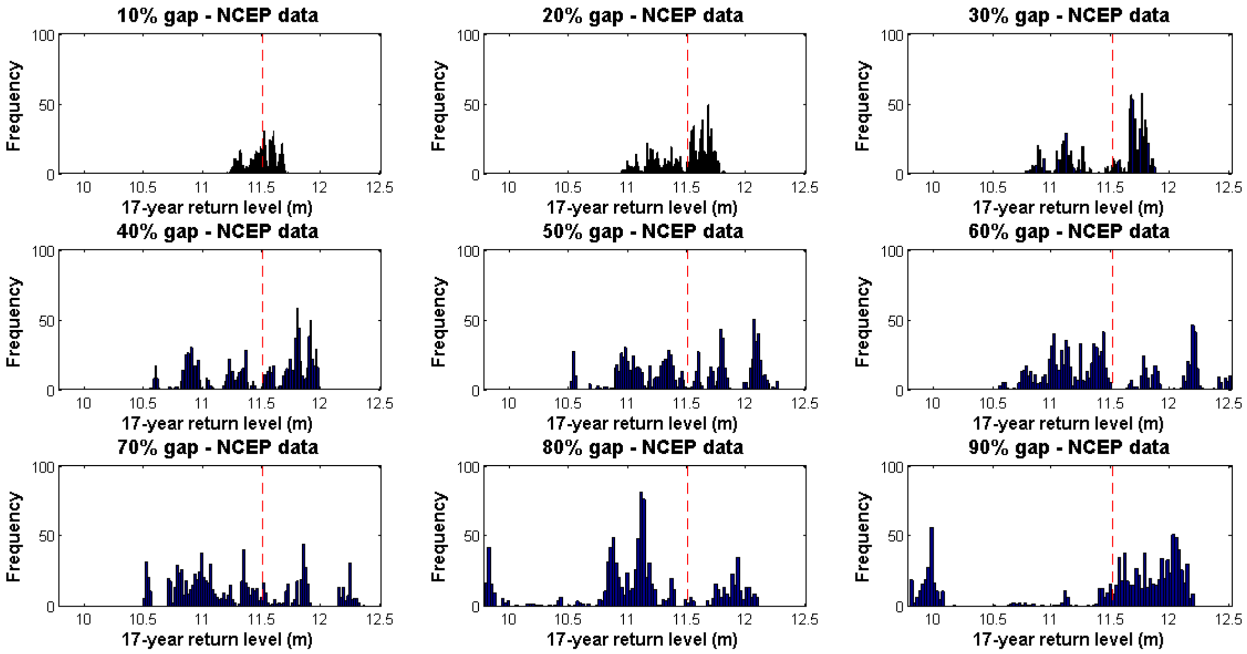


Figure 6.2.6: Histograms of 1000 estimated 17-year return levels when different percentage 'chunks' of data values are removed from the dataset. The red, dashed lines indicate the estimated 17-year return level based on the complete NCEP-dataset.

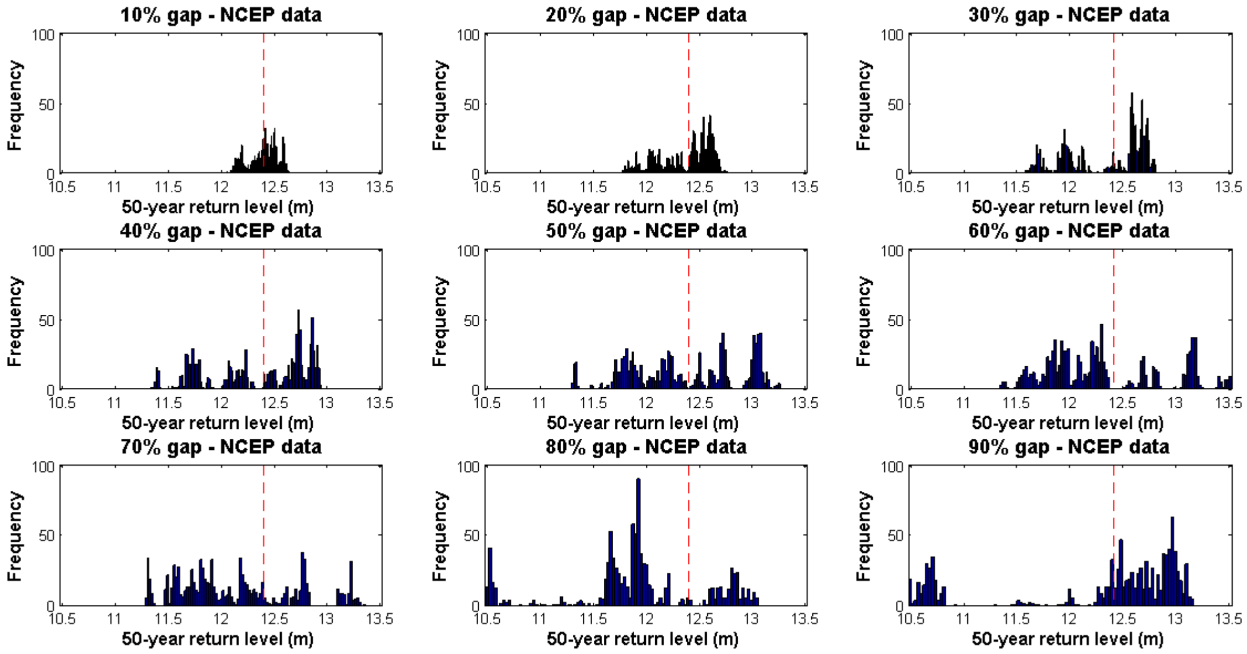


Figure 6.2.7: Histograms of 1000 estimated 50-year return levels when different percentage 'chunks' of data values are removed from the dataset. The red, dashed lines indicate the estimated 50-year return level based on the complete NCEP-dataset.

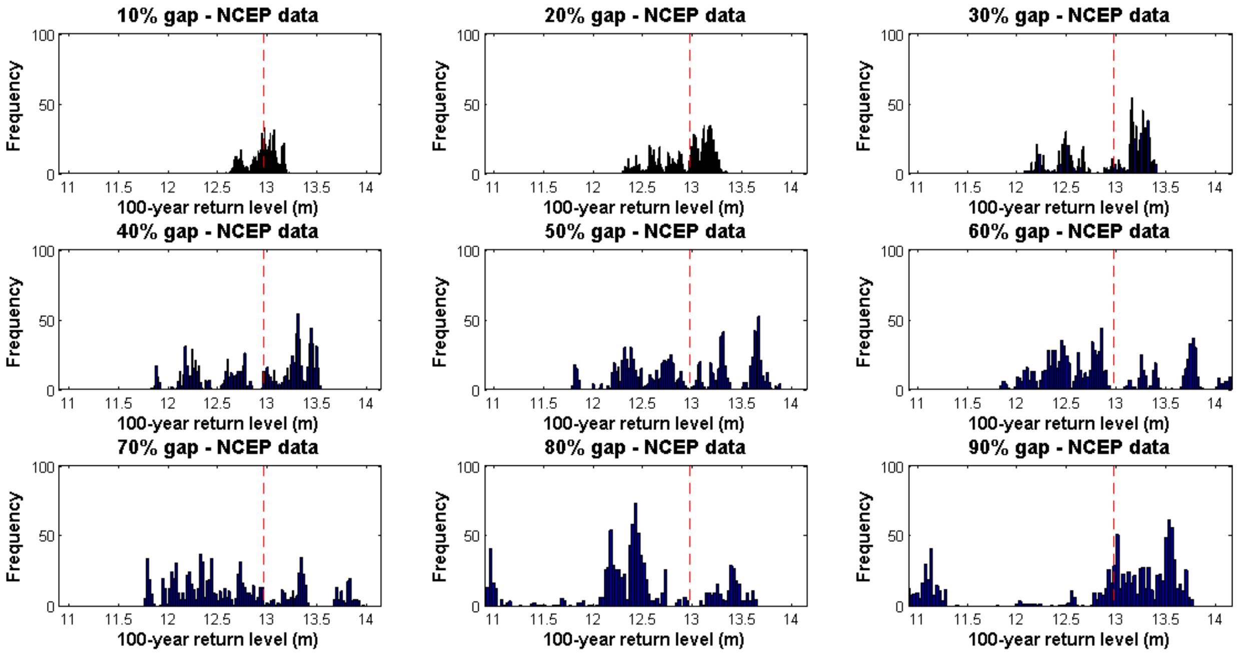


Figure 6.2.8: Histograms of 1000 estimated 100-year return levels when different percentage 'chunks' of data values are removed from the dataset. The red, dashed lines indicate the estimated 100-year return level based on the complete NCEP-dataset.

To aid interpretation of the spread of the return levels in the histograms in Figures 6.2.4 to 6.2.8 a few plots of the Weibull distribution (parameters determined by the method of maximum likelihood) together with the actual data values, as well as plots of the gradient of the Weibull distribution are made for three cases where a $p\%$ block of the data is removed. It is made for when the first $p\%$ of the data values in the dataset is removed, when the middle $p\%$ is removed, and when the last $p\%$ is removed ($p = 10, 30, 50, 70,$ and 90). This is shown in Figures 6.2.9 to 6.2.13. In each of these figures, the H_{mo}-plot in the top row corresponds to the return level and gradient plots directly below it.

In all of these figures, it can be seen that when the first $p\%$ of the data values are removed from the dataset (refer to the left-hand panels in Figures 6.2.9 to 6.2.13), the gradients (indicating the pace of increase in return periods for an increase in the return levels) of the Weibull distributions based incomplete datasets are lower than its gradient based complete dataset for return levels greater than 10 m. Also, the greater the size of the gap becomes, the lower the gradients based on the incomplete sets become. The smaller the gradient, the lower the pace of increase in return periods for an increase in the return levels, i.e., the higher the return levels associated with specified return periods. This leads to the conclusion that when the first $p\%$ of data values are removed, over-estimations of return levels based on the incomplete datasets are likely, as well as that the greater the gap-size becomes, the greater the over-estimations. The exact opposite is however true when the last $p\%$ (refer to the right-hand panels in Figures 6.2.9 to 6.2.13) of data values are removed from the dataset. The gradients of the Weibull distribution based on the incomplete datasets are then all higher than its gradient based on the complete dataset for return levels higher than 10 m, and the gradients increase with an increase in gap size. Hence, this leads to the conclusion that when the last $p\%$ of data values are removed, under-estimations of return levels based on the incomplete datasets are likely, as well as that the greater the gap-size becomes, the greater the under-estimations become.

When the middle $p\%$ (refer to the middle panels in Figures 6.2.9 to 6.2.13) of data values are removed from the dataset, the gradients of the Weibull distributions based on the incomplete datasets are similar to its gradient based on the complete dataset for gaps of sizes up to 50%. For a 70% gap, the gradient based on the incomplete dataset is higher than the one based on the complete dataset, and for a 90% gap, the gradient based on the incomplete dataset is only slightly lower than the one based on the complete dataset. There is no trend in the gradient when the middle $p\%$ of the data values are removed.

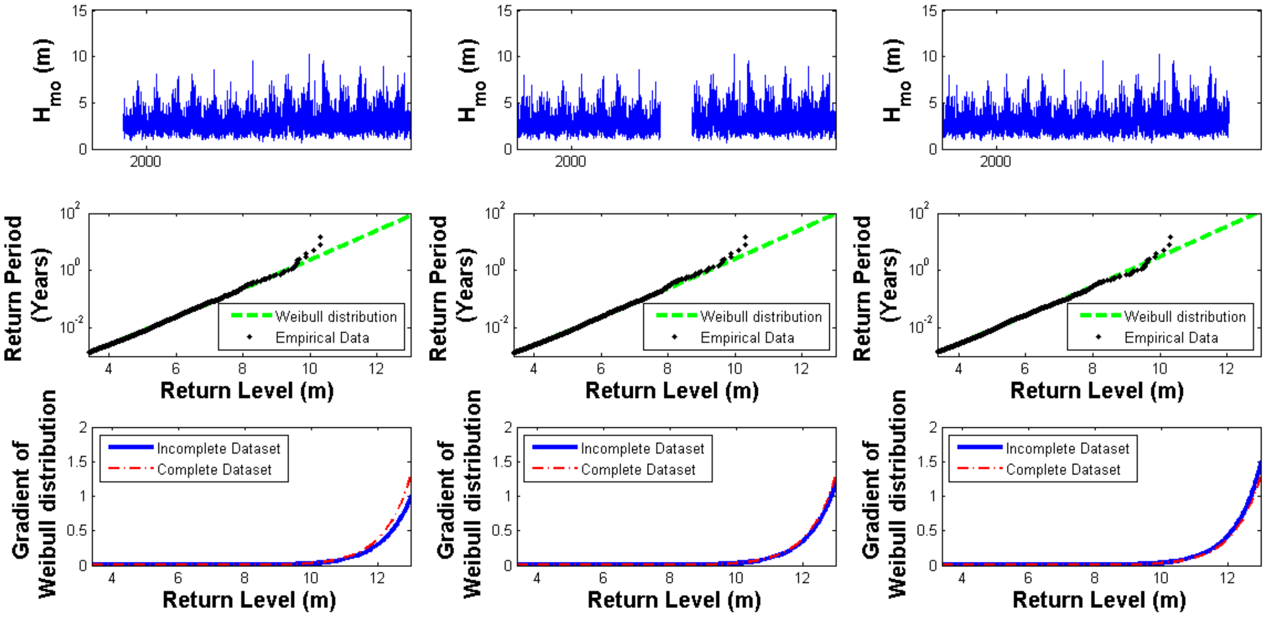


Figure 6.2.9: LEFT: First 10% of data values removed from dataset; MIDDLE: middle 10% of data values removed from dataset; RIGHT: last 10% of data values removed from dataset.

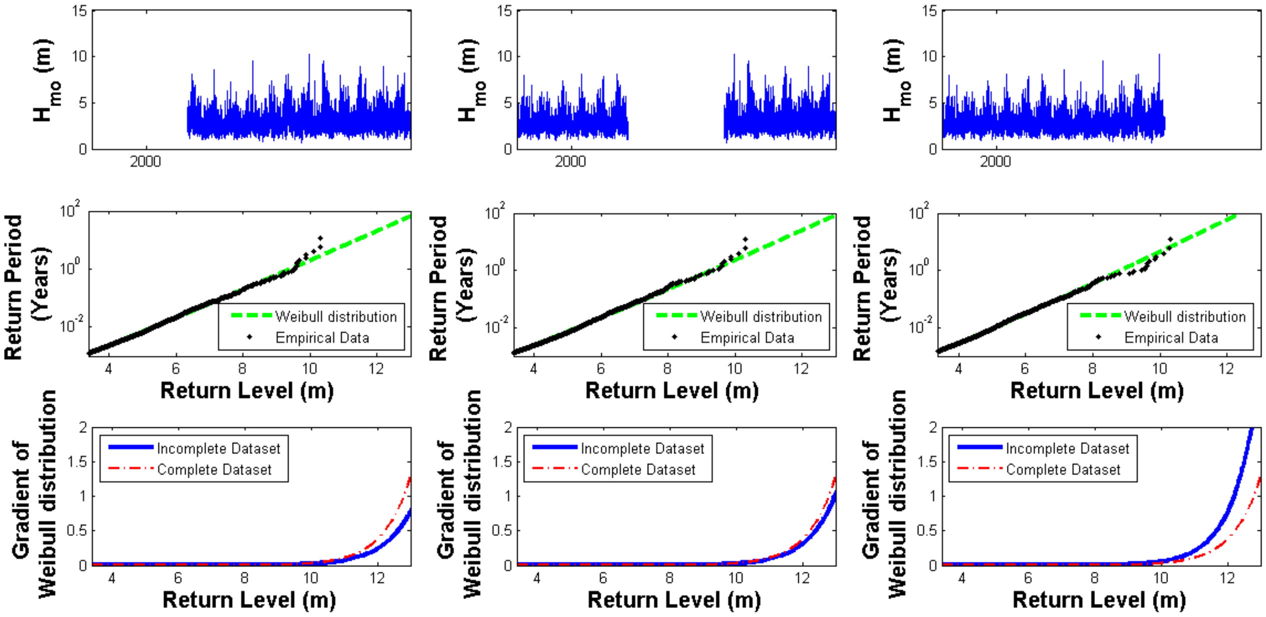


Figure 6.2.10: LEFT: First 30% of data values removed from dataset; MIDDLE: middle 30% of data values removed from dataset; RIGHT: last 30% of data values removed from dataset.

A conculsion that can be drawn based on observations made from Figures 6.2.9 to 6.2.13, is that the average of the points over threshold of the last few years covered by the NCEP data, are overall higher than the average of the points over threshold of the first few years. This is justified by the fact that when the first p% of data values were removed from the NCEP dataset, return levels based on the incomplete datasets over-estimated the return levels based on the complete dataset, whereas when the last p% of data values were removed from the NCEP dataset, return levels based on the incomplete datasets under-estimated those return levels. In other words, it seems as though more extreme values were observed in the later years covered by the NCEP data than in the earlier years. This finding is, in fact, proven true when calculating the average of the removed POTs when the

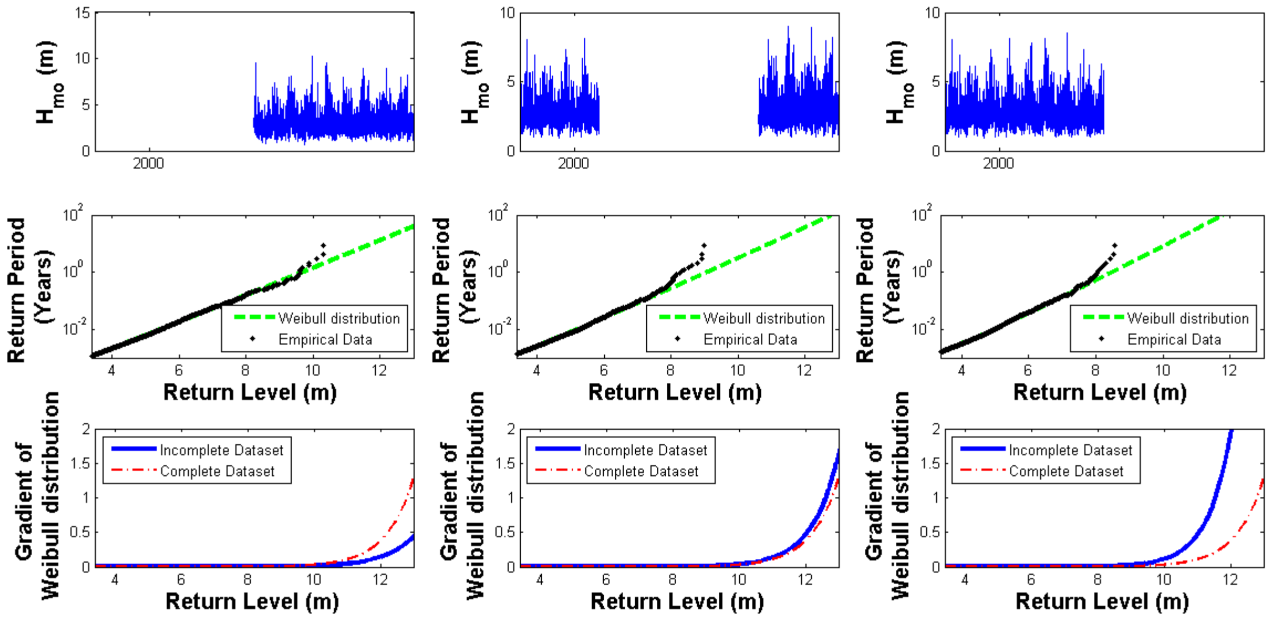


Figure 6.2.11: LEFT: First 50% of data values removed from dataset; MIDDLE: middle 50% of data values removed from dataset; RIGHT: last 50% of data values removed from dataset.

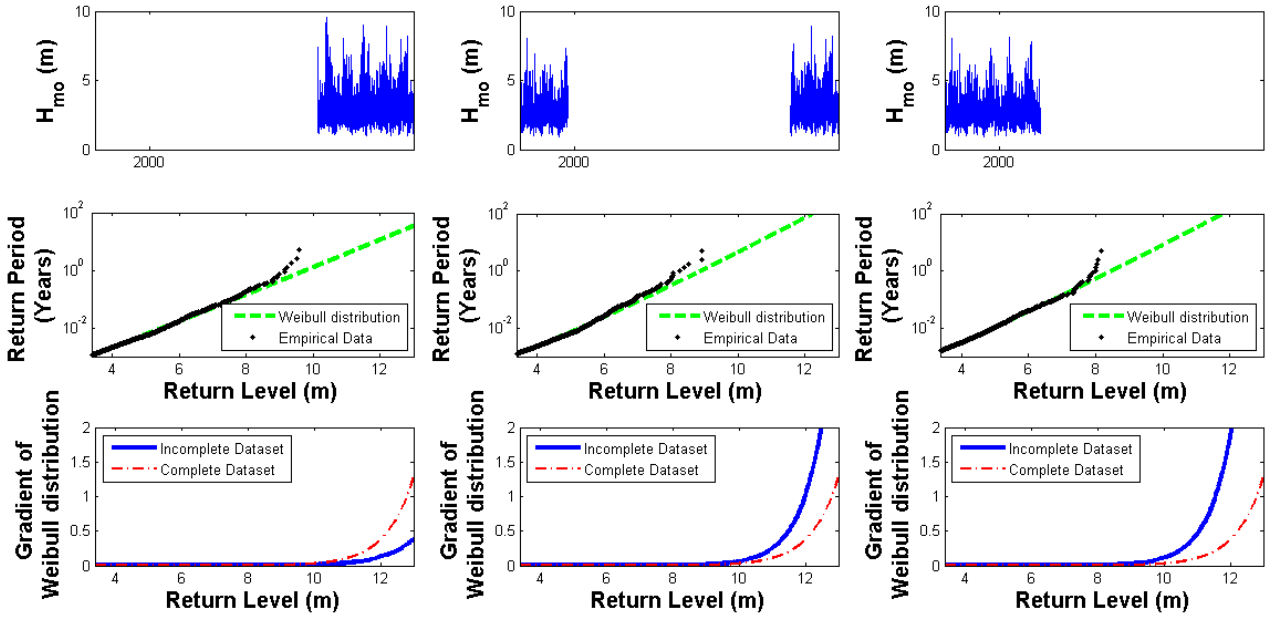


Figure 6.2.12: LEFT: First 70% of data values removed from dataset; MIDDLE: middle 70% of data values removed from dataset; RIGHT:last 70% of data values removed from dataset.

first p% of the data values are removed and comparing it to the average of the removed POTs when the last p% of the data values are removed, as shown in Table 6.5.

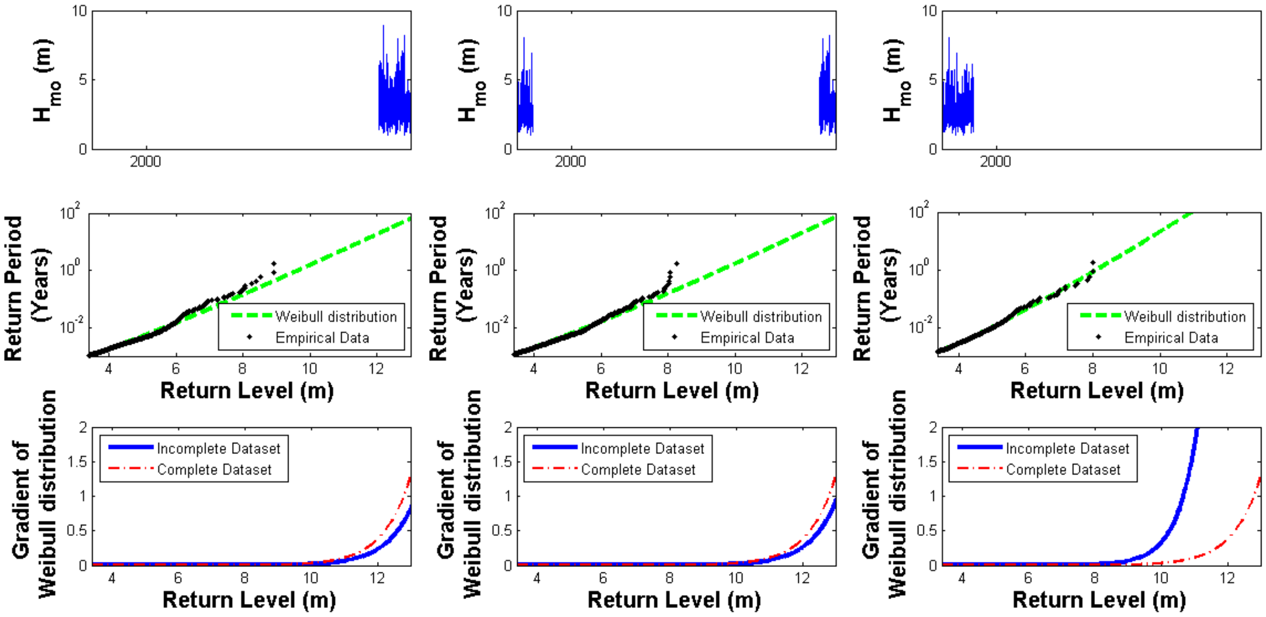


Figure 6.2.13: LEFT: First 90% of data values removed from dataset; MIDDLE: middle 90% of data values removed from dataset; RIGHT: last 90% of data values removed from dataset.

Table 6.5: Average of removed POTs when first and last p% of data values, respectively, are removed from NCEP data.

p	Average of Removed POTs	
	First p% of data values removed	Last p% of data values removed
10	4.2301	4.4254
30	4.2473	4.3652
50	4.2580	4.3456
70	4.2777	4.3294
90	4.2917	4.3169

The Weibull distribution based on the complete NCEP-dataset, together with the actual data values of the complete dataset are plotted in Figure 6.2.14.

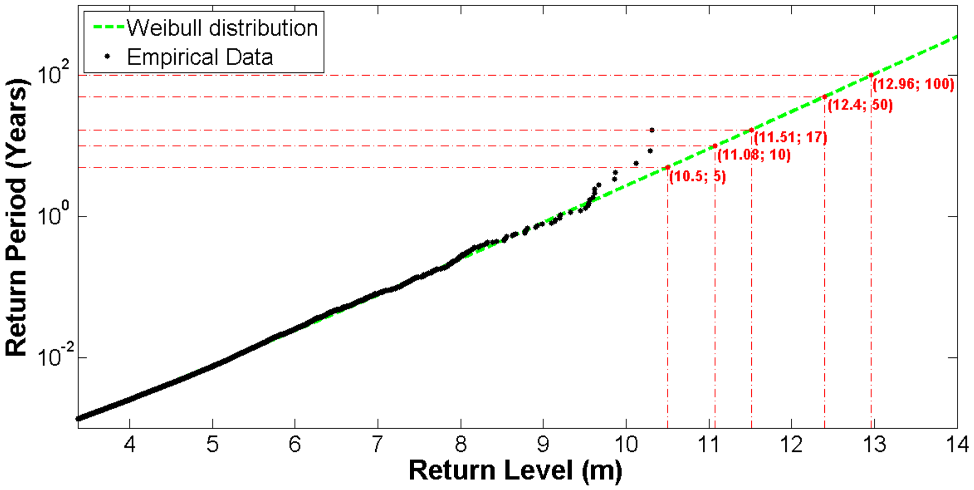


Figure 6.2.14: Weibull distribution based on the complete NCEP-dataset, together with the actual data values of the complete dataset.

Figure 6.2.15 shows the averages of the 1000 estimated 5-year, 10-year, 17-year, 50-year and 100-year return levels, respectively (corresponding to the histograms in Figures 6.2.4 to 6.2.8), based on datasets with gaps ranging from 10% up to 90%. In other words, the following is plotted for a dataset with a p% gap ($p = 10, 20, 30, \ldots, 90$): The average of 1000 estimated 5-year return levels; the average of 1000 estimated 10-year return levels; the average of 1000 estimated 17-year return levels; the average of 1000 estimated 50-year return levels; and the average of 1000 estimated 100-year return levels. An overall trend that can be seen in this figure, is that for larger gap sizes, the average return levels are lower (the return levels based on the dataset with a 90% gap is an exception, but this is not of much interest, since a 90% gap is very unrealistic and unlikely in practice). This is indicative of the fact that when there are gaps in a dataset, making detrimental under-estimations of return levels is likely.

The 5-year, 10-year, 17-year, 50-year, and 100-year estimated return levels based on the complete dataset are also plotted (these are not averages). All estimations are done by the POT method along with the Weibull distribution and the Method of Maximum Likelihood.

Figure 6.2.16 shows the variance in the 1000 estimated 5-year, 10-year, 17-year, 50-year, and 100 year return levels, respectively, against the size of the 'chunk' gaps.

Figure 6.2.17 shows the following quantity, for 'chunk' gaps ranging from 0% to 90% (it is again noted that all estimations are done by the POT method along with the Weibull distribution and the method of maximum likelihood): *adapted variance* $= \frac{1}{1000}\sum_{i=1}^{1000}(y_i - y_N)^2$, where the y_is are the 1000 estimated N-year return levels, based on an incomplete dataset (with the gap size shown on the horizontal axis), and y_N is

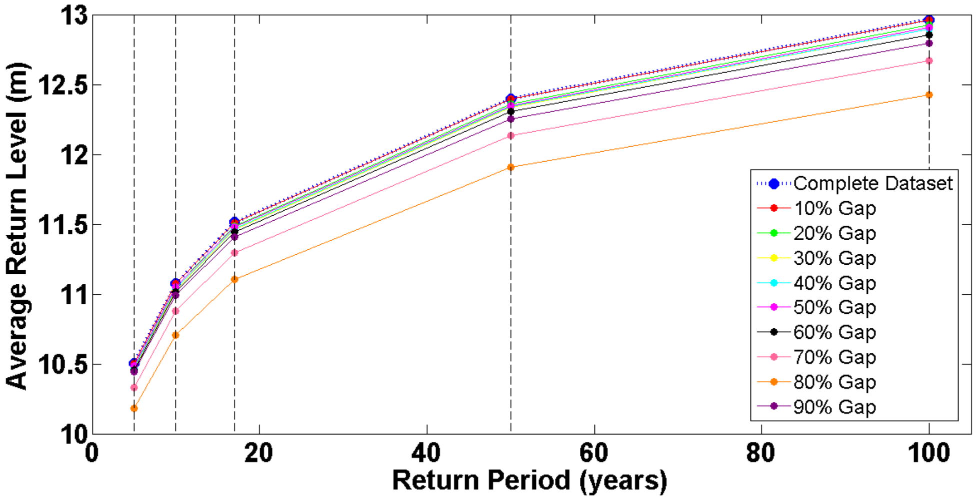


Figure 6.2.15: The averages of 1000 estimated 5-year, 10-year, 17-year, 50-year, and 100-year return levels, respectively, based on datasets with gaps ranging from 10% up to 90%. The 5-year, 10-year, 17-year, 50-year, and 100-year estimated return levels based on the complete dataset are also plotted (these are not averages). The 5-year, 10-year, 17-year, 50-year, and 100-year positions are indicated by the vertical black dashed lines.

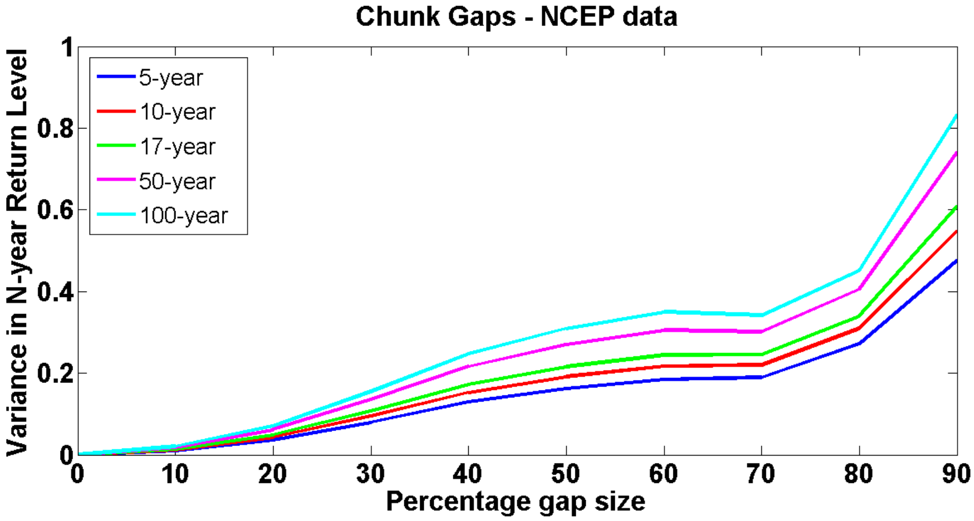


Figure 6.2.16: The variance in the 1000 estimated 5-year, 10-year, 17-year, 50-year, and 100 year return levels, respectively, against the size of the 'chunk' gaps

the N-year return level ($N = 5, 10, 17, 50, 100$), estimated from the complete dataset. In other words, it is similar to the variance (variance $= \frac{1}{1000}\sum_{i=1}^{1000}(y_i - \bar{y})^2$), but instead of subtracting the mean ($\bar{y}$) from each of the estimated return levels, the estimated return level based on the complete dataset (y_N) is subtracted. It therefore plots the sum of the squared differences between the estimated return levels based on incomplete datasets and the estimated return level based on the complete dataset (whereas the variance plot, in Figure 6.2.16, shows plots of the sum of the squared differences between the estimated return levels based on incomplete datasets and the means of these estimated return levels).

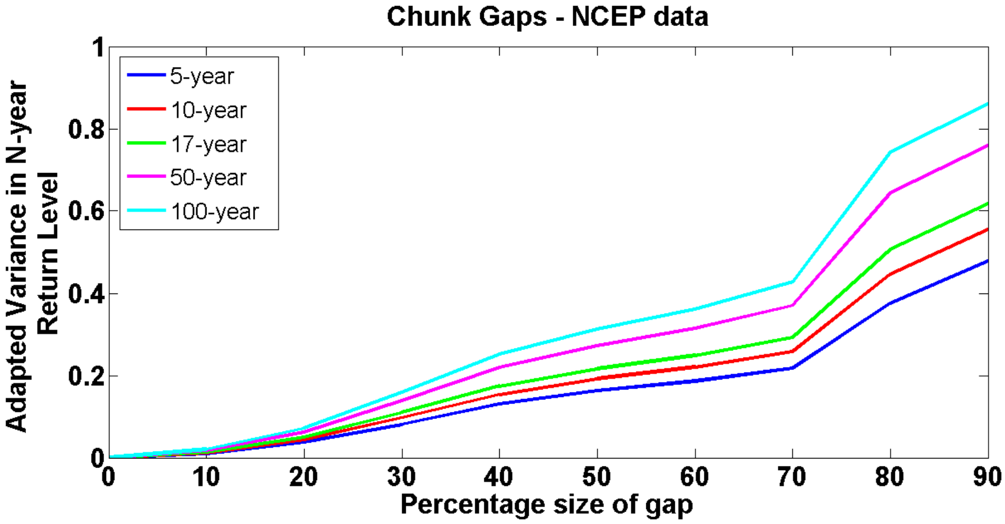


Figure 6.2.17: The adapted variance $= \frac{1}{1000}\sum_{i=1}^{1000}(y_i - y_N)^2$, where the y_is are the 1000 estimated N-year return levels, based on an incomplete dataset (with size of gap shown on horizontal axis), with gaps in the form of connected blocks, and y_N is the N-year return level ($N = 5, 10, 17, 50, 100$), estimated from the complete dataset

Comparing Figures 6.2.16 and 6.2.17, the values on the vertical axis are relatively similar for corresponding gap sizes ranging up to 70%. Thereafter there is quite a sharp rise in Figure 6.2.17, whereas a sharper rise in Figure 6.2.16 occurs from a gap size of 80%. This is however not of as much interest, since datasets with gaps of such great sizes are unrealistic to be used in practice. The focus is therefore placed on the trends of the graphs in Figures 6.2.16 and 6.2.17 for the gaps of smaller size.

A conclusion to be made from Figures 6.2.16 and 6.2.17 (for smaller gap sizes), is that the means of the estimated return levels are very similar to the estimations based on the complete datasets for the corresponding return levels. Hence, in this case, a small variance (i.e., below 0.2) is also indicative of a small deviation in the estimated return levels based on the incomplete datasets from the estimated return levels based on the complete dataset.

Conclusion on Single Block Gaps

Return level estimations made from datasets with gaps in the form of single blocks/ 'chunks' do not seem to yield very reliable results, since these estimations vary rather unpredictalby (as seen in the histograms in Figures 6.2.4 to 6.2.8). The variations of these estimations are already relatively large for a gap of size 10% and only increase with an increase in gap-size.

6.2.2 Random 1-Week Period Gaps

Figures 6.2.18 to 6.2.22 show histograms generated as follows: The first histogram in each of the figures (in the top, left hand corner, entitled '10% gap - NCEP data') displays 1000 return levels associated with a return period of 5 years, 10 years, 17 years, 50 years, and 100 years, respectively. These estimations are again made by means of the POT method and the Weibull distribution along with the method of maximum likelihood (used for parameter estimation) based on the NCEP dataset, but here 10% of the data values are removed from the set in the form of randomly spread 1-week periods. A more detailed explanation of how this is done, along with Matlab code for the removal of random 1-week period gaps, is given in Appendix E.2. The red, dashed lines on the histograms indicate the estimated specified return level, based on the complete NCEP-dataset. The rest of the histograms in Figures 6.2.4 to 6.2.8 are generated in the same manner as explained previously, except for gaps increasing from 20% up to 90%.

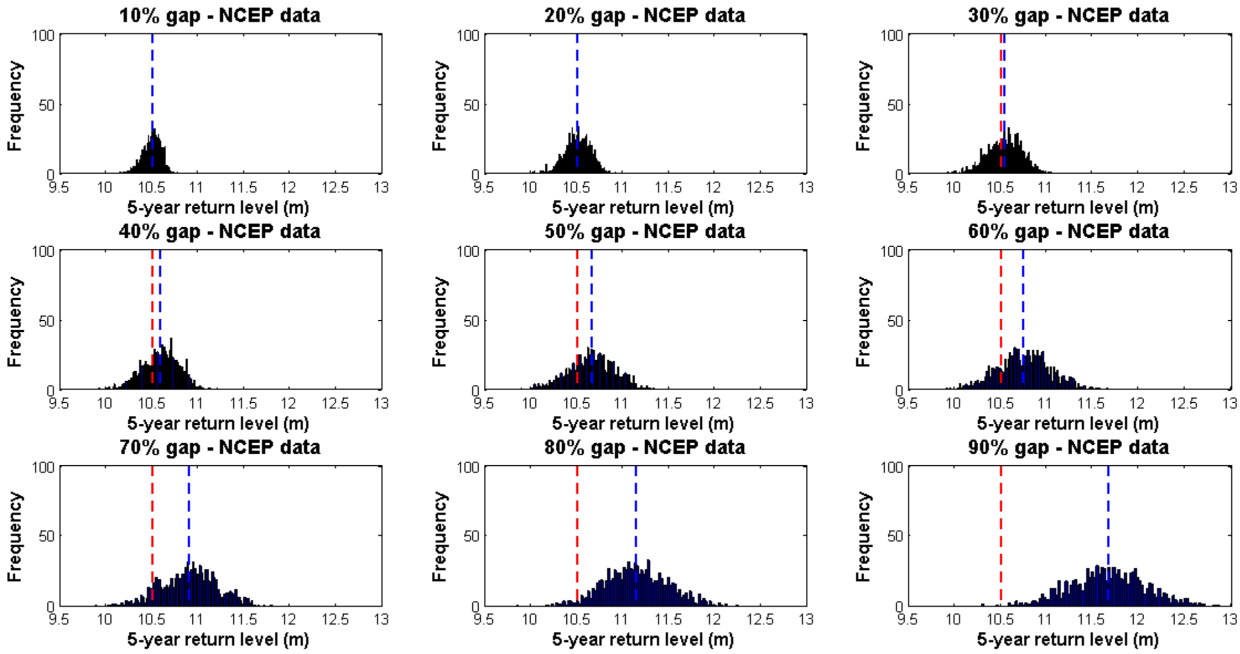


Figure 6.2.18: Histograms of 1000 estimated 5-year return levels when different percentage gaps are created in the dataset by removing the data values of random 1-week periods. The red, dashed lines indicate the estimated 5-year return level based on the complete NCEP-dataset, and the blue, dashed lines indicate the histogram means.

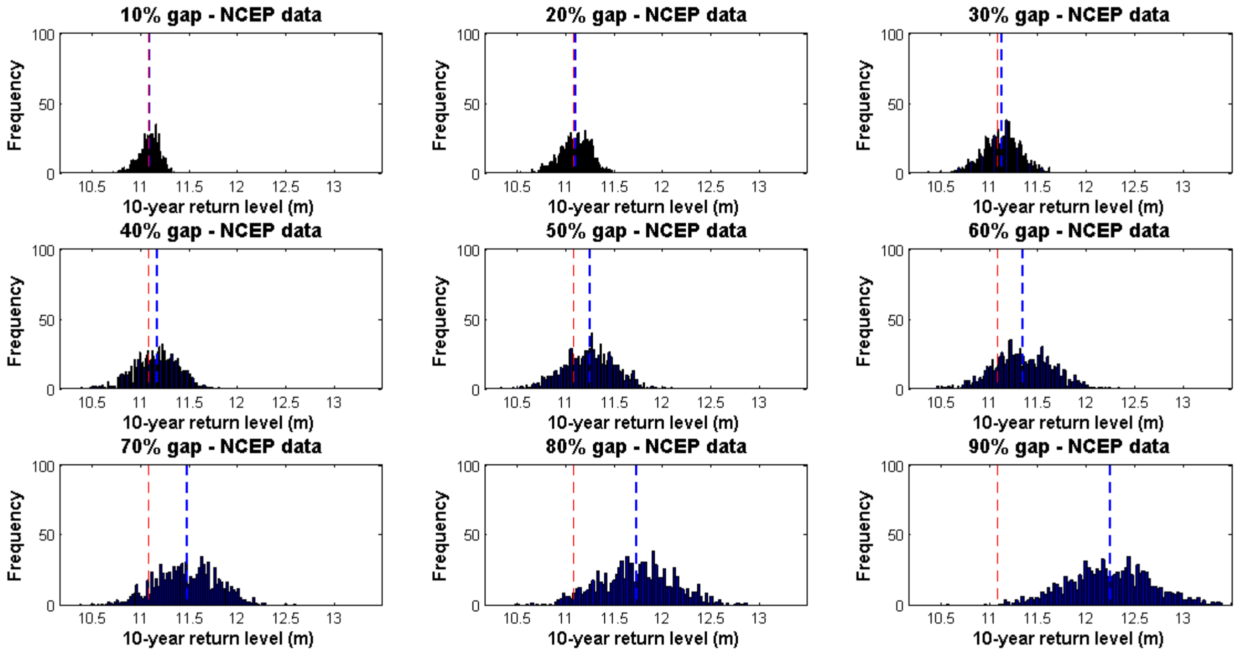


Figure 6.2.19: Histograms of 1000 estimated 10-year return levels when different percentage gaps are created in the dataset by removing the data values of random 1-week periods. The red, dashed lines indicate the estimated 10-year return level based on the complete NCEP-dataset, and the blue, dashed lines indicate the histogram means.

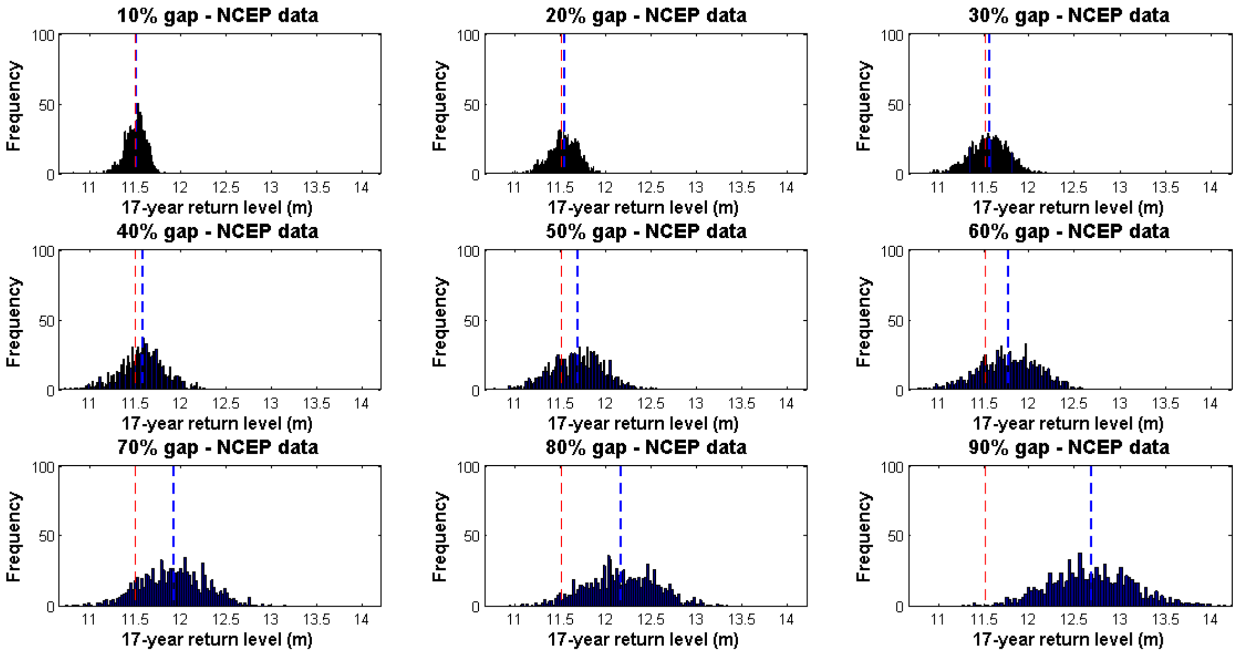


Figure 6.2.20: Histograms of 1000 estimated 17-year return levels when different percentage gaps are created in the dataset by removing the data values of random 1-week periods. The red, dashed lines indicate the estimated 17-year return level based on the complete NCEP-dataset, and the blue, dashed lines indicate the histogram means.

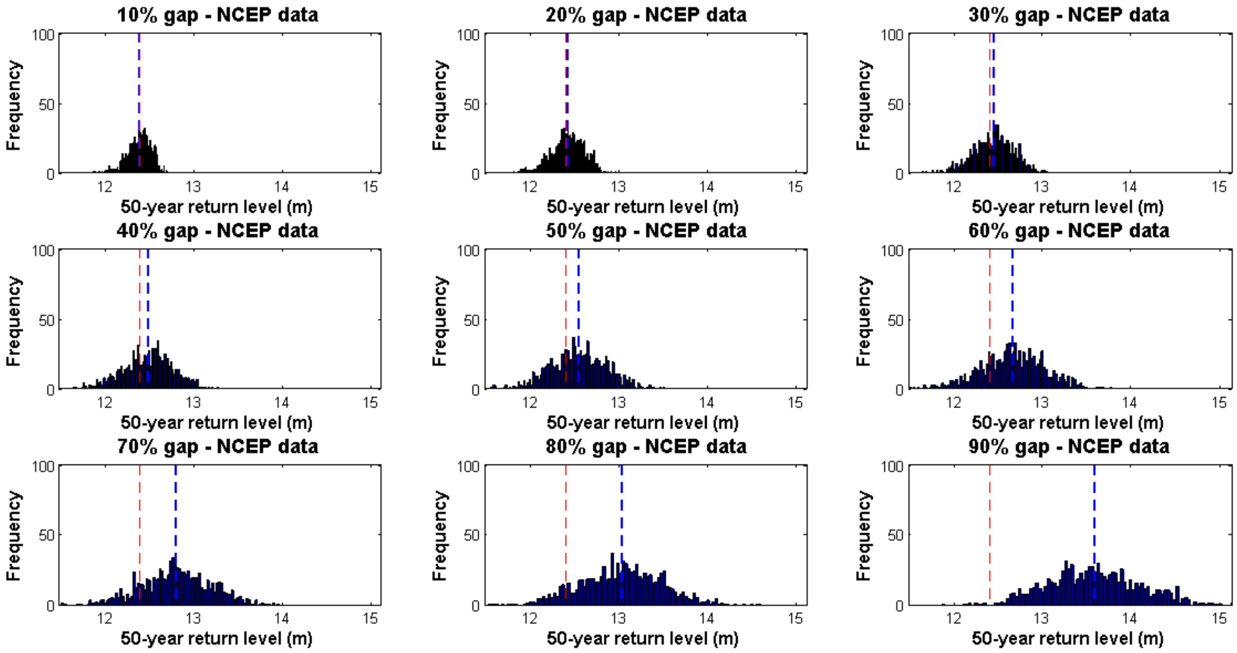


Figure 6.2.21: Histograms of 1000 estimated 50-year return levels when different percentage gaps are created in the dataset by removing the data values of random 1-week periods. The red, dashed lines indicate the estimated 50-year return level based on the complete NCEP-dataset, and the blue, dashed lines indicate the histogram means.

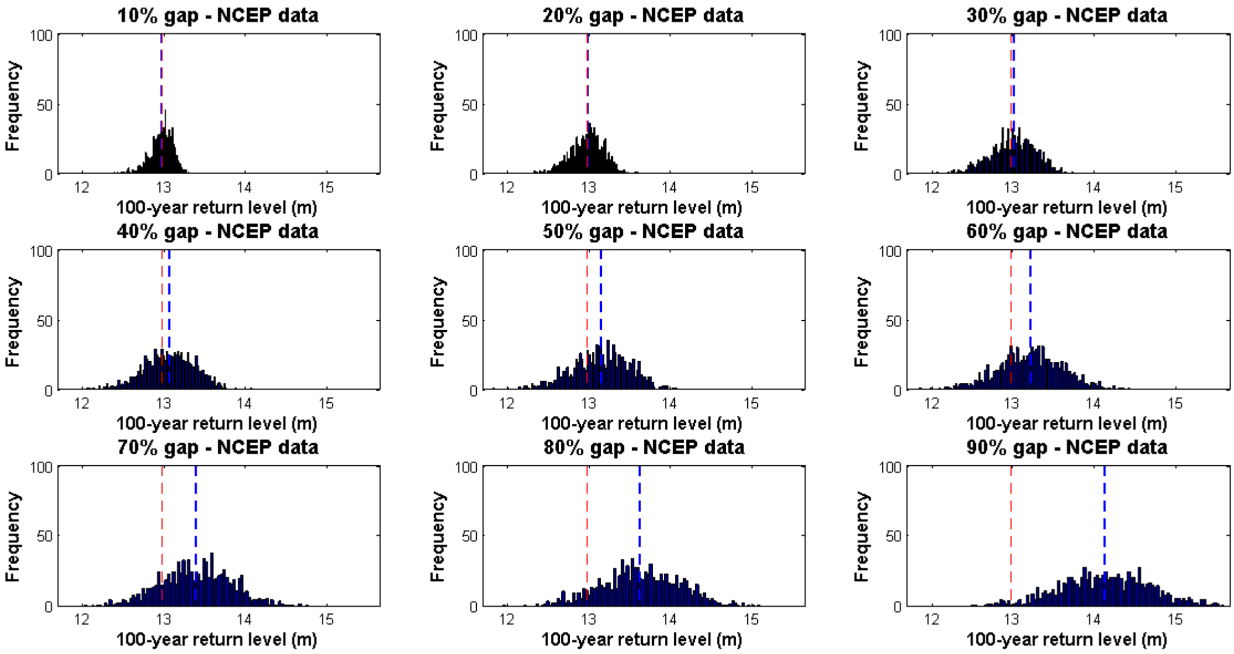


Figure 6.2.22: Histograms of 1000 estimated 100-year return levels when different percentage gaps are created in the dataset by removing the data values of random 1-week periods. The red, dashed lines indicate the estimated 100-year return level based on the complete NCEP-dataset, and the blue, dashed lines indicate the histogram means.

Figures 6.2.18 to 6.2.22 all show a shift in the return levels to the right (i.e., an overall increase in the estimated return levels) as well as a spreading out of the return levels as the size of the gap increases. The observation of the rightward shift can be explained as follows: The return level associated with the non-exceedance probability p (determined by the POT method and the Weibull distribution along with the method of maximum likelihood) is given by the following equation (refer to equation (3.1.11) in Section 3.1.2):

$$x_p = u + \alpha \left[-\ln(1-p)\right]^{\frac{1}{\xi}} , \tag{6.2.1}$$

where u is the threshold (which is 3.3718 m), and α and ξ are the two other parameters of the Weibull distribution (estimated by the method of maximum likelihood). The non-exceedance probability, p, is defined as $p = F_W(x_p)$ (refer to equation (3.1.8) in Section 3.1.2), where F_W is the cumulative distribution function of the Weibull distribution. In other words, p is the probability of observing a wave with a height less than or equal to x_p.

In order to analyse the removed data, Figure 6.2.23 shows the means of the removed data values for different percentage gaps. For each gap size 1000 means are determined (by generating 1000 samples, each without the specified percentage data values) and then plotted.

All of the means in Figure 6.2.23 are below the threshold of 3.3718 m, which is indicative of the fact that more "low" than "high" data values are removed – "low" referring to values below the threshold and "high" referring to those above the threshold. Since the means are all relatively far below the threshold it is reasonable to say that very few of the data values above the threshold are removed (in comparison to those removed above the threshold) in all of the 1000 iterations. This implies that the relation of the number of "high" points to the number of "low" points increases, whereas the length of the period covered by the data decreases (since a percentage of the data values is removed). When it is of interest to determine the N-year return level, the non-exceedance probability can be calculated from the equation

$$p = 1 - \frac{1}{\text{return period}} \tag{6.2.2}$$

(refer to Section 2.1), where the return period is the expected number of POTs per N years (i.e., it is the number of POTs that is expected between the occurence of two waves of height x_p m). Therefore,

$$\text{return period} = \frac{\text{total number of POTs in the entire dataset}}{\text{length of dataset in years}} \times N. \tag{6.2.3}$$

Since the numerator decreases minimally, but the denominator decreases with the percentage size of the gap created, the return period will increase. This leads to an increase in the non-exceedance probability, p, which, in turn, leads to an increase in the return level, x_p. This explains the rightward shift of the histograms in Figures 6.2.18 to 6.2.22.

The observation of the spreading out of the return levels as the size of the gap increases is attributed to the increase in the variance of the return levels with an increase

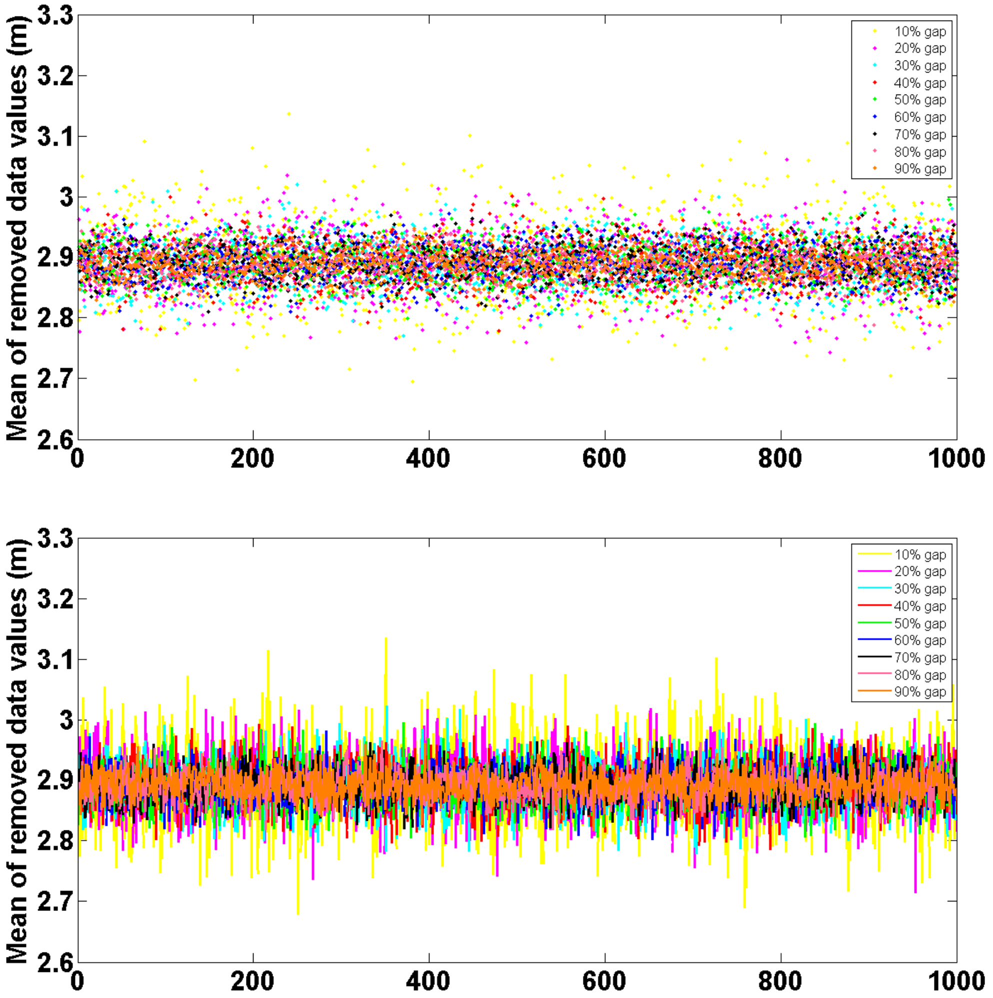


Figure 6.2.23: Means of removed data values for different percentage gaps. 1000 means are determined and plotted for each gap size.

in gap size. This increase in variance is shown in Figure 6.2.24, where the variance in the 5-year, 10-year, 17-year, 50-year, and 100-year return levels, respectively, are plotted against gap size.

Figure 6.2.25 shows plots of the adapted variance, for random 1-week gaps ranging from 0% to 90% (it is again noted that all estimations are done by the POT method along with the Weibull distribution and the method of maximum likelihood): $\frac{1}{1000}\sum_{i=1}^{1000}(y_i - y_N)^2$, where the y_is are the 1000 estimated N-year return levels, based on an incomplete dataset (with the gap size shown on the horizontal axis), and y_N is the N-year return level ($N = 5, 10, 17, 50, 100$), estimated from the complete dataset. It plots the sum of the squared differences between the estimated return levels based on incomplete datasets

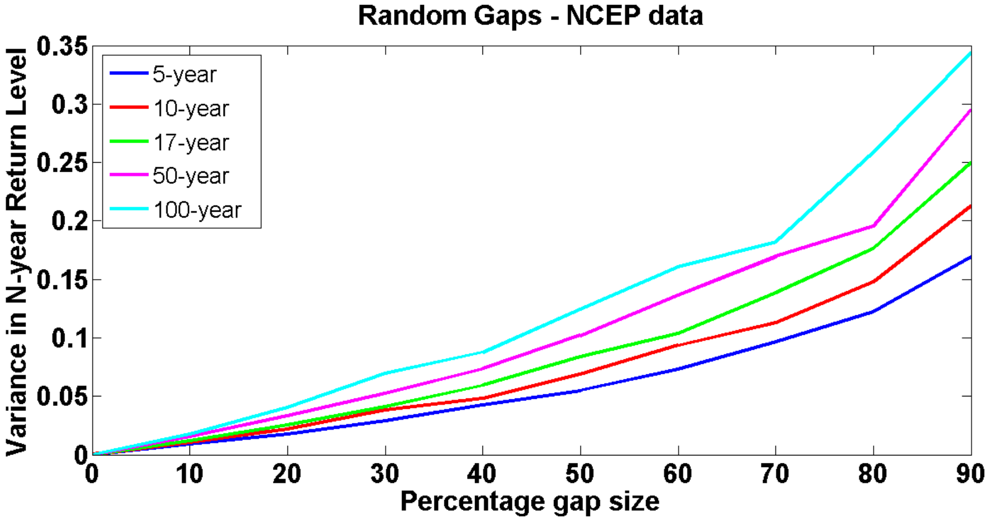


Figure 6.2.24: Variance in the 5-year, 10-year, 17-year, 50-year, and 100-year return levels, respectively, plotted against gap size.

and the estimated return level based on the complete dataset. In other words, Figure 6.2.25 is similar to Figure 6.2.17, except that it is for gaps of random 1-week periods.

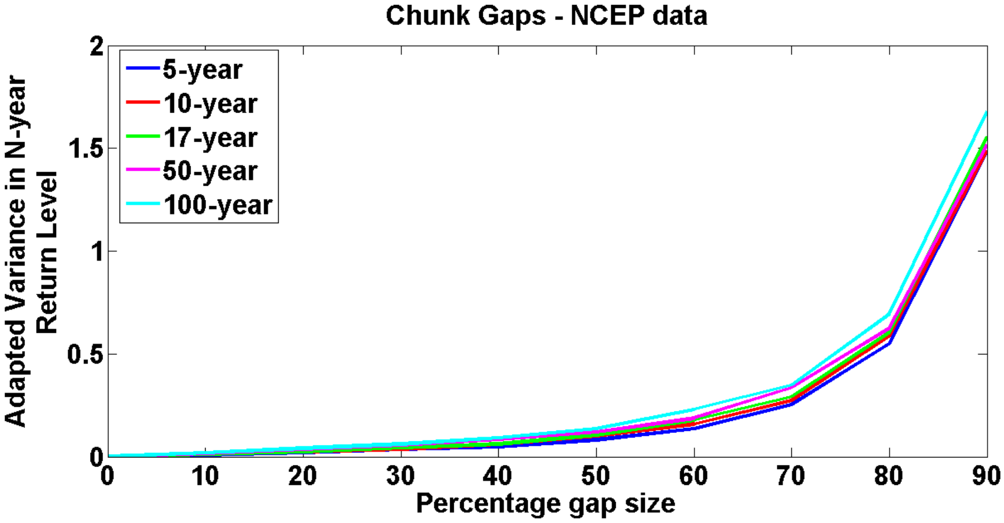


Figure 6.2.25: The adapted variance $= \frac{1}{1000}\sum_{i=1}^{1000}(y_i - y_N)^2$, where the y_is are the 1000 estimated N-year return levels, based on an incomplete dataset (with size of gap shown on horizontal axis), with gaps in the form of random 1-week periods, and y_N is the N-year return level ($N = 5, 10, 17, 50, 100$), estimated from the complete dataset.

When comparing Figures 6.2.24 and 6.2.25, the values on the vertical axis differ significantly for large gap sizes. The adapted variances in Figure 6.2.25 are much higher than the variances in Figure 6.2.24 for large gaps of corresponding sizes. This is indicative of the fact that when the percentage size of the gaps (made up by random 1-week periods) is large (i.e., $\geq 70\%$), the estimations of return levels deviate significantly from the estimations made based on the complete dataset, whereas the deviations of the estimations from their means are not nearly as dramatic.

A conclusion to be made from this, is that even though the variances of return level estimations are relatively low, it does not at all imply that the estimations are reliable, since the deviations from the estimations based on the complete dataset can still be high. For example, when 80% of the data values are removed in the form of 1-week periods, the variance in the 100-year return level is a relatively low 0.2588, whereas the corresponding adapted variance in Figure 6.2.25 is 0.6938. This is quite a significant difference.

Conclusion on Random 1-Week Period Gaps

Even though the variance of estimated return levels based on datasets with large gaps made up of randomly spread 1-week periods is low, it does not imply that the estimations are accurate, since the deviations of the estimated return levels from the estimations made based on the complete dataset can still be high. Hence, the variance is not a reliable measure to determine the reliability of return level estimations based on an incomplete dataset containing gaps in the form of 1-week periods.

6.2.3 Block Gaps Versus Random 1-Week Period Gaps

Figure 6.2.26 shows a comparison between the variance in return levels when the gaps are in the form of ‘chunks’/blocks versus when the gaps are made up of randomly spread 1-week periods. It is a merge of Figures 6.2.16 and 6.2.24.

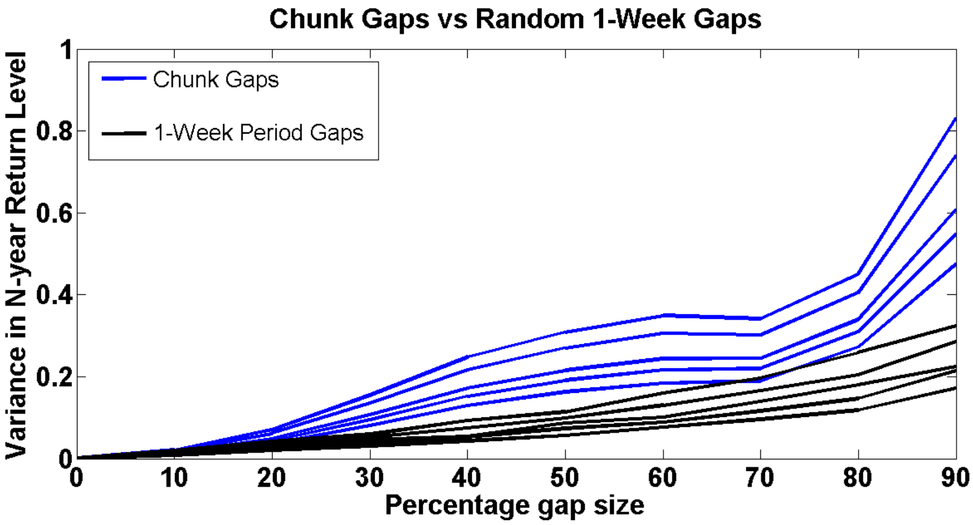


Figure 6.2.26: The variance in the return levels when there are ‘chunk’/block gaps in the dataset (blue lines) versus when then are random 1-week period gaps (black lines).

The variance when 'chunk' gaps are used, are overall much higher than when randomly spread 1-week period gaps are used. This fact indicates that the estimations of return levels based on incomplete datasets deviate less from the means when the dataset contains gaps in the form of 1-week periods in comparison to when the gaps are in the form of 'chunks'.

Figure 6.2.27 shows a comparison between the quantity $\frac{1}{1000}\sum_{i=1}^{1000}(y_i - y_N)^2$ (i.e., the adapted variances) when the gaps are in the form of 'chunks'/blocks (blue lines) versus when the gaps are made up of randomly spread 1-week periods (black lines). It is a merge of Figures 6.2.17 and 6.2.25.

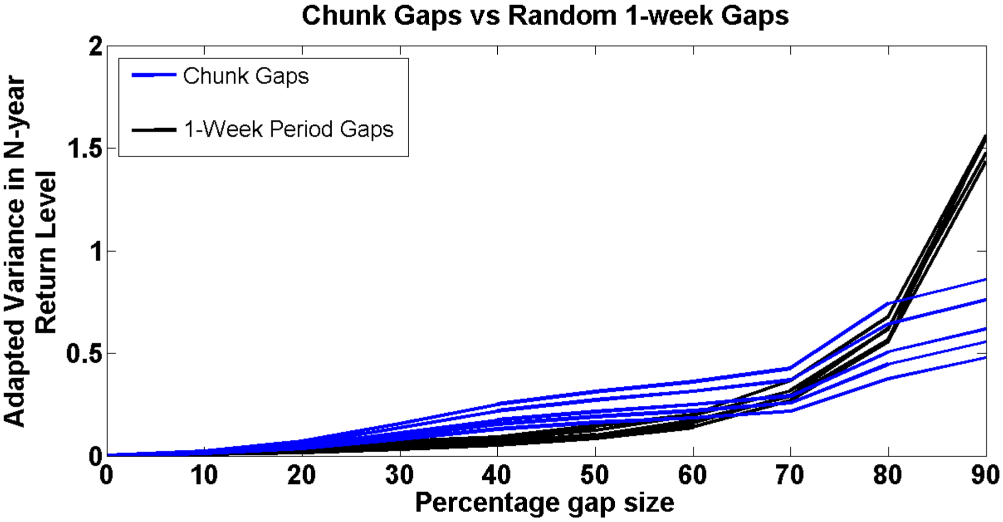


Figure 6.2.27: The adapted variance = $\frac{1}{1000}\sum_{i=1}^{1000}(y_i - y_N)^2$, where the y_is are the 1000 estimated N-year return levels, based on an incomplete dataset (with size of gap shown on horizontal axis), and y_N is the N-year return level ($N = 5, 10, 17, 50, 100$), estimated from the complete dataset. The blue lines are when there are 'chunk'/block gaps in the dataset and the black lines are when the gaps are in the form of random 1-week periods.

Conclusion on Block Gaps Versus Random 1-Week Period Gaps

Estimations regarding return levels based on an incomplete dataset with data missing in the form of random 1-week periods in comparison to estimations based on an incomplete dataset with data missing in the form of connected blocks/'chunks' are overall more reliable. This can be said, since the overall variances of the return levels are lower for the case of 1-week periods gaps, as well as the overall deviations of the estimated return levels based on incomplete datasets from the estimations based on the complete dataset (this is for smaller gap-sizes, i.e, $< 70\%$).

Chapter 7

Application of Findings to Slangkop Data

7.1 Gaps in Slangkop Dataset

As stated earlier, in Chapter 6, the Slangkop dataset has missing data values. Also as mentioned before, all the analyses performed on this dataset therefore far, were done with the replacement of the absent data by zeros. This is equivalent to assuming that no waves higher than the threshold occured at the times where data values are missing, which is an assumption not based on any specific, substantial argument. Hence, the specific values of the estimations previously determined, based on this dataset, are not as accurate in practice, since the gaps were not filled in a reliable manner.

The next objective is to determine the percentage of the Slangkop dataset made up by absent data values as well as to determine te approximate spread of these missing values. The term *approximate* spread is used, since the goal is to determine whether the spread of the missing data is most similar to a single block/'chunk' gap or to gaps in the form of random 1-week periods. These are the two spreads of absent data values that were considered in Chapter 6.

7.1.1 Spread of Slangkop Gaps

The gaps in the Slangkop dataset make up 8.29% of the set. Figures 7.1.1 and 7.1.2 display plots of the Slangkop H_{mo} measurements from the year 1994 to 2012. The gaps in the dataset (i.e., missing data values) are indicated by the red vertical lines. It is clear from these plots that the spread of the gaps are most similar to randomly spread 1-week period gaps (rather than single block gaps).

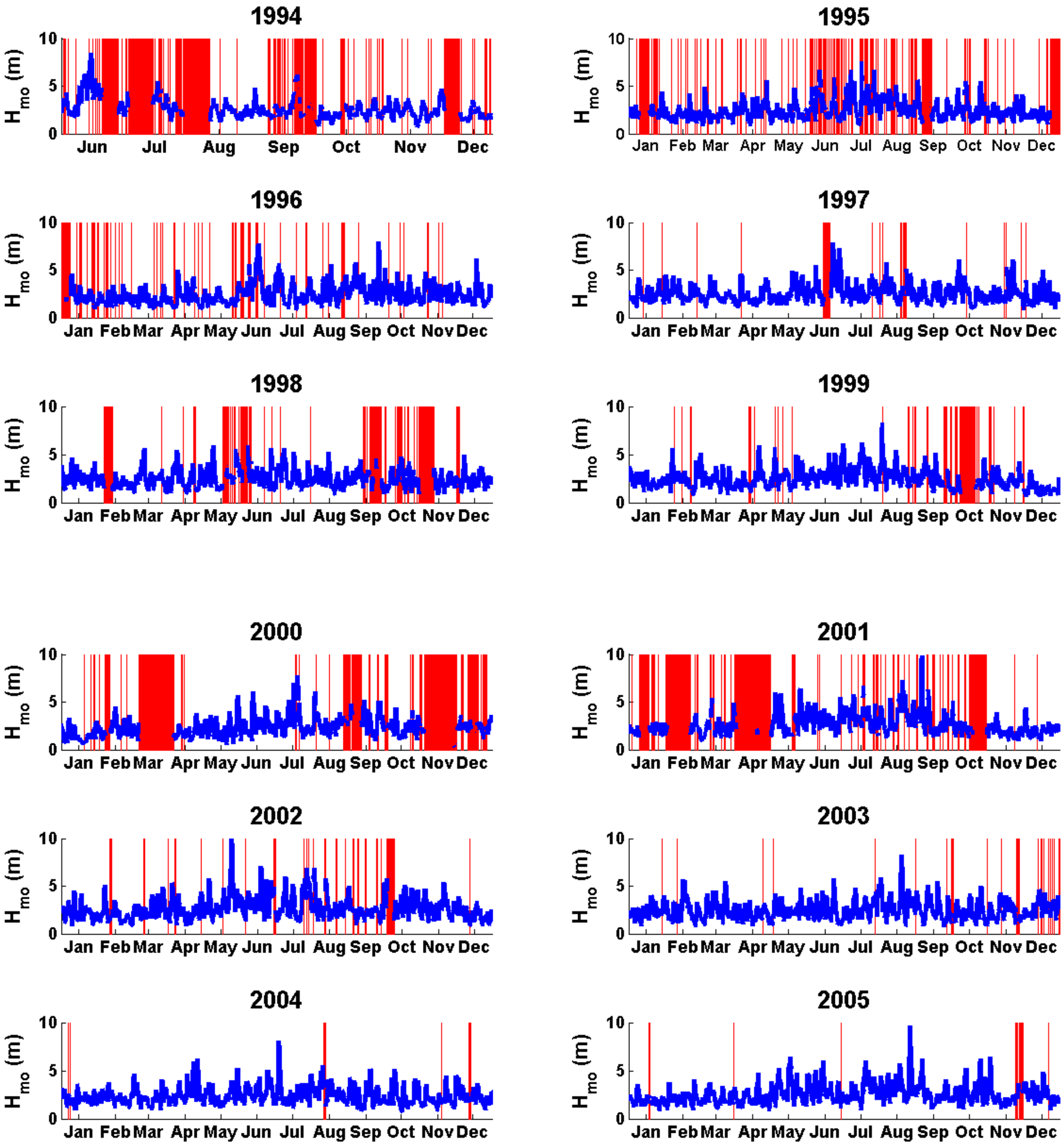


Figure 7.1.1: Plots of the H_{mo} measurements from the Slangkop dataset for the years 1994 to 2005. The red lines indicate missing data values. Note that the scale on the horizontal-\time-axis of 1994 differs from those of the other years, since only the months of June to December are covered during this year.

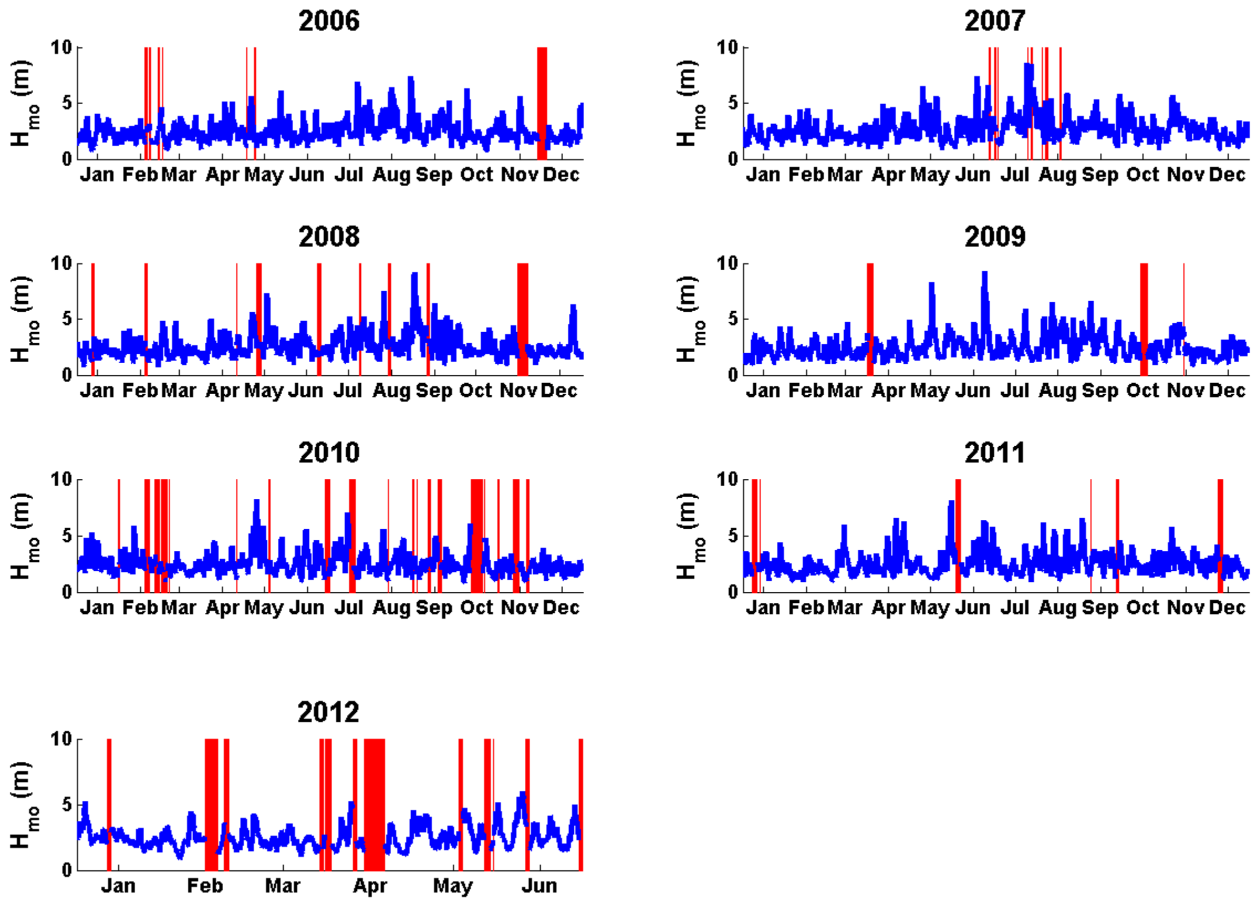


Figure 7.1.2: Plots of the H_{mo} measurements from the Slangkop dataset for the years 2000 to 2012. The red lines indicate missing data values. Note that the scale on the horizontal-\time-axis of 2012 differs from those of the other years, since only the months of January to June are covered during this year.

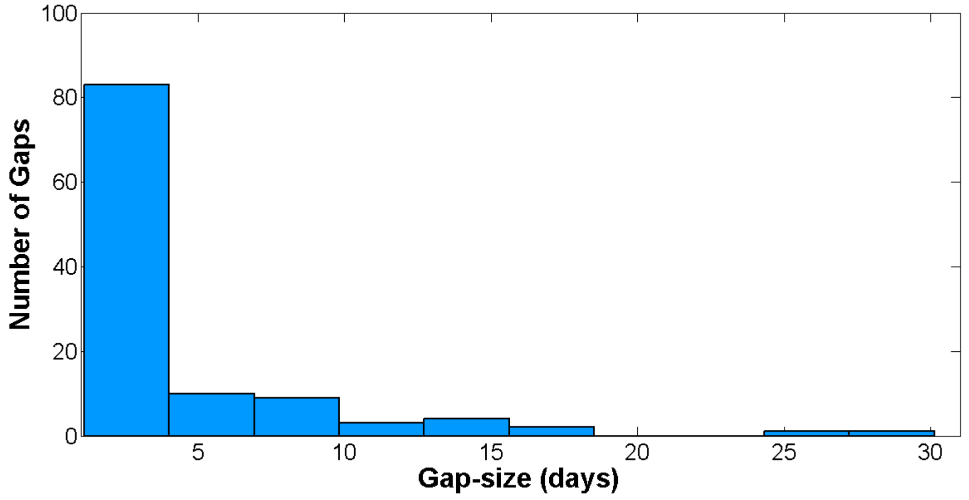


Figure 7.1.3: Histogram displaying the spread of the Slangkop gap-sizes for gaps greater than one day in length.

The size of the smallest gap in the Slangkop dataset is 1 missing measurement, and the size of the largest gap is 241 consecutive missing 3-hourly measurements, which is equivalent to 30.125 days. Figure 7.1.3 is a histogram displaying the spread of the Slangkop gap-sizes for gaps greater than one day in length. The majority of the gaps is not displayed in this histogram, since they consist of missing values making up less than one day in length.

7.2 Treatment of Gaps

The next decision to be made is on how to treat the gaps in this dataset. There are one of two options: remove the gaps and shorten the length of the dataset (i.e, casewise deletion, as discussed in Chapter 6, Section 6.1.3) or fill the gaps. Methods discussed in Chapter 6 for filling gaps, are imputation (Section 6.1.3) and mean substitution (Section 6.1.2), which is a simple example of imputation. Mean substitution will not be implemented in this study, since it has too many disadvantages. In the next section, an imputation method for generating a complete dataset from an incomplete dataset is discussed.

7.2.1 Generated Dataset

It firstly has to be noted that the method described below seems scientific, but in actual fact has no statistical basis supporting it. It is merely a thought-out method, which attempts to fill in missing data values in a reliable manner. The reliablity of this method is evaluated later on. More sophisticated gap-filling techniques inlcude non-linear regression methods, a dual unscented Kalman filter approach, artificial neural networks, look-up tables, marginal distribution sampling, a mean diurnal variation approach, a multiple imputation method, and a terrestrial biosphere model (Moffat et al. (2007)).

To obtain a complete dataset from an incomplete set, the following procedure is followed: gaps consisting of 4 or less missing measurements (which is the equivalent of 12 hours) in the incomplete dataset are replaced by the use of linear interpolation. Gaps consisting of 5 or more missing measurements are replaced by the data values from the first year following the current year where there are no missing values in the time period corresponding to the gap in the current year. For example, in Table 6.3, consider the first gap which consists of 9 missing values (from 1994\06\08 06:00 to 1994\06\09 06:00). In order to fill this gap, the values of 1995\06\08 06:00 to 1995\06\09 06:00 are considered. If there are no missing values in the latter time interval, the 1995 values are used to complete the 1994 values. If this time interval does, however, contain missing values the same time interval in 1996 is considered. This process is repeated until the first year is found where this time interval contains no missing values and those values are then used to replace the 1994 values. When considering gaps in the last year of the dataset (i.e., 2012), the corresponding values in the first year of the dataset (i.e., 1994) is firstly considered. If also incomplete, the second year is considered, etc. This procedure is repeated until there are no missing values remaining in the dataset. The Matlab code for implementing this procedure is given in Appendix F.

As stated previously, the above method cannot be substantiated scientifically, since wave measurements made in different years are completely independent. Imputation in this manner was simply chosen based on the assumption that wave heights follow seasonal trends. Hence, replacing missing values in a specific month by data values from the same month in another year ensures that the values used for replacement fall in the same season as the missing values. Another method to fill in missing values might have been to choose a month that is statistically the closest to the month that contains the missing values. This could be done, for instance, by calculating the mean and variance of the H_{mo}s for each month and then substituting the values from the month with mean and variance closest to the mean and variance of the month of which the missing values have to be filled in. If the time interval from the month to be used for the imputation does, however, contain missing values, the month that is statistically second closest to the month with the missing values is considered. This process is repeated until the first month is found which does not have missing values in the same time interval as where the missing values have to be filled in. This method was not implemented in this study.

To evaluate how reliable the above method to fill gaps is, gaps similar to those in the Slangkop dataset are created in the NCEP dataset (used in Chapter 6). In other words, 8.29% of the NCEP data values are removed in the form of randomly spread 1-week periods. The procedure described above is then implemented to create a complete generated dataset from the incomplete NCEP data. Finally, the complete generated dataset is compared to the original NCEP dataset in order to draw a conclusion on how well the applied method filled the gaps. Figures 7.2.1 to 7.2.3 show plots of the H_{mo}s of the original NCEP dataset as well as the generated dataset for the years 1997 to 2014.

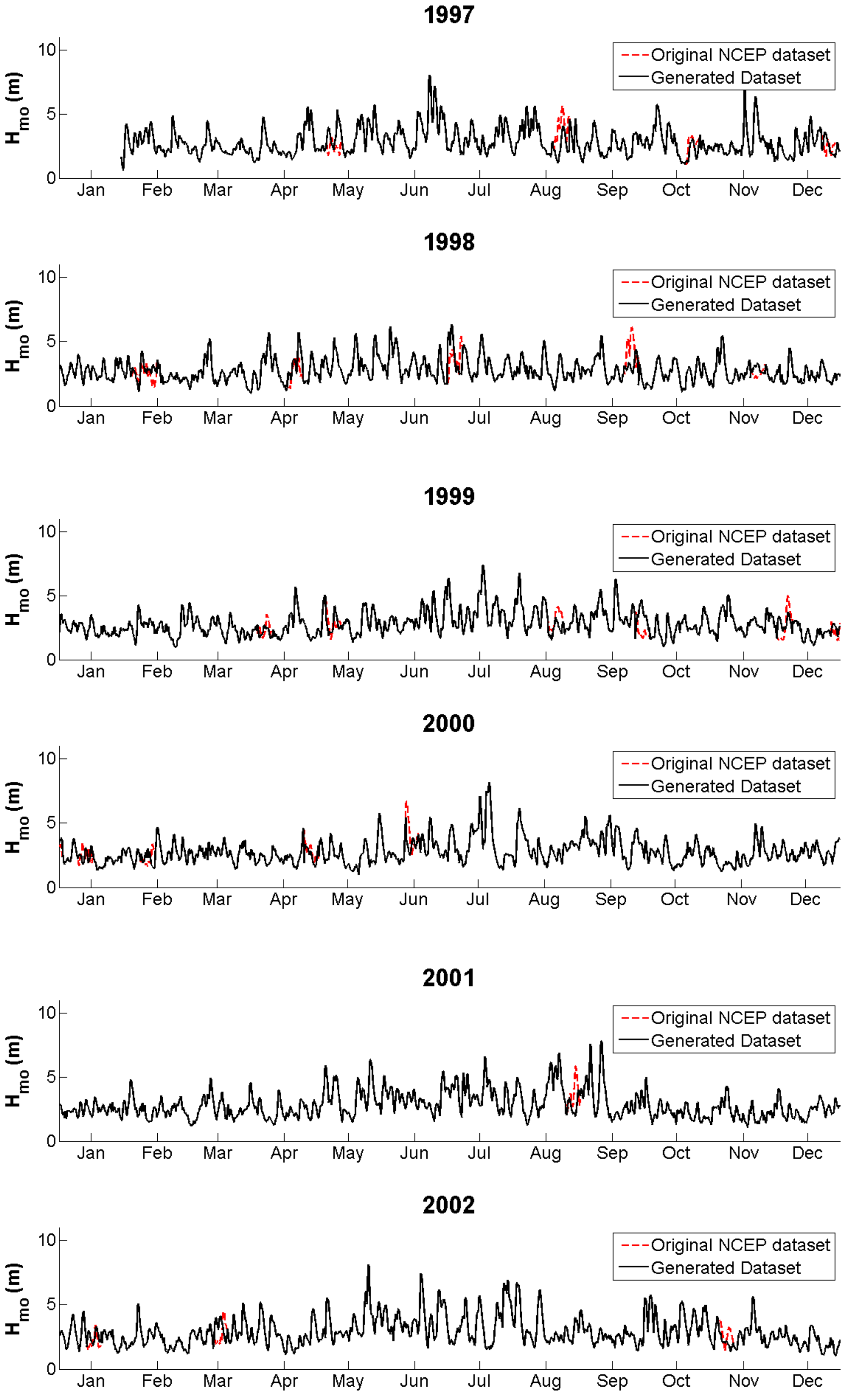


Figure 7.2.1: Plots of the H_{mo}s of the original NCEP dataset and the generated dataset for the years 1997 to 2002.

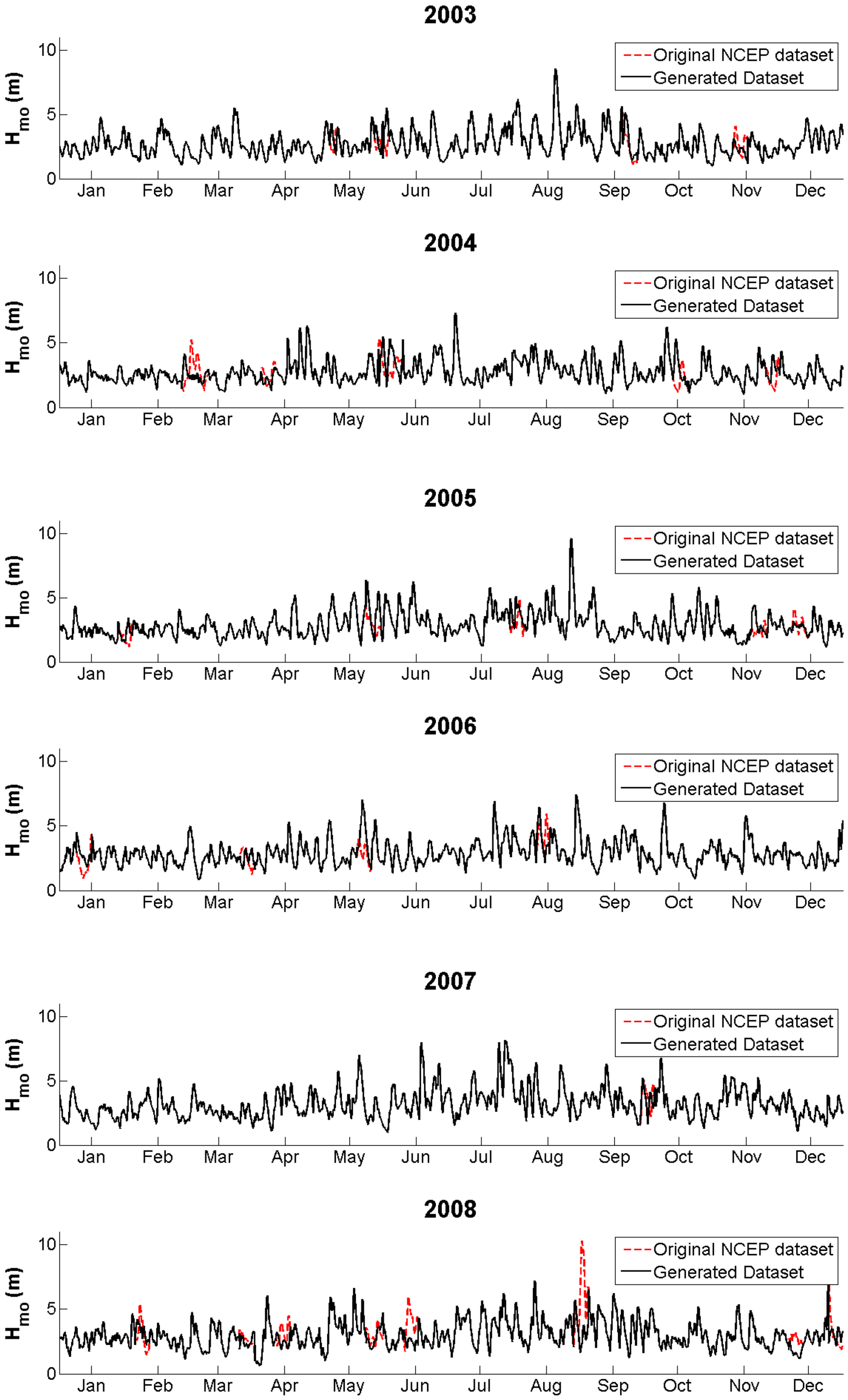


Figure 7.2.2: Plots of the H_{mo}s of the original NCEP dataset and the generated dataset for the years 2003 to 2008.

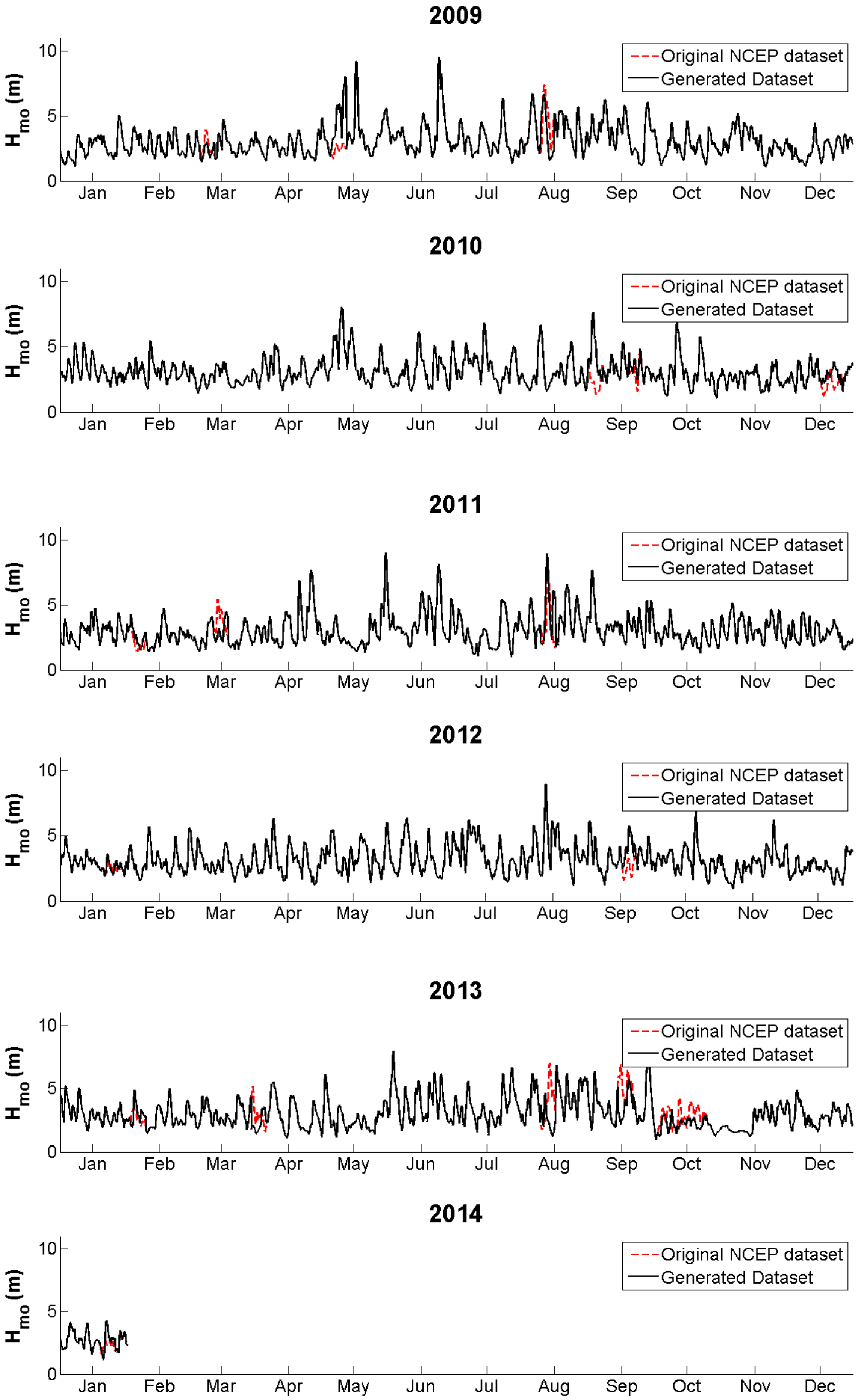


Figure 7.2.3: Plots of the H_{mo}s of the original NCEP dataset and the generated dataset for the years 2009 to 2014.

Figure 7.2.4 shows a plot of the estimated return levels as determined by the Weibull distribution (along with the method of maximum likelihood for parameter estimation) based on both the original and generated NCEP datasets. It can be seen that the estimations based on the two different datasets differ only slightly. The generated dataset leads to slight over-estimations of return levels when compared to the estimations based on the original NCEP dataset. Based on the generated dataset, the 5-, 10-, 17-, 50-, and 100-year return levels are over-estimated by 0.57%, 0.63%, 0.70%, 0.73%, and 0.84%, respectively. That is an average over-estimation of the five specified return levels (i.e., the 5-, 10-, 17-, 50-, and 100-year) of 0.69%.

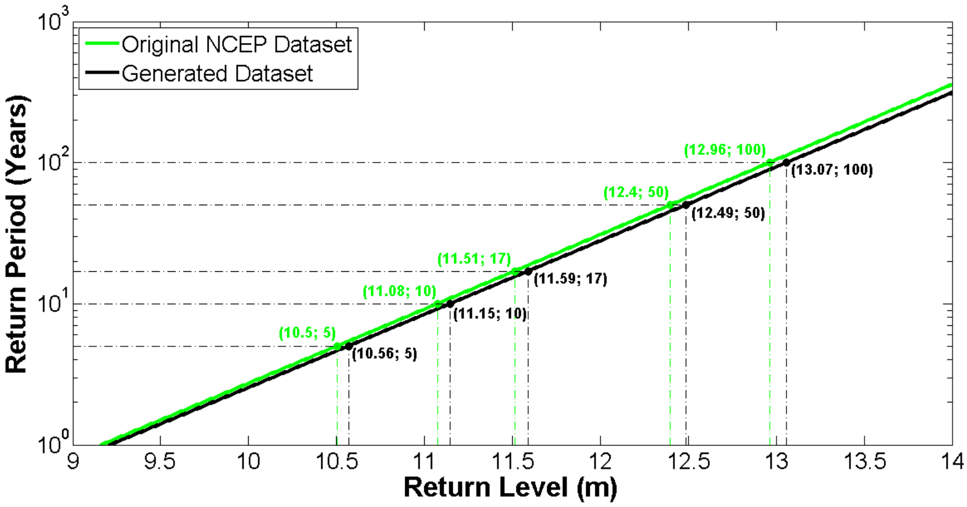


Figure 7.2.4: Estimated return levels determined by the Weibull distribution (and method of maximum likelihood for parameter estimation) based on the original and generated NCEP datasets, respectively. The generated dataset was created by filling up randomly spread, 1-week period gaps in the NCEP dataset (making up a total of 8.29% of the dataset) by the procedure described in Section 7.2.1.

In the following section, the second alternative for the treatment of gaps in the dataset is considered, namely removing the gaps and shortening the length of the dataset.

7.2.2 Removal of Gaps

In this case, the missing values are removed from the incomplete NCEP dataset (with 8.29% of it's values absent) and the length of the dataset is decreased by 8.29%. In other words, the length of the original dataset is 17.016438356164382, hence the length of the sortened dataset is 17.016438356164382(1 − 8.29%) years.

Figure 7.2.5 shows a plot of the estimated return levels as determined by the Weibull distribution (along with the method of maximum likelihood for parameter estimation) based on both the original and shortened datasets. Once again it is evident that the estimations based on the two different datasets differ minimally. The use of the shortened dataset leads to slightly higher over-estimations when compared to estimations based on the original NCEP dataset than in the case of the generated dataset. These over-estimations are 1.14%, 1.08%, 1.13%, 1.13%, and 1.23% for the 5-, 10-, 17-, 50-, and 100-year return levels, respectively. This leads to an average over-estimation of the five specified return levels (i.e., the 5-, 10-, 17-, 50-, and 100-year) of 1.14%.

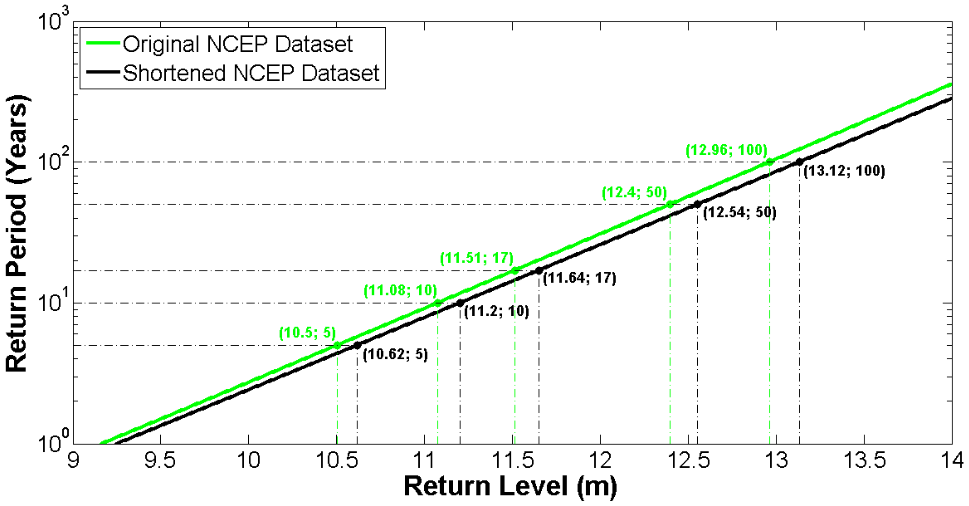


Figure 7.2.5: Estimated return levels determined by the Weibull distribution (and method of maximum likelihood for parameter estimation) based on the original NCEP and shortened datasets, respectively. The shortened dataset was obtained by creating randomly spread, 1-week period gaps in the original NCEP dataset (making up a total of 8.29% of the dataset) and then removing all the absent values and, hence, shortening the length of the dataset.

7.2.3 Conclusion on Treatment of Gaps

As observed in Sections 7.2.1 and 7.2.2, the use of both the generated and shortened NCEP datasets lead to slightly higher estimations of return levels than the original NCEP dataset. The generated dataset produces an average over-estimation of 0.69%, whereas the shortened dataset produces an average over-estimation of 1.14%. It is therefore reasonable to assume that in order to attempt to minimise the risk of the under-estimation of return levels, the use of the shortened dataset is safer than the use of the generated dataset.

From Sections 6.2.1 and 6.2.2 it is also known that a dataset with a 10% gap in the form of a single connected block yields more unreliable results than a dataset with 10% of its data values missing in the form of randomly spread 1-week periods. In the latter case the estimations (i.e., "results") are acually quite reliable (consider the histograms entitled "10% gap - NCEP data" in Figures 6.2.18 to 6.2.22). Since less than 10% of Slangkop's data values are missing (only 8.29%), and the missing values are scattered, reaffirms that the shortened dataset (with all the absent values removed) should yield reliable results.

7.3 Analyses of Slangkop Dataset

In this section, three different methods are implemented for analyses of the Slangkop dataset, namely the *points over threshold*, *peaks over threshold*, and *block maxima* methods. The Weibull distribution, along with the method of maximum likelihood for parameter estimation, will be used for estimations for the points over threshold and peaks over threshold methods. The generalized extreme value (GEV) distribution, once again along with the method of maximum likelihood, will be used for estimations in the case of the block maxima method (refer to Appendix A.1 for the reasoning behind why this distribution is used in this case).

7.3.1 Treatment of Slangkop Gaps

Since the use of a shortened dataset (with all missing data values removed) proves to be a relatively safe and reliable method for the treatment of gaps in order to prevent the under-estimation of return levels, the Slangkop dataset will be treated in this manner. All of the missing values in the dataset will be removed and, hence, the length of the dataset shortened.

7.3.2 Analyses By Means of Points Over Threshold

In the points over threshold method, all values higher than the threshold of 3.3718 m are taken into account. This is illustrated in Figure 7.3.1 for the year 1997 of the Slangkop dataset. The red dots indicate all of the data values above the threshold of 3.3718 m.

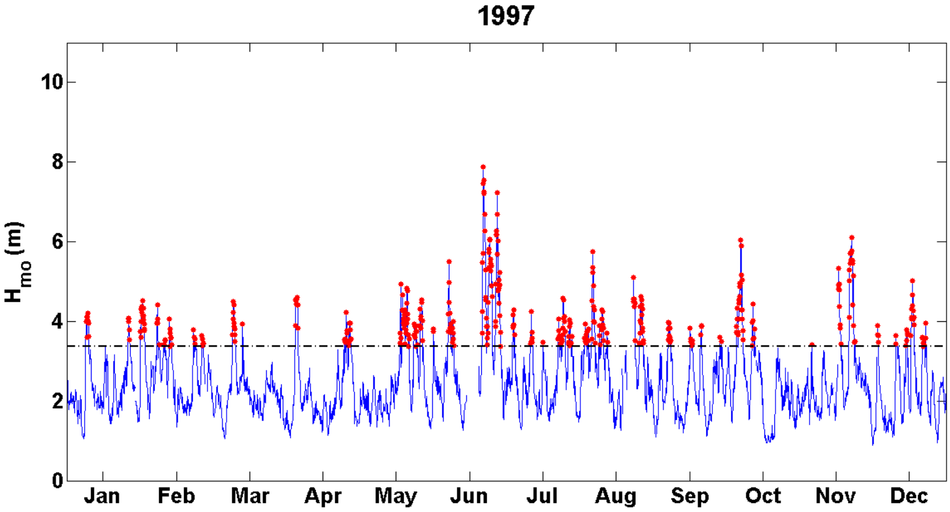


Figure 7.3.1: H_{mo}s of the Slangkop dataset for the year 1997. The red dots indicate all of the points above the threshold of 3.3718 m (threshold is indicated by the black, dashed-dotted line).

All of the points exceeding the threshold are used to determine the Weibull parameters α and ξ by means of the method of maximum likelihood. The estimated return levels are then determined from a Weibull distribution with these parameters. This was applied to the incomplete NCEP dataset in Section 7.2.2 and Figure 7.2.5. The plot of the estimated return levels when this is implemented on the Slangkop dataset is shown in Figure 7.3.2.

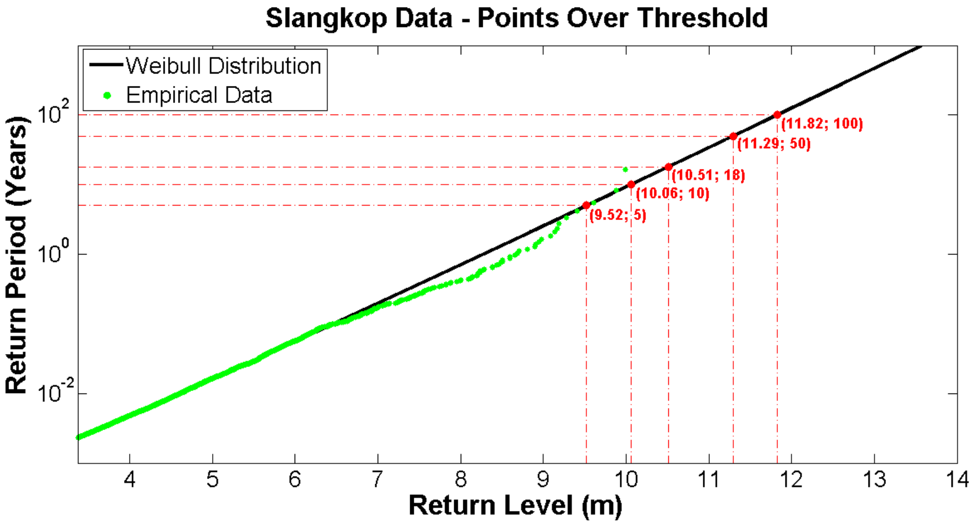


Figure 7.3.2: Estimated return levels determined by the Weibull distribution (and method of maximum likelihood for parameter estimation) based on the Slangkop dataset. All of the absent values in the dataset were removed and, hence, the length of the dataset shortened.

7.3.3 Analyses By Means of Peaks Over Threshold

In this case, only *peaks* over the threshold of 3.3718 m are considered. Peaks are defined as independent exceedances over the threshold. Here, independence of exceedances are ensured by defining an interevent time criterion, Δt_c, and defining two successive exceedances (or events) to be independent if the time between the two events is larger than Δt_c (DHI (2003)) (refer to Appendix B.1 for more information on independent and identically distributed random variables). The peaks over the threshold of 3.3718 m in the Slangkop data for 1997, with an imposed interevent time criterion of one week, are illustrated in Figure 7.3.3. This interevent time criterion (also, inter-arrival time) was not selected on a statistical basis. More sophisticated selections of interevent time criteria can be made by the use of, for example, autocorrelation techniques (Northrop et al. (2015), Mann et al. (2002)).

Only the peaks over the threshold are used to determine the Weibull parameters α and ξ by means of the method of maximum likelihood. The estimated return levels are then, once again, determined from a Weibull distribution with these parameters. The plot of the estimated return levels when this method is implemented on the entire Slangkop dataset is shown in Figure 7.3.4.

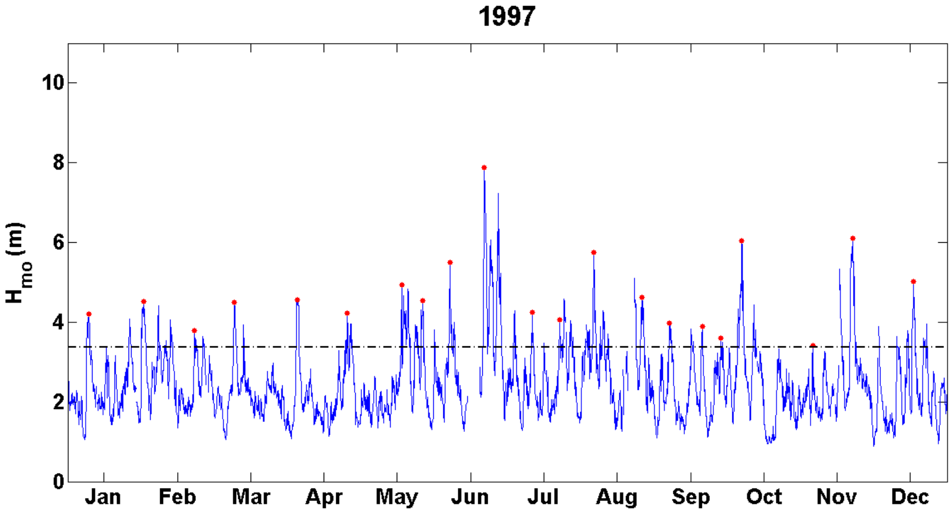


Figure 7.3.3: H_{mo}s of the Slangkop dataset for the year 1997. The red dots indicate all of the peaks above the threshold of 3.3718 m (indicated by the black, dashed-dotted line). An interevent time criterion of one week is used.

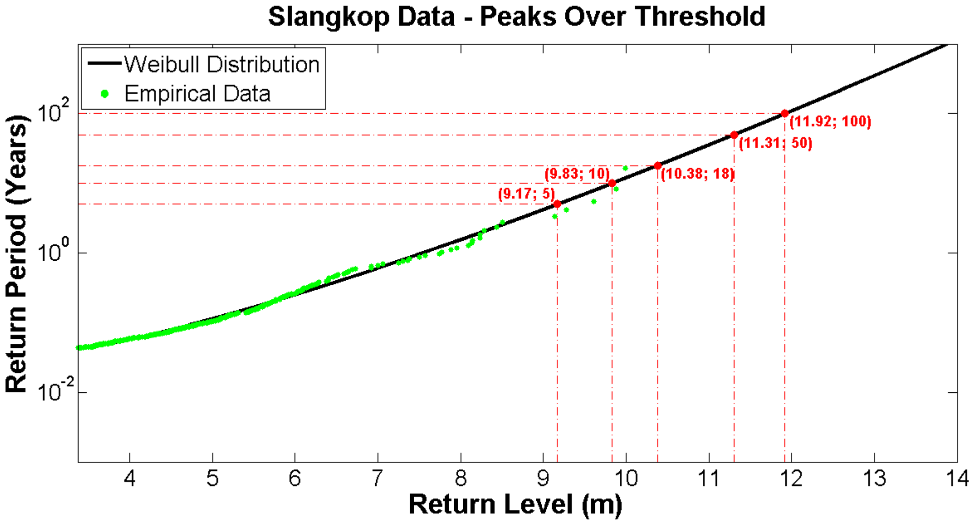


Figure 7.3.4: Estimated return levels determined by the Weibull distribution (and method of maximum likelihood for parameter estimation) based on the Slangkop dataset. All of the absent values in the dataset were removed and, hence, the length of the dataset shortened.

7.3.4 Analyses By Means of Block Maxima

In the case of the block maxima method, only a single value is considered within every "block". A block refers to a pre-defined period of time, and here it will be defined as a month. In other words, the maximum H_{mo} value will be extracted for each of the twelve calender months for the time period covered by the Slangkop dataset. The monthly maxima of the Slangkop data for 1997 are shown in Figure 7.3.5.

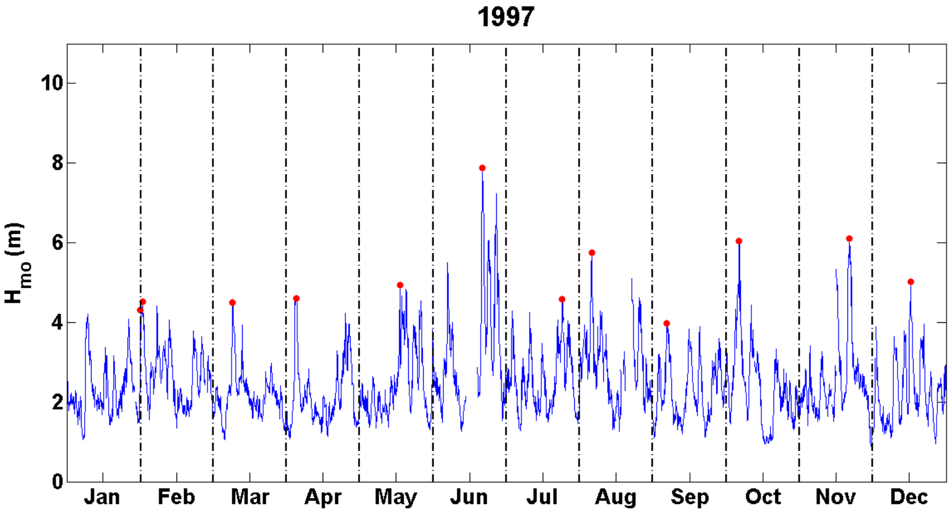


Figure 7.3.5: H_{mo}s of the Slangkop dataset for the year 1997. The red dots indicate the monthly maxima.

The monthly maxima are used to determine the parameters of the GEV distribution, namely ϕ, α, and ξ (refer to Section 2.2.1), by using the method of maximum likelihood. The plot of the estimated return levels when this method is implemented on the Slangkop dataset is shown in Figure 7.3.6

7.3.5 Comparison of Points Over Threshold, Peaks Over Threshold, and Block Maxima Methods

Finally, the estimations made by the points over threshold, peaks over threshold, and block maxima methods can be compared. Figure 7.3.7 is a merge of Figures 7.3.2, 7.3.4, and 7.3.6. It displays a comparison of return level estimations when the three different methods are used for analyses of the Slangkop dataset. The Weibull distribution is used in the cases of the points and peaks over threshold methods, and the GEV distribution is used in the case of the block maxima method. The parameters of the distributions are estimated by the method of maximum likelihood in all three cases.

Overall, the block maxima method gives the lowest estimations of return levels. For return periods up to just under 50 years, the points over threshold method gives higher estimations for return levels than the peaks over threshold method. At a return period of 50 years, the estimated return levels for these two methods are approximately similar (11.29 m for the points over threshold method, and 11.31 m for the peaks over threshold method). For return periods greater than 50 years, the peaks over threshold method gives higher return level estimations than the points over threshold method.

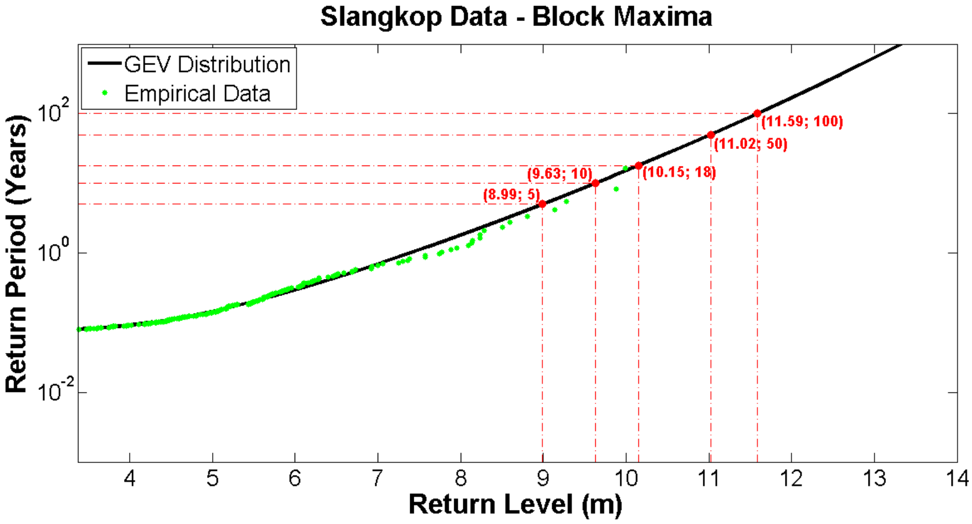


Figure 7.3.6: Estimated return levels determined by the GEV distribution (and method of maximum likelihood for parameter estimation) based on the Slangkop dataset. All of the absent values in the dataset were removed and, hence, the length of the dataset shortened.

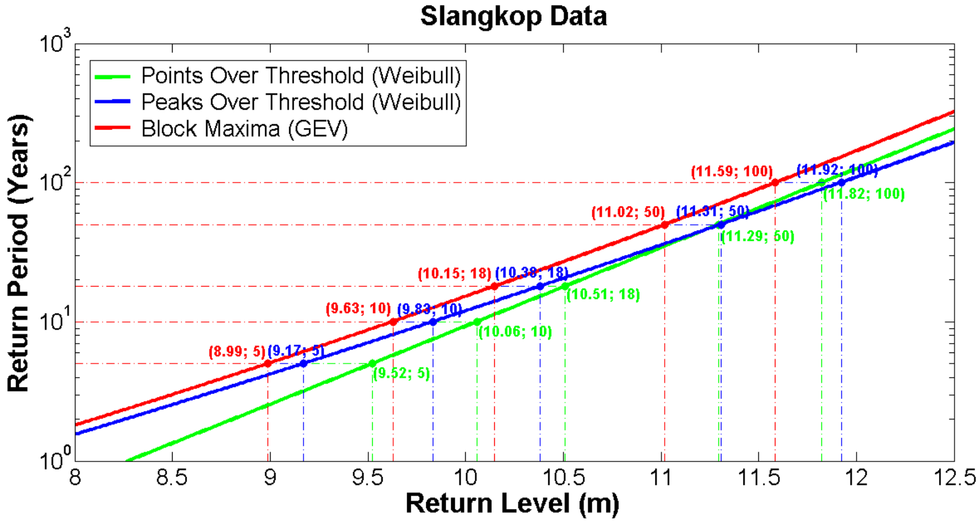


Figure 7.3.7: A comparison of the estimations regarding return levels when the points over threshold, peaks over threshold, and block maxima methods, respectively, are used to analyse the Slangkop data. The Weibull distribution is used in the cases of the points and peaks over threshold method, and the GEV distribution is used in the case of the block maxima method. The method of maximum likelihood is used for parameter estimation in all three cases.

Based on the facts that, for the Slangkop dataset, the block maxima method gives the lowest return level estimations overall, and that this method has the risk of omitting significant data, since only one H_{mo}-value is considered within each month (refer to Section 1.2), the conclusion can be made that this method may not produce reliable results.

The points over threshold and peaks over threshold methods produce relatively similar results for all return periods considered in Figure 7.3.7. The difference in estimations for return levels with return periods of 50- and 100-years are minimal (with only a difference in return level estimation of 0.02 m for a 50-year return period, and a difference of 0.1 m for a 100-year return period).

Chapter 8

Conclusions and Recommendations

It was concluded that the use of the method of maximum likelihood together with the Weibull distribution is efficient and reliable for making estimations regarding return levels. This conclusion was reinforced by the findings in Chapters 4, 4.2, and 5. Hence, this was the combination of parameter estimation method and probability distribution that was selected to be used again in later chapters of this study.

The effect of gaps of different spreads (single connected blocks/'chunks' versus randomly spread 1-week period gaps) as well as different sizes (ranging from 10% to 90% of the entire dataset) were investigated in Chapter 6. It was found that estimations regarding return levels based on an incomplete dataset with data missing in the form of random 1-week periods overall yield more reliable results, in comparison to estimations based on an incomplete dataset with data missing in the form of one single, connected block.

In the case of gaps made up of randomly spread 1-week periods, gap sizes of up to 30% seemed to yield fairly acceptable results (refer to Figures 6.2.18 to 6.2.22), whereas estimations based on incomplete datasets with gaps in the form of single, connected blocks already vary relatively significantly from such a small percentage as only 10% (refer to Figures 6.2.4 to 6.2.8). The latter case does not seem to yield reliable results, since the estimations vary quite unpredictably. The overall variances of the return level estimations, as well as the overall deviations of the estimated return levels based on incomplete datasets from the estimations based on the complete dataset are lower in the case of 1-week period gaps, than in the case of single connected block gaps (for gaps of sizes $< 70\%$).

Finally, in Chapter 7, the findings of the study was applied to the Slangkop dataset. The spread of the gaps in this dataset was found to be more similar to random 1-week period gaps, than a gap in the form of one single, connected block/'chunk', and was found to make up 8.29% of the entire set. Unlike earlier in the study, where these gaps were just replaced by zeros, all of the missing data values were removed from the dataset and, hence, the length of the dataset shortened.

The shortened Slangkop dataset was analysed based on three methods, namely the points over threshold, peaks over threshold, and block maxima methods. For the first two methods, the Weibull distribution was used for the estimation of return levels, whereas for the latter method, the GEV distribution was used. In al three cases, the method of maximum likelihood was used for the estimation of the probability distribution parameters. The block maxima method was found to yield the lowest estimations of return levels overall, and since this method also creates the risk of omitting significant data, it was concluded that this method does not produce reliable results. The points and peaks over threshold methods produced similar results. For return levels with shorter return periods (i.e., up to just under 50 years), the points over threshold method gives higher estimations than the peaks over threshold method. However for return levels with higher return periods (i.e., 50 years and more), this phenomena is interchanged and the peaks over threshold method is the method that gives slightly higher estimations. Based on these findings, and the fact that over-estimation of return levels are always safer than under-estimation, the peaks over threshold method is recommended above the points over threshold method.

In summary, the recommendations for extreme value analyses of a dataset, based on this study, are listed below:

1. remove all missing data values and shorten the length of the entire dataset;
2. select an appropriate threshold based on the method described in Section 3.1.1;
3. impose an interevent time criterion and determine the peaks over the chosen threshold;
4. estimate the parameters, α and ξ, of the Weibull distribution by using the method of maximum likelihood;
5. use the Weibull distribution, with it's estimated parameters, in order to estimate H_{mo} return levels.

The above conclusions and recommendations are all subjectively based on the findings in this study. It can therefore not be exclusively confirmed as the optimal way for the analyses of extreme values in general.

Appendix A

Statistical Theory

All of the content in this appendix is adapted from Coles (2001).

A.1 Fisher-Tippett / Extremal Types Theorem

If M_n is the maximum of a sequence of independent and identically distributed random variables, $X_1, X_2, \ldots, X_n$, with common (unknown) distribution function F, then the distribution of M_n is given by

$$\begin{aligned} P(M_n \leq z) &= P(X_1 \leq z, X_2 \leq z, \ldots, X_n \leq z) \\ &= P(X_1 \leq z) \times P(X_2 \leq z) \times \ldots \times P(X_n \leq z) \quad \text{(due to independence)} \\ &= [F(z)]^n \,. \end{aligned} \tag{A.1.1}$$

In order to estimate statistical models for F^n on the basis of extreme data, the behaviour of F^n is considered as $n \to \infty$. If z_+ is the smallest value of z such that $F(z) = 1$ (i.e., $F(z) = 1$ for all $z \geq z_+$), then for any $z < z_+$, $F^n(z) \to 0$ as $n \to \infty$. It follows that the distribution of M_n degenerates to a point mass on z_+. It is therefore necessary to renormalize M_n as follows:

$$M_n^* = \frac{M_n - b_n}{a_n}, \tag{A.1.2}$$

with $\{a_n > 0\}$ and $\{b_n\}$ appropriate choices of sequences of constants. This stabilizes the location and scale of M_n^* as n increases and ensures that the distribution of M_n^* converges to a non-degenerate distribution function.

The Extremal Types Theorem states that, when M_n can be stabilized with appropriate sequences $\{a_n\}$ and $\{b_n\}$, the normalized variable, M_n^*, has a limiting distribution that is one of the three members of the GEV distribution, i.e., the Gumbel, Fréchet, or Weibull distribution. These three are the only possible limits for the distributions of the M_n^* . (The proof will not be considered, since it is beyond the scope of this study.) In other words, it follows that

$$P\left(\frac{M_n - b_n}{a_n} \leq z\right) \approx F_{\text{GEV}}(z), \tag{A.1.3}$$

for large enough n, where F_{GEV} is a member of the GEV family. This is equivalent to

$$P(M_n \leq z) \approx G\left(\frac{z-b_n}{a_n} \leq z\right) = F^*_{\text{GEV}}(z), \tag{A.1.4}$$

where F^*_{GEV} is another member of the GEV family. This shows that if the distribution of M^*_n can be approximated by a member of the GEV family, then the distribution of M_n can also be approximated by a member of the GEV family (Coles (2001)).

A.2 Generalized Pareto Distribution

Coles (2001) states that if block maxima have an approximating distribution, G, which is a member of the GEV family, then the threshold exceedances have a corresponding approximate distribution within the GPD. A short, outlining argument for this statement follows.

By the Extremal Types Theorem, in Section A.1, $[F(z)]^n = F^n(z)$ in equation (A.1.1) can be approximated by a member of the GEV family for large enough n. Hence, from equation (2.2.1),

$$F^n(z) \approx e^{-\left[1+\xi\left(\frac{z-\phi}{\alpha}\right)\right]^{-\frac{1}{\xi}}}. \tag{A.2.1}$$

It follows that

$$n \log F(z) \approx -\left[1+\xi\left(\frac{z-\phi}{\alpha}\right)\right]^{-\frac{1}{\xi}}. \tag{A.2.2}$$

A Taylor expansion for large z yields

$$\log F(z) \approx -\left[1-F(z)\right]. \tag{A.2.3}$$

Substituting equation (A.2.3) into equation (A.2.2) and rearranging leads to

$$1-F(u) \approx \frac{1}{n}\left[1+\xi\left(\frac{u-\phi}{\alpha}\right)\right]^{-\frac{1}{\xi}}, \tag{A.2.4}$$

for large u, and for $y > 0$

$$1-F(u+y) \approx \frac{1}{n}\left[1+\xi\left(\frac{u+y-\phi}{\alpha}\right)\right]^{-\frac{1}{\xi}}. \tag{A.2.5}$$

The CDF of exceedances is given by

$$\begin{aligned} P(X < u+y|X > u) &= 1 - P(X > u+y|X > u) \\ &= 1 - \frac{P(X > u+y)}{P(X > u)} \\ &= 1 - \frac{1-F(u+y)}{1-F(u)}. \end{aligned} \tag{A.2.6}$$

Hence, substituting equations (A.2.4) and (A.2.5) yields

$$\begin{aligned} P(X < u + y | X > u) &\approx 1 - \frac{\frac{1}{n}\left[1 + \xi\left(\frac{u+y-\phi}{\alpha}\right)\right]^{-\frac{1}{\xi}}}{\frac{1}{n}\left[1 + \xi\left(\frac{u-\phi}{\alpha}\right)\right]^{-\frac{1}{\xi}}} \\ &= 1 - \left[1 + \frac{\xi\left(\frac{u+y-\phi}{\alpha}\right)}{1 + \xi\left(\frac{u-\phi}{\alpha}\right)}\right]^{-\frac{1}{\xi}} \\ &= 1 - \left[1 + \frac{\xi y}{\alpha + \xi(u - \phi)}\right]^{-\frac{1}{\xi}}. \end{aligned} \tag{A.2.7}$$

Substituting y by $x - u$ (which represents the exceedances over a threshold, u) and $\alpha + \xi(u - \phi)$ by α_u yields the CDF of the GPD as given in equation (2.2.3):

$$F_{\text{GPD}}(x; u, \alpha_u, \xi) = 1 - \left[1 + \frac{\xi(x - u)}{\alpha_u}\right]^{-\frac{1}{\xi}}, \quad \xi \neq 0.$$

From the above substitutions it is evident that the GPD and GEV shape parameters, ξ, are identical and their scale parameters, α_u and α, respectively, are related by the equation $\alpha_u = \alpha + \xi(u - \phi)$.

Appendix B

POT Approach

B.1 Independent and Identically Distributed Random Variables

A condition that has to hold in order for exceedances above a specified threshold to be modelled by the GPD is that the observed values have to be independent and identically distributed. This means that the random variables representing each observation have the same probability distribution and that none of the observations influences any of the other observations. In other words, the occurence of one event is independent of the occurence of another event, i.e., the occurence of one does not influence the probability of another (Wackerly et al. (2008)).

In order to illustrate the concept of independence, an example is considered. Figure B.1.1 shows the maximum daily temperatures of Cape Town during February 2014. A threshold of 30°C is also indicated. It is observed that the threshold exceedances occur in groups, which implies that one extremely hot day is followed by another. This clustering causes dependence in the observations. According to Coles (2001), the most widely adopted method for dealing with dependant exceedances is called *declustering*. This method involves defining clusters by the use of an empirical rule, identifying the maximum exceedance within each cluster, and finally fitting the GPD to the cluster maxima (by assuming independance of cluster maxima). An empirical rule to be used for defining a cluster is to consider exceedances above the threshold to be part of the same cluster until r consecutive values fall below the threshold, where r is a pre-specified value (Coles (2001)). Consider, for example, imposing $r = 1$ in the example in Figure B.1.1. This leads to obtaining three clusters, whereas imposing $r = 2$ leads to obtaining only two clusters. This is shown in Figure B.1.2.

An equivalent way for ensuring independent events is by defining an interevent time criterion, Δt_c, and defining two succesive events to be independent if the time between the two events is larger than Δt_c (DHI (2003)).

Another possible criterion that can be imposed for ensuring independance of consecutive events is an interevent level criterion, p_c $(0 < p_c < 1)$. Two succesive events are then defined to be independant if the level between the events become smaller than p_c

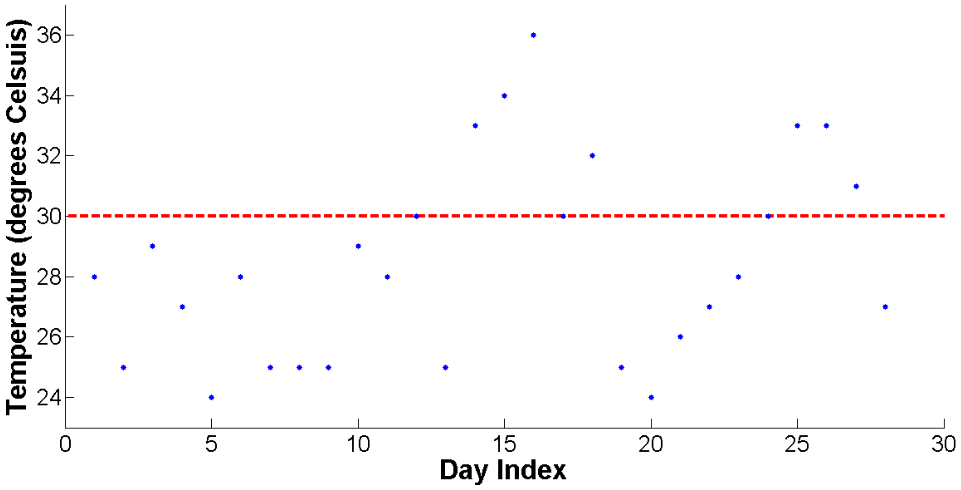


Figure B.1.1: Graph representing the maximum daily temperatures of Cape Town during February 2014 [Source of temperatures: `http://www.accuweather.com/en/za/cape-town/306633/month/306633?monyr=2/01/2014`].

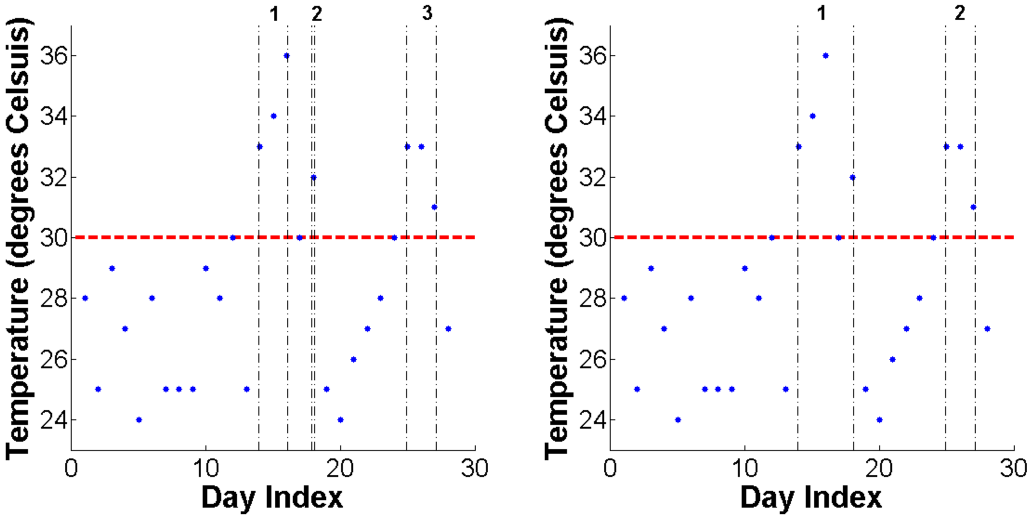


Figure B.1.2: Graphs illustrating cluster grouping if $r = 1$ and $r = 2$, respectively. When $r = 1$, three clusters are obtained, and when $r = 2$ only two clusters are obtained.

times the lower of the two events (DHI (2003)).

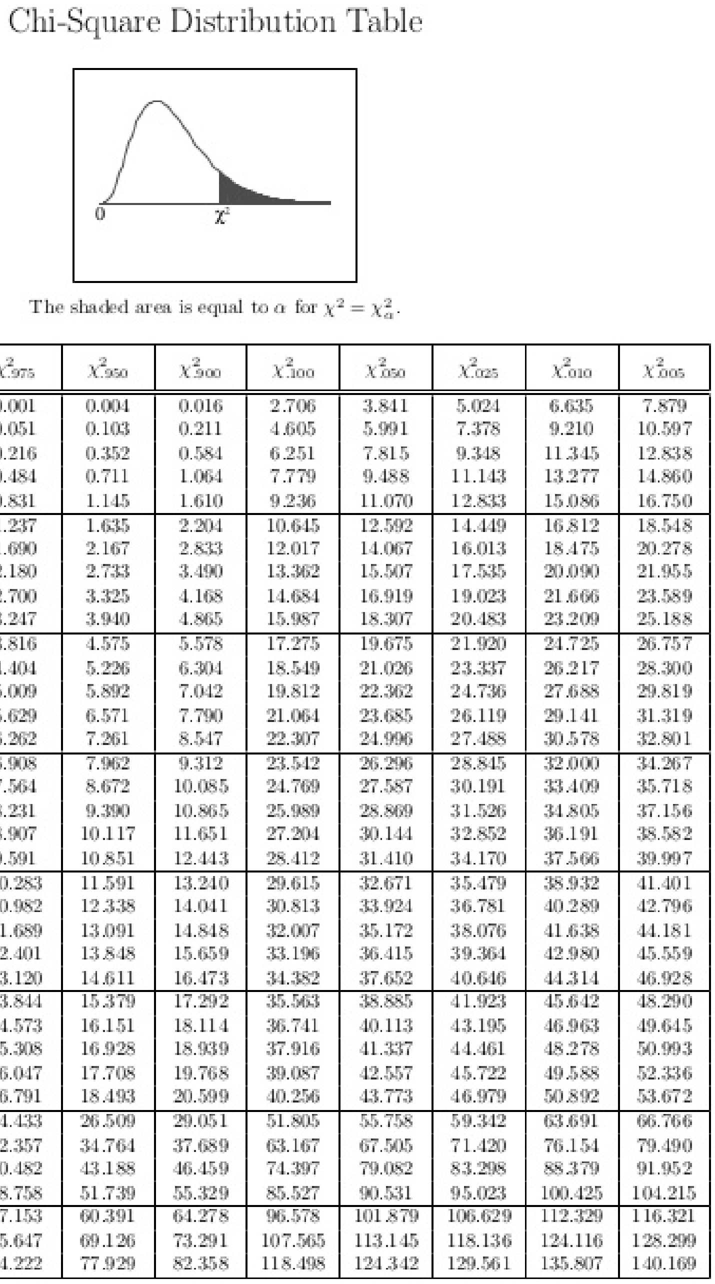
B.1.1 Independence Assumptions in This Study

For the sake of simplicity, three-hourly H_{mo} measurements are assumed to be independent in this study. Hence, since the dataset supplies three-hourly measurements, no additional time criterion or other method to ensure the independence of exceedances over the specified threshold are imposed.

Appendix C

χ^2-Test

C.1 χ^2-Table



Chi-Square Distribution Table

The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure C.1.1: [Source: `http://www.slideshare.net/dennimardomingo/chi-square-table`]

Appendix D

Least Squares Approximation

D.1 The Basic Linear Least Squares Problem

MATLAB CODE:

```
function [c, m] = lin_least_sqrs(x_values, y_values)
% lin_least_sqrs takes as input an array of x-values (x_values) and
% associated y-values (y_values) and gives as output the least squares
% appproximations of m and c of the best approximating line y = mx + c.

n = length(x_values);      % n is the number of (x, y) data pairs

Y_sum = sum(y_values);      % the sum of all the y-values

X_sum = sum(x_values);      % the sum of all the x-values

X2_sum = sum((x_values).^2);      % the sum of the squares of all the
                                  % x-values

XY_sum = sum((x_values).*(y_values));      % the sum of the product of
                                           % all the corresponding x-
                                           % and y-values

c = (X2_sum.*Y_sum - XY_sum.*X_sum)./(n.*X2_sum - X_sum.^2);

m = (n.*XY_sum - X_sum.*Y_sum)./(n.*X2_sum - X_sum.^2);
```

D.2 Newton-Rhapson Iteration

MATLAB CODE:

```
function a = NR(f, df, a0)
% NR takes as input a function f and its derivative df as well as an
% initial guess, a0, and performs Newton Rhapson iteration on it until a
% max number of 20 iteration steps are performed OR until convergence to a
% tolerance level of 1e-8 is achieved

N0 = 20;  TOL = 1e-8;        % Max number of steps and error tolerance

T = [0 a0];                  % T (for table) is used to save the data at
                             % each step

for n = 1:N0
    a = a0-f(a0)/df(a0);     % Newton's method
    T = [T; n a];            % Append values to the table
    if abs(a-a0)<TOL;        % Stopping criterion
        break;               % If satisfied, stop
    end
    a0 = a;
end
```

Appendix E

Gaps in the Dataset

E.1 Removing a $p\%$ 'Chunk' From a Dataset

The Matlab script below is a function that removes a $p\%$ chunk from a given dataset and then predicts a specified y-year return level. This is done by means of the POT method and the Weibull distribution along with the Method of Maximum Likelihood (used for parameter estimation) based on the incomplete dataset. The estimated return level is stored in an array. It then removes a new $p\%$ chunk from the original, complete dataset by moving on $N\left(1-\frac{p}{100}\right)/s$ (where N is the length of the dataset, without the $p\%$ chunk, and s is the number of y-year return levels to be estimated) measurements from the starting point of the previous $p\%$ chunk. In other words, if the previous $p\%$ chunk started at index (measurement) i, the next $p\%$ chunk starts at index $i+N\left(1-\frac{p}{100}\right)/s$. This process is repeated until the end of the dataset is reached. The function takes as input the following parameters:

- `dataset` – an array containing H_{mo}-values
- `p` – the percentage chunk to be removed from the dataset
- `s` – the number of return levels to be estimated (i.e., the number of times the process of removing a $p\%$ chunk from the dataset is repeated)
- `u` – the threshold level in order to determine the exceedances over the threshold
- `y` – the return period for which the return levels are to be determined
- `years` – the length of the dataset in years after the $p\%$ chunk has been removed

MATLAB CODE:

```
function ret_levs = remove_chunk1(dataset, p, s, u, y, years)
N = length(dataset);      % determine the number of measurements in the
                          % dataset
ret_levs = [];      % create an empty array for storing the estimated return
```

```
                    % levels
for i = 1:uint32((N-N*p/100)/s):uint32((N-N*p/100)/s)*s
           % loop through the dataset, each time removing a new p% chunk
    dataset_wout_chunk = dataset;
    for j = i:i+uint32(N*p/100)-1
        dataset_wout_chunk(j) = nan;       % remove a p% chunk from the
                                           % original dataset, starting at
                                           % index i
    end
    [alpha, xi] = Weibull_param(dataset_wout_chunk, u, 1);
             % determine parameters of the Weibull distribution by means
             % of the Method of Max Likelihood based on the incomplete
             % dataset
    ret_lev = Weibull_ret_lev_new(dataset_wout_chunk, u, y, alpha, xi, years);
            % determine the y-year return level
    ret_levs = [ret_levs;
                ret_lev];      %store the return level in an array
end
end
```

In order to illustrate how the code works a 10% gap will be used as example. The NCEP-dataset consists of 49856 measurements. This means that the for-loops in the code look as follows:

```
for i = 1:45:45000      % the 45 is rounded from 44.8704
...
    for j = i:i+4985
```

The 45 is rounded up from 44.8704. The rounding error is therefore $45 - 44.8704 = 0.1296$, which means that each of the first 999 10% chunks are 0.1296 measurements larger than they are supposed to be due to rounding. This, in turn, causes the last 10% chunk to be $0.1296 \times 999 = 129$ measurements smaller than it is supposed to be. This can be seen when $i = 45000$, since then the second for-loop runs from 45000 to $45000 + 4985 = 49985$. However, the original dataset contains only 49856 measurements, which means that the last chunk to be removed is $49985 - 49856 = 129$ measurements smaller than the previous 999 chunks due to rounding. This is illustrated in Figure E.1.1.

E.2 Removing $p\%$ From a Dataset in the Form of Random 1-Week Periods

The Matlab script below is a function that removes $p\%$ of the data values from a dataset in the form of random 1-week periods, and then predicts a specified y-year return level.

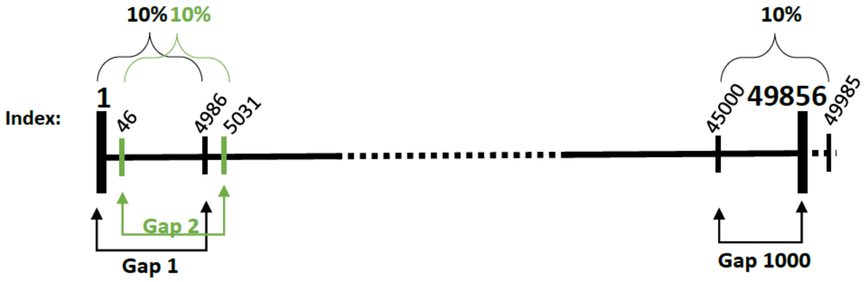


Figure E.1.1: Illustration of 10% ‘chunk’ gaps and smaller 1000th gap due to rounding.

This is done by means of the POT method and the Weibull distribution along with the Method of Maximum Likelihood (used for parameter estimation) based on the incomplete dataset. The estimated return level is stored in an array. This process is repeated a specified number of times. The function takes as input the following parameters:

- `dataset` – an array containing 3-hourly H_{mo}-values
- `p` – the percentage of data values to be removed from the dataset
- `s` – the number of return levels to be estimated (i.e., the number of times the process of removing a p% chunk from the dataset is repeated)
- `u` – the threshold level in order to determine the exceedances over the threshold
- `y` – the return period for which the return levels are to be determined
- `years` – the length of the dataset in years after the p% of the data values have been removed

MATLAB CODE:

```
function ret_levs = remove_gaps_1week(dataset, p, s, u, y, years)
N = length(dataset);      % determine the number of measurements in the
                          % dataset
ret_levs = [];      % create an empty array for storing the estimated return
                    % levels
for i = 1:s      % s return levels are to be estimated, each based on a
                 % dataset with p% of data values removed
    dataset_w_gaps = dataset;
    indices = randi(N-8*7, uint32(N*p/100/(8*7)), 1);
                 % generate random indices for starting points of 1-week
                 % periods to make up p% of dataset
                 % Last index is one week before end of dataset: N-8*7
                 % (8 = number of measurements per day)
    for j = 1:length(indices)      % create 1-week gaps in dataset, with
                                   % starting points given in ‘indices’
        dataset_w_gaps(indices(j), 1) = nan;
        for k = 1:8*7-1
            dataset_w_gaps(indices(j)+k, 1) = nan;
        end
    end
    [alpha, xi] = Weibull_param(dataset_w_gaps, u, 1);
             % determine parameters of the Weibull distribution by means
             % of the Method of Max Likelihood based on the incomplete
             % dataset
    ret_lev = Weibull_ret_lev_new(dataset_w_gaps, u, y, alpha, xi, years);
             % determine the y-year return level
    ret_levs = [ret_levs;
                ret_lev];      % store the return level in an array
end
end
```

Appendix F

Fabricated Dataset

The Matlab script below is a function that fabricates a complete dataset from an incomplete dataset by using the procedure decribed in Section 7.2.1. The function takes as input the following parameters:

- `dataset` – the incomplete dataset, in the form of an array containing 3-hourly H_{mo}-values. Missing values are indicated by the value 999.999
- `dates` – a matrix containing two columns with the dates of the measurements in the dataset in the first column (in the form “YYYYMMDD”) and the times of the measurements in the second column (in the form “HHMMSS”)
- `years` – the length of the dataset in years

MATLAB CODE:

```
function Hm0_fabr = fabricate_dataset(dataset, dates, years)
T = [number2string(dates(:,1),8) number2string(dates(:,2),6)];
mtime = datenum(T,'yyyymmddHHMMSS');
Hm0 = dataset;
Hm0_fabr = Hm0;
indices = (1:length(mtime))';

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% If a gap consists of 4 or less missing measurements, fill in by means of
% linear interpolation:
indices_wout_gaps = indices;
indices_wout_gaps(Hm0 > 999) = [];  % store only indices containing
                                    % measurements (remove all indices with
                                    % missing measurements)
for i = 2:length(indices_wout_gaps)
    i1 =  indices_wout_gaps(i-1);
    i2 =  indices_wout_gaps(i);
    if i2-i1>=1 && i2-i1<=4   % if 4 or less measurements are missing, do
                              % linear interpolation between points (i1;
                              % Hm0(i1)) and (i2; Hm0(i2)) to fill in Hm0
                              % values at indices between i1 and i2
        m = (Hm0(i2)-Hm0(i1))/(i2-i1);  % gradient
        c = Hm0(i2) - m*i2;  % y-intercept
        for j = 1:i2-i1-1
            Hm0_fabr(i1+j) = m*(i1+j)+c;    % do linear interpolation
        end
    end
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Fill remaining gaps (i.e., gaps consisting of 5 or more missing
% measurements) by replacing a gap with the Hm0-values from corresponding
% time in another year, where no measurements are missing:
for n = 1:18
indices_wout_gaps2 = indices;
indices_wout_gaps2(Hm0_fabr > 999) = [];
for i = 2:length(indices_wout_gaps2)
    i1 =  indices_wout_gaps2(i-1);
    i2 =  indices_wout_gaps2(i);
    for j = 1:i2-i1-1
        if i1+j+2920<=length(Hm0_fabr)  % if gap is not in last year,
                                        % replace gap by measurements
                                        % during correponding time in
                                        % following year
            Hm0_fabr(i1+j) = Hm0_fabr(i1+j+2920);   % 2920 = number of
                                                    % 3-hour periods in one
                                                    % year (i.e., 8*365)
        else    % if gap is in last year, replace gap by measurements
                % during corresponding time in first year
            Hm0_fabr(i1+j) = Hm0_fabr(i1+j-(years-1)*365*8);
                                                 % (years-1)*365*8 = number
                                                 % of 3-hour periods in
                                                 % years-1

        end
    end
end
end
```

Bibliography

Adan, I. J. B. F. & Kulkarni, V. G. (2003). Single-Server Queue with Markov Dependent Inter-Arrival and Service Times.

Balkema, A. A. & De Haan, L. (1974). Residual Life Time at Great Age. *The Annals of Probability*, *2*, 792–804.

Bobeé, B. & Rabitaille, R. (1975). Correction of Bias in the Estimation of the Coefficient of Skewness. *Water Resources Research*, *11*(6), 851–854.

Boos, D. D. (2003). Introduction to the Bootstrap World. *Statistical Science*, *18*, 168–174.

Burden, L. & Faires, J. D. (2001). *Numerical Analysis, 9th edition*. Brooks/Cole.

Burke, S. (1998). *Scientific Data Management*, *2*(2), 32–40.

Burke, S. (2001). Missing Values, Outliers, Robust Statistics & Non-parametric Methods. *Scientific Data Management, Europe online supplement*, 19–24.

Casella, G. & Berger, R. L. (2002). *Statistical Inference, 2nd edition*. Thomson Learning.

Chong, S. F. & Choo, R. (2011). Introduction to Bootstrap. *Proceedings of Singapore Healthcare*, *20*, 236–240.

Coles, S. (2001). *Introduction to Statistical Modeling of Extreme Values*. Springer.

Corbella, S. & Stretch, D. (2012). The Wave Climate on the Kwazulu-Natal Coast of South Afirca. *Journal of the South African Institution of Civil Engineering*, *52*(2), 45–54.

David, H. A. (1981). *Order Statistics, 2nd edition*. Wiley.

DHI (2003). *Extreme Value Analysis Reference Manual*.

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Fisher, R. A. & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, (pp. 180–190). Cambridge Univ Press.

Gilli, M. & Këllezi, E. (2006). An Application of Extreme Value Theory for Measuring Financial Risk. *Computational Economics*, *27*(2-3), 207–228.

Greenwood, P. E. & Nikulin, M. S. (1996). *A Guide to Chi-Squared Testing*. Wiley.

Hansen, L. P. (2007). Generalized Method of Moment Estimation. *Palgrave Dictionary of Economics*.

Holthuijsen, L. H. (2007). *Waves in Oceanic and Coastal Waters*. Cambridge University Press.

Hosking, J. R. M. (1990). L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *J. Royal Statis. Soc. B*, *52*(1), 105–124.

Hosking, J. R. M. (1991). Fortran Routines for Use With the Method of L-Moments. *IBM Research*.

Hosking, J. R. M., Wallis, J. R., & Wood, E. F. (1985). Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments. *Technometrics*, *27*(3), 251–261.

Landwehr, J. M. & Matatlas, N. C. (1989). Probability Weighted Moments Compared With Some Traditional Techniques in Estimating Gumbel Parameters and Quantiles. *Water Resources Research*, *15*(5), 1055–1064.

Lawson, C. L. & Hanson, R. J. (1974). *Solving Least Squares Problems*. Prentice-Hall.

Li, Y., Simmonds, D., & Reeve, D. (2008). Quantifying Uncertainty in Extreme Values of Design Parameters with Resampling Techniques. *Ocean Engineering*, *35*, 1029–1038.

Little, R. J. A. & Ruben, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.

Mann, I., McLaughlin, S., Henkel, W., Kirkby, R., & Kessler, T. (2002). Impulse Generation with Appropriate Amplitude, Length, Inter-Arrival, and Spectral Characteristics. *Journal on Selected Areas in Communications*, *20*.

Martins, E. & Stedinger, J. (2000). Generalized Maximum-Likelihood Generalized Extreme-Value Quantile Estimators for Hydrologic Data. *Water Resources Research*, *36*(3), 737–744.

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G., Desai, A. R., ad J H. Gove, E. F., Heimann, M., Hui, D., Jarvis, A. J., Karrge, J., Noormets, A., & Stauch, V. J. (2007). Comprehensive Comparison of Gap-Filling Techniques for Eddy Covariance Net Carbon Fluxes. *Agricultural and Forest Meteorology*, *147*.

Mooney, C. Z. & Duval, R. D. (1993). *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Chapman & Hall.

Northrop, P. J., Attalides, N., & Jonathan, P. (2015). Cross-Validatory Extreme Value Threshold Selection and Uncertainty with Application to Ocean Storm Severity.

Pickands III, J. (1975). Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, *3*, 119–131.

Rind, D. M. (2014). Proof, p-values, and Hypothesis Testing. *UpToDate, Inc.*

Ross, S. (2010). *A First Course in Probability, 8th edition*. Pearson.

Scarrott, C. & MacDonald, A. (2012). A Review of Extreme Value Threshold Esimation and Uncertainty Quantification. *REVSTAT – Statistical Journal*, *10*, 33–60.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Smith, R. L. (1989). Extreme Value Analysis of Environmental Time Series: An Application to Trend Detection in Ground-Level Ozone. *Statistical Science*, *4*(4), 367–393.

Snodgrass, F. (1951). Wave recorders. In *Proceedings of Conference on Coastal Engineering*, (pp.$\tilde{6}$9). American Society of Civil Engineers, United Engineering Center.

Sverdrup, H. U. & Munk, W. H. (1946). Empirical and Theoretical Relations Between Wind, Sea, and Swell. *Transactions of the American Geophysical Union*, *27*(6), 823–827.

Tanaka, S. & Takara, K. (2000). A Study on Threshold Selection in the Pot Analysis of Extreme Floods. *The Extremes of the Extremes: Extaordinary Floods*, *271*.

Teena, N., Kumar, V. S., Sudheesh, K., & Sajeev, R. (2012). Statistical Analysis on Extreme Wave Heights. *Natural Hazards*, *64*, 223–236.

Thompson, P., Cai, Y., Reeve, D., & Stander, J. (2009). Automated Threshold Selection Methods for Extreme Wave Analysis. *Coastal Engineering*, *56*, 1013–1021.

Wackerly, D., Mendenhall, W., & Scheaffer, R. (2008). *Mathematical Statistics with Applications, 7th edition*. Thomson.