Uni- and bivariate statistical analysis of long-term wave climates

A Repko
Delft, August 1998
Delft University of Technology

Preface

This report is written as a part of my study for the MSc. degree at the Faculty of Civil Engineering at the Delft University of Technology, Hydraulic engineering group.

During the research I have been adviced and assisted by prof. drs. ir. J.K. Vrijling, ir. P.H.A.J.M. van Gelder, ir. H.G. Voortman and dr. ir. L.H. Holthuijsen. For this, I would like to express my gratitude to them.

Finally, I would like to acknowledge ir. A.P. Roskam of RIKZ for providing the North Sea data set.

Delft, August 1998

Albert Repko

Contents

Appendixes

1 Introduction

For the probability-based design and assesment of marine structures interacting with sea waves, a reliable knowledge of the long-term wave climate is required. Wave climate data are commonly presented in the form of histograms of spectral wave parameters. The severity of a sea state is usually expressed in terms of significant wave height H_s and corresponding wave period T.

From the earlier stages of the development of a statistical approach to wave climate, the advantage of an analytical representation of empirical distributions of data through parametric models was recognized. The compactness of analytical description, the standardization of the representation, and the filling of information gaps, led researchers to use specific marginal and bivariate parameter models, suitable for the description of wave height and wave period statistics.

Previous case studies
With regard to the marginal distribution functions of H_s and T, a large amount of case studies is present in the literature. The most extensive studies that have been found were examined by Mathiesen et al. (1993) and Maes et al. (1994). These two studies provide a detailed analysis of the influence of the data selection procedure, the parameter estimation method and the chosen distribution function for the estimation of return values (the significant wave height or wave period corresponding to a return period of for example 50 years.)

With regard to the bivariate distribution functions of H_s and T, one of the first publications is made by Houmb and Overvik (1976). They proposed to use a marginal distribution for H_s and a conditional distribution for T. Ochi (1978) utilized the bivariate Log-normal distribution for the bivariate model. Fang and Hogben (1982) proposed a development of the bivariate Log-normal distribution, in which a correction for the skewness of the significant wave height distribution is included. It is the above two basic approaches (Log-normal model and conditional distribution approach) that have been used, almost exclusively up to present, for the joint wave height and wave period statistics (Haver (1985), Burrows et al .(1986), Mathisen et al. (1990) and Chung-Chu Teng et al. (1996)).

Besides the above mentioned approaches, bivariate models that are based on the marginal distribution functions are present in literature. These joint distribution functions contain a parameter ψ describing the dependence between H_s and T. The parameter ψ is then defined by some complex formulae which is called the dependence structure between H_s and T. Such models were first posed by Fréchet in 1951, and, accordingly, the class of solutions of this problem is called the Fréchet class. Mainly in more mathematical orientated literature ((Athanassoulis et al (1993), Morton and Bowers (1997), Mardia (1970), Johnson and Kotz (1972), Johnson (1981), and Metcalfe (1997)), this type of approach is found. The models are more sophisticated than the two above mentioned approaches. For standard applications in civil engineering practice these models might probably be too complicated and might not fit the wave data.

Aim of present study
The aim of the present study is to find a particular bivariate distribution function for H_s and T, which provides a close fit to long-term (extreme) wave data presenting a deep water wave field. Several types of joint distribution function for H_s and T are compared with reference to measured data. The comparison is based on the utility of the distribution functions for predictions of extreme sea states. The report is thus concerned with the estimation of extreme significant wave heights and wave periods (zero-up-crossing periods or spectral peak periods).

The present study of bivariate functions is similar to the above mentioned case studies of marginal distributions. It provides a detailed analysis of the influence of the data selection procedure, the parameter estimation method and the chosen distribution function on the estimation of bivariate return values.

In total five bivariate probability models are tested for the joint statistics of H_s and T. These are:
- the bivariate Log-normal distribution

- the bivariate Log-normal distribution with correction for skewness (the Fang and Hogben distribution)
- the bivariate distribution constructed from a marginal distribution for H_s and a conditional distribution for T
- the bivariate distribution based on a marginal distribution for H_s and a marginal distribution for the (deepwater) wave steepness s
- the bivariate distribution with given marginals developed by Morton and Bowers (1997)

The fourth model is proposed by Vrijling (1996). It is based on the assumption that the significant wave height (H_s) and the wave steepness (s) are independent. With in the calculations, first the bivariate distribution of H_s and s is computed by simply taking the product of the marginals of H_s and s. Then the bivariate distribution of H_s and T is determined by transforming the joint model of H_s and s.

The fifth model is a distribution of the Fréchet class. Morton and Bowers (1997) have published an article in which a detailed description is given about the application of the model to extreme wave height and windspeed observations. They obtained good results. No further tests of the model are known to the author. Therefore, the model is included in the present study.

Outline of the report

The report is divided in two parts. The first part describes each component of the statistical wave analysis. These are

- the selection of extreme wave data (Chapter 2)
- the selection of uni- and bivariate probability distributions (Chapter 3)
- the estimation of distribution parameters (Chapter 4)
- the assesment of the goodness of fit of the probability models to the data (Chapter 5)
- the calculation of return values and confidence bands for return values (Chapter 6)

Furthermore it contains a description of the computer program (Chapter 7) that has been written for the calculations.

The second part, i.e. chapter 8, contains two case studies. Marginal and bivariate distributions are fitted to wave data measured at the coast of India (Karwar) and at the Euro platform located in the North Sea.

Finally, chapter 9 presents the conclusions and recommendations of the study.

References

[1] Vrijling, J.K., *Probabilistic design in hydraulic engineering. (in Dutch).* Delft: Delft University of Technology, Faculty of Civil Engineering, 1996

Case studies considered with the marginal distribution of H_s

[1] Cavaleri, L., De Filippi, P.L., Grancini, G.F., Lovenitti, G.L., and Tosi, R., *"Extreme wave conditions in the Tyrrhenian sea."*. Ocean Engineering, 1986, Vol. 13, no. 2, p 157-180

[2] Cristopoulos, S., Solomonidis, C., *"Wave climate assesment in the South Aegean shelf"*. Orlando, Proc. 26th Int. conf. coastal. eng., 1996, pp. 689-702

[3] Deo, M. C., Burrows, R., *"Extreme wave prediction using directional data"*. Proc. 20th Int. conf. coastal. eng., 1986, pp. 236-149

[4] Hatada, Y., Yamaguchi, M., *"Eight years wave hindcast and analysis of wave climate"*. Venice: Proc. 23th Int. conf. coastal. eng., 1992, pp. 116-127

[5] Kato, H., Nobuoka, H., *"Estimation of persistence of the waves observed on Japanese coast in the light of recent studies"*. Orlando, Proc. 26th Int. conf. coastal. eng., 1996, pp. 794-807

[6] Leyden, V.M., Dally, W. R., *"Probabilistic modelling of long-term wave climate".* Orlando, Proc. 26th Int. conf. coastal. eng., 1996, pp. 807-821

[7] Maes, M. A., and Gu, G.Z.,*"Techniques used to determine extreme wave heights from the NESS Data set"*. Gaithersburg: Proc. of the conf. on extreme value theory and applications, volume 2, 1994, pp. 435-444

[8] Mathiesen, M. et al., *"Intercomparison of extreme wave analysis: a comparitive study".* New Orleans: Proc. WAVES '93 conf., july 1993, pp 963-977

[9] Rossouw, J., 1988, *"Design waves and their probability density functions".* Malaga: Proc. 21th int. conf. coastal eng., 1988, pp. 137-139

[10] Rossouw, J., Medina J.R., *"Stability of design wave estimates".* Orlando, Proc. 26th Int. conf. coastal. eng., 1996, pp. 328-339

[11] Yamaguchi, M., *"Intercomparison of parameter estimation methods in extremal wave analysis".* Orlando, Proc. 26th Int. conf. coastal. eng., 1996, pp. 900-913

Numerical simulation studies considered with the marginal distribution of H_s

[1] Burcharth, H.F., and Liu,Z, *"On the extreme wave height analysis".* Yosuka: Proceedings of the International Conference on Hydro-Technical Engineering for Port and Harbour Construction, 1994

[2] Goda, Y.,*"On the methodology of selecting design wave height".* Malaga: Proc. 21th Int. conf. coastal eng., 1988, pp. 135-136

[3] Goda,Y.,Kobune,K., *"Distribution function fitting for storm wave data".* Delft: Proc. 22th Int. conf. coastal. eng., 1990, pp. 18-31

[4] Mathiesen, M. et al., *"Intercomparison of extremal wave analysis methods using numerically simulated data".* New Orleans: Proc. WAVES '93 conf., july 1993, pp 963-977

[5] Van Gelder, P.H.A.J.M., Vrijling, J.K., *"A comparative study of different parameter estimation methods for statistical distributions functions in civil engineering applications".* Tokyo : ICOSSAR'97, 7th International Conference on Structural Safety and Reliability Kyoto, 1997

Case studies considered with the bivariate distribution of H_s and T

[1] Athanassoulis, G. A., Skarsoulis, E.K., Belibassakis, K.A., *"Bivariate distributions with given marginals with an application to wave climate description".* Applied ocean research, 1994, v. 16, p 1-17

[2] Battjes, J. A.,*"Long-term wave height distributions at seven stations around the British Isles.".* N.I.O. Report No. A 44., 1970

[3] Burrows, R., Salih, B. A., *"Statistical modelling of long-term wave climates".* Proc. 20th Int. conf. coastal. eng., 1986, pp. 42-56

[4] Fang, Z.S., and Hogben, N., *"Analysis and prediction of long-term probability distributions of wave height and periods".* London: Technical report, National Maritime Institute, 1982 (*)

[5] Haver, S., *"Wave climate off northern Norway".* Applied ocean research, 1985,v 7, p 85-92

[6] Mathisen, J., and Bitner-Gregersen, E., *"Joint distributions for significant wave height and wave zero-up-crossing period".* Applied Ocean Research, 1990, Vol. 12, no. 2, pp 93-103

[7] Morton, I.D., Bowers, J., *"Extreme value analysis in a multivariate offshore environment".* Applied Ocean Research, 1997, v 18, pp. 303-317

[8] Ochi, M. K., *"Wave statistics for the design of ships and structures".* New York: Proc. SNAME Conf., 1978 (*)

[9] Tang, C., and Palao, I.M., *"Wave height and period distributions from long-term wave measurement".*: Proc. 25th Int. conf. coastal. eng., 1996, pp. 368-378

Literature in which the theory of bivariate distributions with given marginals is described

[1] Johnson, N.L., and Kotz, S., *Distributions in statistics: continuous multivariate distributions.* New York: John Wiley, 1972.

[2] Johnson, M.E., *Multivariate statistical simulation.* New York: John Wiley, 1987

[3] Mardia, K.V., *Families of bivariate distributions.* London: Griffin, 1970

[4] Metcalfe, A. V., *Statistics in Civil Engineering.* New York: John Wiley & Sons Inc, 1997

N.B:

(*) These articles have been mentioned in one of the other articles. The marked articles have not been studied by the present author.

2 Data selection

2.1 Introduction

In civil engineering practice the following sources of wave data are currently used:

- visual observations
- instrumentally observed data
- hindcasted data

Visual observations are routinely made from many ocean weather ships and land based stations.The accuracy of the observations are wholly dependent on the experience and skill of the observer. An example of measured data are sets containing 3-hourly wave buoy observations. The accuracy of such measurements is relatively high. In the case of hindcasted wave data, historical wind data are used in numerical wave models to calculate the wave conditions in the area of interest. The accuracy of the hindcasted data depends on the type of wave growth model that is used.

As mentioned in the introduction, the present study is concerned with the analysis of extreme wave data. Extreme data must fulfil the following demands:

- the data set is homogeneous,
- the observations are independent,
- the data set fully represents the maxima of the physical proces.

The first demand means that all data points with in the data set belong to one parent population with an unknown distribution. The stationarity of the proces has to be studied. For wave data, this means that the influence of different wave directions (stationarity in space) and the influence of seasonal changes and inter-annual changes (stationarity in time) have to be analyzed. The homogeneity of the data sets is dealt with in section 2.2.

Observations of significant wave heights are assumed to be independent when they correspond to seperate storm events. To obtain a set of independent observations of H_s, the (mean) duration of a storm event in the wave field considered must be known. If for example a (mean) storm lasts about 30 hours, one knows that when the time interval between two successive observations of H_s is smaller than 30 hours, the two observations are correlated.
In the bivariate case, the selection procedure is more complex since it is possible for the variables to have their maxima occuring at different time intervals. In section 2.3, this problem is discussed.

The third demand requires a selection procedure for extreme storm events. In general, three of such procedures are in use. The first two approaches are statistical methods: *the peak over treshold method* and *the annual maxima method*. The third method is a physical approach: extreme observations are obtained by making distinction between storms of different nature. In section 2.4, these methods are described.

2.2 The homogeneity of the data set

2.2.1 Omnidirectional and directional data

Usually omnidirectional data is used for data fitting. The estimates made in this way do not account explicitly for the directions of the wave or the winds generating them. An alternative is to categorize the data into different direction sectors and fit for each sector theoretical probability distribution functions. The total sample is thus divided into smaller samples. The distribution is then given by

$$P(H_s) = \sum_{all\theta} P(H_s \mid \theta) W(\theta) \tag{2.1}$$

where

$P(H_s)$	=	distribution function of H_s
$P\ (H_s \vert\ \theta)$	=	conditional distribution function of H_s for given θ;
$W\ (\theta)$	=	weighting function representing proportion of sea states along direction θ in The entire population.

The problem of the above approach is that the statistical uncertainty of the fittings proces increases when the sample size decreases. In a case study (Burrows and Salih; 1986), it appeared that predictions of extreme wave heights from certain worst direction sectors exceeded the equivalent prediction using the complete data set.

2.2.2 Seasonal and inter-annual climate variability

In many cases, most heavy storms occur during some specific period of a year. In such situations, one can choose to use only observations corresponding to that period of time. However, one must always be aware of the possibility that important storm events may occur during the censored part of the year.

The change of weather between successive years, also called long-term weather changes or inter-annual climate variability, can be studied by analyzing historical metereological data covering several decades. Long-term trends that are significant should be considered in the wave analysis.

2.2.3 Missing data

If there is a large amount of missing data, one has to check the occurrence of important storm events during the missing time intervals. One can choose to fill the gaps of the sample with data from other sources, for example visual observations. However, one must always be aware of the possibility that data coming from a different source is of a different degree of accuracy and probably measured at a different location.

For marginal distributions, Castillo et al. (1994) proposed to raise the estimated cumulative distribution function to the power (n+r)/n, where n and r are the number of known and unknown peaks, respectively.

2.3 The independency of extreme observations

Observations of extreme significant wave heights are considered to be independent when they represent seperate storm events. In general, a storm event lasts more than 3 hours. This implies that if a data set covers several storms and consists of 3-hourly measurements some of the data points in the sample are correlated.

In order to obtain independent observations representing local maxima the minimum time between successive storm events must be determined. This time interval, which is also called the cluster interval, depends on the wave climate that is considered. A study of wave heights off the coast of Norway (Mathiesen et al. (1993)) used an interval of 18 h. Another study considered with cyclones in the mid-latitudes (Oke (1987)) worked with a time interval of 1 day.

The declustering of data is illustrated in an application to a sequence of significant wave heights in figure 2.1. This figure and figure 2.2 are taken from a study made by Morton and Bowers (1997). They used 3-hourly wave height and wind speed data measured in the northern North Sea and adopted a time interval of 30 h.

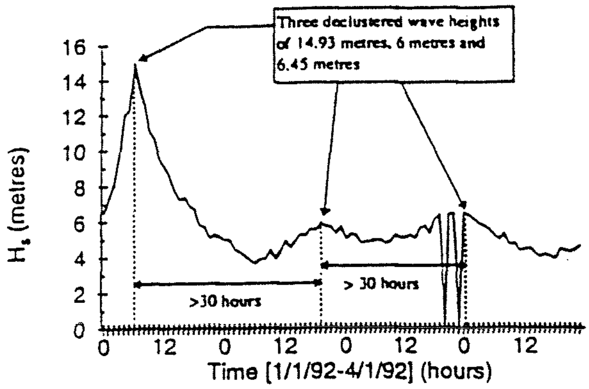


Fig 2.1 Declustered wave heights (Taken from Morton and Bowers (1997))

The declustering of bivariate data is not immediately obvious. This is illustrated in figure 2.2. While the same cluster interval of 30 h should be applicable, unique local maxima can no longer be defined. In this report, the data points corresponding to the local maxima of H_s are assumed to represent the local maxima of H_s and T.

Appendix [1] provides some details of the computer program that has been written for the declustering of the presently used wave data sets.

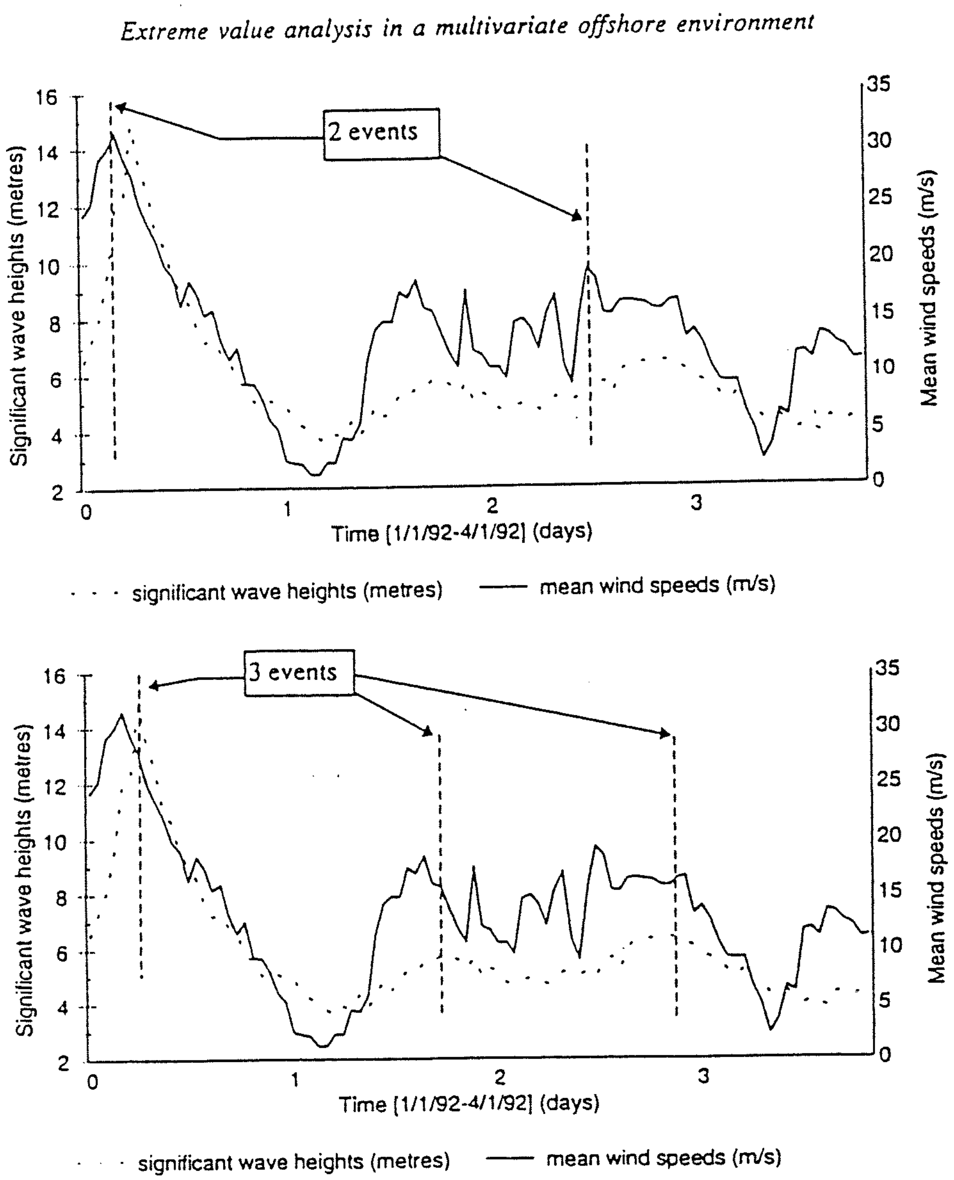


Fig 2.2 Extremes and concomitants. (a) Maximum mean wind speeds and concomitant significant wave heights. (b) Maximum significant wave heights and concomitant mean wind speeds. (Taken from Morton and Bowers (1997))

2.4 The selection of extreme observations

The Peak Over Treshold (POT) method

The peak over treshold (POT) method is used to detect significant storm periods. For the marginal analysis of significant wave heights, a common technique is to apply a fixed treshold value to identify storm periods comprising a sequence of sea-states with wave heights all exceeding the given treshold. It is then common practice to select the highest wave in a storm period, which is denoted as the peak or extreme significant wave height.

The POT method is subjective, since the height of the treshold has to be chosen. Several studies (Mathiesen et al. (1993), Maes et al. (1994)) have attempted to determine a standard method for the choice of the treshold level. Though, an appropriate procedure has not yet been found. In general, a low treshold has the disadvantage that data points, which not represent extreme events, are involved, whereas a high treshold implies relative large statistical uncertanties.

An application of the POT method in the bivariate case, at least for wave height and wind speed maxima, is presented in the paper of Morton and Bowers (1997). For the determination of the joint treshold, they used procedures, which are specified to the bivariate distribution function with given marginals. According to the article, similar procedures were used in several other studies (Joe et al (1992), Coles and Tawn (1994)). Applications of the followed approach to other types of bivariate models seem difficult. The method is described in appendix [3.3], together with the description of the bivariate model of Morton and Bowers.
In the present report, the significant wave height is considered as the key parameter for the joint treshold. Since the wave period and wave height observations are coupled, the treshold level of the significant wave height can be used for the selection of extreme pairs of H_s and T.

When the POT method is used, the final set of data for the extremal analysis is determined by the *minimum duration between successive storm events*, i.e. the declustering of the observations (section 2.3), and the height of the treshold level.

The annual maxima (AM) method
The annual maxima method selects the highest significant wave height per year. The annual maxima method has been critized by several authors (Castillo et al. (1994)). Their objection against the method is that it discards large significant wave heights, when they occur in years with large storms, but includes relative small significant wave heights that are maxima of calm years. (Inter-annual climate variability, section 2.2.2).

No application of the annual maxima method has been found in the bivariate case.

On physical consideration
This approach classifies and groups storm waves of different nature (waves generated by hurricanes, monsoons and frontal systems, for example). Further it identifies and separates waves generated by wind fields of different directions. In this way the hypothesis, that all the data come from the same type of event, is better.

The difference with the statistical approach is that the selection of storm events is not fully based on the set of wave data. In addition, other sources of information are used to detect the typical features of the wave field considered. For instance, the local geography, the local bottom geography and the pattern of local winds are studied in order to obtain realistic estimates of extreme storm events.

Example of the physical approach: Extreme wave conditions in the Thyrrenian Sea
An example of the physical approach is described by Cavaleri et al. (1986). This hindcast study was concerned with the prediction of extreme wave heights in the Tyrrhenian Sea (Mediterranean Sea). To classify the different local storm events they started from the daily classification of the European weather used by the Deutscher Wetterdienst. This classification is based on a pattern scheme made by Hess and Brezowsky (1969). It includes 29 different patterns. For practical purpose, Caveleri et al. reduced this number to 4 (named A,B,C and D) as shown in the table below.

Table 2.1 Relationship between the four basic types of storms identified in the Thyrrhenian Sea and the European weather definitions by Deutscher Wetterdienst (Taken from Cavaleri et al. (1986))

Storm type	Deutscher Wetterdienst definition
A	NA,SA,SWA,SWZ,NWA,HM,HNA,HB,HFA
B	NEA,NEZ,NZ,NWZ,TRM,HFZ,HNFZ,BM
C	HNZ,WS,WZ,WA,WW,SZ,SEA
D	TM,SEZ,TRW,TB,HNFA

On basis of 30 years of meteorological (wind) data they selected 80 storm events and classified them to the four different classes of storms. In this way, four homogeneous sets of extreme data were obtained, representing four different types of storms.

The number of storms (80) that has been chosen is subjective. In the case of a hindcast study it depends basically on two factors, (a) the costs of the computations, (b) the accuracy of the final results.

Another example of the physical approach is presented in a paper written by Van Gelder and Vrijling (1998). In their article, the homogeneity of wave data sets is discussed. Also the delta report, which is dealt with the analysis of (North-Westerly) storms in the southern North Sea, provides a good example of the "physical approach".

2.5 References

[1] Burrows, R., Salih, B. A., *"Statistical modelling of long-term wave climates".* Proc. 20th Int. conf. coastal. eng., 1986, pp. 42-56

[2] Cavaleri, L., De Filippi, P.L., Grancini, G.F., Lovenitti, G.L., and Tosi, R., *"Extreme wave conditions in the Tyrrhenian sea.".* Ocean Engineering, 1986, Vol 13, no. 2, p 157-180

[3] Castillo, E.,Sarabia, J. M., *"Extreme value analysis of wave heights".* Gaithersburg: Proc. of the conf. on extreme value theory and applications, volume 2, 1994, p. 445-454

[4] Coles, S.G., and Tawn, J.A., *"Modelling extreme multivariate events".* Journal of the royal Statistical Society, 1991, 53(2), p 377-392 (*)

[5] Herbich, J. B., *Handbook of coastal and ocean engineering, Volume: 1 Wave Phenomena and Coastal Structures.* Houston: Gulf publishing Company, 1990

[6] Hess, P., and Brewosky, H. , *Catalogue of European large-scale weather patterns.* Dt. Wetterd., Ber. 113(15), 54, 1969 (*)

[7] Joe, H., Smith, R.L., and Weissman, I., *"Bivariate tresholds methods for extremes".* Journal of the Royal Society, B, 1992, 54(1), p 171-183 (*)

[8] KNMI, Report of the Delta committee (1960)

[9] Mathiesen, M., Peltier, E., Thompson, E., Van Vledder, G., Goda, Y., Hawkes, P., Mansard, E., Martin, M.J. *"Case studies of extreme wave analysis: a comparative analysis".* Proc. WAVES '93 conf., New Orleans: July 1993, pp. 963-977

[10] Maes, M. A., and Gu, G.Z., *"Techniques used to determine extreme wave heights from the NESS Data set".* Gaithersburg: Proc. of the conf. on extreme value theory and applications, volume 2, 1994, pp. 435-444

[11] Morton, I.D., Bowers, J., *"Extreme value analysis in a multivariate offshore environment".* Applied Ocean Research, 1997, v 18, pp. 303-317

[12] Oke, T.R., *"Boundary layer climates".* London: Methuen, 1987 (*)

[13] Van Gelder, P.H.A.J.M., and Vrijling, J.K. *"Homogeneity aspects in statistical analysis of coastal engineering".* ICCE, 1998

N.B:

(*) These articles have been mentioned in one of the other articles. The marked articles have not been studied by the present author.

3 Probability density functions

3.1 Introduction

Section 3.2 provides background information about the selection of marginal distributions for an extreme wave analysis. First marginals are described which are based on the extreme value theory. Second, marginals are discussed which have much been applied and recommended in earlier case studies. In section 3.3 the presently used marginals are given.

The used bivariate distributions have already been introduced in chapter 1. These are:

- the bivariate Log-normal distribution
- the bivariate Log-normal distribution with correction for skewness
- the bivariate distribution based on the product of a marginal and a conditional distribution
- the bivariate distribution proposed by Vrijling (1996)
- the bivariate distribution with given marginals developed by Morton and Bowers (1997)

The bivariate probability models are described in section 3.4.

3.2 Candidate marginal distribution functions

3.2.1 The three asymptotic distributions of extreme values

Let the probability density function and the cumulative distribution function of a random variable X be denoted by f(x) and F(x), respectively. These functions are called the initial probability density function and the initial cumulative distribution function, respectively. The largest value expected to occur in n observations, denoted by Y_n, is also a random variable and follows its own probability law, which is different from that applicable for the random variable X. Let the probability density function and the cumulative distribution function of this variable be written as $g(y_n)$ and $G(y_n)$, respectively. The probability functions f(x), F(x), $g(y_n)$, and $G(y_n)$ have mathematical relationships. Therefore, the extreme values can be evaluated precisely from knowledge of the initial probability distribution.

These relationships can be derived by use of order statistics. Let a set of observations $(x_1,x_2,...x_n)$ be a random sample of size n from a distribution with probability density function f(x). When the elements of this random sample are rearranged in ascending of magnitude such that $y_1< y_2 < ....<y_n.$, then $(y_1,y_2,..,y_n)$ is called the ordered sample of size n, and Y_j is called the jth order statistics (j = 1,2...n). (Y_j is a stochast, y_j is the "actual value").

As discussed in the previous chapter, the random variables $X_1,X_2,..X_n$ are statistically independent and have all the same probability density function f(x). On the other hand, the random variables $Y_1,Y_2..,Y_n$ are statistically independent and each has its own probability density function.

The joint probability density function of $Y_1,Y_2,...,Y_n$, denoted by $g(y_1,y_2,...,y_n)$, is given by

(3.1) $$g(y_1,y_2,..,y_n)=n!\,f(y_1)f(y_2)...f(y_n)$$

The probability density function of the largest value, Y_n, can be obtained from equation (3.1) in terms of the probability density function f(x) and the cumulative distribution function F(x) as follows:

(3.2) $$g(y_n)=\int_{-\infty}^{y_n}....\int_{-\infty}^{y_3}\int_{-\infty}^{y_2}n!\,f(y_1)f(y_2)..f(y_n)dy_1dy_2..dy_{n-1}$$

Carrying out successive integration, the following equation can be derived

(3.3) $$g(y_n)=n!\,f(y_n)\frac{\{F(y_n)\}^{n-1}}{(n-1)!}=nf(y_n)\{F(y_n)\}^{n-1}$$

The cumulative distribution function of Y_n becomes,

(3.4) $$G(y_n)=\int_{-\infty}^{y_n}g(y_n)dy_n=\{F(y_n)\}^n$$

Rewritting equation (3.4) gives:

(3.5) $$G(y_n)=\{F(y_n)\}^n=e^{n\log(F(y_n))}$$

It can be seen in the above equation that $\log(F(y_n)) \to 0$ for $n \to \infty$. Hence, the value of G(y$_n$) depends entirely on the asymptotic behaviour of the initial distribution function F(x) towards the extreme value.
If the tail of the initial (cumulative) distribution is of the exponential type, the distribution asymptotically converges to the *Type 1 asymptotic extreme value distribution.* The initial distribution is unlimited towards the extreme value and all moments exist. Examples are the exponential, normal, log-normal, chi-square, and so on.

If the tail of the initial distribution is given by

$$(3.6) \qquad F(x) = x^{-1/b},$$

in which b is a constant value, the distribution asymptotically converges to the *Type 2 asymptotic extreme value distribution.* The initial distribution is unlimited towards the extreme value but only a finite number of moments exist.

If the tail of the initial distribution is linear, the distribution asymptotically converges to the *Type 3 asymptotic extreme value distribution.* The initial distribution is limited at the upper and/or lower bounds.

Type I: the Exponential family
The Exponential family is also called the Fisher-Tippett-type 1. For fitting of wave data, the main distribution of this family is the Gumbel distribution.

The cumulative distribution function of the Gumbel is given by

$$(3.7) \qquad F(x) = \exp\left[-\exp\left(-\frac{(x-\lambda)}{\delta}\right)\right]$$

The Exponential family of distributions permits unlimited values of the variables. The area under the tail of the distribution curve must converge to zero for large values of the variable at least as strongly as the tail of the Exponential function, Exp(-x). For the members of this family al moments exist.

Type II: the Cauchy distributions
The Cauchy distribution is also called the Fisher-Tippet-type 2. This family of distributions has no moments beyond a certain level order. The distributions have very long tails so that they converge less strongly than those of the Exponential family do. For the statistical analysis of wave data, the main distribution of this family is the Frechet distribution.

The cumulative distribution function of the Frechet is given by

$$(3.8) \qquad F(x) = \exp\left[-\left(\frac{\delta}{x-\lambda}\right)^{\beta}\right]$$

The Cauchy distribution has no variance. The sample variance will increase in an unbounded way as the sample size gets large. (Thus the variance of the distribution does not converge to the sample variance as the sample size increases).

Type III: the Weibull family
The Weibull distribution is also called the Fisher-Tippet-type 3. The members of the family are distributions with an upper or lower limit. The largest or smallest extreme value is thus bounded; the limit is a parameter of the extreme value distribution.
Examples of this family are the Weibull distribution and the Beta distribution.

The cumulative distribution function of the Weibull is given by

$$F(x) = \exp\left[-\left(\frac{x-\lambda}{\delta}\right)^{\beta}\right] \tag{3.9}$$

3.2.2 Distribution functions recommended in earlier case studies

Besides the above mentioned functions, in case studies also other distributions are used for extreme wave analysis. Examples are the Weibull distribution, the Log-normal distribution, the Pareto distribution, the Logistic distribution and the Exponential distribution.

In most of these studies (see references of chapter 1), the Gumbel, the Weibull and the Log-normal distribution provided the best fit. The good fit of the Weibull and the Log-normal distribution might be explained by their relative long tails. According to the earlier case studies, these two functions provide a better fit to extreme wave data than the Frechet, which is theoretically justified.

3.3 Marginal distributions of present report

Apparently, a large number of distribution functions are used for extreme wave analysis. It would be too extensive to test each of these distributions. Therefore, only the most recommended and applied distributions are tested in the present report. These are:
- the Exponential distribution
- the Gumbel distribution;
- the three-parameter Frechet distribution;
- the three-parameter Weibull distribution;
- the two-parameter Log-normal distribution.

Some parameters of the above distribution functions are listed in table 3.2.

Table 3.2 Some parameters of the chosen marginal distributions (Taken from Castillo (1988))

	Gumbel	Weibull	Frechet
F(x)	$\exp\left[-\exp\left(-\frac{(x-\lambda)}{\delta}\right)\right]$	$1-\exp\left[-\left(\frac{x-\lambda}{\delta}\right)^{\beta}\right]$	$\exp\left[-\left(\frac{\delta}{x-\lambda}\right)^{\beta}\right]$
f (x)	$\frac{1}{\delta}\exp\left[-\frac{(x-\lambda)}{\delta}\right]\exp\left[-\exp\left(-\frac{(x-\lambda)}{\delta}\right)\right]$	$\frac{\beta}{\delta}\left[\frac{x-\lambda}{\delta}\right]^{\beta-1}\exp\left[-\left(\frac{x-\lambda}{\delta}\right)^{\beta}\right]$	$\frac{\beta}{\delta}\left[\frac{\delta}{x-\lambda}\right]^{\beta+1}\exp\left[-\left(\frac{\delta}{x-\lambda}\right)^{\beta}\right]$
Range	$-\infty < x < \infty, 0 < \delta < \infty, -\infty < \lambda < \infty$	$X \leq \lambda$ $, 0 < \delta < \infty, 0 < \beta < \infty$	$x \geq \lambda, \delta > 0, \beta > 2$
Mean	$\lambda + 0.57772\delta$	$\lambda + \delta\Gamma\left(1+\frac{1}{\beta}\right)$	$\lambda + \delta\Gamma\left(1-\frac{1}{\beta}\right)$
Variance	$\frac{\pi^2\delta^2}{6}$	$\delta^2\left[\Gamma\left(1+\frac{2}{\beta}\right)-\Gamma^2\left(1+\frac{1}{\beta}\right)\right]$	$\delta^2\left[\Gamma\left(1-\frac{2}{\beta}\right)-\Gamma^2\left(1-\frac{1}{\beta}\right)\right]$

	Exponential	Log-normal
F(x)	$1-\exp\left[-\frac{(x-\lambda)}{\delta}\right]$	$\frac{1}{\sqrt{2\pi\delta}}\int_0^x \frac{1}{x}\exp\left[-\frac{1}{2}\left(\frac{\log(x)-\lambda}{\delta}\right)^2\right] \quad dx$
f (x)	$\frac{1}{\delta}\exp\left[-\frac{(x-\lambda)}{\delta}\right]$	$\frac{1}{\sqrt{2\pi\delta}}\exp\left[-\frac{1}{2}\left(\frac{\log(x)-\lambda}{\delta}\right)^2\right]$
Range	$X \geq \lambda$ $, 0 < \delta < \infty$	$-\infty < x < \infty, -\infty < \lambda < \infty, -\infty < \delta < \infty$
Mean	$\lambda + \delta$	$\exp\left[\lambda + \frac{1}{2}\delta^2\right]$
Variance	δ^2	$\exp(2\lambda)\exp(\delta^2)(\exp(\delta^2)-1)$

3.4 Bivariate distribution functions

Before the theoretical distributions are introduced, first some typical features of bivariate wave observations are discussed. Figure 3.1 shows an example of observations of H_s and T_p at one location. It is nearly always found that the area containing the observations is bounded on the upper side by a line of constant wave steepness (Tucker (1991)). In the illustration, a wave steepness of 5% has been used. Assuming a deep water wave field and starting from the linear wave theory, the wave steepness is defined as

(3.10) $$s_p = \frac{H_s}{L_0} = \frac{H_s}{\left[\frac{gT_p^{\,2}}{2\pi}\right]}$$

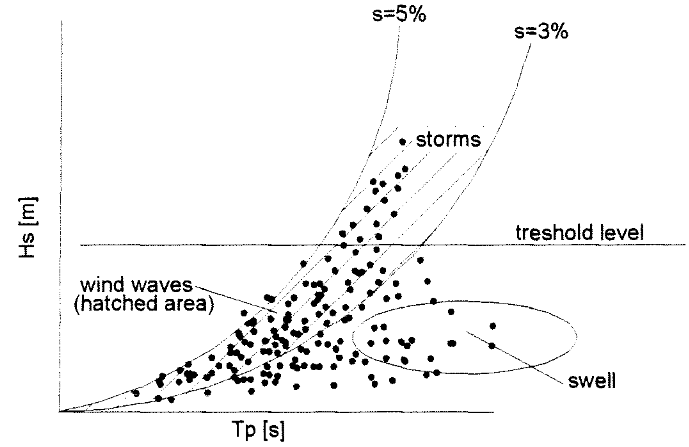


Fig. 3.1 Example of observations of H_s and T_p at one location

The upper figure represents a wave climate under the influence of both wind waves and swell. The correlation between the significant wave height and peak periods is insignificant. For an extreme wave analysis, the observations of swell should be censored and only the observations above a relative high treshold level (the choice of the treshold level is subjective (chapter 3)) should be selected. The correlation between the extreme observations, shown in figure 3.2, is relative strong.

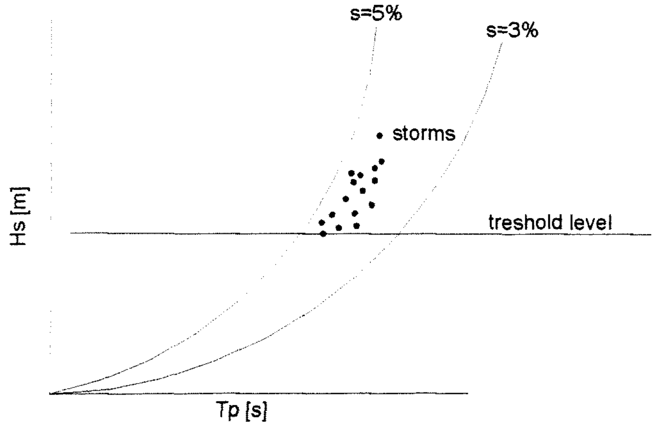


Fig 3.2 Extreme observations of H_s and T_p (corresponding to figure 3.1)

The location parameter of this conditional function is equal to the conditional expectation of T_p given H_s=c and may be written as (see appendix [2])

(3.17)
$$g(H_s) = E(T_p \mid H_s) = \exp\left[\lambda_{T_p} + \rho\left(\frac{\delta_{H_s}}{\delta_{T_p}}\right)\left(\log H_s - \lambda_{H_s}\right)\right]$$

The above expression is called *the regression function of T_p on H_s* . This regression function describes the wave period as a function of the significant wave height and gives an indication in which way the model relates H_s and T_p. (See fig 3.3). Apparently, the bivariate model describes the correlation between H_s and T_p with an exponential function. Note that the scale parameter of the conditional function is constant.

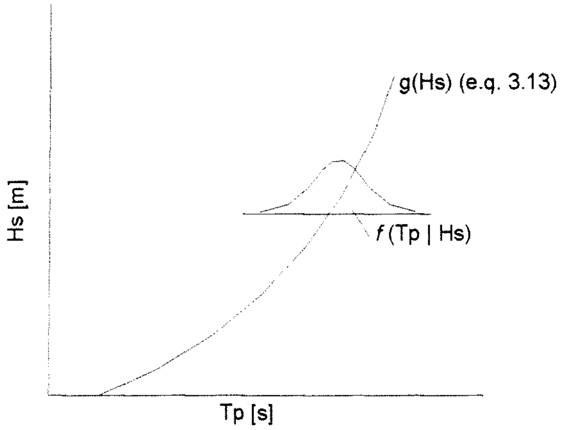


Fig 3.3 Illustration of regression function of T_p on H_s (MODEL 1)

3.4.2 Model 2: the bivariate Log-normal distribution of H_s and T_p with a correction factor for skewness

In the wave study made by Ochi (1978) it appeared that the fit of the bivariate Log-normal distribution, applied to various sets of data, was good for the bulk of the probability mass. However the tails, in particular that of H_s, were not well matched beyond a probability of about 0.99. An attempt to improve the model has been made by Fang and Hogben (1982). They included a measure of skewness in a term modifying the Log-normal form of the marginal distribution of H_s.

The probability density function of this bivariate model, also called the Fang and Hogben distribution, is given by

(3.18)
$$f(H_s,T_p) = \frac{0.5}{H_s T_p \pi \delta_{H_s} \delta_{T_p} \sqrt{1-\rho^2}} *$$
$$* \exp\left\{-\frac{0.5}{1-\rho^2}\left[\frac{(\log T_p - \lambda_{T_p})^2}{\delta_{T_p}^{\ 2}} - \frac{2\rho(\log T_p - \lambda_{T_p})(\log H_s - \lambda_{H_s})}{\delta_{H_s}\delta_{T_p}} + \frac{(\log H_s - \lambda_{H_s})^2}{\delta_{H_s}^{\ 2}}\right]\right\}\left(1 - \frac{\kappa_{H_s}}{6}\left[3(\log H_s - \delta_{H_s}) - (\log H_s - \delta_{H_s})^3\right]\right)$$

It is interesting to study in which way the following parametric models describe the correlation between H_s and T_p. To obtain a realistic approach of the joint probability of occurrence of H_s and T_p, the bivariate function of H_s and T_p should be based on physical laws.

One could suggest describing the relationship between H_s and T_p with a formula similar to the wave steepness equation. In that case, it is given by

(3.11) $$H_s = aT_p{}^2$$

or

(3.12) $$T_p = aH_s{}^{1/2}$$

in which a is some constant value.

3.4.1 Model 1: the bivariate Log-normal distribution of H_s and T_p

Ochi (1978) introduced the use of a bivariate Log-normal distribution for the joint distribution of the significant wave height and the wave period. The probability density function of this model may be written as

(3.13) $$f(H_s,T_p) = \frac{0.5}{T_p H_s \pi \delta_{H_s} \delta_{T_p} \sqrt{1-\rho^2}} * \\ * \exp\left\{ -\frac{0.5}{1-\rho^2} \left[\frac{(\log T_p - \lambda_{T_p})^2}{\delta_{T_p}{}^2} - \frac{2\rho(\log T_p - \lambda_{T_p})(\log H_s - \lambda_{H_s})}{\delta_{H_s}\delta_{T_p}} + \frac{(\log H_s - \lambda_{H_s})^2}{\delta_{H_s}{}^2} \right] \right\}$$

The estimators of the parameters $\lambda_{T_p}, \lambda_{H_s}, \delta_{T_p}$, and δ_{H_s} are similar to those of the marginal Log-normal distribution. (See chapter 4).

Relationship between H_s and T_p

The parameter ρ is a linear correlation coefficient between the two variables, and may be written as

(3.14) $$\rho = \frac{Cov(\log(T_p), \log(H_s))}{\delta_{T_p}\delta_{H_s}}$$

Using the equation

(3.15) $$f(T_p \mid H_s) = \frac{f(H_s,T_p)}{f(H_s)}$$

the conditional distribution of T_p given H_s = c ,(c>=0, c=constant), follows as

(3.16)

$$f(T_p \mid H_s) = \frac{f(H_s,T_p)}{f(H_s)} = \frac{1}{T_p \delta_{T_p} \sqrt{2\pi}(1-\rho)} \exp\left(-\frac{1}{2} \left\{ \frac{\log T_p - \left[\lambda_{T_p} + \rho\left(\frac{\delta_{T_p}}{\delta_{H_s}}\right)(\log H_s - \lambda_{H_s}) \right]}{\delta_{T_p}\sqrt{1-\rho^2}} \right\}^2 \right)$$

where κ_{Hs} is the coefficient of skewness for log H_s. (The remaining parameters are similar to those of the bivariate Log-normal distribution.)

Relationship between H_s and T_p
The characterisation of the relationship between H_s and T_p is similar to those of the bivariate Log-normal distribution (See fig 3.3).

3.4.3 Model 3: T_p conditional on H_s

This model consists of a marginal distribution for the significant wave height (H_s) and a conditional distribution for the wave period (T_p). It is based on the expression,

(3.19)
$$f(H_s,T_p) = f(H_s)f(T_p \mid H_s)$$

In earlier case studies (Mathisen et al. (1990), Haver (1985)), the Weibull and Log-normal distribution were used for the marginal and the conditional distribution of H_s and T_p, respectively. In the present study, the selection is based on a preceding marginal analysis of H_s and T_p. (The conditional distribution of T_p is modelled by the distribution that provides a close fit to the wave period observations of all classes of H_s.)

In the considered bivariate function, the parameters of the conditional distribution are defined as a function of the significant wave height.
At first, the parameters of the conditional distribution are estimated for each class of H_s by using one of the parameter estimation methods described in chapter 4. This gives a discrete version of the conditional function, which may be written as

(3.20)
$$\begin{aligned} f(H_s \mid T_p)_{i=1..n} &= f(T_p \mid H_{s1};\hat{\delta}_1,\hat{\lambda}_1(,\hat{\beta}_1)) + f(T_p \mid H_{s2};\hat{\delta}_2,\hat{\lambda}_2(,\hat{\beta}_2)) + ..f(T_p \mid H_{sn};\hat{\delta}_n,\hat{\lambda}_n(,\hat{\beta}_n)) \\ &= \sum_{n=1}^{i} f(T_p \mid H_{s_i};\hat{\delta}_i,\hat{\lambda}_i(,\hat{\beta}_i))) \end{aligned}$$

where $\hat{\delta}_i, \hat{\lambda}_i$ and $\hat{\beta}_i$ are the scale, location, and shape estimators for each class of H_s. (The shape estimator is only valid for three parameter distributions).

Secondly, for each parameter of the conditional distribution function, an empirical regression function is selected, which defines the relationship between the parameters and the significant wave height. In earlier case studies (Mathisen et al. (1990), Haver (1985)), various empirical regression models were applied. These functions were similar to expressions like

(3.21)
$$g(H_s)_i = aH_s + b$$

$$g(H_s)_i = aH_s^{\,2} + bH_s + c$$

$$g(H_s)_i = a\exp(bH_s) + c\exp(dH_s)$$

in which

$g(H_s)_i$ = estimators for the parameters of the conditional distribution as a function of H_s
(i=1: scale parameter;i=2 location parameter; i=3 shape parameter)
a, b and c = parameters of the regression functions.
y = significant wave height

The above functions are applied in the case studies of chapter 8. The parameters of the regression functions (a,b(,c)) are estimated with the (non) linear least squares method.

Finally, the fitted regression lines are used to compose a continuous conditional distribution function, which then follows as

(3.22) $f(T_p \mid H_s) = f(T_p \mid H_s; g(H_s)_1, g(H_s)_2 (, g(H_s)_3))$

in which the parameters of the function are formulated by $g(H_s)_1, g(H_s)_2$ and $g(H_s)_3$.

Relationship between H_s and T_p
The function describing the location parameter of the conditional distribution as a function of H_s, i.e. the function denoted by $g(H_s)_2$, is similar to eq. (3.17). It is therefore also called *the regression function of T_p on H_s*. The function indicates in which way the two variables are related in the bivariate model.

The three above mentioned regression functions are purely empirical. In order to include the physical proces, one could suggest to use a regression function for the *location* parameter, which is similar to equation 3.12. Thus,

(3.23) $g(H_s)_2 = a(H_s)^{1/2}$

in which

$g(H_s)_2$	= estimator for the location parameter of the conditional distribution of T_p as a function of H_s
a	= parameter of the regression function.
H_s	= significant wave height

In fig 3.4 the different regression lines for the expected value of T_p given H_s are shown.

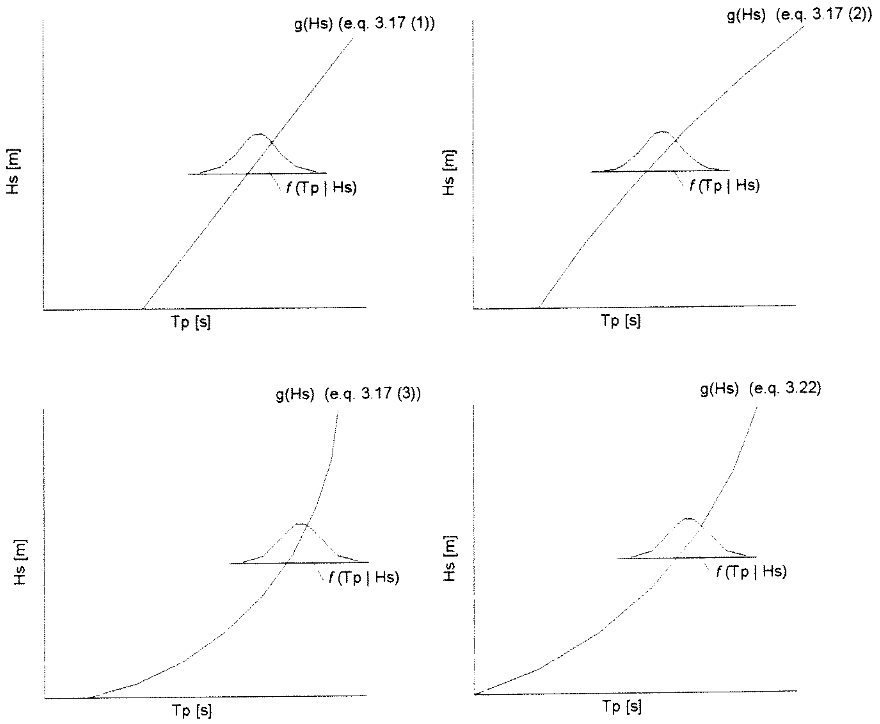


Fig 3.4 Illustration of regression functions of T_p on H_s (MODEL 3)

3.4.4 Model 4: Joint distribution of H_s and T_p derived by transforming $P(H_s,s)$

This method has been proposed by Vrijling (1996). It is based on the assumption that the significant wave height (H_s) and the wave steepness (s) (e.q. 3.10) are independent. Since the wave steepness (s) and the significant wave height (H_s) are assumed to be independent, the joint probability density function of these variables can be expressed as

(3.24) $$f(H_s,s) = f(H_s)f(s)$$

The joint probability function of the significant wave height (H_s) and the wave period (T) can be derived from the above equation by using the following transformation procedure

(3.25) $$f(y_1,y_2) = f(x_1,x_2)\,|\,J\,|$$

with

(3.26) $$|\,J\,| = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} \end{vmatrix}$$

where J is the Jacobian of the transformation. Thus, when

(3.27) $$x_1 = H_s;\quad x_2 = s;\quad y_1 = H_s;\quad y_2 = \left(\frac{2\pi x_1}{g x_2}\right)^{1/2}$$

the partial derivatives of equation (3.26) follow as

(3.28) $$\frac{\partial x_1}{\partial y_1} = 1;\quad \frac{\partial x_2}{\partial y_1} = \frac{2\pi}{g(y_2)^2}$$
$$\frac{\partial x_1}{\partial y_2} = 0;\quad \frac{\partial x_2}{\partial y_2} = \frac{-4\pi y_1}{g(y_2)^3}$$

Therefore,

(3.29) $$|\,J\,| = \begin{vmatrix} 1 & 0 \\ \dfrac{2\pi}{g(y_2)^2} & \dfrac{-4\pi y_1}{g(y_2)^3} \end{vmatrix} = \frac{4\pi y_1}{g(y_2)^3}$$

and thus,

(3.30) $$f(H_s,T_p) = f(H_s,s)\,|\,J\,| = f(H_s,s)\frac{4\pi y_1}{g(y_2)^3}$$

Note that this derivation is only valid for the deep water wave field.

The relationship between H_s and T_p

In the preceding paragraphs the relationship between H_s and T_p has been modelled by an expression called *the regression function of T_p on H_s*. For model 4, this regression function is similar to the parabolic curve $s_{50\%}$, which represents the wave steepness value that is not reached by 50% of the waves.

Example:

If the wave steepness is described by the Gumbel distribution, the wave steepness that is not reached by 50% of the waves follows from

$$P(s_p < s) = F_{s_p}(s) = \exp(-\exp(-(s-\lambda)/\delta)) = 0.5$$

or

$$s_{p(50\%)} = \lambda - \delta \log(-\log(0.50))$$

The above mentioned parabolic curve $s_{50\%}$ is then described as

$$H_s = aT_p{}^2$$

with

$$a = \frac{s_{p(50\%)*9.81}}{2*\pi}$$

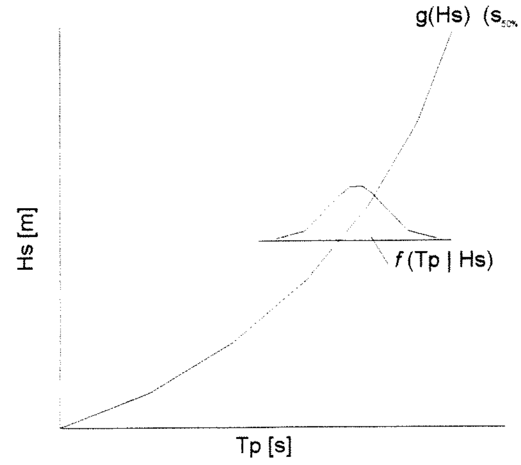


Fig 3.5 Illustration of regression function of T_p on H_s (MODEL 4)

3.4.5 Model 5: a bivariate model of H_s and T_p developed by Morton and Bowers

In this section the joint probability function (pdf) of H_s and T_p will be determined according to a method that is proposed by Morton and Bowers (1997). In a recent published paper they illustrated a bivariate analysis which included a point proces model of extreme events together with a function describing the joint dependence of the two variables. Their study was concerned with the joint distribution of extreme wind speeds and significant wave heights, without any consideration of their directions. (The corresponding sets of data consisted of environmental data for 1990-1994 from the Shell UK Exploration and Production North Cormorant Platform in the northern North Sea).

According to Morton and Bowers (1997) the construction of a pdf describing the variables' joint extreme behaviour consists of four key elements:

1. determining which observations are "extreme" (data selection);
2. modelling the extreme observations with marginal distributions;
3. modelling the dependence between the marginal variables;
4. combining the marginal and the dependence models to provide a final model of the multivariate extremes.

The methodology is summarized in figure 3.6 with the four key elements A,B,C, and D marked.

The first two steps, the key elements A and B, have already been described previously. The third step is the modelling of the dependence between the marginal variables. This proces, which is shown in box C of fig. 3.6, involves a transformation to unit Frechet space. It then requires a further transformation which combines the seperate variables into "pseudo polar" coordinates before fitting an appropriate function to describe the dependencies between the multiple variables.
When the marginal distributions and the depence model are computed, the joint pdf can be constructed, (box D). The complete joint pdf (box D) combines information from the marginal distributions and the dependence structure

A detailed description of the method is given in appendix [3]. The joint distribution function is given by:

(3.31) $$\frac{\partial^2}{\partial x_1 \partial x_2} P(X_1 > x_1, X_2 > x_2)$$

where

(3.32) $$P(X_1 > x_1, X_2 > x_2) = \exp(-V(z))$$

Therefore,

(3.33) $$\frac{\partial^2}{\partial x_1 \partial x_2} P(X_1 > x_1, X_2 > x_2) = \left(V_1(z)V_2(z) - V_{12}(z)\right)\frac{\partial z_1}{\partial x_1}\frac{\partial z_2}{\partial x_2}\exp(-V(z))$$

where

(3.34) $$V(z) = \left(\left(\frac{1}{z_1}\right)^{\varphi} + \left(\frac{1}{z_2}\right)^{\varphi}\right)^{\frac{1}{\varphi}}$$

is the logistic model and

(3.35) $$V_1(z) = \frac{\partial V}{\partial z_1} = (-z_1^{-\varphi-1})\left(\left(\frac{1}{z_1}\right)^{\varphi} + \left(\frac{1}{z_2}\right)^{\varphi}\right)^{\frac{1}{\varphi}-1}$$

(3.36) $$V_2(z) = \frac{\partial V}{\partial z_2} = (-z_2^{-\varphi-1})\left(\left(\frac{1}{z_1}\right)^{\varphi} + \left(\frac{1}{z_2}\right)^{\varphi}\right)^{\frac{1}{\varphi}-1}$$

(3.37) $$V_{12}(z) = \frac{\partial^2 V}{\partial z_1 \partial z_2} = (-z_1^{-\varphi-1})(-z_2^{-\varphi-1})(1-\varphi)\left(\left(\frac{1}{z_1}\right)^{\varphi} + \left(\frac{1}{z_2}\right)^{\varphi}\right)^{\frac{1}{\varphi}-2}$$

In the above equations, x_1 and x_2 are the significant wave height (H_s) and the wave period (T_p), respectively. z_1 and z_2 represent H_s and T_p transformated into the Frechet space.

Relationship between H_s and T_p

As mentioned above the dependency between the variables is defined by a logistic depence function (e.q. 3.34). This function contains the parameter φ which is a correlation coeffient between the two variables.

It must be noticed that although the model might provide good results, the theoretical background of the model is quite complicated. For civil engineers this model will probably be considered as a black box.

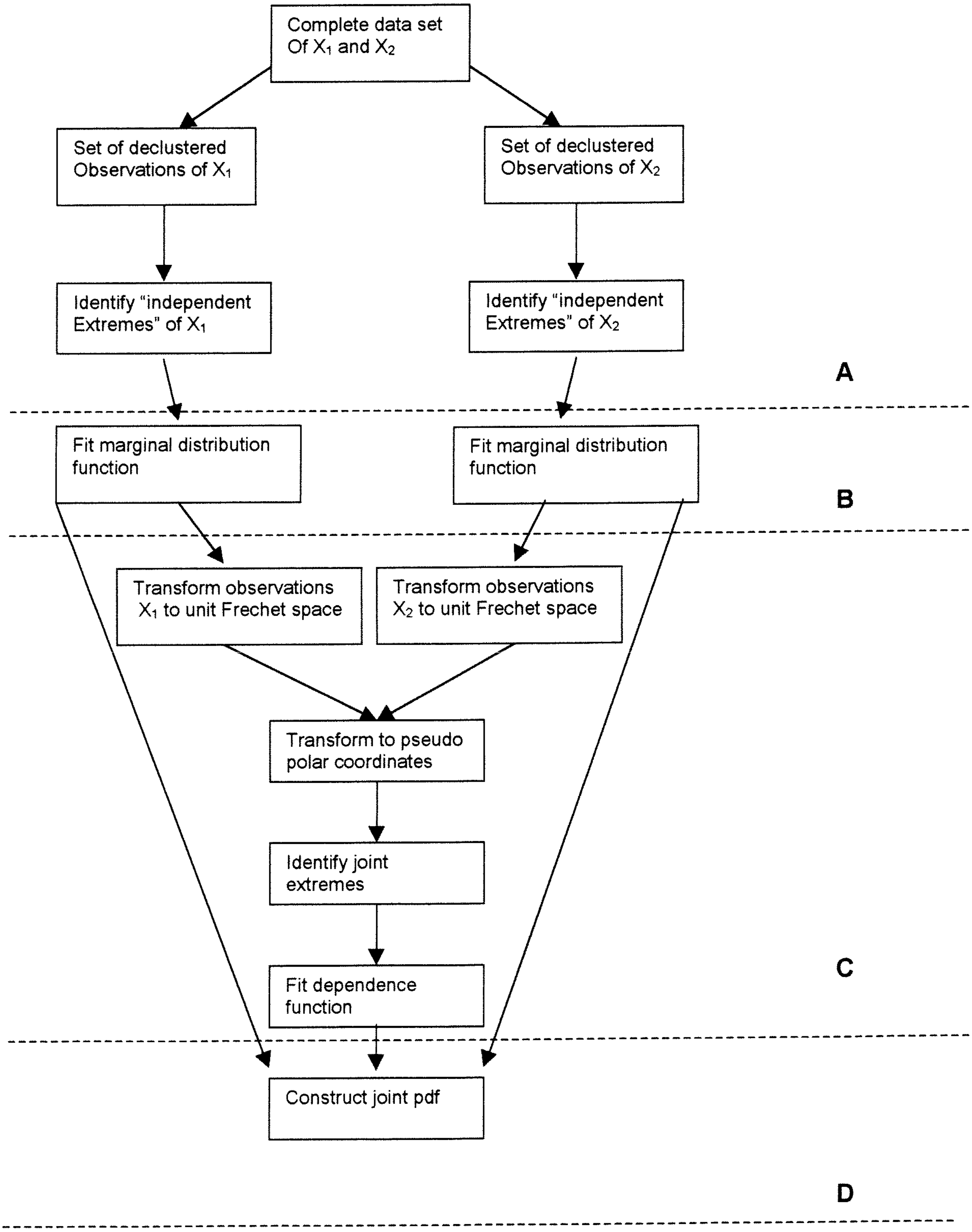

Complete data set
Of X_1 and X_2
Set of declustered Observations of X_1
Set of declustered Observations of X_2
Identify "independent Extremes" of X_1
Identify "independent Extremes" of X_2
A
Fit marginal distribution function
Fit marginal distribution function
B
Transform observations X_1 to unit Frechet space
Transform observations X_2 to unit Frechet space
Transform to pseudo polar coordinates
Identify joint extremes
Fit dependence function
C
Construct joint pdf
D

Fig 3.6 Schematization of the bivariate model developed by Morton and Bowers (1997)
(Bivariate distribution with given marginals)

3.5 References

[1] Almering ,J.H.J., Bavinck, H., Goldbach, R.W., *Mathematics (in Dutch),* Delft: Delftse Uitgevers Maatschappij, 1993

[2] Anderson, C.W., Nadarajah,S., *"Environmental factors affecting reservoir safety",* in *Statistics for the Environment.* Londen: Wiley, 1993, pp. 163-182 (*)

[3] Balakcrishnan, N., Clifford cohen, A., *Order statistics and interference, estimation methods.* San Diego: Academic Press Inc, 1991

[4] Battjes, J. A.,*"Long-term wave height distributions at seven stations around the British Isles.".* N.I.O. Report No. A 44., 1970

[5] Buckley, W.H., *"Extreme and climatic wave spectra for use in the structural design of ships.".* Naval Engineers Journal, 1988, pp. 36-58 (*)

[6] Burrows,R., Salih, B. A., *"Statistical Modelling of long-term wave climats".* Proc. 20th Int. conf. coastal. eng., 1986, pp. 42-56

[7] Castillo, E., *Extreme value theory in engineering.* San Diego: Academic Press INC, 1987

[8] Cavanie, A. *"Joint occurence of extreme wave heights and wind gusts during severe storms on the Frigg field".* Workshop on the application of joint probability of metocean phenomena in the oil industry's structural design work. London: E&P Forum, 1985 (*)

[9] Coles, S. G., Tawn, J.A, *"Statitical modelling for multivariate extremes: An application to strutural design".* Applied statistics, 1994, 43(I), pp. 1-48 (*)

[10] CUR research comittee E10 "Risk analysis", *Risk analysis in civil engineering, Vol. 1: probabilistic design in theory (Dutch).* Gouda: CUR (Centre for civil engineering research and codes), maart 1997

[11] Fang, Z.S., and Hogben, N., *"Analysis and prediction of long-term probability distributions of wave height and periods".* London: Technical report, National Maritime Institute, 1982 (*)

[12] Gumbel, E.J., *Statistics of Extremes.* New York and London: Colombia University Press,1958 (*)

[13] Haver, S.H., *"Wave climate off northern Norway".* Applied Ocean Research, vol.7, no. 2, 1985 pp. 85 - 92

[14] Mathisen,J., and Bitner-Gregersen, E., *"Joint distributions for significant wave height and wave zero-up-crossing period".* Applied Ocean Research, 1990, Vol. 12, no. 2, pp 93-103

[15] Morton, I.D., Bowers, J., *Applied ocean research 18; Extreme value analysis in a multivariate offshore environment.* Great Britain: Elsevier Sience B.V., 1997, pp. 303-317

[16] Ochi, M. K., *"Wave statistics for the design of ships and structures".* New York: Proc. SNAME Conf., 1978 (*)

[17] Tang, C., and Palao, I.M., *"Wave height and period distributions from long-term wave measurement".*: Proc. 25th Int. conf. coastal. eng., 1996, pp. 368-378

[18] Tucker, M.J. *Waves in ocean engineering. Measurement, analysis, interpretation.* London: Ellis Horwoord, 1991

[19] Vrijling, J.K.,*Probabilistic design in hydraulics. (Dutch).* Delft: Delft University of Technology, Faculty of Civil Engineering, 1996

(*) These articles have been mentioned in one of the other articles. The marked articles have not been studied by the present author.

4 Parameter estimation methods

4.1 Introduction

In this chapter, several parameter estimation methods are described for the marginal distributions. The estimation of parameters of the bivariate model which do not correspond to the parameters of the marginal functions, for instance the dependence parameter of the bivariate Log-normal distribution, have been discussed in the preceding chapter.

The following estimation methods are used in the present study:

- The method of moments
- The linear least squares method
- The non linear least squares method
- The maximum likelihood method

In section 4.2 to 4.4 the above estimation methods are described. In the last section of this chapter, the bias and efficiency of the estimators are discussed.

4.2 The method of moments

The method of moments is the most widely used fitting technique, because of its simplicity. The method works by equating the first m statistical moments of the target distribution to the moments derived from the observations. The number of statistical moments that needs to be used is equal to the number of parameters of the target distribution.

The estimated values of the parameters are expressed in terms of $\bar{x}$, $\bar{x}^2$ and $\bar{x}^3$ as indicated in table 4.1. Here $\bar{x}$, $\bar{x}^2$ and $\bar{x}^3$ are obtained directly from the data and are defined as follows:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.1}$$

$$\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{4.2}$$

$$\bar{x}^3 = \frac{1}{n}\sum_{i=1}^{n} x_i^3 \tag{4.3}$$

The Weibull and the Frechet distribution involve three parameters and of these, the third parameter, i.e. the shape parameter β, is estimated by equating the skewness of the sample to that of the model. The remaining parameters can then be estimated from the first and second moments. The skewness of the sample is

$$skewness = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\bar{x}^3 - 3\overline{x x}^2 + 2(\bar{x})^2}{(\bar{x}^2 - (\bar{x})^2)^{3/2}} \tag{4.4}$$

in which μ_2 and μ_3 are the second and third central moments of the distribution.

The reliability of the estimated parameter values diminishes with the 'order' of the statistic, especially for samples of limited size. In consequence, the computed values of skewness, being a third order statistic, shows much greater scatter than that shown by the sample mean and standard deviation which are determined from the first and second order moments, respectively (Burrows, Salih; 1986).

Table 4.1 Parameters of distributions as estimated by the method of moments

Distribution	Estimated parameters		
	$\hat{\beta}$	$\hat{\delta}$	$\hat{\lambda}$
Exponential	-	$(\bar{x}^2 - (\bar{x})^2)^{1/2}$	$\bar{x} - \hat{\delta}$
Gumbel	-	$\frac{\sqrt{6}}{\pi}(\bar{x}^2 - (\bar{x})^2)^{1/2}$	$\bar{x} - \gamma\hat{\delta}$
Weibull	*skewness*	$\left[\frac{\bar{x}^2 - (\bar{x})^2}{\Gamma(1+2/\hat{\beta}) - \Gamma^2(1+1/\hat{\beta})}\right]^{1/2}$	$\bar{x} - \hat{\delta}\Gamma(1+1/\hat{\beta})$
Frechet	*skewness*	$\left[\frac{\bar{x}^2 - (\bar{x})^2}{\Gamma(1-2/\hat{\beta}) - \Gamma^2(1-1/\hat{\beta})}\right]^{1/2}$	$\bar{x} - \hat{\delta}\Gamma(1-1/\hat{\beta})$
Lognormal	-	$\sqrt{\log\left(\frac{\sum x_i^2}{N\bar{x}^2}\right)}$	$\log(\bar{x}) - \frac{1}{2}\log\left(\frac{\sum x_i^2}{N\bar{x}^2}\right)$

4.3 The linear least squares method

The principle of the least-squares method is based on minimizing the sum of squares of the differences between the cumulative distribution function ($F(x;\lambda,\delta(,\beta))$) and the (empirical) data values (y_i). The method is based on the minimization, with respect to the parameters of the chosen distribution, of the distance

(4.5) $$\sum_{i=1}^{n}(y_i - F(x_i;\lambda,\delta(,\beta)))^2$$

or

(4.6) $$\chi^2 = \min\sum_{i=1}^{n}(y_i - F(x_i;\hat{\lambda},\hat{\delta}(,\hat{\beta})))^2$$

in which

n	=	total number of observations in the data set.
i	=	index number
x_i	=	observation i (abscissa) . (The observations are rearranged in ascending order with the smallest value being assigned the order number i=1))
y_i	=	empirical non-exceedance probability of observation i (emp. ordinate)
$F(x_i,\lambda,\delta(,\beta))$	=	theoretical non-exceedance probability of observation i (theor. ordinate)
λ,δ,β	=	location, scale, and shape parameter of the cumulative distribution function
$\hat{\lambda},\hat{\delta},\hat{\beta}$	=	estimator of location, scales, and shape parameter of the cumulative distribution function (shape par./est. only valid for the Weibull and the Frechet distribution)

Extreme value distributions are non-linear functions. The least squares method can be applied after transforming the selected distribution in such way that when the cumulative distribution function is plotted against the data points a straight line is obtained. This approach is found in various extreme wave analyses (Mathiesen et al. 1993; Maes et al. 1994). The estimation method is called the *linear* least squares method and is based on minimization of the distance

(4.7) $$\chi^2 = \min \sum_{i=1}^{n} (y_i - (A + Bx_i))^2$$

in which

n = total number of observations in data set.
i = index number
x_i = observations i (abscissa) .
(The observations are rearranged in ascending order with the smallest being assigned the order number i=1))
y_i = transformed empirical non-exceedance probability of observation i (emp. ordinate), see table 4.2.
A = slope of linear regression line
B = intercept of linear regression line

The slope (A) and intercept (B) of the best-fitted linear regression line are determined by finding the roots of the partial derivatives $\frac{\partial \chi^2}{\partial A}$ and $\frac{\partial \chi^2}{\partial B}$ of equation 4.7. The derivation of the parameters A and B is given in appendix [4]. The estimators of the distribution parameters can be derived from the slope and intercept of the linear regression line. (See table 4.2).

Table 4.2 Scale relationships for probability distributions

Distribution	Abscissa scale (x)	Ordinate scale (y)	Slope(A)	Intercept(B)
Exponential	x	$-\log(1-F(x))$	$1/\hat{\delta}$	$-\hat{\lambda}/\hat{\delta}$
Gumbel	x	$-\log(-\log(F(x)))$	$1/\hat{\delta}$	$-\hat{\lambda}/\hat{\delta}$
Weibull	$\log(x-\hat{\lambda})$	$\log(-\log(1-F(x)))$	$\hat{\beta}$	$-\hat{\beta}\log\hat{\delta}$
	x	$(-\log(1-F(x)))^{1/\hat{\beta}}$	$1/\hat{\delta}$	$-\hat{\lambda}/\hat{\delta}$
Frechet	$\log(x-\hat{\lambda})$	$-\log(-\log(F(x))$	$\hat{\beta}$	$-\hat{\beta}\log\hat{\delta}$
	x	$(-\log(1-F(x)))^{-1/\hat{\beta}}$	$1/\hat{\delta}$	$-\hat{\lambda}/\hat{\delta}$

With the linear regression method only two parameters can be estimated. Therefore, the third parameter of the three-parameter functions will have to be estimated with a trial and error procedure.

In table 4.2 two methods for linearization of the Weibull and the Frechet distribution function are listed. The first method requires an estimation of the location parameter (δ), the second requires an estimation of the shape parameter (β). A first approximation of the location parameter(λ) is the lowest value of the sample –0.1. Least squares estimators for the Log-normal distribution have not been derived in the present study.

When least squares methods are used, one is faced with the problem of assigning proper probability levels y_i to the observed data values x_i, i.e. choosing the plot position such that the bias in the parameter estimate is minimized. This is a vital problem, as an improper choice of the plotting position formula may lead to highly biased parameter estimates.

For each cumulative distribution function (cdf), the optimal choice of the plotting position formula is different. In the present study, the plotting formulas recommended by Goda (1988) will be used. The recommended plotting formula for each distribution is shown in table 4.3.

Table 4.3. Plotting position formulas
(n=total number of data points; n_t =total number of storm events ; k=scale parameter cdf)

Distribution function	Name plotting position formula	Plotting position formula (emp. ordinate)
Exponential	Gumbel / Weibull	$\frac{i}{n+1}$
Log-normal	Blom	$\frac{i-0.375}{n+0.25}$
Gumbel	Gringorten	$\frac{i-0.44}{n+0.12}$
Weibull	Petruaskas and Aagaard	$\frac{i-0.49-0.50/k}{n+0.21+0.32/k}$
Frechet	Goda and Kobune	$\frac{i-0.11-0.52/k}{n_t+0.23-0.22/k}$

4.4 The non linear least squares method

Least squares estimating equations are obtained by finding the root of the partial derivates of equation (4.6) with respect to the parameters of the distribution function:

(4.8) $$\frac{\partial\chi^2}{\partial\delta}=0\,;\frac{\partial\chi^2}{\partial\lambda}=0\,;\left(\frac{\partial\chi^2}{\partial\beta}=0\right)$$

The above set of equations can be solved analytically. Here the non-linear least squares estimators are determined by the direct minimization of equation (4.6). This iterative procedure requires an initial estimate of the parameters. For this estimators calculated with another fitting technique, for example the linear regression method, can be used.

4.5 The maximum likelihood method

This method is based on maximizing the likelihood function of data with respect to the parameters. The central idea consists of assuming that the sample comes from a population with a distribution belonging to a parametric family and choosing the parameter values that maximize the probability of occurence of the sample data.

The parameters of the target probability density function $f(x_i;\delta,\lambda(,\beta))$ should be chosen such that the likelihood function

(4.9) $$L(x_1,x_2,...,x_N)=\prod_{i=1}^{n} f(x_i;\delta,\lambda(,\beta))$$

is maximized by the choice of the model parameters. *L* can be considered as the probability of getting the particular set of data values x_i. The maximum of *L* is achieved by the model parameters for which the partial derivatives are equal to zero:

(4.10) $$\frac{\partial L}{\partial \delta}=0;\frac{\partial L}{\partial \lambda}=0;\left(\frac{\partial L}{\partial \beta}=0\right)$$

In practice, it is more convenient to work with the logarithm of the likelihood *L*, which attains its maximum for the same values of the model parameters as *L* itself.

The loglikelihood function can be written as

(4.11) $$\log L(x_1,x_2,...,x_N)=\sum_{i=1}^{n}\log f(x_i;\delta,\lambda,\beta))$$

The maximum of log *L* is achieved with a set of equations similar to eq. 4.10, with the difference that *L* is replaced by Log *L*. The derivation of the maximum likelihood estimators for the presently used distributions is presented in appendix [5].

The maximum likelihood estimators of the distribution parameters are listed in table 4.4. Note that the estimators for the parameters of the Weibull and the Frechet distribution have to be determined iteratively.
Instead of using the formulas of table 4.4, one can also obtain maximum likelihood estimators of the distribution parameters by direct maximization of the loglikelihood function.

Table 4.4 Parameters of distributions as estimated by the maximum likelihood method

Distribution	Estimated parameters		
	$\hat{\beta}$	$\hat{\delta}$	$\hat{\lambda}$
Exponential	-	$\bar{x} - x_1$ (*)	x_1 (*)
Lognormal	-	$\left[n^{-1} \sum_{j=1}^{n} (Z_j - \bar{Z})^2 \right]^{1/2}$ (**)	$\bar{Z}$ (**)
Gumbel	-	$n^{-1} \sum_{i=1}^{n} x_i - \frac{\sum_{i=1}^{n} x_i \exp\left[-\frac{x_i}{\hat{\delta}} \right]}{\sum_{i=1}^{n} \exp\left[-\frac{x_i}{\hat{\delta}} \right]}$	$-\hat{\delta} \log(n^{-1} \sum_{i=1}^{n} \exp\left[-\frac{x_i}{\hat{\delta}} \right]$
Weibull	$\sum_{i=1}^{n} (x_i - \lambda)^{\hat{\beta}} = \delta^{\beta}$ $\left[\frac{\sum_{i=1}^{n} (x_i - \hat{\lambda})^{\hat{\beta}} \log(x_i - \hat{\lambda})}{\sum_{i=1}^{n} (x_i - \hat{\lambda})^{\hat{\beta}}} - \frac{1}{\hat{\delta}} \right] - \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\lambda}) = 0.$		
Frechet	$\frac{\partial \log L}{\partial \delta} = \frac{n}{\beta} + (\beta + 1) \frac{n}{\delta} - \sum_{i=1}^{n} \left(\frac{\delta}{(x_i - \lambda)} \right)^{\beta} \frac{\beta}{\delta} = 0$ $\frac{\partial \log L}{\partial \lambda} = (\beta + 1) \sum_{i=1}^{n} \frac{1}{(x_i - \lambda)} - \sum_{i=1}^{n} \left(\frac{\delta}{(x_i - \lambda)} \right)^{\beta} \frac{\beta}{(x_i - \lambda)}$ $\frac{\partial \log L}{\partial \beta} =$ $= -\frac{n}{\delta} + n \log(\delta) + \sum_{i=1}^{n} - \log(x_i - \lambda) +$ $+ \sum_{i=1}^{n} - \log(x_i - 1) - \sum_{i=1}^{n} \left(\frac{\delta}{(x_i - \lambda)} \right)^{\beta} \log\left(\frac{\delta}{(x_i - \lambda)} \right) = 0$		

(*) $x_1 = \min(x_i)$

(**) $Z_i = \log(x_i)$

4.6 Bias and efficiency of estimators

The previous sections provided several methods for the estimation of the parameters of a (marginal) distribution function. This section has been included in order to mention the random character of the estimators.

A parameter of a distribution function is estimated from a set of observations. The observations are considered to be independent extractions of a random variable. (In this study the random variables are the significant wave height and the wave period). Since an estimator of a distribution parameter is a function of the sample data, the estimator is a random variable.

Each estimator can thus be defined with a probability density function (pdf). When the expected value of the pdf of the estimator $\hat{\theta}$ is equal to the true value of the unknown parameter θ, then the estimator of the parameter is unbiased. Thus, if

(4.12) $$E(\hat{\theta}) = \theta$$

The bias of an estimator is the expected value of the difference between the parameter value and the estimator value:

(4.13) $$Bias = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

The value of the standard deviation of the estimator with respect to the parameter value is defined as the efficiency of the estimator:

(4.14) $$Efficiency = E[(\hat{\theta} - \theta)^2]$$

Since for the determination of the bias and efficiency of an estimator the true distribution must be known, for wave statistics these properties of an estimator can only be studied by using numerical simulated data.
Beem (1992) and Slijkhuis (1996) have studied the efficiency and bias of the applied estimation methods. To this end, 14 sets of 10 data points were generated from an Exponential distribution function with location parameter A=1.96 and scale parameter B=0.33. For each data set, the parameters of the distribution function were estimated. With the outcome of the estimation proces, the efficiency and bias of the three estimation methods were computed. The results of their studies are listed in table 4.5.

Table 4.5 Unbiasedness and efficiency (Exponential distribution function (A=1.96;B=0.33)) (Taken from Slijkhuis(1996));

	Method of moments	Least squares method (linear)	Least squares method (non linear)	Maximum likelihood method
Unbiasedness A	0.002	0.037	-0.016	-0.087
Unbiasedness B	0.02	-0.015	0.048	0.165
Efficiency A	0.089	0.037	0.056	0.125
Efficiency B	0.126	0.087	0.125	0.187

The table shows that in this case the method of moments provides the least biased estimators, where as the maximum likelihood estimators appear to be the most efficient estimators.

4.7 References

[1] Beem, R., *A comparative study of different parameter estimation methods for statistical distributions functions (in Dutch).* Rijkswaterstaat, 1992 (*)

[2] Burrows,R., Salih, B. A., *"Statistical Modelling of long-term wave climats".* Proc. 20th Int. conf. coastal. eng., 1986, pp. 42-56

[3] Castillo, E., *Extreme value theory in engineering.* San Diego: Academic Press INC, 1987

[4] Cohen, A. C., *Maximum likelihood estimation in the Weibull distribution based on complete and censored samples.* Technometrics 1, 1965, pp 217-237 (*)

[5] CUR research comittee E10 "Risk analysis", *Risk analysis in civil engineering, Vol. 1: probabilistic design in theory (Dutch).* Gouda: CUR (Centre for civil engineering research and codes), maart 1997

[6] Goda, Y.,*"On the methodology of selecting design wave height".* Malaga: Proc. 21th Int. conf. coastal eng., 1988, pp. 135-136

[7] Goda,Y.,Kobune, K., *"Distribution function fitting for storm wave data".* Delft: Proc. 22th Int. conf. coastal. eng., 1990, pp. 18-31

[8] Groeneboom, P., Kraaikamp, C., Kopuhaä, H.P, and van der Weide, J.A.M., *Probability and statistics (in Dutch).* Delft: Delft University of Technology, Faculty of Civil Engineering, 1995

[9] Herbich, J. B., *Handbook of coastal and ocean engineering, Volume: 1 Wave Phenomena and Coastal Structures.* Houston: Gulf publishing Company, 1990

[10] Hines, W.H., and Montgomery, D.C., *Probability and statitics in Engineering and Management Science.* New York: John Wiley and Sons, 1990

[11] Maes, M. A., and Gu, G.Z.,*"Techniques used to determine extreme wave heights from the NESS Data set"* . Gaithersburg: Proc. of the conf. on extreme value theory and applications, volume 2, 1994, pp. 435-444

[12] Mathiesen, M. et al., *"Intercomparison of extreme wave analysis: a comparitive study".* New Orleans: Proc. WAVES '93 conf., july 1993, pp 963-977

[13] Miller, I.R., Freund, J.E., Johnson, R., *Probability and statistics for engineers.* Englewood Cliffs: Prentice-Hall, 1990

[14] Ochi, M.K., *Applied probability & stochastic processes, in engineering and physical sciences.* New York: John Wiley & Sons, 1990

[15] Petruaskas, C., and Aagaard, P.M., "*Extrapolation of historical storm data for estimating design wave heights".* Prepr. 2nd Annual Offshore Techn. Conf., 1970 (*)

[16] Slijkhuis, K.A.H., *The influence of statistical uncertainty on the determination of crest heights of dikes (in Dutch).* Delft: Delft University of Technology, Faculty of Civil Engineering, 1996

[17] Van Gelder, P.H.A.J.M., Vrijling,J.K., *"A comparative study of different parameter estimation methods for statistical distributions functions in civil engineering applications".* Kyoto : ICOSSAR'97, 7th International Conference on Structural Safety and Reliability Kyoto, 1997

(*) These articles have been mentioned in one of the other articles. The articals have not been studied by the present author.

5 Methods for the judgement of the goodness of fit

5.1 Introduction

The rejection or acceptance of a fitted distribution is usually based on a qualitative (visual) inspection and quantitative judgements of the fits. In literature, a large number of tests are used to quantify the goodness of fit, to reject or accept distributions and to choose between various fitted distributions.

A selection of these tests is described in this chapter. For the marginal case, the following tests are presented:
- Visual judgement
- Kolmogorov-Smirnov test
- Chi-square test
- Rejection criteria based on outlier (Goda (1990))
- Linear correlation coefficient
- Rejection criteria based on (linear) correlation coefficient (Goda (1990))

The two rejection criteria are fully empirical. It must be noted that in literature, a lot of such criteria are present. They are attempts to establish steadfast guidelines for the selection of one particular probability function. However, up to present, none of these criteria appears to be applicable for all wave data.

For the bivariate probability distributions mostly goodness of fit criteria are used, which are based on the scatter diagram of the two random variables. The tests presented are
- Visual judgement
  (Comparison of "theoretical" scatter diagram (based on statistical model) with "empirical" scatter diagram (based on wave data))
- 2-dimensional chi-square test
- A test developed by Mathisen and Bitner–Gregersen (1990)

5.2 Goodness of fit of marginal distributions

5.2.1 Visual judgement

In practice, the fit of a distribution to a set of data is often visually judged. With this method, the shape of the distribution and the quality of the fit are examined. The main drawback of visual judgement is its subjectivity.

Visual checks are easy to perform but they need an experienced eye. Mainly when large quantities of fits are to be judged, visual checks are less suitable.

5.2.2 Chi-square test

With the chi-square test a comparison is being made between the hypothesized distribution function and the histogram composed from the observations. It is based on the statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_i^{(o)} - n_i^{(e)}\right)^2}{n_i^{(e)}} \qquad (5.1)$$

in which

$n_i^{(o)}$ Number of observed values in cel i of the histogram

$n_i^{(e)}$ Number of expected values in cel i of the histogram. This number is computed with the hypothesized distribution function, using the expression

$$n_i^{(e)} = N \cdot f(x) \cdot dx$$

in which N is the total number of observations

It can be shown that χ^2 approximately follows the chi-square distribution with k-p-1 degrees of freedom, where p denotes the numbers of parameters of the hypothesized distribution. This approximation improves as the number of observations increases. Hence, χ^2 given in eq. (5.1) is used as test statistic for making the decision to accept or reject the hypothesized distribution. That is, a level of significance α is assigned for the test and the critical value $\chi^2_{k-r-1}(\alpha)$ is determined.

This implies that the larger the χ^2 value, the more the hypothesized distribution should be descredited. Therefore, if the χ^2 value evaluated by eq. (5.1) is greater than the critical value, the hypothesized distribution may be rejected; otherwise, the function may be accepted. The most usual value to assign α is 0.05, that is, the test is conducted with 5% risk.

5.2.3 Kolmogorov-Smirnov test

This test concentrates on the deviations between the hypothesized cumulative distribution function (cdf) $F(x)$ and the observed cumulative histogram

(5.2) $$G(x(i)) = \frac{i}{N}$$

in which x(i) is the largest observed value in the random sample of size N. The test statistics used can be expressed as

(5.3) $$D = \max_{i=1}^{n}\left[|G(x(i)) - F(x(i))|\right] = \max_{i=1}^{n}\left[|\frac{i}{N} - F(x(i))|\right]$$

In words, D is the largest of the absolute values of the N differences between the hypothesized cdf and the observed cumulative histogram evaluated at the observed values.

The goodness of fit criteria is that

(5.4) $$D < \frac{\alpha}{\sqrt{N}} \qquad \text{for N>5}$$

The test can be performed at various prescribed significance levels (α):

Significance level	
$\alpha = 1.23$	10%
$\alpha = 1.36$	5%
$\alpha = 1.63$	1%

5.2.4 Rejection criterion based on outliers (DOL criterion)

Let x_1 be the value of the largest value among data. Then its magnitude is measured with the following dimensionless deviation ξ:

(5.5) $$\xi = (x_1 - \bar{x}) / \sigma$$

in which $\bar{x}$ and σ are the mean and standard deviation of sample data.

Goda and Kobune (1990) recommended that the ξ value of a given sample be compared with the values of $\xi_{5\%}$ and $\xi_{95\%}$, which they have computed for various theoretical distributions. If the ξ value of a sample occupies a location at either the upper or lower tail of the cumulative ξ-curve of the distribution function being tested for fitting, the chance that the sample belongs to that population is slim and that distribution could be rejected. The 5% and 95% level were tentatively chosen as the treshold values for rejection.

In the paper of Goda and Kobune, an empirical formula has been derived from the simulation results. This formula is given by

(5.6) $$\xi_{5\%}, \xi_{95\%} = a + b \ln N + c(\ln N)^2$$

The coefficients a,b, and c are expressed as the function of the censoring parameter v, which is defined as

(5.7) $$\text{v} = \frac{N}{N_T}$$

where N is the number of analyzed data, and N_T is the total number of data during the period of analysis. The empirical coefficients a, b and c of equation (5.6) are presented in table 5.1 and 5.2.

Table 5.1 Empirical coefficients for the lower DOL criterion $\xi_{5\%}$

Distribution	Coefficient a	Coefficient b	Coef. c
FT-I	0.257+0.133v^2	0.452-0.118v^2	0.032
Ft-II (β = 2.5)	1.481-0.126$v^{1/4}$	-0.331-0.031v^2	0.192
Ft-II (β = 3.3)	1.025	-0.077-0.050v^2	0.143
Ft-II (β = 5.0)	0.700+0.060v^2	0.139-0.076v^2	0.100
Ft-II (β = 10.0)	0.424+0.088v^2	0.329-0.094v^2	0.061
Weibull (β = 0.75)	0.534-0.162v	0.277+0.095v	0.065
Weibull (β = 1.0)	0.308	0.423	0.037
Weibull (β =1.4)	0.192+0.126$v^{3/2}$	0.501-0.081$v^{3/2}$	0.018
Weibull (β = 2.0)	0.050+0.182$v^{3/2}$	0.592-0.139$v^{3/2}$	0

Table 5.2 Empirical coefficients for the upper DOL criterion $\xi_{95\%}$

Distribution	Coefficient a	Coefficient b	Coef. c
FT-I	-0.579+0.468v	1.496-0.227v^2	-0.038
Ft-II (β = 2.5)	4.653-1.076$v^{1/2}$	-2.047+0.307$v^{1/2}$	0.635
Ft-II (β = 3.3)	3.217-1.216$v^{1/4}$	-0.903+0.294$v^{1/4}$	0.427
Ft-II (β = 5.0)	0.599-0.038v^2	0.518-0.045v^2	0.210
Ft-II (β = 10.0)	-0.371+0.171v^2	1.283-0.133v^2	0.045
Weibull (β = 0.75)	-0.256-0.632v^2	1.269+0.254v^2	0.037
Weibull (β = 1.0)	-0.682	1.600	-0.045
Weibull (β =1.4)	-0.548+0.452$v^{1/2}$	1.521-0.184v	-0.065
Weibull (β = 2.0)	-0.322+0.641$v^{1/2}$	1.414-0.326v	-0.069

5.2.5 Linear correlation coefficient

The linear correlation coefficient assesses the degree of linearity of the reduced variates and is defined as

(5.8) $$r = \left[\frac{\sum_{i=1}^{n} \left((x_i - \mu_x)(y_i - \mu_y) \right)^2}{\sum_{i=1}^{n} (x_i - \mu_x)^2 (y_i - \mu_y)^2} \right]^{\frac{1}{2}}, \quad \left(|r| \leq 1 \right)$$

where x and y represent the reduced variates of the abscissa and the ordinate, respectively. This statistic can be used as goodness of fit criterion because it also explains the amount of variance of the dependent variable accounted for by the independent variable. The higher the correlation coefficient, the better the fit.

The drawback of the linear correlation coefficient is that it provides a measure of fit of *transformed* distribution functions. For the linearization of various distributions, different transformation formula is needed. (See table 4.2). It is therefore not fair to compare the degree of fitting of various distributions to a sample by means of the absolute value of the correlation coefficient.

5.2.6 Rejection criterion based on correlation coefficient

Based on the correlation coefficient, Goda and Kobune (1990) proposed the REC-criterion (REsidue of Correlation coefficient) to test the goodness-of-fit of a particular distribution.

The REC-criterion tests the goodness-of-fit of a particular distribution. The REC-criterion considers the residue of r , i.e. $\Delta r = 1\text{-}r$. The cumulative distribution of Δr has been obtained through extensive numerical simulations.
For quantitative analysis, the 95% exceedance value is taken as the treshold value and analyzed from the simulation data. In the paper by Goda and Kobune (1990) an empirical formulation has been derived for $\Delta r_{95\%}$, i.e.,

(5.9) $$\Delta r_{95\%} = \exp[a + b\ln N + c(\ln N)^2]$$

The coefficient a, b, and c are expressed as the function of the censoring parameter (eq. 5.9) for each distribution as listed in table 5.3. The 95% exceeded value of the residue of correlation coefficient can be utilized as a reference for the rejection of the considered distribution.

Table 5.3 Empirical coefficients for $\Delta r_{95\%}$ in the REC Criterion

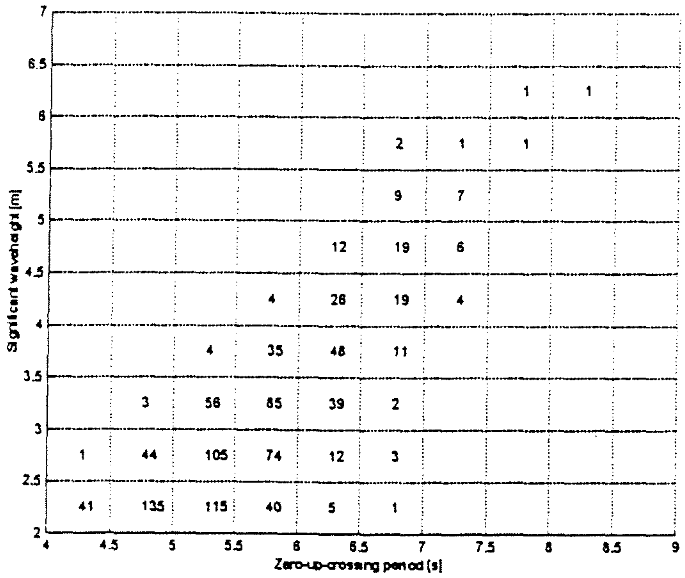
Distribution	Coefficient a	Coefficient b	Coef. c
FT-I	-1.444	$-0.2733-0.414v^{5/2}$	-0.045
Ft-II (β = 2.5)	$-1.122-0.037v$	$-0.3298+0.0105v^{1/4}$	0.016
Ft-II (β = 3.3)	$-1.306-0.105v^{3/2}$	$-0.3001+0.0404v^{1/2}$	0
Ft-II (β = 5.0)	$-1.463-0.107v^{3/2}$	$-0.2716+0.0517v^{1/4}$	-0.018
Ft-II (β = 10.0)	$-1.490-0.073v$	$-0.2299-0.0099v^{5/2}$	-0.034
Weibull (β = 0.75)	$-1.473-0.049v^{2}$	$-0.2181+0.0505v^{2}$	-0.041
Weibull (β = 1.0)	-1.433	-0.2679	-0.044
Weibull (β =1.4)	-1.312	$-0.3356-0.0449v$	-0.045
Weibull (β = 2.0)	$-1.188+0.073v^{1/2}$	$-0.4401-0.0846v$	-0.039

5.3 Goodness of fit of bivariate distributions

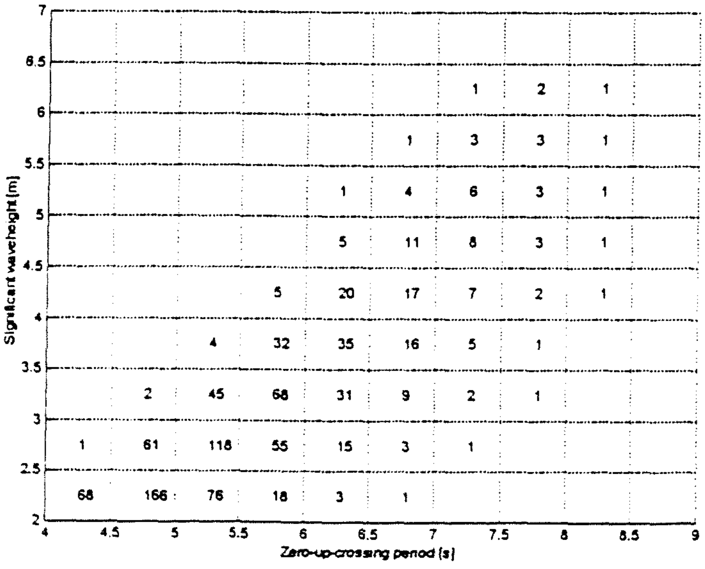
5.3.1 Visual judgement

For two-dimensional probability density functions, mostly goodness of criteria are used which are based on the scatter diagram of the two random variables.

Here, the scatter diagram that is composed from wave observations is called the "empirical" scatter diagram. With the fitted theoretical model, a similar scatter diagram can be composed. This scatter diagram can be considered as a "theoretical" scatter diagram. A visual judgement of the fit of the bivariate model to the wave data can be obtained by comparing these scatter diagrams. This is illustrated in figure 5.1.



							1	1	
					2	1	1		
					9	7			
				12	19	6			
			4	28	19	4			
		4	35	48	11				
	3	56	85	39	2				
1	44	105	74	12	3				
41	135	115	40	5	1				



						1	2	1	
					1	3	3	1	
				1	4	6	3	1	
				5	11	8	3	1	
			5	20	17	7	2	1	
		4	32	35	16	5	1		
	2	45	68	31	9	2	1		
1	61	118	55	15	3	1			
68	166	76	18	3	1				

Fig 5.1 Empirical (the upper figure) and theoretical scatter diagram. (The corresponding theoretical model is bivariate model 4 being composed of two Weibull distributions. The used sample is the North Sea data set)

Consider just one cell (with index (i, j)) of a scatter diagram. An observed data point (x,y) falls in this cell if

$$x_i < x \le x_{i+1} \qquad ; \qquad y_j < y \le y_{j+1} \tag{5.10}$$

where x_i, i=1,2,..,l and y_j , j=1,2..,m, are the boundaries of the cells. Under the hypothesised distribution $F(x,y)$, the expected probability $p_{ij}^{(e)}$ that any observed data point falls within this cell can easily be evaluated

$$p_{ij}^{(e)} = F(x_{i+1},y_{j+1}) - F(x_{i+1},y_j) - F(x_i,y_{j+1}) + F(x_i,y_j) \tag{5.11}$$

When the expected probability $p_{ij}^{(e)}$ is multiplied with the total number of observations (N), the expected number of data points in that cell is obtained:

$$n_{ij}^{(e)} = Np_{ij}^{(e)} \tag{5.12}$$

The *expected* number of data points $n_{ij}^{(e)}$ can be calculated for each cell of the scatter diagram. In this way, a *theoretical* diagram is obtained. The *empirical* scatter diagram is built from values which present the number of *observed* data points $n_{ij}^{(o)}$ in each cell.

5.3.2 2-dimensional Chi-square test

The 2-dimensional chi-square test is similar to the 1-dimensional version. The test statistic used is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(n_{ij}^{(o)} - n_{ij}^{(e)}\right)^2}{n_{ij}^{(e)}} \tag{5.13}$$

in which

$n_{ij}^{(o)}$	Number of observed values in cel i of the 2-dimensional histogram
$n_{ij}^{(e)}$	Number of expected values in cel i of the 2-dimensional histogram. This number is computed with the hypothesized bivariate distribution function (section 5.3.1.)

In earlier case studies (Burrows and Salih (1986), Athanassoulis et al. (1996), Mathisen et al (1990)) , it appeared that the standard Chi-square is less suitable for the test of fit of bivariate models to wave data. In general, the number of classes did not conform to the requirements of the standard test. Therefore, the resulting values of χ^2 could not be related to appropriate levels of significance in the usual way. Burrows and Salih (1986) proposed to use the numerical values of χ^2 as only qualitative indicators of the goodness of fit. This approach has also been followed in the present study.

5.3.3 Method developed by Mathisen and Bitner Gregersen

The test developed by Mathisen and Bitner-Gregersen (1990) is similar in principle to the chi-square test. The difference is that the chi-square provides an overall indication of the fit of the hypothesized distribution whereas this test provides an indication of the fit for each cell of the scatter diagram.

The test is based on the binomial distribution. With the binomial distribution the expected standard deviation for the number of data points falling in the cell can be derived as:

(5.14) $$\sigma_{ij} = \sqrt{Np_{ij}^{(e)}(1-p_{ij}^{(e)})}$$

Hence, the normalized deviation d_{ij} between the observed number of data points falling in a cell $n_{ij}^{(o)}$ and the expected number $n_{ij}^{(e)}$ is given by

(5.15) $$d_{ij} = \frac{n_{ij}^{(o)} - n_{ij}^{(e)}}{\sigma_{ij}^{(e)}} = \frac{n_{ij}^{(o)} - Np_{ij}^{(e)}}{\sqrt{Np_{ij}^{(e)}(1-p_{ij}^{(e)})}}$$

This deviation provides an indication of the goodness of fit.

An application of the method is presented in the figure below. It contains contour lines presenting the deviations d_{ij} =+1,+2,-1,-2 and –3.

The plot gives an indication of the fit of model 4 (consisting of two marginal Weibull distributions) to the North Sea data set (See chapter 8). Note that the figure gives a similar indication of the goodness of fit as the earlier described visual inspection with the scatter diagrams: model 4 overpredicts sea states with a relative small wave steepness (T_z=6.5..7.5 s, H_s=3.5..4,5 m) and underpredicts waves with a mean wave steepness value, especially just above the treshold level (T_z=5..6 s, H_s=2..2.5 m). Also the overprediction within the cell (T_z=4..4.5, H_s=2..2.5 m) is similar.

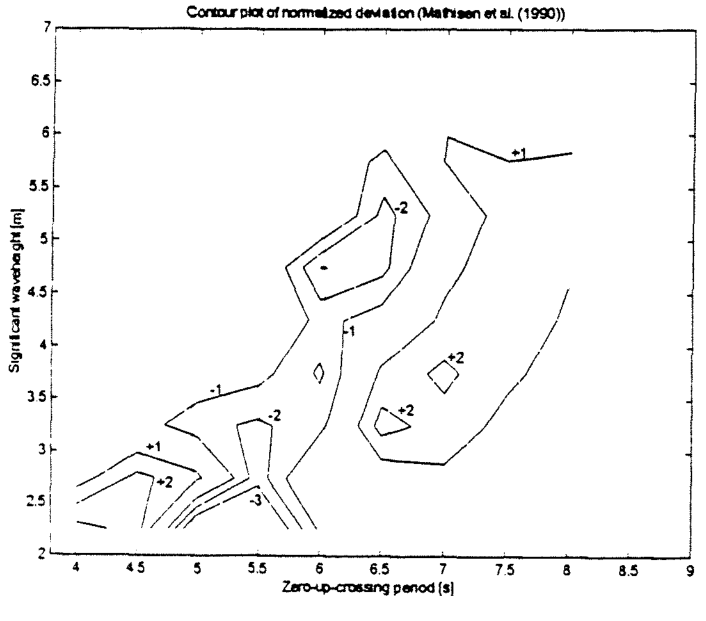


Fig 5.2 Contour plot of normalized deviation for model 4 (the model proposed by Vrijling, consisting of two marginal Weibull distributions) relative to the North Sea data (H_s>2m)

5.4 References

[1] Athanassoulis, G. A., Skarsoulis, E.K., Belibassakis, K.A., *"Bivariate distributions with given marginals with an application to wave climate description"*. Applied ocean research, 1994, v. 16, p 1-17

[2] Burrows, R., Salih, B. A., *"Statistical Modelling of long-term wave climats"*. Proc. 20th Int. conf. coastal. eng., 1986, pp. 42-56

[3] Goda,Y.,Kobune,K., *"Distribution function fitting for storm wave data"*. Delft: Proc. 22th Int. conf. coastal. eng., 1990, pp. 18-31

[4] Mathisen,J., and Bitner-Gregersen, E., *"Joint distributions for significant wave height and wave zero-up-crossing period"*. Applied Ocean Research, 1990, Vol. 12, no. 2, pp 93-103

6 Extrapolation beyond the set of data

6.1 Introduction

Once the extreme data are fitted to a distribution function and the parameters are estimated, the probability function can be used to extrapolate beyond the set of data. Usually one is interested in the once per 50 or 100 years return period value.

This chapter deals with the computation of such return values. In section 6.2, the determination of marginal return values is presented. The bivariate return values are dealt with in section 6.3.

6.1 Return values of marginal distributions

The computation of return periods for a specific peak seastate is based on the encounter frequency of extreme events and the exceedance probability of a given level in a single event. For the Peak over Treshold (POT) method, the return period R, expressed in years, for a specific level of exceedance x_p, is computed as (Mathiesen et al. (1993))

(6.1) $$R = \frac{1}{k(P(x > x_p))}$$

in which k is the mean number of extreme wave events per year. From equation (6.1) it follows that

(6.2) $$P(x \le x_p) = 1 - \frac{1}{kR}$$

Thus, the mean value of the 1/R year prediction is computed as

(6.3) $$x_p = P^{-1}\left[1 - \frac{1}{kR}\right]$$

where P^{-1} denotes the inverse of $P(x \le x_p)$. For methods based on year maxima the parameter k is set to one.

In table 6.1, for each of the tested marginal distribution functions, the formulas are presented by which the 1/R year prediction can be calculated.

Table 6.1 Formulas for calculation of return values of the selected marginal distributions

Distribution	Formula x_p
Exponential	$x_p = \lambda + \delta(-\log(1 - P(x \le x_p)))$
Gumbel	$x_p = \lambda - \delta \log(-\log(P(x \le x_p))$
Weibull	$x_p = \lambda + \delta(-\log(1 - P(x \le x_p)))^{1/\beta}$
Frechet	$x_p = \dfrac{\delta + \lambda(-\log(P(x \le x_p))^{1/\beta}}{(-\log(P(x \le x_p))^{1/\beta}}$

For the Log-normal distribution there exists no analytical expression for the cumulative distribution function. Therefore also an expression for the inverse distribution cannot be determined. For this distribution, return values are computed numerically.

Confidence intervals
The reliability of return values is usually indicated with confidence bands, computed through various methods. The purpose of confidence intervals is to take into account the influence of the large number of uncertanties that are involved in extreme wave statistics.

In this section, three different confidence bands are discussed. The first type includes statistical uncertainty. The remaining two confidence bands are two attempts to include sample variability within the width of the bands. Only the first type is used in the present case studies.

Type 1: Confidence bands, which include statistical uncertainty
This type of confidence bands can be determined analytically and numerically.

The analytical solution is illustrated for the Gumbel distribution in appendix [6]. The variance of the return value is derived from the variance of the estimators of the distribution parameters. This procedure is not further used in the case studies.

The numerical approach is based on the bootstrap technique (Groeneboom et al. (1995)). From an available sample,

(6.4) $$x_1, x_2 \ldots\ldots, x_n$$

data points are drawn random, with replacement, until a new set of data is composed with the same number of data points as the original set of data. The set of bootstrap data is given by

(6.5) $$x_1^*, x_2^*, \ldots\ldots\ldots, x_n^*$$

Repeating this procedure a large number of times, for instance 500 times, 500 sets of bootstrap data are composed. For each set, the parameters of the selected distribution function are estimated and the corresponding return value is estimated. This gives a set of 500 return value estimates

(6.6) $$x_p(1), x_p(2), \ldots\ldots\ldots, x_p(500)$$

From this set, the mean value and variance of the return value can be determined.
When the variance of a return value is known, the corresponding confidence intervals can be calculated. Considering a 95% confidence interval, the expression follows as (Groeneboom et al. (1995))

(6.7) $$\left\{\hat{x}_p; E(\hat{x}_p) - 1.96\sqrt{\mathrm{var}(\hat{x}_p)} < E(\hat{x}_p) < E(\hat{x}_p) + 1.96\sqrt{\mathrm{var}(\hat{x}_p)}\right\}$$

in which

$E(\hat{x}_p)$	mean value of return value
$\mathrm{var}(\hat{x}_p)$	variance of return value

Type 2:Confidence bands that are composed on basis of "artificial" sample variability (Maes et al. (1993))

In this method, "artificial" sample variability is created by adding random errors to the data set.

Random errors are generated using a Monte Carlo simulation assuming a normal deviation with a chosen standard deviation. (The choice of the standard deviation is subjective.,e.g. 1.0 m). Next, the data set containing the random errors is fitted to the chosen distribution function. After for example 10,000 of such simulations, a set of 10,000 estimates of the return value (x_p) is obtained. From this sample the average and standard deviation of x_p is known. With equation 6.5 the confidence bands can be determined.

A weak point of this approach is ofcourse the subjectivity that is involved by the choice of the standard deviation of the normal distributed error.

Type 3: Confidence bands that are based on an empirical standard deviation of the return value
Goda (1988;1990) calculated the weighted mean values of standard errors of return values for uncensored and censored data (υ=1.0,0.5,0.25 ; definition υ: see equation 5.3), based on his simulated data. The results of the weighted mean errors have been expressed by empirical formulas. The first empirical formula is appropriate for the Gumbel and the Weibull distribution function with the shape parameter fixed at β=0.75,1.0,1.4 and 2.0:

(6.8) $$\sigma[x_R] = (\sigma_x / \sqrt{N})[1.0 + A_s(y_R + \alpha \log \nu)^q]$$

(The formulae for x_R and y_R have already been shown in table 4.2). σ_x is the unbiased standard deviation of the sample data. The coefficient A_s and α and the exponent q are given the following values depending on the best-fitting distribution function:

Gumbel:

(6.9) $$A_s = \begin{cases} 0.24 + 0.36(\log_{10} N/80)^2 : \nu = 1.0 \\ 0.46 + 0.14(\log_{10} N/50)^2 : \nu = 0.5;0.25 \end{cases}$$
$$\alpha = 0.9; q = 1.6$$

Weibull(β=0.75):

(6.10) $$A_s = \begin{cases} 0.57 + 0.18(\log_{10} N/20)^2 : \nu = 1.0 \\ 0.41 + 0.22(\log_{10} N/20)^2 : \nu = 0.5;0.25 \end{cases}$$
$$\alpha = 2.7; q = 1.2$$

Weibull(β=1.0):

(6.11) $$A_s = \begin{cases} 0.55 + 0.15(\log_{10} N/15)^2 : \nu = 1.0 \\ 0.38 + 0.17(\log_{10} N/20)^2 : \nu = 0.5;0.25 \end{cases}$$
$$\alpha = 1.0; q = 1.7$$

Weibull(β=1.4):

(6.12) $$A_s = \begin{cases} 0.37 + 0.08(\log_{10} N/1000)^2 : \nu = 1.0 \\ 0.46 + 0.09(\log_{10} N/20)^2 : \nu = 0.5;0.25 \end{cases}$$
$$\alpha = 0.35; q = 3.2$$

Weibull (β=2.0):

(6.13)
$$A_s = \begin{cases} 0.30 + 0.36(\log_{10} N/80)^2 : \nu = 1.0 \\ 0.56 + 0.20(\log_{10} N/100)^2 : \nu = 0.5;0.25 \end{cases}$$
$$\alpha = 0.35; q = 3.2$$

The second empirical formula is appropriate for the Frechet distribution function with the shape parameter fixed at β=2.5,3.33,5.0 and 10.0:

(6.14)
$$\sigma[x_R] = (\sigma_x / \sqrt{N})[1.0 + A_s(y_R - c + \alpha \log \nu)^2]$$

$$A_s = A_1 \exp\left\{ A_2\left(\log\left(\frac{N\nu^{0.5}}{N_0}\right)\right)^2 - \kappa\left(\log\left(\frac{\nu}{\nu_0}\right)\right)\right\}$$

The corresponding empirical parameters are listed in table 6.2

Table 6.2: Empirical coefficients for the standard deviation of the Frechet return value

β	A_1	A_2	N_0	κ	υ0	c	α
2.5	1.27	0.12	23	0.24	1.34	0.3	2.3
3.33	1.23	0.09	25	0.36	0.36	0.2	1.9
5.0	1.34	0.07	35	0.41	0.45	0.1	1.6
10.0	1.48	0.06	60	0.47	0.34	0	1.4

With eq. (6.14), the standard deviation value of the return value can be determined. The corresponding confidence bands follow from eq. (6.7).

6.2 Return values of bivariate distributions

For the bivariate case, the return period R, expressed in years, for specific levels of exceedance x_p and y_p, is computed as

(6.15) $$R = \frac{1}{k(P(x > x_p, y > y_p))}$$

where x represents the wave period T, and y represents the significant wave height H_s. Again k is the mean number of extreme wave events per year. (As mentioned in appendix 1, in the present report the *extremes wave events* are the storm events corresponding to the extremes of H_s. The observations of T_p that are used for the extreme analysis are associated with the extremes of H_s).

From equation (6.15) it follows that

(6.16) $$P(x > x_p, y > y_p) = \frac{1}{kR}$$

The probability of exceedance the given values x_p and y_p is

(6.17) $$P(x > x_p, y > y_p) = \int_{y_p}^{\infty}\int_{x_p}^{\infty} f(x, y)dxdy$$

With the two above equations, for a given return period R, pairs of return values (x_p ,y_p) can be calculated.

For a constant value of the probability of exceedance, various combinations of x_p and y_p are possible. These pairs of return values (x_p,y_p) form a line in the joined probability space. In figure 6.1 the lines corresponding to a return period of 1, 10 and 50 years are superimposed on the contour plot of the joint probability density function. In the figure, the start point of the lines is a pair of H_s and T_p corresponding to a wave steepness of 1.5%. The endpoint of the lines is a pair of H_s and T_p corresponding to a wave steepness of 0.5%. The lines represent thus a constant value of the probability of exceedance (See fig 6.2). (In the marginal case, a constant value of the probability of exceedance is a point in the 1-dimensional probability space).

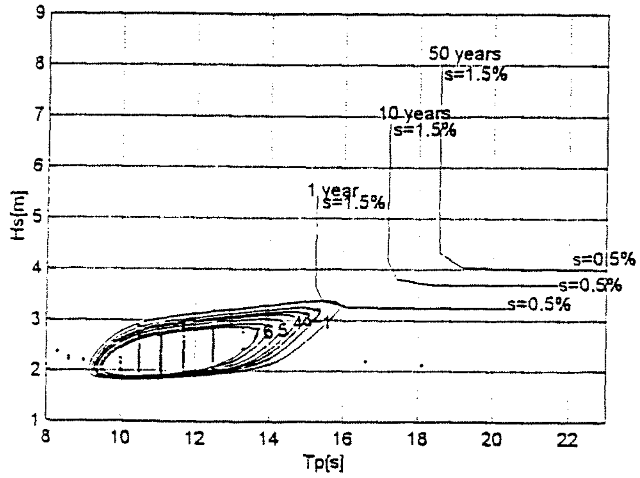


Fig. 6.1 Lines of return values with a return period of 1,10 and 50 years for bivariate model 5 (s=[0.5%..1.5%],)

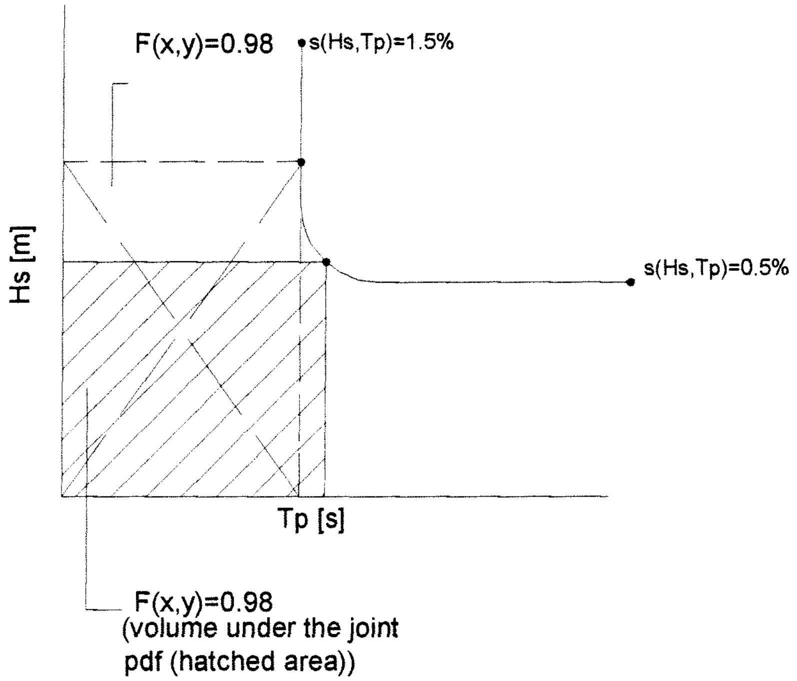


Fig 6.2 Illustration of line corresponding to non-exceedance probability F(x,y)=0.98 (R=50 years, k=1)

6.3 References

[1] Goda, Y.,*"On the methodology of selecting design wave height"*. Malaga: Proc. 21th Int. conf. coastal eng., 1988, pp. 135-136

[2] Goda,Y.,Kobune,K., *"Distribution function fitting for storm wave data"*. Delft: Proc. 22th Int. conf. coastal. eng., 1990, pp. 18-31

[3] Groeneboom, P., Kraaikamp, C., Kopuhaä, H.P, and van der Weide, J.A.M., *Probability and statistics (in Dutch)*. Delft: Delft University of Technology, Faculty of Civil Engineering, 1995

[4] Maes,M. A.,and Gu,G.Z.,*"Techniques used to determine extreme wave heights from the NESS Data set"*. Gaithersburg: Proc. of the conf. on extreme value theory and applications, volume 2, 1994, pp. 435-444

[5] Mathiesen, M. et al., *"Intercomparison of extremal wave analysis methods using numerically simulated data"*. New orleans: Proc. WAVES '93 conf., july 1993, pp 963-977

7 Computer program for statistical analysis

7.1 Introduction

Previous procedures for the selection of data, probability functions, parameter estimation methods, and goodness of fit criteria have been programmed. For this, the computer language Matlab (Matlab 5) is used.

The outline of the program is given in section 7.2. Further the following parts of the program are explained and illustrated in section 7.3 to 7.6:

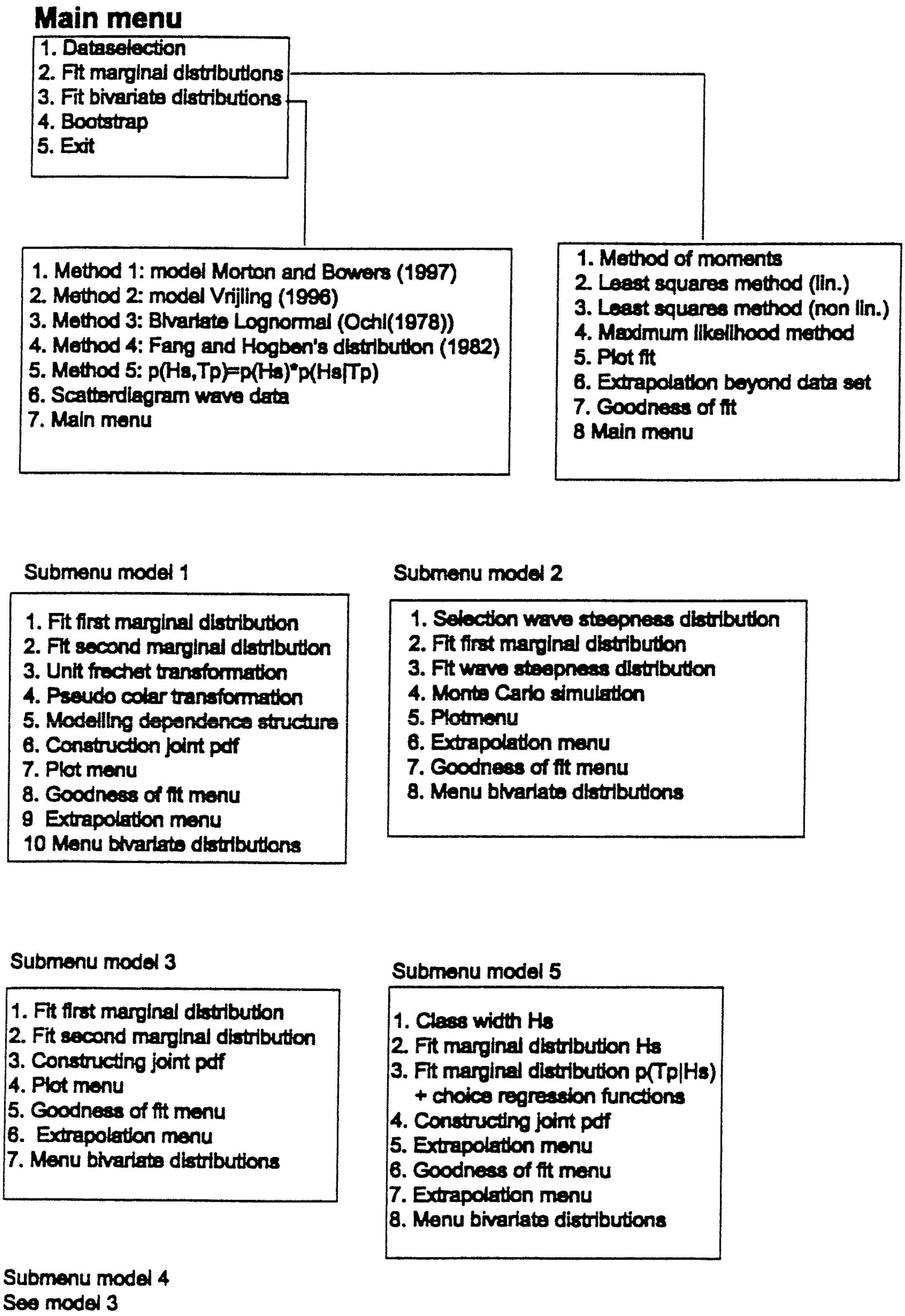
- the computation of the "empirical" scatter diagram (sec. 7.3)
- the iteration procedure for the maximum likelihood method (sec. 7.4)
- the iteration procedure for the non linear least squares method (sec. 7.5)
- the calculation of return-values for the bivariate models (sec. 7.6)

7.2 Outline of the program

As can be seen from fig 7.1, the program starts with the selection of data. In this submenu, it is possible to load a sample via the floppy disk station of the pc. If this sample consists of 3-hourly measurements, a selection of extremes can be made by using a declustering procedure (section 2.3) and by applying the POT method (section 2.4).

Then the probability distribution functions and parameter estimation methods are selected. These are similar to those described previously. After the fit of the functions to the data, the goodness of fit (chapter 5) can be determined and return values (chapter 6) can be calculated.

The last option of the main menu contains a bootstrap procedure, which can be used to compute confidence intervals for marginal return values (section 6.2).



Main menu

1. Dataselection
2. Fit marginal distributions
3. Fit bivariate distributions
4. Bootstrap
5. Exit

1. Method 1: model Morton and Bowers (1997)
2. Method 2: model Vrijling (1996)
3. Method 3: Bivariate Lognormal (Ochi(1978))
4. Method 4: Fang and Hogben's distribution (1982)
5. Method 5: p(Hs,Tp)=p(Hs)*p(Hs|Tp)
6. Scatterdiagram wave data
7. Main menu

1. Method of moments
2. Least squares method (lin.)
3. Least squares method (non lin.)
4. Maximum likelihood method
5. Plot fit
6. Extrapolation beyond data set
7. Goodness of fit
8 Main menu

Submenu model 1

1. Fit first marginal distribution
2. Fit second marginal distribution
3. Unit frechet transformation
4. Pseudo colar transformation
5. Modelling dependence structure
6. Construction joint pdf
7. Plot menu
8. Goodness of fit menu
9 Extrapolation menu
10 Menu bivariate distributions

Submenu model 2

1. Selection wave steepness distribution
2. Fit first marginal distribution
3. Fit wave steepness distribution
4. Monte Carlo simulation
5. Plotmenu
6. Extrapolation menu
7. Goodness of fit menu
8. Menu bivariate distributions

Submenu model 3

1. Fit first marginal distribution
2. Fit second marginal distribution
3. Constructing joint pdf
4. Plot menu
5. Goodness of fit menu
6. Extrapolation menu
7. Menu bivariate distributions

Submenu model 5

1. Class width Hs
2. Fit marginal distribution Hs
3. Fit marginal distribution p(Tp|Hs)
 + choice regression functions
4. Constructing joint pdf
5. Extrapolation menu
6. Goodness of fit menu
7. Extrapolation menu
8. Menu bivariate distributions

Submenu model 4
See model 3

Fig 7.1 Structure of the program

7.3 The computation of the "empirical" scatter diagram

A scatter diagram is a bivariate histogram that is composed from two sets of data. In this case, the data consists of observations of H_s and T. The first step of the computation of the empirical scatter diagram is the rearranging of the H_s-values in ascending order, with the lowest value of the sample being the first data point. Since the two variables are coupled, the wave period values are replaced similar with the wave height values.

Example:

[(3.45 m, 7.20 s), (2.90 m, 6.80 s), (5.90 m, 11.00 s), (2.05 m, 8.00 s)]

Rearranging coupled sets of data gives:

[(2.05 m, 8.00 m), (2.90 m, 6.80 s), (3.45 m, 7.20 s), (5.90 m, 11.00 s)]

The cell sizes of the scatter diagram can be varied. The class width of H_s and T lies within the interval [0,1 m] and [0,1 s], respectively. The class widths are defined for the complete sample range of H_s and T; it is not possible to compose a scatter diagram with for instance two different class widths for H_s.

In figure 7.2, a schematization of the next stages of the calculation procedure is shown. First, the class width of H_s and T are entered by the user of the program. The parameters are defined as "classwidthHs" and "classwidthT". Then for both H_s and T the number of classes is computed. ("nclassHs" and "nclassT").

Next, the number of observations is counted for each cell of the diagram. This stage is built of two loops. The outer loop is considered with the horizontal boundaries of the cell, and the inner loop is dealt with the vertical boundaries of the cell.

In the first loop the equations (1), (2), (3) and (4) are involved. With equation (1), wave height values beyond the left boundary of the cell are stored in array A. With equation (2) wave height values beyond the right boundary are stored in array B. ("Floor" in eq. (1), and (2) stands for the Matlab command that rounds off values towards minus infinity (5.6 → 5.0 ; 5.3 → 5.0)). With equation (3) the index numbers of the wave height values in the considered cell are stored in array C. These index numbers are used to detect the corresponding wave period values in the wave period sample. (eq. (4)) The result of the first loop is a sample of wave period values, stored in array D.

Example:

Initial data set:

[(2.50 m,8.00 s),(2.85 m,6.80 s),(2.90 m, 6.60 s),(2.95 m,6.55 s),(3.25 m,.5.80 s),(3.50 m,6.80 s),(4.50 m,6.55 s)]

Boundaries of cell that is considered: Left : H_s=2.75 m; right : H_s=3.00 m ; top: 6.75 s; bottom: 6.50 s

Array A = [(2.85 m),(2.90 m),(2.95 m),(3.25),(3.50),(4.50 m)]
Array B = [(3.25 m),(3.50 m),(4.50 m)]
Array C = [2 3 4]
Array D = [(6.70 s),(6.80 s),(6.55s)]

Within the second loop, the top and bottom boundary of the cell is taken into account. In this example, two values fall with in the cell boundaries:
T = 6.70 s and T = 6.55 s

These values are stored in an array. By computing the length of this array, the number of observations in the cell is known. This value is stored in the matrix "scatter". Thus:

Scatter (1,1)=2 (In this example, the first cell of the diagram is considered)

By repeating the above procedure for each cell, the scatter diagram is composed. For each cell, the number of observation is stored with in the matrix "scatter". This matrix can be presented in the (H_s,T)-space. (See figure 3.1).

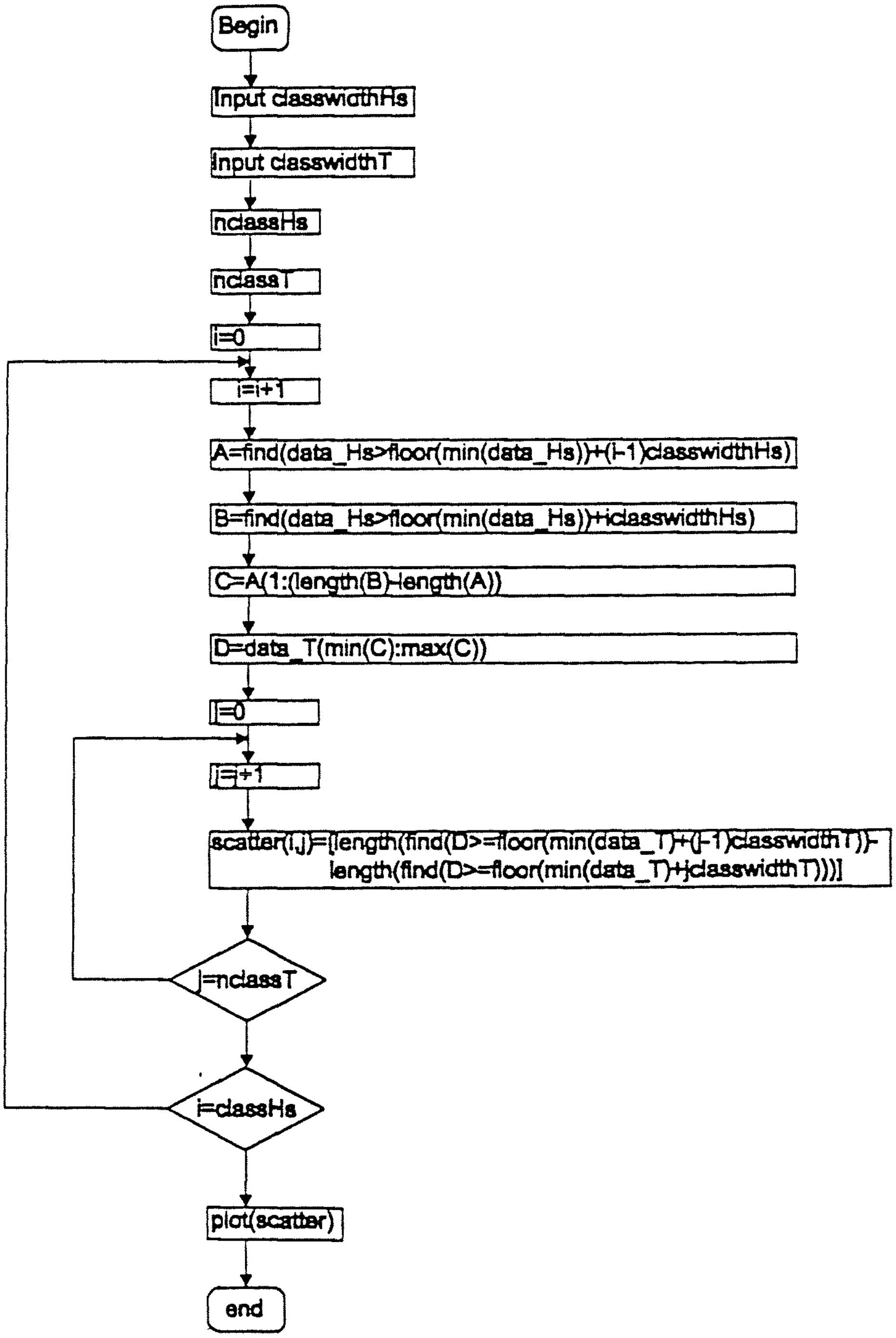

Begin
Input classwidthHs
Input classwidthT
nclassHs
nclassT
i=0
i=i+1
A=find(data_Hs>floor(min(data_Hs))+(i-1)classwidthHs)
B=find(data_Hs>floor(min(data_Hs))+iclasswidthHs)
C=A(1:(length(B)-length(A))
D=data_T(min(C):max(C))
j=0
j=j+1
scatter(i,j)=[length(find(D>=floor(min(data_T)+(j-1)classwidthT))-
length(find(D>=floor(min(data_T)+jclasswidthT)))]
j=nclassT
i=classHs
plot(scatter)
end

Fig 7.2. Flow chart of calculation of scatter diagram

7.4 The iteration procedure for the maximum likelihood method

In chapter 4, the analytical derivation of maximum likelihood estimators has been presented for each parameter of the selected distribution functions. These estimators were derived by finding the roots of the partial derivatives of the (log) likelihood function.

In the computer program, this method has been employed for the estimation of the parameters of the Exponential and the Log-normal distribution. The maximum likelihood estimators for the parameters of the remaining probability functions, i.e. the Gumbel, the Weibull, and the Frechet distribution, are determined by direct maximization of the loglikelihood function.

With the latter method, the estimators are computed numerically by use of some iteration procedure. The presently used iterative calculation method is illustrated in figure 7.3.

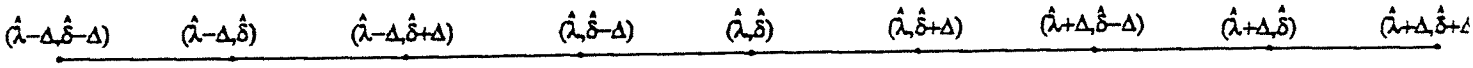


Fig 7.3 Iteration grid for numerical determination of maximum likelihood estimators

The figure shows a grid of estimator values. As can be seen from the figure, two distribution parameters are involved in the iteration calculations. For the three-parameter distribution functions, i.e. the Weibull and the Frechet distribution, the location parameter (λ) is estimated by the minimum value of the considered sample – 0.1. (In the case of the Gumbel distribution, the shown scheme contains its location (λ) and scale parameter (δ) , for the Weibull and the Frechet distribution the scheme contains their scale and shape parameter (β).)

The grid consists of nine grid points and at each point, a different combination of estimator values is present. For each parameter (δ), three values are included, i.e. $\hat{\delta} - \Delta, \hat{\delta}$ and $\hat{\delta} + \Delta$. "Δ" is a step size, which decreases during the iteration proces.

For each grid point, the (log) likelihood function is computed. The grid point that contains the largest value of the (log) likelihood function is stored. When this grid point is equal to point 5, the two estimators without an added step size Δ $(\hat{\lambda}, \hat{\delta})$, provide the largest value of the (log) likelihood function. In that case, the step size is halved.

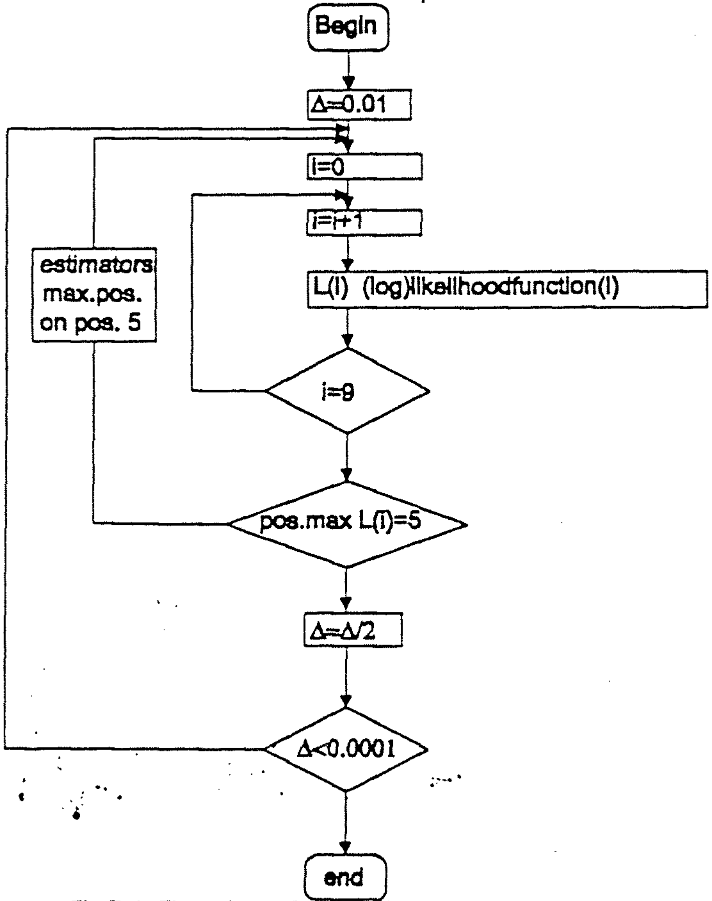

Begin
Δ=0.01
i=0
i=i+1
L(i) (log)likelihoodfunction(i)
i=9
pos.max L(i)=5
estimators max.pos. on pos. 5
Δ=Δ/2
Δ<0.0001
end

Fig 7.4. Flow chart of calculation of maximum likelihood estimators

When another grid point (X) contains the largest value of the loglikelihood function, then the grid is shifted. Grid point (X) is then placed at grid point 5. This procedure is repeated until again grid point 5 contains the largest value of the (log) likelihood function. The above iteration procedures are repeated until the step size is sufficiently small. In figure 7.4, a flow chart of the iteration procedure is shown.

The iteration proces can only be started, when begin values of the estimators are present. In the computer program, least squares estimators and estimators calculated with the method of moments are used as begin values.

7.5 The iteration procedure for the non linear least squares method

Non linear least squares estimated are computed numerically. The calculation method is based on the following expression:

$$\chi^2 = \min \sum_{i=1}^{n} (y_i - F(x_i; \hat{\lambda}, \hat{\delta}(, \hat{\beta})))^2 \tag{7.1}$$

The iteration proces is similar to the one that has been described for the maximum likelihood method. Instead of calculating the (log) likelihood function, here the sum of squares is calculated at each point. The grid point that contains the smallest value of the sum of squares is stored. When this point is equal to point 5, the step size is reduced, otherwise the grid is shifted as described previously.

For the start values of the proces, different parameter estimation methods were used. During tests of the program it appeared that various start values resulted in different values of the non-linear least squares estimators. In order to minimize the influence of the start values, begin values are calculated with the method of moments, the linear least squares method and the maximum likelihood. This results in three different non-linear least squares estimators. The one that provides the smallest sum of squares is considered as the best non-linear least squares estimator.

7.5 Estimation of return values for the bivariate models

Simultaneous with the two previous iteration procedures, a grid scheme is set up for the calculations. Only in this case, the estimators of distribution parameters are replaced by estimation of return values. The grid is shown in figure 7.5.

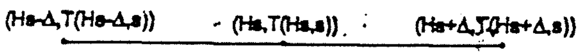


Fig 7.5. Iteration grid for numerical determination of bivariate return values

Return values are computed with the following equations (See chapter 6):

$$P(x > x_p, y > y_p) = \frac{1}{kR} \tag{7.2}$$

and

$$P(x > x_p, y > y_p) = \int_{y_p}^{\infty}\int_{x_p}^{\infty} f(x,y)dxdy \tag{7.3}$$

in which x is the wave period and y represents the significant wave height. With the two above equations, for a given return period R, pairs of return values (x_p ,y_p) can be calculated. Various combinations of x_p and y_p are possible: these pairs of return values (x_p,y_p) form a region in the joined probability space.

In order to fix the relation between x and y, the wave steepness is used. Assuming a deepwater wave field, the wave steepness is given by

$$(7.4) \qquad s(x,y)=\frac{y}{\left(\frac{gx^2}{2\pi}\right)}$$

The input parameters for the calculation method are thus the return period (1, 10 or 50 years) and a value for the wave steepness ([1...8%]).

Example: return period = 50 years
wave steepness (s) =4 %
maximum significant wave height value in data set (max (sample))=6.40 m
Δ (step size)= 0.5 m
k=1

The iteration procedure is started with the following begin values:

grid point 1: H_s (return period)=max(sample)-Δ
grid point 2: H_s (return period)=max(sample)
grid point 3: H_s (return period)=max(sample)+Δ

Thus,

grid point 1: $H_s(50)$=5.90 m and T(50)=9.72 s (eq. (4))
grid point 2: $H_s(50)$=6.40 m and T(50)=10.12 s
grid point 3: $H_s(50)$=6.90 m and T(50)=10.51 s

With eq. (7.2), the exceedance probability of the return value is computed:

$$P(x>x_p,y>y_p)=\frac{1}{kR}=\frac{1}{50}=0.02$$

With eq. (7.3), the non exceedance probability of the grid point values are computed:

grid point 1: $P(x>9.72s,y>5.90m)=\int_{5.90}^{\infty}\int_{9.72}^{\infty}f(x,y)dxdy=0.030$

grid point 2: $P(x>10.12s,y>6.40m)=\int_{6.40}^{\infty}\int_{10.12}^{\infty}f(x,y)dxdy=0.018$

grid point 3: $P(x>10.51s,y>6.90m)=\int_{10.51}^{\infty}\int_{6.90}^{\infty}f(x,y)dxdy=0.012$

The grid points provide an estimate of the exceedance probability of the return value. The exceedance probability calculated with eq (7.3), (0.02), is the true exceedance probability. In this example, grid point 2 provides the best estimate (0.018).

The iteration procedure is similar to the one that has been described for the maximum likelihood and the non-linear least squares method. When the best estimate of the non-exceedance probability is found at grid point 2, i.e. the grid point with estimators of H_s (50) and T (50) without an added step size Δ, then the step size is halved. Otherwise, the grid is shifted until again the best estimates are found at point 2.

The exceedance probability of the grid points is calculated by integration of the bivariate probability function (eq. (7.3)). In the computer program, the non-exceedance probability of these values is calculated numerically. The exceedance probability is determined by using the following expression

(7.5) $$P(x > x_p, y > y_p) = 1 - P(x \le x_p, y \le y_p)$$

The numerical integration of the bivariate probability density functions is illustrated with figure 7.6.

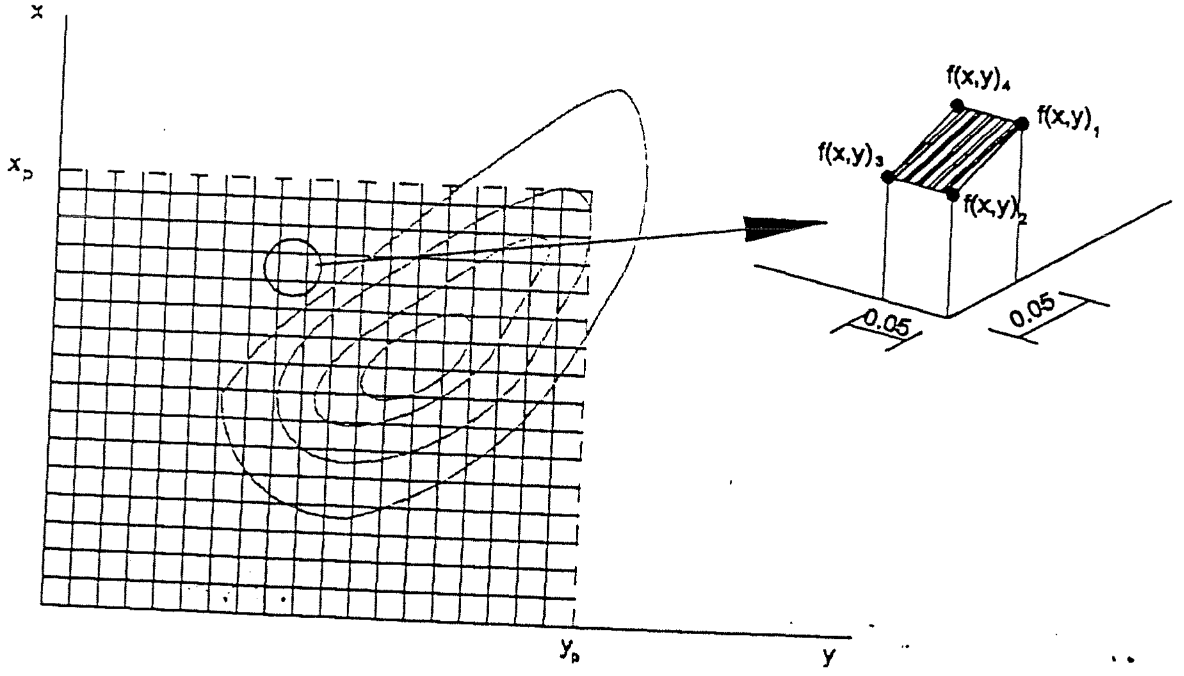


Fig. 7.6 Illustration of numerical calculation

In the figure, some contour lines of a bivariate probability density function are presented in the (H_s,T)-space. Integration of the probability density function means that the volume under the hatched surface is computed.

In order to obtain an accurate computation of the probability, the hatched area is subdivided into small cells (0.05 s x 0.05 m). For each corner of a cell, the value of f(x,y) is computed. From these four values, the mean value is computed,

$$f(x,y)_{cell(i)} = \frac{f(x,y)_{corner1} + f(x,y)_{corner2} + f(x,y)_{corner3} + f(x,y)_{corner4}}{4} \tag{7.6}$$

The probability of the cell is then calculated by

$$p_{cell(i)} = f(x,y)_{cell(i)} * 0.05 * 0.05 \tag{7.7}$$

The non-exceedance probability of the return value follows then as

$$P(x \leq x_p, y \leq y_p) = 1 - \int_0^{y_p}\int_0^{x_p} f(x,y)dxdy = \sum_i^n f(x,y)_{cell(i)} * 0.05 * 0.05 \tag{7.8}$$

in which n denotes the number of cells in the hatched area.

As mentioned above, for most of the cases, the dimensions of the cells were 0.05 m x 0.05 s. It must noted, however, that sometimes, especially when a relative high treshold level is used, this grid is not sufficient. A finer grid is then needed in order to obtain an accurate calculation. (The accuracy can be checked by intergrating the bivariate function over of a very large area: when the value of the probability mass is equal to 1 (or 0.99999999), the grid size is sufficient. When the value is higher then 1, the bins of the cells must be diminished).

7.6 References

Hanselman, D., Littlefield, B. , *The student edition of Matlab, version 5. The language of technical computing.* New Jersey: Prentice-Hall Inc / The Math Works Inc., 1997

8 Case studies

8.1 Introduction

The preceding chapters provided theoretical information about the statistical tools that are needed for an extremal wave analysis. In this chapter, the theory is applied in two case studies.

The case studies follow the phases of the analysis successively. First, sets of extreme observations are composed from the initial sets of data. For this, the statistical approach is used. In section 8.2, the initial sets are introduced and sets of extremes are obtained by declustering of the local maxima and by selecting extremes with the POT method.

Second, a marginal analysis of the extreme observations is made in section 8.3. The candidate distribution functions proposed in section 3.3 are fitted to the data, using the four parameter estimation methods described in chapter 4. The goodness of fit (chapter 5) is judged with visual inspection of the fits, and at least in the first case study, by comparing the linear correlation coefficients (section 5.2.5) and by using the two rejection criteria of Goda (section 5.2.4 and 5.2.6). Further marginal return values (section 6.2) corresponding to a return period of 50 years are computed.

Third, a bivariate analysis is examined. The marginal distributions recommended in the foregoing marginal analysis are used as the marginal components of the bivariate functions. The goodness of fit is being judged by visual comparison of the empirical and the theoretical scatter diagrams (section 5.3.1) and by evaluating the computed (2-dimensional) chi-square values of the models. Further bivariate return values (section 6.3) are computed.

It must be noted that the outcome of the following case studies has not been checked on physical grounds. The change of the climate between successive years has not been studied and also the physical limitations of the considered wave fields have not been analyzed. Further, the reliability of the data sets has not been checked extensively.

The applied bivariate probability density functions are illustrated with a contour plot. Each of these plots contains the same set of lines, which makes it possible to compare the models with each other. Table 8.1 shows this set of lines.

Table 8.1 Contour lines

Levels	(1/sec.m)
1	0.005
2	0.010
3	0.015
4	0.020
5	0.030
6	0.050
7	0.070

8.2 The selection of extreme observations

8.2.1 The data sets of Karwar

The first case study is dealt with the wave climate at the southwest coast of India. This climate is characterized by monsoon periods. During three months, the south-westerly monsoon is blowing, causing a wave field with an average of approximately H_s=2.0 m. During the other months of the year the sea is very calm.

The first data set of Karwar (291 data points) contains 3-hourly observations measured during the south-west monsoon (june-july 1988). It contains both wind waves and swell formed in the south of the Indian Ocean ("Roaring Forties").

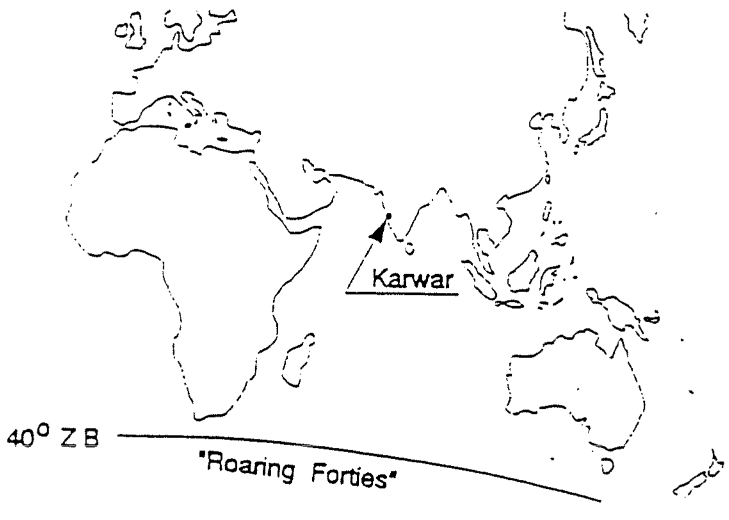


Fig 8.1 Geographical map (Taken from Vrijling (1996))

In figure 8.2 the data set is shown, together with lines representing a constant wave steepness value. (The wave period values of the set are spectral peak periods T_p. The given deepwater wave steepness is thus defined by H_s and T_p).

The relatively small value of the wave steepness of the observations (s= 0.8–1 %) might be explained by the long distance that the waves travel before they reach the coast of India. The weak correlation between the observed wave height and wave period values might be due to the fact that the observations representing both wind waves and swell. (The linear correlation (r) coefficient between H_s and T_p is 0.27).

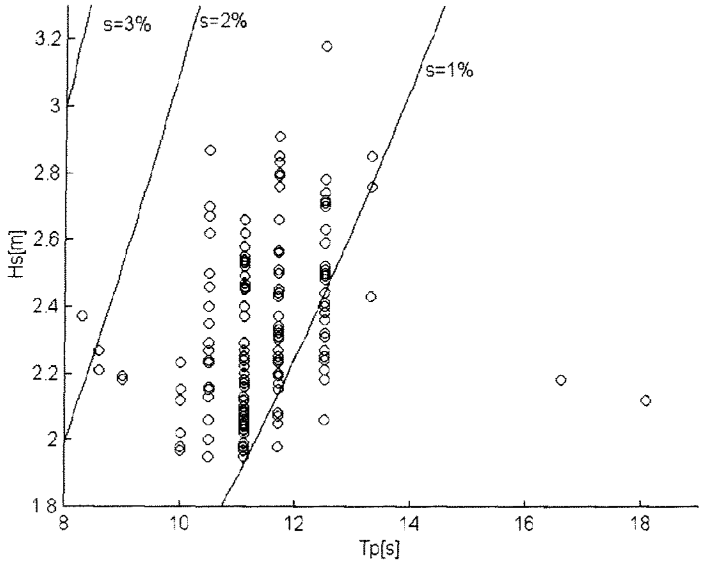


Fig 8.2 The first data set (291 points) of Karwar, together with lines of constant wave steepness (H_s>1.55 m)

On the average, one time a year the coast (Vrijling (1996)) is being hit by a hurricane. The second data set of Karwar that is analyzed, consists of 25 hindcasted significant wave heights and peak periods of hurricanes. The set is shown in figure 8.3.

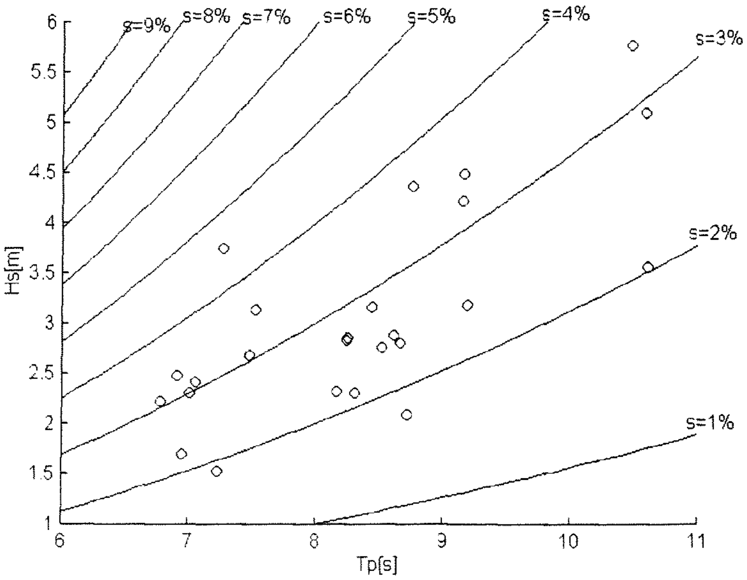


Fig 8.3 The hurricane observations (25 points) of Karwar, together with lines of constant wave steepness (The observations are denoted by small circles)

To determine the extreme conditions during the monsoon and the hurricane, both data sets have to be analyzed. The first data set, i.e. the observations of the south-west monsoon, consists of 3-hourly measurements. Since the mean duration of a storm lasts more than 3-hours, some observations in the set are correlated. To obtain a set of extreme and independent observations of H_s and T_p, the declustering of data described in section 2.3 should be carried out. The problem is, however, that the points of time, at which the observations have been measured, are not known. The location of

possible data gaps or missing data in the data set is not known so that the declustering cannot be examined. Therefore, it is decided to pass over the declustering of the observations and to select extremes of H_s and T_p only with the POT-method. It is thus assumed that the data points above the treshold level are independent.
As mentioned in section 2.4, the choice of the treshold level is subjective. By trial and error it was found that a treshold of H_s=1.95 m gave the closest fit of the candidate distributions to the data.

In table 8.2, both sets of data are listed.

Table 8.2 The data sets of Karwar used in the following analysis

Set nr	Description of observations	Treshold level [m]	Number of observations
1	South west monsoon	1.95	167
2	Hurricane	-	25

8.2.2 The data set of the North Sea

In the second case study, the wave climate in the North Sea is considered. The initial data set used has been measured at the Euro platform and consists of 3-hourly measurements (37.951 observations) covering a period of 12 years (1979 – 1991). The data was made available by ir A.P. Roskam from the RIKZ in The Hague.

From this set, three sets of extreme observations are drawn by following the statistical approach described in section 2.3 and 2.4. The first step of this approach is to identify independent storm events in the initial sample. For this, a time interval of 25 h, i.e. the minimum time interval between successive storms, is taken. The second step is to select storms with the POT method using the treshold levels H_s=2.00 m , 4.50 m and 5.00 m.

Table 8.3 The data sets of the North Sea

Set nr	Treshold level [m]	Number of observations
1	2.00	971
2	4.50	59
3	5.00	22

Figure 8.4 shows the observations of the first data set. The wave steepness of the data points vary from s=3 % to s=9%. (The wave steepness is defined on the zero-up-crossing period T_z).

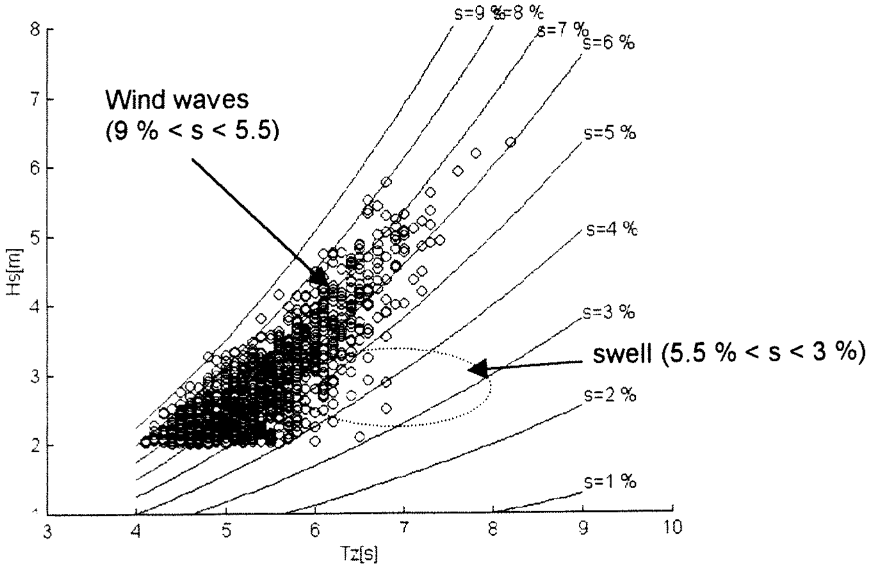


Fig.8.4 The observations of the North Sea data set (H_s>2.00 m)

It is assumed that the waves with a steepness between s=3% and s=5.5 % are swell. These waves are censored (section 3.4) for the extreme analysis, as shown in fig 8.5. The correlation between the remaining observations of wind waves is strong: the linear correlation coefficient (ρ) between the bivariate data points is 0.92.

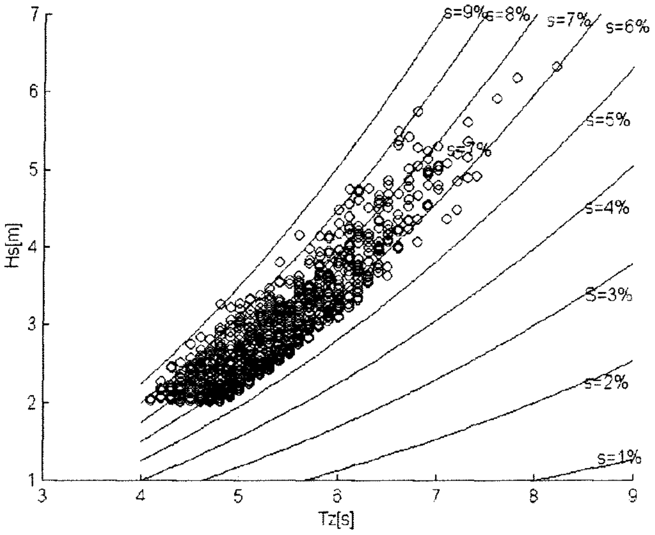


Fig. 8.5 The observations of the North Sea data set (H_s> 2m). The shown observations are wind waves. (ρ = 0.92)

8.3 Fit of marginal distribution functions to wave data

8.3.1 The data sets of Karwar

Data set 1; The observations of the monsoon wave climate

Marginal analysis of H_s
Figure 8.6 to 8.9 show the fit of candidate distributions to the observations of H_S. A visual inspection of the empirical cumulative distribution function (cdf) suspects inhomogeneous data. Halfway the plot, the curve of the empirical function increases strongly. Further, it is seen that the highest observation of H_s appears to be inconsistent with the other observations. Trying to model the probability distribution of the significant wave heights leads to large deviations in the extreme part of the fit. All the distributions tend to overpredict the extremes of H_s.

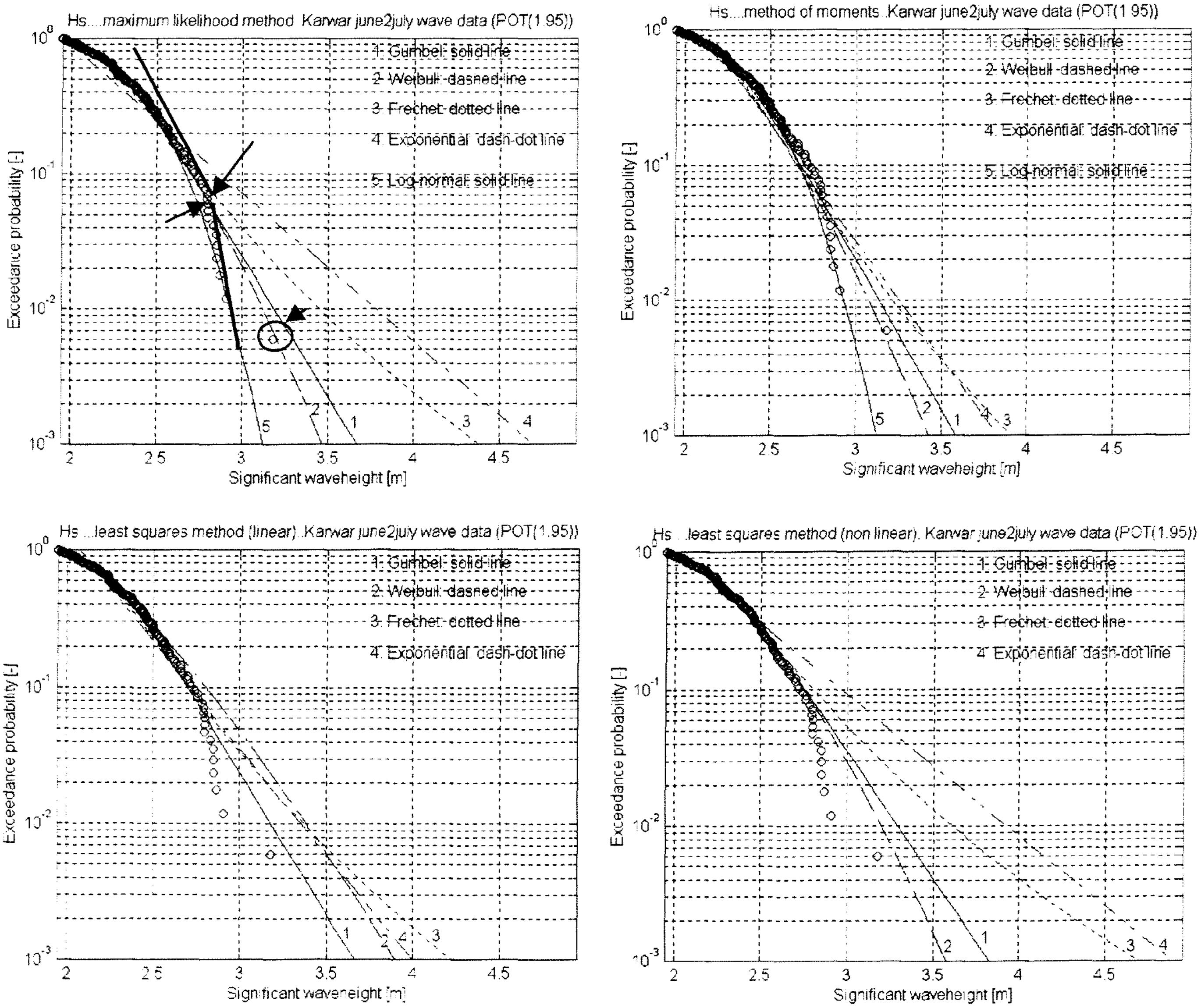


Fig 8.6 to 8.9 The candidate distributions fitted with the four parameter estimation methods to the observations of the south west monsoon (H_s>1.95 m , 167 points)

The inhomogeneity of the data points is probably due to the fact that the set contains observations of both wind waves and swell. The direction of both wave types is the shame, which makes it hard to distinguish them. (See fig. 8.2).

Goodness of fit

Observing the above figures, it is seen that the Gumbel, Weibull and Log-normal distribution provide the most reasonable fits to the data. This also follows from the linear correlation coefficient (r). The Weibull distribution fitted with maximum likelihood method gives the highest r-value : r=0.9929. Table 8.4 shows the estimators for the parameter of the Gumbel, Weibull and Log-normal distribution.

Table 8.4 Estimators for the parameters of the Gumbel, Weibull and Log-normal distribution

Data : H_s	Parameter	Symbol	Estimated values			
			MAX	MOM	Lin. LS	Non lin. LS
Gumbel	location	λ	2.228	2.234	2.230	2.227
	scale	δ	0.209	0.196	0.207	0.232
Weibull	location	λ	1.94	1.94	1.94	1.94
	scale	δ	0.451	0.410	0.469	0.466
	shape	β	1.588	1.500	1.546	1.347
Log-normal	location	λ	0.848	0.848	-	-
	scale	δ	0.107	0.107	-	-

Table 8.5 presents the results of the application of the REC and DOL criteria. As already mentioned in section 5.2, these empirical criteria have been developed by Goda only for the distribution functions listed below. Note that the Gumbel distribution and the Weibull distribution with a shape parameter (β) value of 1.4, i.e. the Weibull function which conforms mostly to the estimated ones (table 7.1: $\beta \approx 1.5$), are not rejected.

Table 8.5 : the REC and DOL criteria

Distribution	Rejected (r) / Not rejected (n)	
	REC	DOL
Gumbel	n	n
Frechet (β = 2.5)	r	r
Frechet (β = 3.3)	r	r
Frechet (β = 5.0)	r	r
Frechet (β = 10.0)	n	r
Weibull (β = 0.75)	r	r
Weibull (β = 1.0)	r	r
Weibull (β =1.4)	n	n
Weibull (β = 2.0)	n	n

Return values of H_s

With the above recommended distributions, the return value corresponding to a return period of 50 years is calculated. The results are shown in table 8.6.

Table 8.6 $H_s(50)$

Distribution	Parameter estimation method	$H_s(50)$ [m]
Gumbel	MOM	4.12
Gumbel	MAX	4.23
Gumbel	Lin LS	4.22
Gumbel	Non lin LS	4.45
Weibull	MOM	3.75
Weibull	MAX	3.72
Weibull	Lin LS	4.32
Weibull	Non lin LS	3.89
Log-normal	MOM	3.43
Log-normal	MAX	3.46

The estimated return values differ significantly. In order to judge which of the listed estimates is the most reliable one, the figures 8.6 to 8.9 must be studied more detailed. As can be seen from the figures, the Gumbel provides the most reliable fit to the extreme part of the observations. Therefore an appropriate value for the return value will be probably be H_s = 4.10 m

Confidence bands for marginal return values of H_s

In section 6.2 a bootstrap procedure has been presented for the computation of confidence bands for return values. An application of this procedure is given in table 8.7. The shown return values are computed with the Gumbel distribution by using the linear least squares method. The distribution has been fitted to extreme observations above the treshold of H_s=1.95 m (167 points).

Table 8.7 95% Confidence interval for $H_s(50)$, (4000 bootstraps; Gumbel fit ; Lin LS method)

Treshold [m]	Lower band [m]	$H_s(50)$ [m]	Upper band [m]	Band width [m]
1.95	3.90	4.07	4.26	0.18

Marginal analysis of T_p

As mentioned in chapter 2, the observations of H_s and T_p are coupled. For the determination of the joint treshold of both variables, the wave height is used as the key parameter. Figure 8.10 shows the maximum likelihood fit of the selected wave period distributions. The functions are fitted to observations of T_p which are associated with observations of H_s above the treshold H_s= 1.95 m.

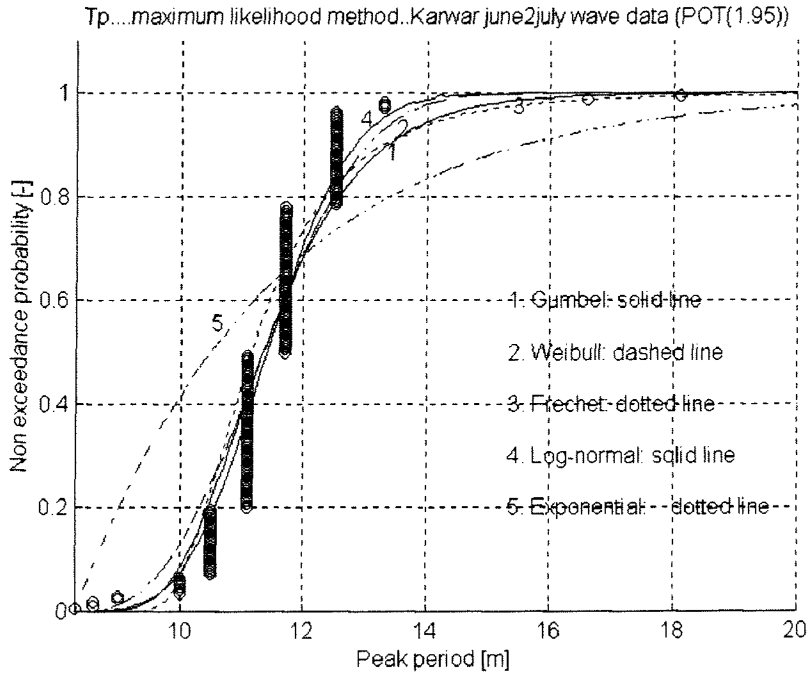


Fig 8.10 Maximum likelihood fit for the selected distributions (T_p)

Table 8.8 presents the parameters of the Log-normal distribution estimated with the maximum likelihood method and the method of moments.

Table 8.8 Estimators for the Log-normal distribution (T_p)

Parameter	Symbol	Estimated values	
		MAX	MOM
location	λ	2.434	2.434
scale	δ	0.091	0.094

Goodness of fit: visual judgement

From fig 8.10 it is clear that the Log-normal distribution gives the best overall fit for T_p.

Marginal return values of T_p

The once per 50 years return period value for T_p is computed with the Log-normal distribution. The value of the calculated return value is 16 s.

Data set 2: the hurricane observations

Marginal analysis of H_s

Figure 8.11 to 8.14 show the fit of candidate distributions to the hurricane data.

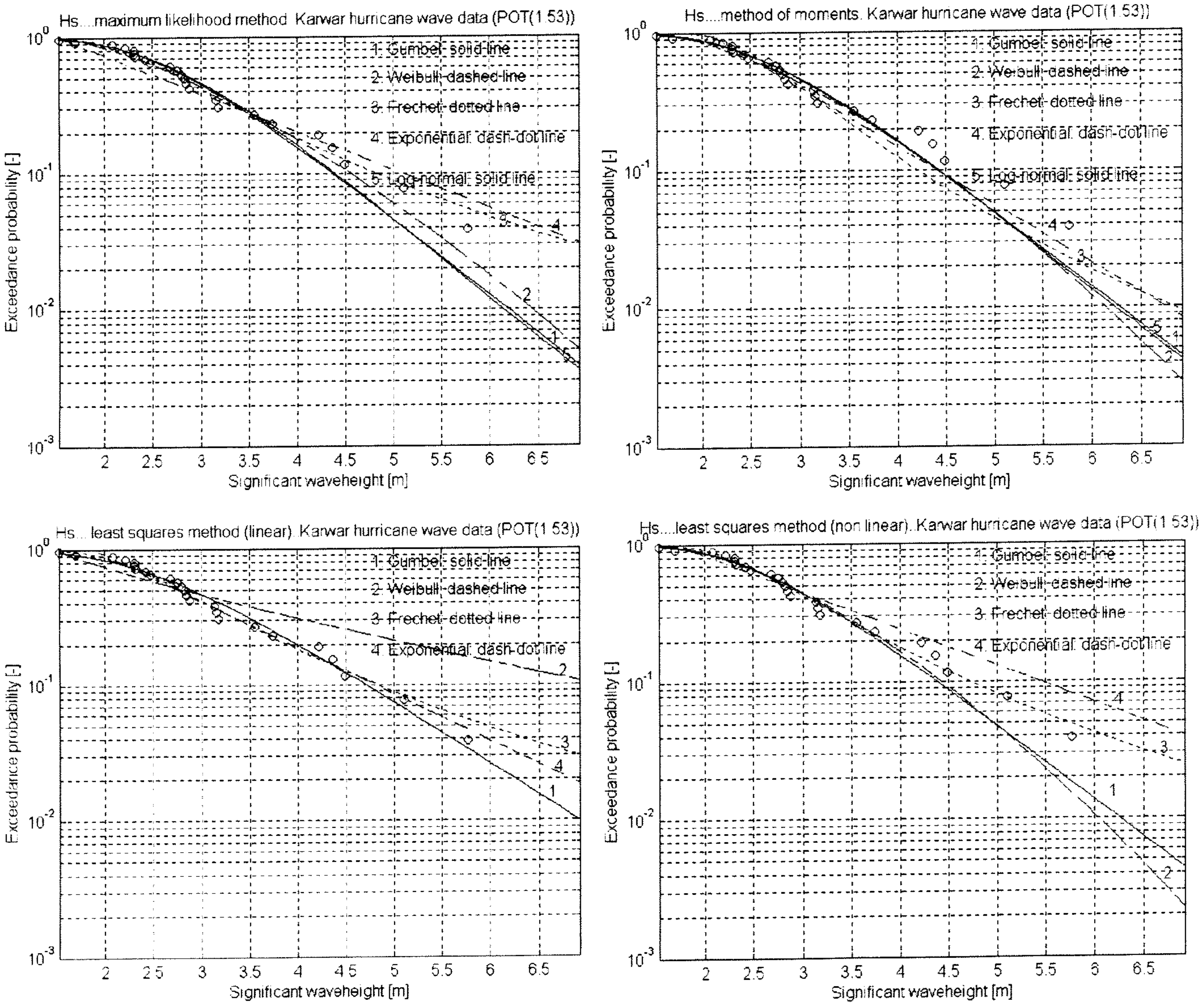


Fig 8.11 to 8.14 The candidate distributions fitted with the four parameter estimation methods to the hurricane observations of Karwar (25 points)

The Lin LS estimators of the parameters of the Gumbel and Exponential distribution are listed in table 8.9.

Table 8.9 Lin LS estimators for the parameters of the Gumbel and Exponential distribution (Hurricane observations of Karwar; 25 points)

Data : H_s	Parameter	Symbol	lin. LS
Gumbel	location	λ	2.573
	scale	δ	0.945
Exponential	location	λ	1.891
	scale	δ	1.263

Goodness of fit: visual judgement
Inspecting the above figures, it is clear that the linear least squares method provides the best fit to the data. The linear least squares fits of the Gumbel and Exponential distribution seem to be the best option.

Return values of Hs
Using the Gumbel and Exponential distribution computed from the Lin LS method, the once per 50 years significant wave height are $H_s(50)=6.26$ m and $H_s(50)= 6.83$ m, respectively. Note that these values are determined *under assumption that each year one hurricane occurs* (see Vrijling (1996), p. IV-37),

Marginal analysis of T_p / Goodness of fit (visual judgement)
The maximum likelihood fit of the Log-normal distribution appears to provide the closest fit to the observed values of T_p (fig 8.15).

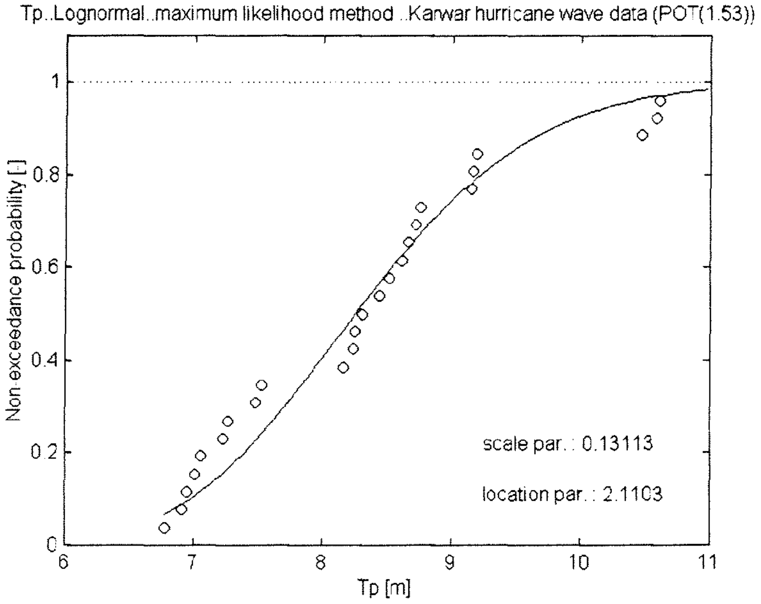


Fig 8.15 The maximum likelihood fit of the Log-normal distribution to the hurricane observations of Karwar

Return value of T_p
Using the above Log-normal distribution, the return value of T_p ,$T_p(50)$, is 11 s. (Again under the assumption that only one hurricane per year occurs !)

8.3.2 The North Sea data set

Marginal analysis of H_s

The figures 8.16 to 8.19 show the fit of the candidate distributions (section 3.3) to the observations of H_s above the treshold of H_s=2.00 m (971 points). Similar with the curve of the monsoon observations of Karwar, the curve of the empirical cumulative distribution function of the North Sea data changes at several places. (See figure 8.16). In general, this gives rise to suspect that the set of data is inhomogeneous.

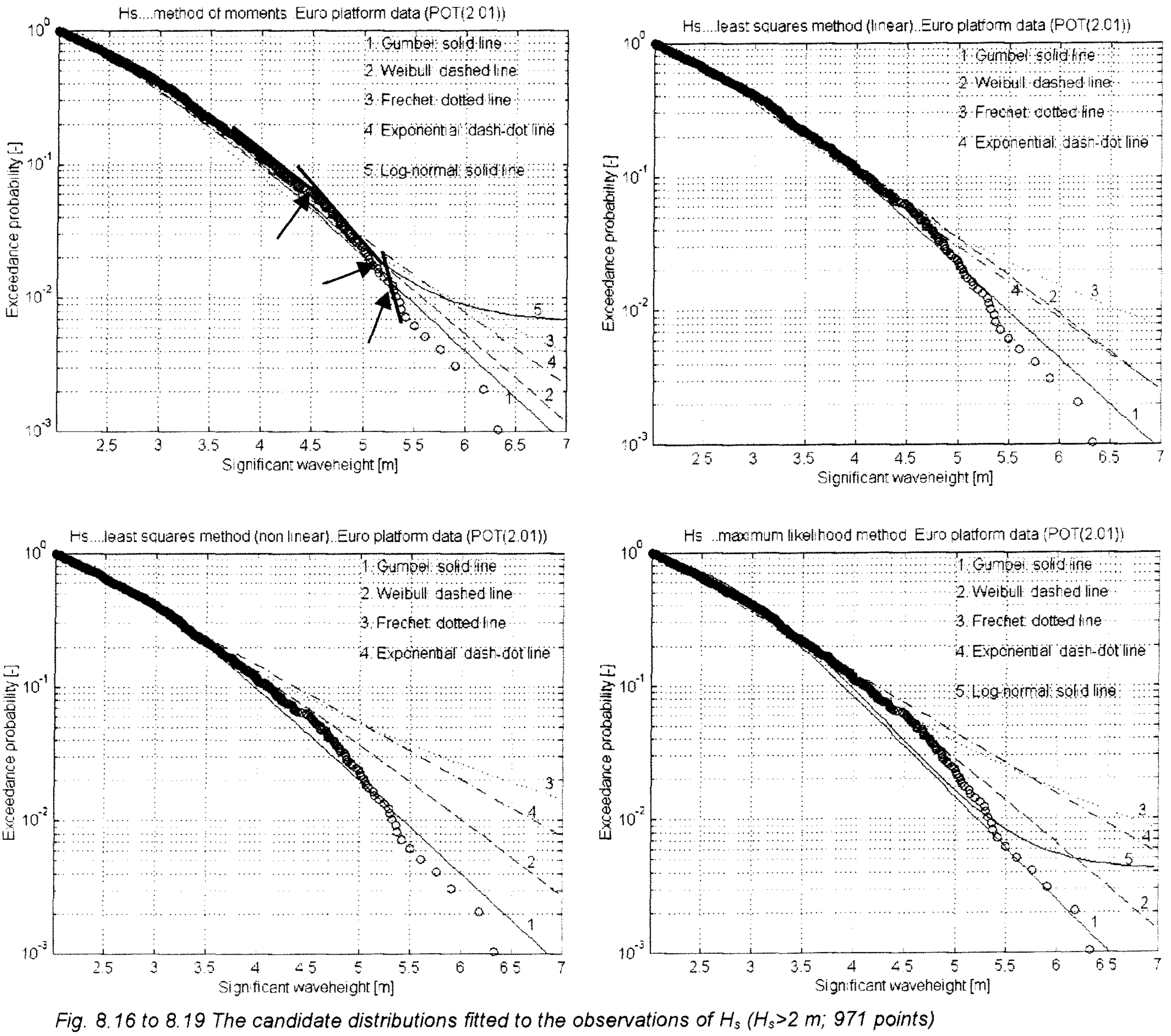


Fig. 8.16 to 8.19 The candidate distributions fitted to the observations of H_s (H_s>2 m; 971 points)

From the above figures it is seen that the extreme part of the observations are largely overpredicted by the distributions. This might be explained by the relative low value of the treshold: the lower observations of H_s , which are not extremes, largely influence the fit of the distributions to the data. The fit of the extreme part of the observations improves when the treshold level is raised. The figures 8.20 to 8.23 show the fit of the distributions to observations above the treshold of H_s=4.5 m (59 points). The figures 8.24 to 8.27 present the fit to observations beyond H_s=5.0 m.

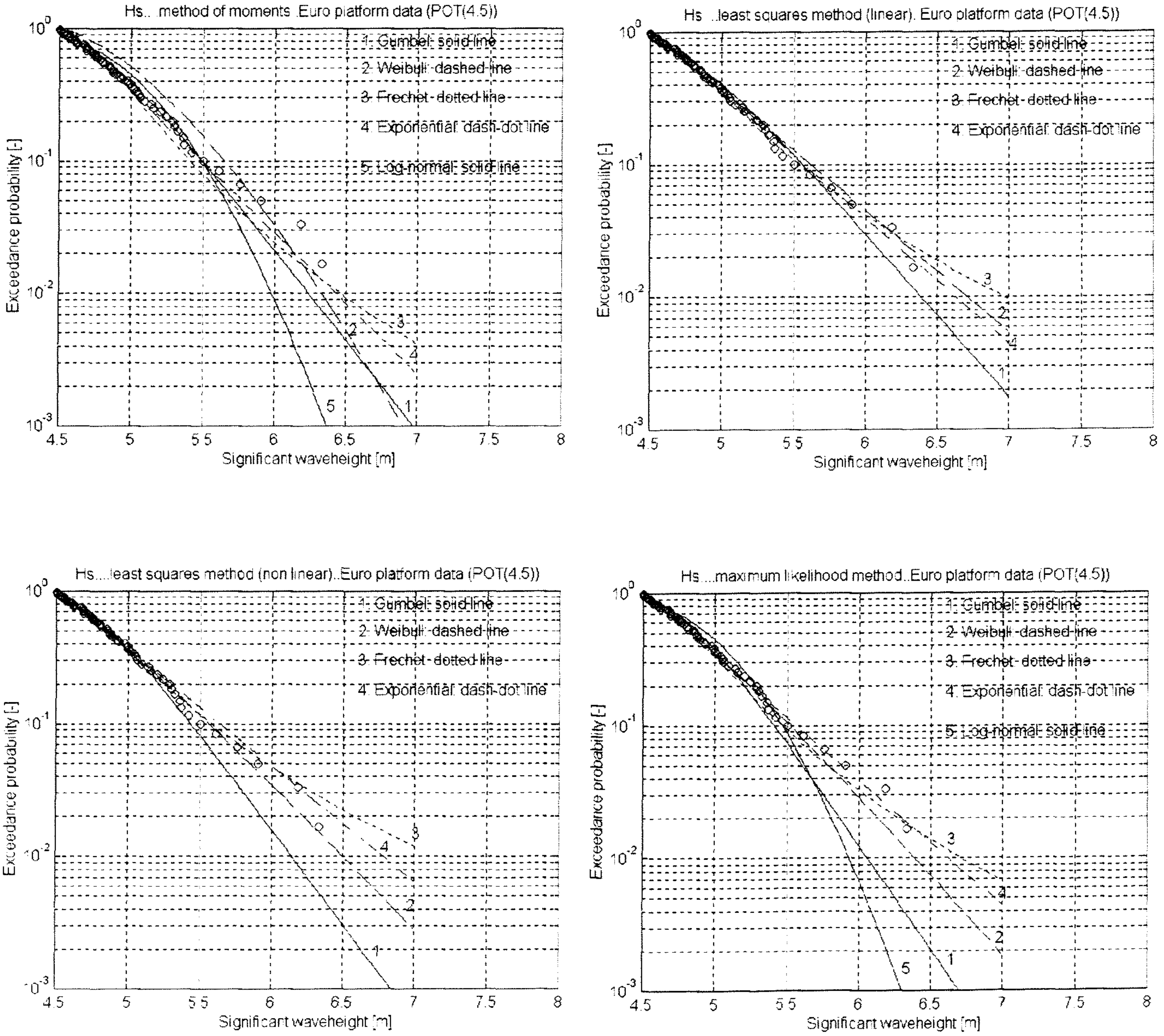


Fig. 8.24 to 8.27 The candidate distributions fitted to the observations of H_s (H_s>4.5 m; 59 points)

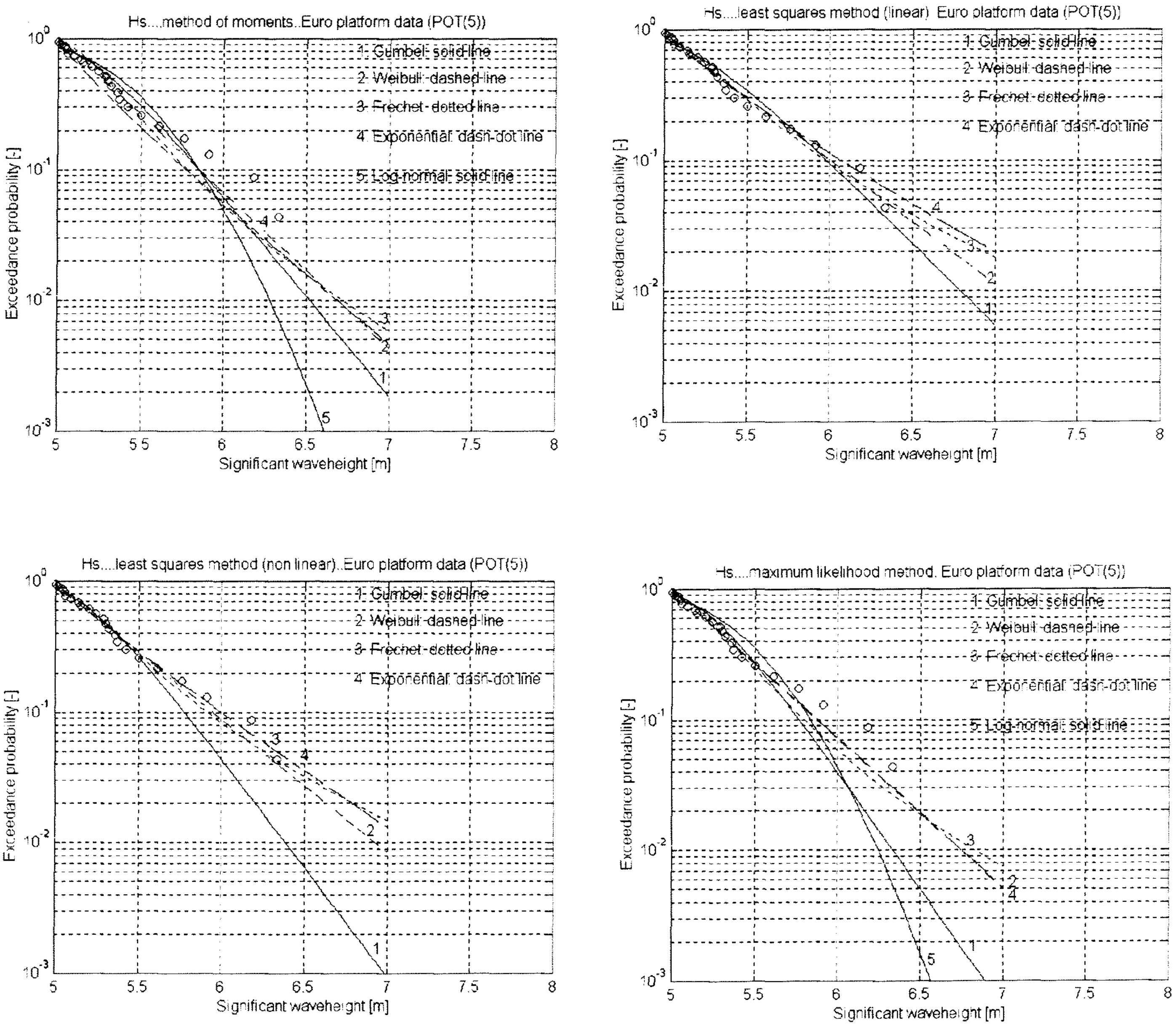


Fig. 8.24 to 8.27 The candidate distributions fitted to the observations of H_s (H_s>5.0 m; 22 points)

On basis of visual inspection of the figures 8.16 to 8.27, the best fitting marginal distributions for H_s have been selected and listed in table 8.10.
As can be seen from the table, the once per 50 years return value of the significant wave height is approximately 7.20 m.

Table 8.10 The best fitting marginal distributions for H_s (North Sea data)

Variable	Treshold level [m]	Parameter Estimation method	Distribution type	Estimated parameter values [-]			Hs(50) [m]
				$\hat{\delta}$	$\hat{\lambda}$	$\hat{\beta}$	
H_s	2.00	MOM	Gumbel	0.614	2.613	-	8.07 (m)
H_s	2.00	Lin LS	Gumbel	0.626	2.608	-	8.03
H_s	2.00	MAX	Gumbel	0.564	2.619	-	7.63
H_s	4.50	LS	Weibull	0.501	4.490	1.030	7.13
H_s	4.50	Non LS	Weibull	0.501	4.490	1.100	6.86
H_s	4.50	LS	Exponential	0.495	4.500	-	7.22
H_s	4.50	MAX	Exponential	0.460	4.500	-	7.04
H_s	5.00	LS	Weibull	0.407	4.990	0.857	7.35
H_s	5.00	Non LS	Weibull	0.406	4.990	0.917	7.10
H_s	5.00	Non LS	Exponential	0.424	4.973	-	6.90

Marginal analysis of T_z

The figures 8.28 to 8.30 show the fits of the candidate distributions to the observations of T_z. The Weibull and the Log- normal provide the best fits to the data. Each of the distributions listed under predicts the largest observation of T_z. Therefore, the once per 50 years return value of the wave period **is underestimated** by the distributions.

Table 8.11 The best fitting marginal distributions for T_s (North Sea data)

Variable	Treshold level [m]	Parameter Estimation method	Distribution type	Estimated parameter values [-]			Tz(50) [s]
				$\hat{\delta}$	$\hat{\lambda}$	$\hat{\beta}$	
T_z	2.00	MAX	Weibull	1.568	4.090	2.170	8.25
T_z	4.50	MAX	Log-normal	0.059	1.913	-	7.92
T_z	5.00	MAX	Weibull	0.462	6.590	1.013	8.64

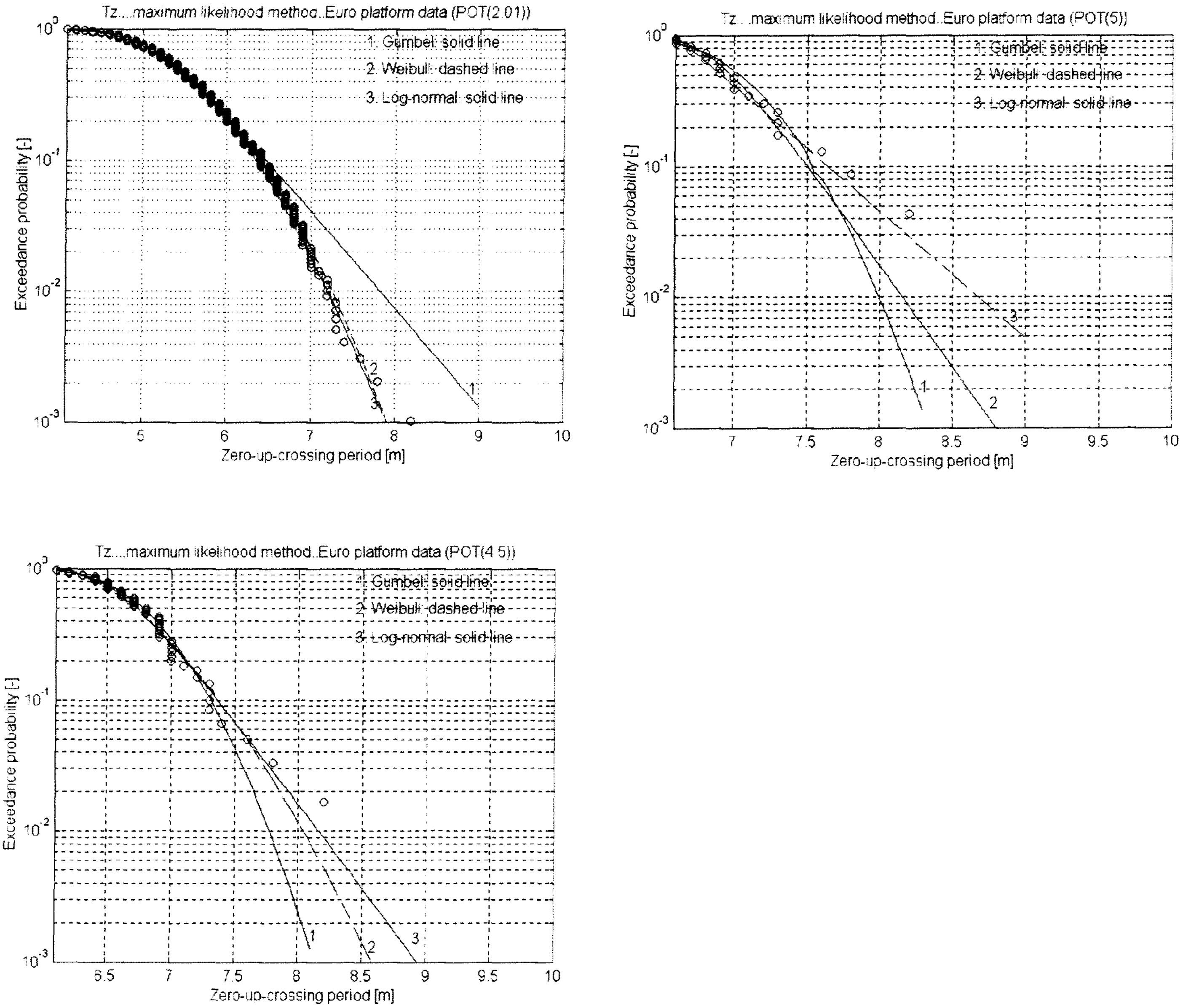


Fig. 8.28 to 8.30 The candidate distributions fitted to the observations of T_z (971p; 59p; 22p)

8.4 Fit of bivariate distribution functions

8.4.1 The data sets of Karwar

Data set 1; The observations of the monsoon wave climate

For the marginal components of the bivariate functions, the best fitting marginals are used (section 8.3.1). The significant wave height is described by a Gumbel, Weibull and Log-normal distribution. The wave peak period is modelled by a Log-normal distribution.

The tested bivariate functions are presented in table 8.12. The estimated values of the parameters of the marginal distributions have already been presented in section 8.3.1. Therefore, only the values of the remaining bivariate parameters are shown. In order to avoid some confusion, these values are only given for the first listed parameter estimation method (the method of moments).
The numbering of the models in the table is equal to the numbering used in chapter 3.

Table 8.12: Tested bivariate models for the Karwar data

Model	Marginal distribution for H_s		Marginal distribution for T_p		The remaining parameters
Nr	Distribution	Parameter estimtaion method	Distribution	Parameter estimation method	
1	Log-normal	MOM/MAX	Log-normal	MOM/MAX	ρ=0.27
3	Gumbel	MOM/MAX/Lin LS/N. Lin. LS	Log-normal	MOM/MAX	regression lines (linear) : scale: a=-0.022 b=0.125 loc. : a=0.090 b=2.234
3	Weibull	MOM/MAX/Lin LS/N. Lin. LS	Log-normal	MOM/MAX	regression lines(linear): scale a=-0.022 b=0.125 loc. a=0.090 b=2.234
4	Gumbel	MOM/MAX/Lin LS/N. Lin. LS	Gumbel (wave steepness d.)	MOM/MAX	par. wavesteep. distr.: δ=0.0017 λ=0.0107
5	Weibull	MOM/MAX/Lin LS/N. Lin. LS	Log-normal	MOM/MAX	dependence parameter: φ=1.66

The bivariate models are illustrated below with figures showing the contour lines of the functions together with lines presenting bivariate return values (section 6.3). The lines presenting return values are called ***quantile lines*** in the following sections. They represent pairs of return values of H_s and T_p. The end point of the vertical part of the quantile line is associated with a wave steepness of s=1.5%. The endpoint of the horizontal part is associated with a wave steepness of s=0.5%.

The goodness of fit of the models is discussed at the end of the section. Theoretical and empirical scatter diagrams are presented for a visual comparison. Further, the chi-square values of the fitted models are listed.

Model 1: The bivariate Log-normal distribution

The correlation between the two wave variables is defined in this model by ρ , the linear correlation coefficient between $\log(H_s)$ and $\log(T_p)$. As can be noticed from fig 8.32, the correlation between the bivariate observations is in this case relative poor (ρ= 0.25).

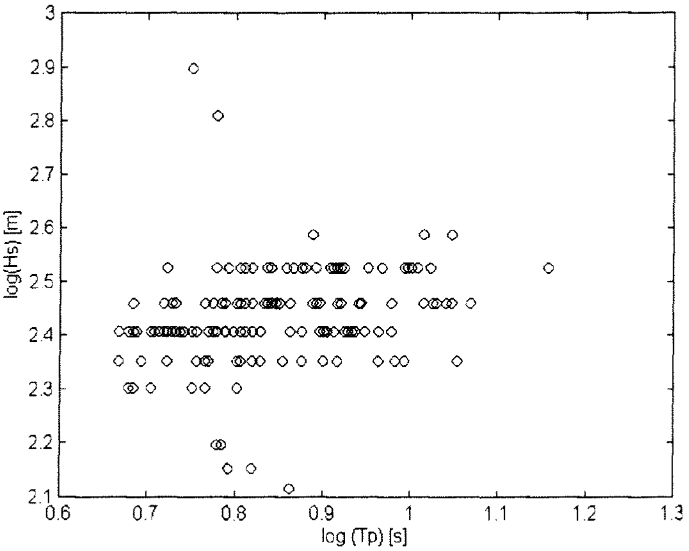


Fig 8.32 Log(T_p) plotted against log(H_s)

Figure 8.33 shows the contour plot of the bivariate Log-normal distribution, together with the above described quantile lines.

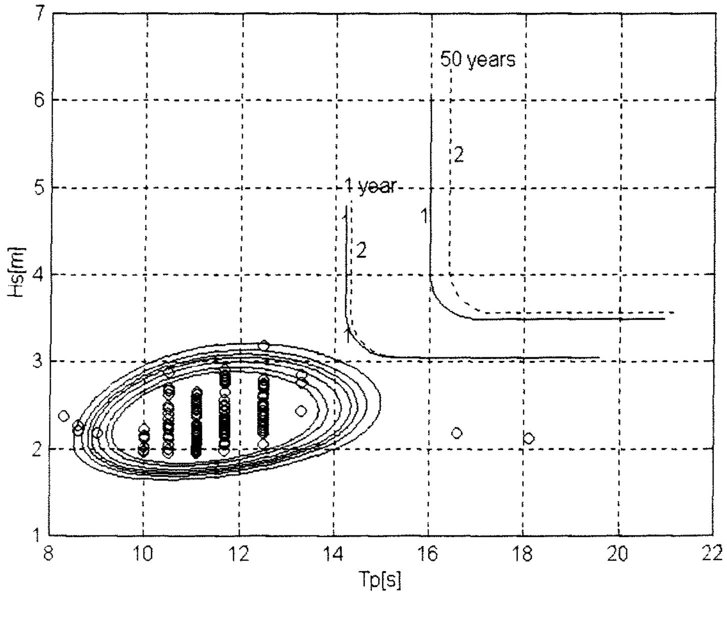


Fig. 8.33 Contour plot of the bivariate Log-normal distribution, together with the quantile lines ((1) MAX (H_s) + MAX (T_p), (2) MOM (H_s) + MOM (T_p)). The observations are denoted by small circles.

Model 2: Fang and Hogben's distribution

This distribution is known as a modified version of the bivariate Log-normal distribution. Application of the model showed however that in this case no improvements were noticed: the shape of both models and the computed quantile lines were almost identical. Therefore, no further illustrations of the model are given here.

Model 3: bivariate model based on a marginal distribution for H_s and a conditional distribution for T_p

The marginal distribution of H_s has been modelled by the Gumbel and the Weibull distribution. The conditional distribution of T_p has been described by the Log-normal distribution. The bivariate models are shown in figure 8.34 and 8.35, respectively.

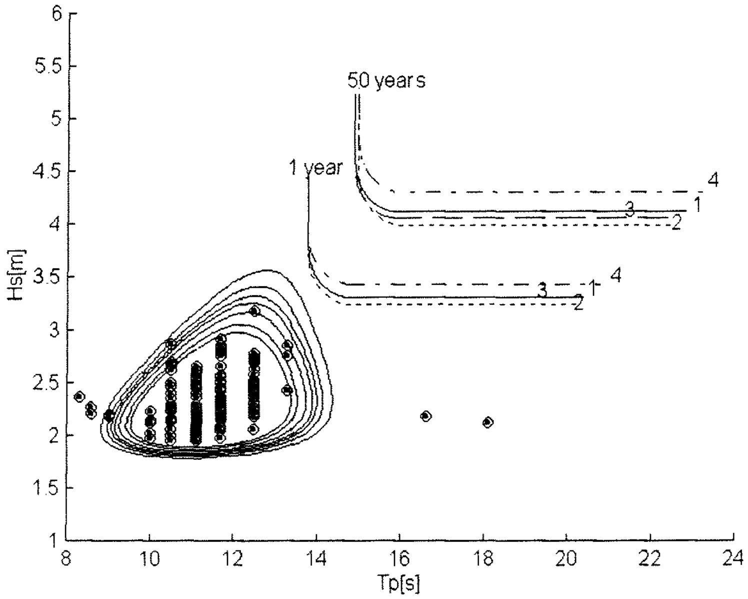


Fig 8.34 Model 3 : Gumbel (H_s) and Log-normal (T_p). (quantile lines : (1) MAX (H_s) + MAX (T_p), (2) MOM (H_s) + MOM (T_p), (3) Lin LS (H_s) + MAX (T_p), (4) Non Lin LS (H_s) + MAX (T_p))

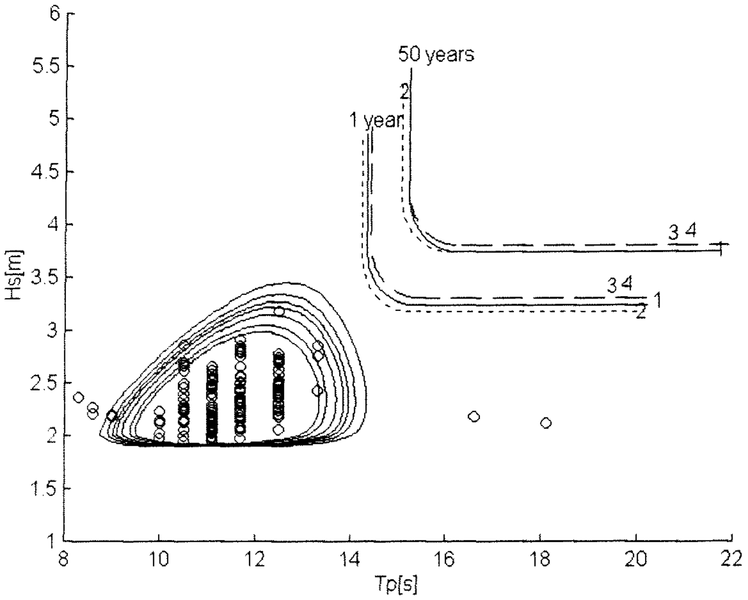


Fig 8.35 Model 3 : Gumbel (H_s) and Log-normal (T_p). (quantile lines : (1) MAX (H_s) + MAX (T_p), (2) MOM (H_s) + MOM (T_p), (3) Lin LS (H_s) + MAX (T_p), (4) Non Lin LS (H_s) + MAX (T_p))

The parameters of the conditional distribution are defined as a function of H_s. Empirical regression lines are used to describe these relationships. The models shown in figure 8.34 and 8.35 contain for both parameters of the Log-normal distribution a linear regression function. As can be seen from figure 8.36 and 8.37, the linear regression lines provide the best fit to the points. Note that the estimator for the scale parameter of H_s-class 3, denoted by an arrow in the figure, differs significantly from the other estimates.

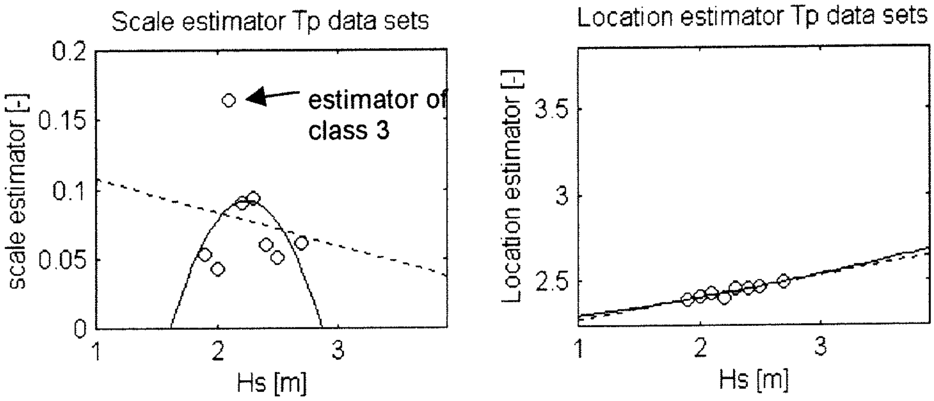


Fig 8.36 and 8.37 Estimators of the scale and the location parameter of the conditional distribution function. The estimators correspond to chosen classes of H_s (see section 3.4.). Used distributions: Gumbel distribution (H_s) and Log-normal distribution (T_p). Shown regression lines: linear (dashed line; eq. 3.21(1)) and parabolic (solid line; eq. 3.21 (2)). The lines defined by eq. 3.21(3) and eq. 3.23 are not shown because they poorly fit the data.

The slope of the linear regression line representing the scale parameter is negative. Apparently, the width of the conditional probability density function of T_p decreases with an increasing H_s

As mentioned in section 3.4.3, the relationship between the significant wave height and the peak period is described in the bivariate model by the location parameter of the conditional distribution. The conditional distribution is the Log-normal distribution. Therefore, the relation between H_s and T_z is described by the following exponential (!) function

(8.1) $$g(H_s) = \exp(aH_s + b)$$

in which a and b are some constant values.

The regression lines are based on parameter estimates corresponding to chosen classes of H_s (See section 3.4.3). The choice of the class width is subjective. Smaller class widths of H_s imply more parameter values, which yields generally to a more accurate estimation of the regression lines. On the other hand, smaller class widths means that less data points are present in each class, which yields to less reliable estimates of the parameters itself. The optimal choice of the class width might be determined by using a goodness of fit criteria: for example the bivariate model which contains the smallest chi-square value could be considered as the model which is computed with the best class width of H_s.

The sensitivity of the return values to the choice of the class widths of H_s (ΔH_s) has been studied, using linear regression lines for the parameter functions. Table 8.13 shows for the class widths ΔH_s=0.10, 0.25, 0,50 and 0.75 the once per 1 and 50 year(s) return period value corresponding to the wave steepness values 0.5,1 and 1.5 %. For the test, the model with the Gumbel (H_s) and the Log-normal (T_p) distribution has been used. The parameter estimation method used is the maximum likelihood method.

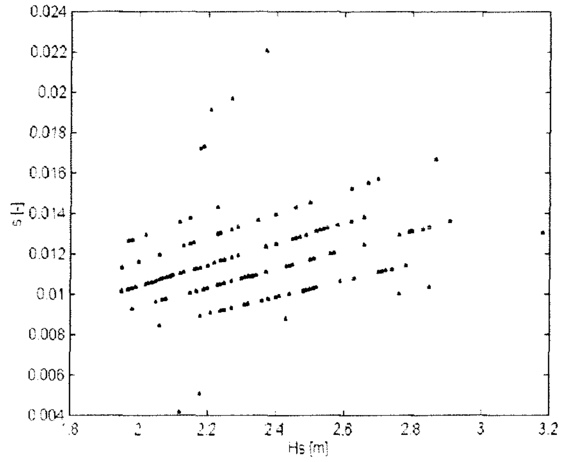
Table 8.13 Return values for various class widths of H_s

ΔH_s [m]	$(H_s(1),T_p(1))$ [m,s]			$(H_s(50),T_p(50))$ [m,s]			Chi-square value [-]
	s=0.5 %	s=1 %	s=1.5%	s=0.5%	s=1.0%	s=1.5%	
0.10	(3.31;20.38)	(3.31;14.41)	(4.68;14.04)	(4.12;22.79)	(4.11;16.12)	(5.56;15.31)	187.294
0.25	(3.31;20.38)	(3.31;14.41)	(4.49;13.75)	(4.12;22.79)	(4.11;16.12)	(5.24;14.87)	205.294
0.50	(3.31;20.38)	(3.36;14.55)	(4.75;14.14)	(4.12;22.79)	(4.11;16.12)	(5.68;15.48)	247.774
0.75	(3.11;19.48)	(3.24;14.27)	(4.90;14.36)	(4.32;14.79)	(3.99;15.86)	(5.34;14.62)	343.949

It is seen from the table that the computed return values are not very sensitive to the variation of the class width of H_s. In the most right column of the table, the corresponding chi-square values are listed. According to this statistic, the model, which is estimated with a class width of 0.10 m, provides the best fit to the observations. Note that the accuracy of the model decreases with an increasing class width.

Model 4: Bivariate distribution based on the marginal distribution of H_s and s

This model is based on the assumption that the observed values of the wave steepness (s) and the significant wave height (H_s) are not correlated. As shown in fig 8.38, for this set of data, the correlation between the variables H_s and s is weak.



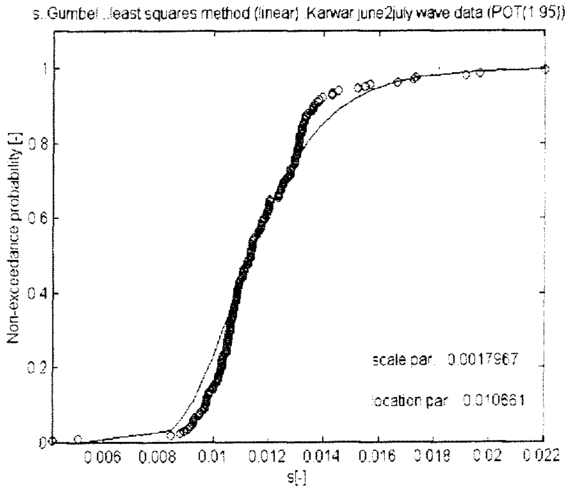


Figure 8.38 H_s against s

Fig. 8.39 Linear least squares fit of the Gumbel distribution to the wave steepness data

Several distributions have been fitted to the wave steepness data. The Gumbel distribution appeared to be the best alternative. However, as seen in fig 8.39 the distribution does not provide a very close fit to the data.

The tested bivariate model consists of two marginal Gumbel distributions (figure 8.40).

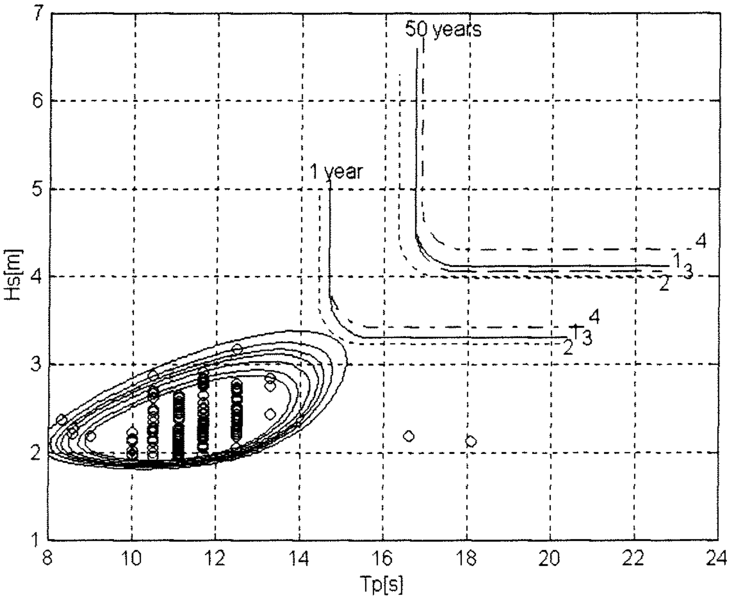


Fig 8.40 Gumbel (H_s) ..Gumbel(s). (quantile lines : (1) MAX (H_s) + MAX (T_p), (2) MOM (H_s) + MOM (T_p), (3) Lin LS (H_s) + Lin LS (T_p), (4) Non lin LS (H_s) + Non lin LS (T_p))

Model 5: Bivariate model with given marginals

During the calculations of this Frechet class model, the estimated marginal distributions are transformed twice. These transformations are illustrated in figure 8.41.

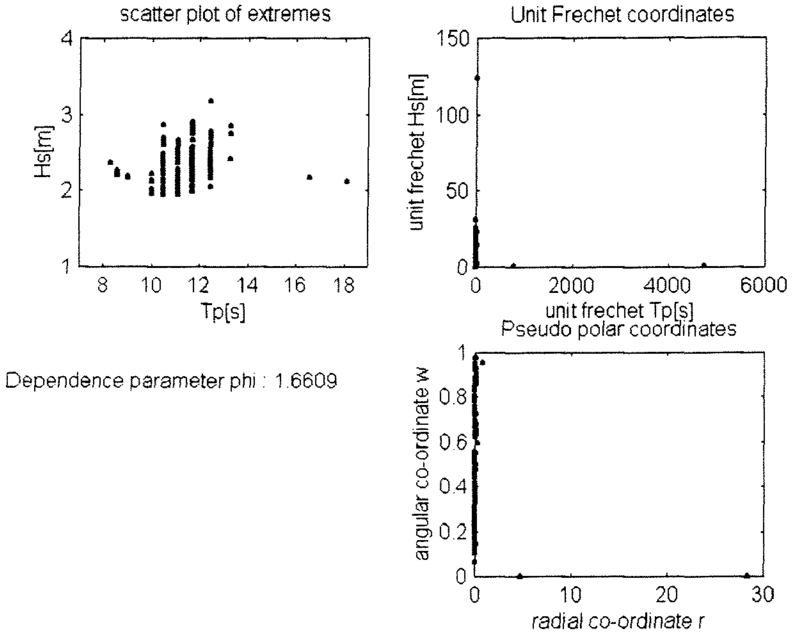


Fig 8.41 Transformation procedures of model 5

The first plot shows the observations before any transformation. The second plot illustrates the process with the margins transformed to unit Frechet. In the third plot, the observations are shown transformed to their pseudo polar equivalents. (See appendix [3]).

According to Morton and Bowers (1997), the first transformation helps to accentuate the extreme events: the less extreme events are scaled to the axes while the extreme observations become more inhomogeneous. The final model only contains this first transformation. The including of the transformation in the model probably implies that the bivariate function is especially determined by the extreme part of the observations. However, as can be seen from the figures, it seems that all observations are scaled to the axis : "the more inhomogeneous extremes" are not observed.

The second transformation is used for calculating the dependence parameter φ: the maximum likelihood estimation of this parameter is based on the presented pseudo co-ordinates.

The analytical background of the model is not very clear. Especially, the transformation of the observations into pseudo polar co-ordinates needed for the calculation of the dependence parameter φ is not very obvious. However, according to Metcalfe (1997), the model is used in civil engineering practice for several purposes. Perhaps a more detailed literature study (Johnson 1987, Mardia 1970) could lead to a better understanding of the model.

Figure 8.42 shows the applied model, which consists of a marginal Gumbel (H_s) and a marginal Log-normal (T_p) distribution.

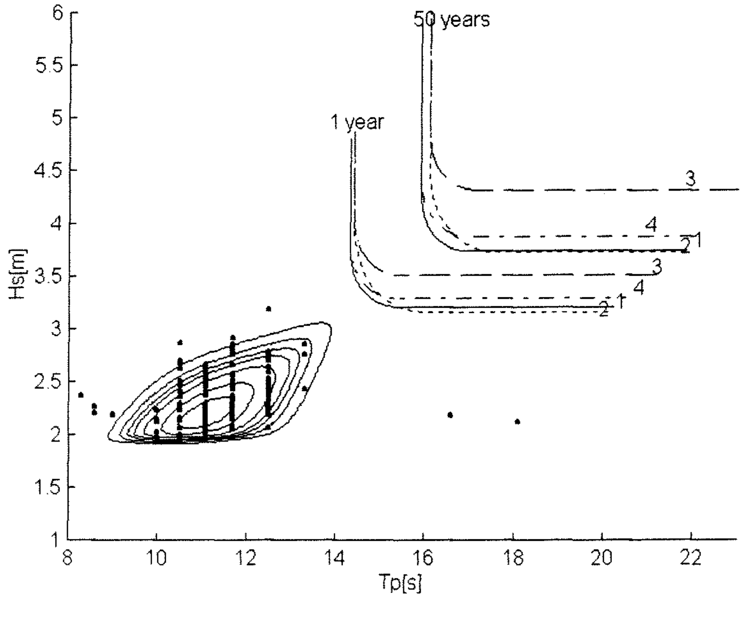


Fig 8.42 Model 5 : Gumbel (H_s) and Log-normal(T_p). (quantile lines : (1) MAX (H_s) + MAX (T_p), (2) MOM (H_s) + MOM (T_p), (3) Lin LS (H_s) + Lin Ls (T_p), (4) Non Lin LS (H_s) + Non Lin LS (T_p))

Goodness of fit

For a qualitative assessment of the goodness of fit of the bivariate models, the chi-square test is used. The computed values are listed in table 8.14. Observing the table, it is seen that the differences between the chi-square values are relative small. According to the statistic, model 4 consisting of two Gumbel marginals is the best option (178).

Table 8.14 Chi-square values of the tested models

Model	Marginal distribution for H_s		Marginal distribution for T_p		Chi-square value
Nr	Distribution	Est. m.	Distribution	Est m.	
1	Log-normal	MAX	Log-normal	MAX	218.27
1	Log-normal	MOM	Log-normal	MOM	205.14
3	Gumbel	MOM	Log-normal	MOM	201.50
3	Gumbel	MAX	Log-normal	MAX	193.07
3	Gumbel	Lin LS	Log-normal	MAX	201.35
3	Gumbel	N. lin LS	Log-normal	MAX	204.00
3	Weibull	MOM	Log-normal	MOM	204.58
3	Weibull	MAX	Log-normal	MAX	201.88
3	Weibull	Lin LS	Log-normal	MAX	212.48
3	Weibull	N. lin LS	Log-normal	MAX	198.96
4	Gumbel	MOM	Gumbel (s)	MOM	201.69
4	Gumbel	MAX	Gumbel (s)	MAX	199.97
4	Gumbel	Lin LS	Gumbel (s)	Lin LS	200.38
4	Gumbel	N. lin LS	Gumbel (s)	N. lin LS	178.34
5	Weibull	MOM	Log-normal	MOM	265.35
5	Weibull	MAX	Log-normal	MAX	210.04
5	Weibull	Lin LS	Log-normal	MAX	180.23
5	Weibull	N. lin LS	Log-normal	MAX	190.14

Further, a theoretical scatter diagram is computed for each bivariate model. These are presented in appendix [7]. They can be used to judge visually the fit of the models by comparing the diagrams with the empirical scatter diagram shown in figure 8.43.

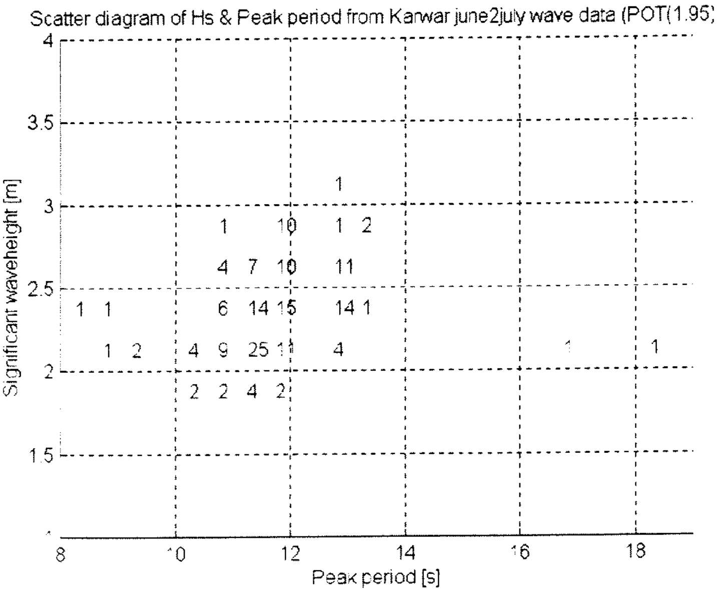


Fig. 8.43 Empirical scatter diagram based on the data set of Karwar (ΔH_s= 0.25 m, ΔT_p = 0.50 s)

Return values

In the preceding figures quantile lines are shown (See section 6.2 ; fig 6.1). These lines are hard to compare with each other. An alternative is to show the return values of H_s and T_p for only one value of the wave steepness. In the table below the values of (H_s(50),T_p(50)) computed by the best fitting models (see table 8.15) are shown for s=1 %. Note that the differences between the values are small.

Table 8.15 (H_s(50),T_p(50)) corresponding to s=1%

Model	Marginal distribution for H_s		Marginal distribution for T_p		(H_s(50).T_p(50)) [m,s]
Nr	Distribution	Est. m.	Distribution	Est m.	
1	Log-normal	MOM	Log-normal	MOM	(4.06,15.99)
3	Gumbel	MAX	Log-normal	MAX	(4.20,16.40)
4	Gumbel	N. lin LS	Gumbel (s)	N. lin LS	(4.41,16.80)
5	Weibull	Lin LS	Log-normal	MAX	(4.30,16.59)

Discussion

Evaluating both goodness of fit criteria, it is clear that the models poorly fit the data. This is probably due to the small correlation between the wave height and wave period observations.

Observing the theoretical scatter diagrams of appendix 7, it is seen that model 3 and 5 correspond mostly to the shape of the empirical scatter plot.

With respect to estimated return values, it appears that the bivariate models tend to follow their marginals. The horizontal and vertical part of the quantile lines corresponds to the marginal return values of H_s and T_p, respectively.

Data set 2: The hurricane observations

The best fitting marginals are used for the bivariate models. These are the Gumbel distribution for H_s and the Log-normal distribution for T_p. (Section 8.3.1). The tested bivariate functions are listed in table 8.16 and shown in the figures below.

The fit of bivariate model 2, the Fang and Hogben distribution, is not shown. Again it appeared that this modified version of the bivariate Log-normal distribution was almost identical to the "normal" version (bivariate model 1). Also bivariate model 3, the model with a conditional distribution for T_p, is not used. The data set (25 points) was far too small to obtain a reasonable fit.

The goodness of fit should be analyzed on basis of visual comparison of the contour plots of the models. Because of the small number of data points, the proposed methods based on the scatter diagram of the observations (section 5.3) can not be applied.

Table 8.16: Tested bivariate models for the Hurricane observations of Karwar

Figure	Model	Marginal distribution for H_s		Marginal distribution for T_p		The remaining parameters
	Nr	Distribution	Parameter estimtaion method	Distribution	Parameter estimation method	
8.44	1	Log-normal	MAX	Log-normal	MAX	ρ=0.82
8.45	4	Exponential	Lin. LS	Normal (wave steepness d.)	MAX	par. wavesteep. distr.: δ=0.0083 λ=0.028
8.48	5	Gumbel	Lin. LS	Log-normal	MAX	dependence parameter: φ=1.87

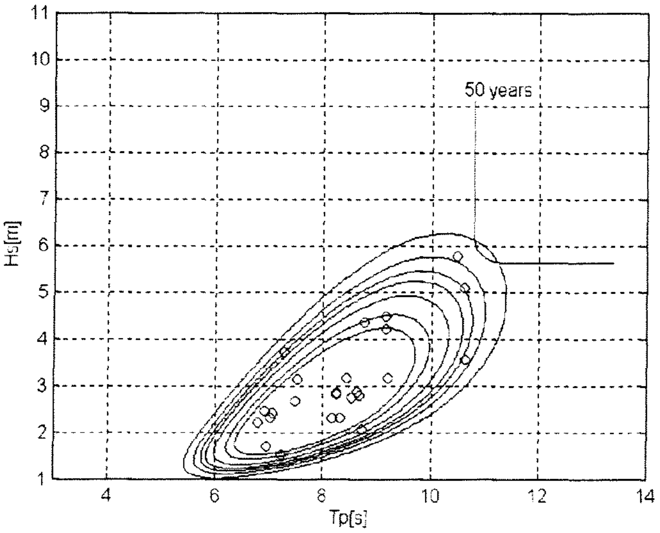


Fig 8.44 Bivariate Lognormal distribution

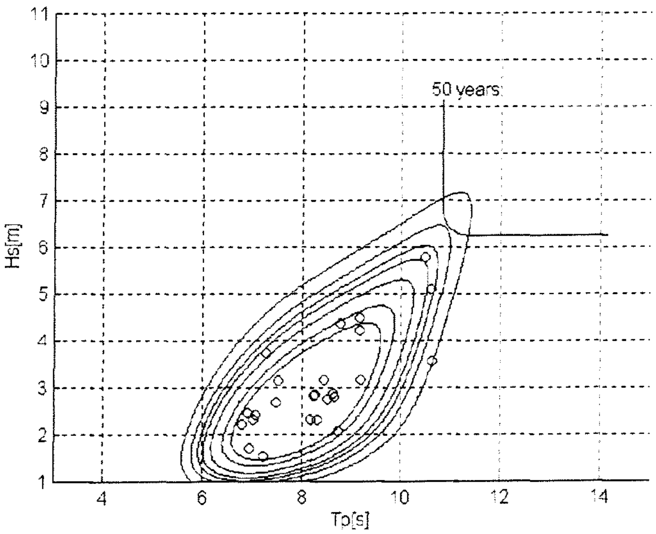


Fig 8.45 Bivariate model 5 consisting of a Gumbel distribution for H_s and a Log-normal distribution for T_p

In bivariate model 4, the wave steepness s is modelled by the Normal distribution:

$$(8.1) \quad f(x)=\frac{1}{\sqrt{2\pi\delta}}\exp\left[-\frac{1}{2}\left(\frac{\log(x)-\lambda}{\delta}\right)^2\right]$$

Fig 8.46 shows the distribution of the deep water wave steepness (s) during the hurricane. In fig 8.47 the data points of s are plotted against the data of H_s. The assumption that the observations of s are independent of the observations of H_s seems to be reasonable. In figure 8.48 the contour plot of model 4 consisting of a Gumbel distribution for H_s and a Normal distribution for s is shown.

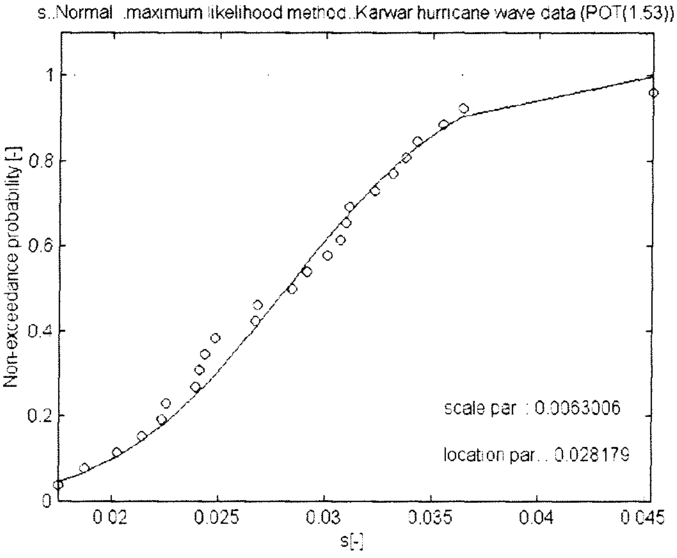


Fig 8.46 The distribution of the deep water steepness

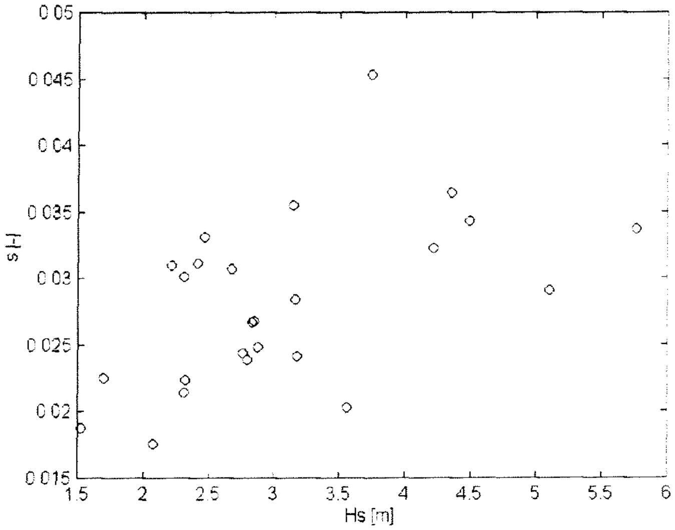
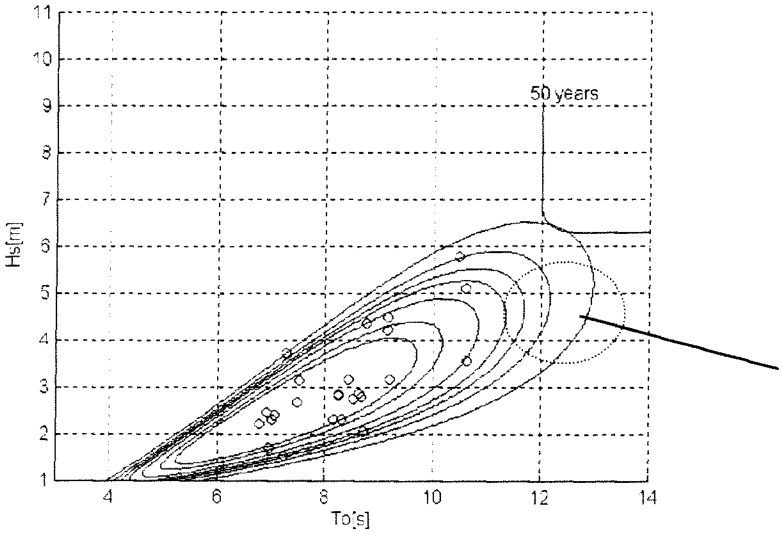


Fig 8.47 The wave steepness observations during the hurricane plotted against the significant wave height observations (25 points)



the model over predicts the extremes in this area

Fig 8.48 Contour plot of model 4 consisting of a Gumbel distribution for H_s and a Log-normal distribution for s

Return values

The values of (H_s(50),T_p(50)) for s=3% are shown in table 8.17.

Table 8.17 (H_s(50),T_p(50)) corresponding to s=3%

Figure	Model	Marginal distribution for H_s		Marginal distribution for T_p		(H_s(50),T_p(50))
	Nr	Distribution	Parameter estimation method	Distribution	Parameter estimation method	
8.26	1	Log-normal	MAX	Log-normal	MAX	(5.70,11.03)
8.30	4	Exponential	Lin. LS	Nórmal (wave steepness d.)	MAX	(6.20,11.51)
8.27	5	Gumbel	Lin. LS	Log-normal	MAX	(6.74,12.00)

Discussion

The bivariate Log-normal distribution underpredicts the extreme part of the observations. This is to be expected since the marginal distribution of H_s also poorly fits the extremes. Further, it is seen that the contour plots of the three models differ significantly.

The joint probability density function of model 4 overpredicts sea states with relative high wave periods as illustrated in figure 8.49. This is due to the fact that the scale parameter of the conditional probability density function of T_p corresponding to this model is constant. (Fig. 8.49).

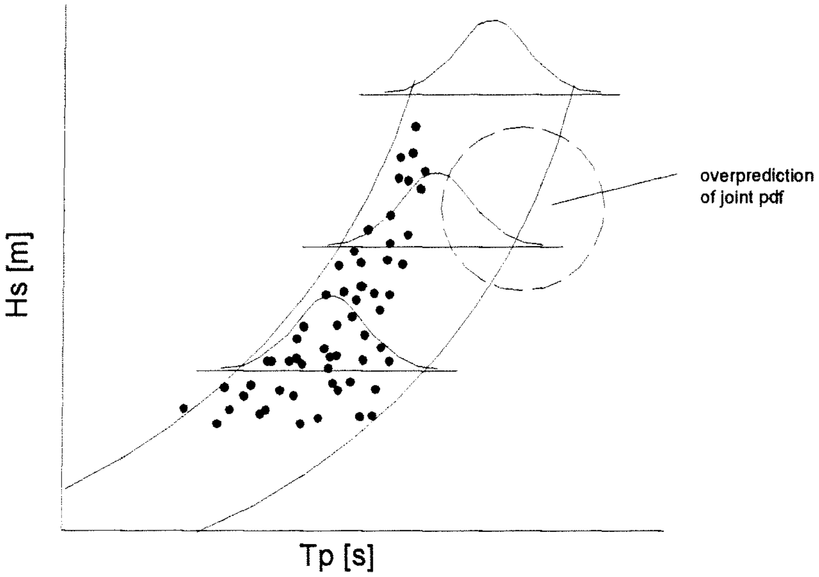


Fig 8.49. The relation between the significant wave height and the peak period. In the figure, the conditional probability density function of H_s corresponding to model 4 has been given for a number of values of H_s.. The location parameter of the conditional function is described as a parabolic curve similar to the wave steepness equation (e.q. 3.10). The scale parameter of the function is constant. This leads to an overestimation of sea states with relative high peak periods (illustration)

8.4.2 The North Sea data set

The tested models are listed in table 8.18. The best fitting marginals (table 8.8) are used for the bivariate functions. The applied parameter estimation methods are the ones that provided the best results in the marginal case.

Table 8.18 Tested bivariate models for the North Sea data

Treshold level [m]	Model	Marginal distribution for H_s		Marginal distribution for T_p	
	Nr	Distribution	Parameter estimation methods	Distribution	Parameter estimation methods.
2.00	1	Log-normal	MAX	Log-normal	MAX
2.00	3	Gumbel	MOM/MAX/Lin LS	Log-normal	MAX
2.00	4	Gumbel	MOM/MAX/Lin LS	Weibull(s)	Lin LS
2.00	5	Gumbel	MOM/MAX/Lin LS	Log-normal	MAX
2.00	5	Gumbel	MOM/MAX/Lin LS	Weibull	MAX
4.50	1	Log-normal	MAX	Log-normal	MAX
4.50	3	Weibull	Lin LS/N. Lin. LS	Log-normal	MAX
4.50	3	Exponential	MAX/Lin LS	Log-normal	MAX
4.50	4	Weibull	Lin LS/N. Lin. LS	Weibull	Lin LS
4.50	4	Weibull	Lin LS/N. Lin. LS	Weibull	Lin LS
4.50	5	Weibull	Lin LS/N. Lin. LS	Log-normal	MAX
5.00	1	Log-normal	MAX	Log-normal	MAX
5.00	4	Weibull	MOM/MAX/Lin LS/N. Lin. LS	Weibull	Lin LS
5.00	4	Weibull	MOM/MAX/Lin LS/N. Lin. LS	Weibull	Lin LS
5.00	5	Weibull	MOM/MAX/Lin LS/N. Lin. LS	Weibull	MAX

The models are illustrated with figures containing a contour plot and a quantile line representing the once per 50 years return period values of H_s and T_z. The endpoint of the vertical part of the quantile line is associated with a wave steepness of s=9%. The endpoint of the horizontal part corresponds to a wave steepness of s=3%.

The quantile lines have not been presented for the first set of data. The preceding marginal analyses have indicated that the extreme observations of H_s are largely overpredicted by the marginal distributions. Since the marginals of H_s are substituted in the bivariate models, it is clear that the joint probability density functions will also overpredict the extremes.

Data set 1: The observations above H_s = 2.00 m

Model 1: the bivariate Log-normal distribution
The contour plot of the bivariate model is shown in figure 8.50.

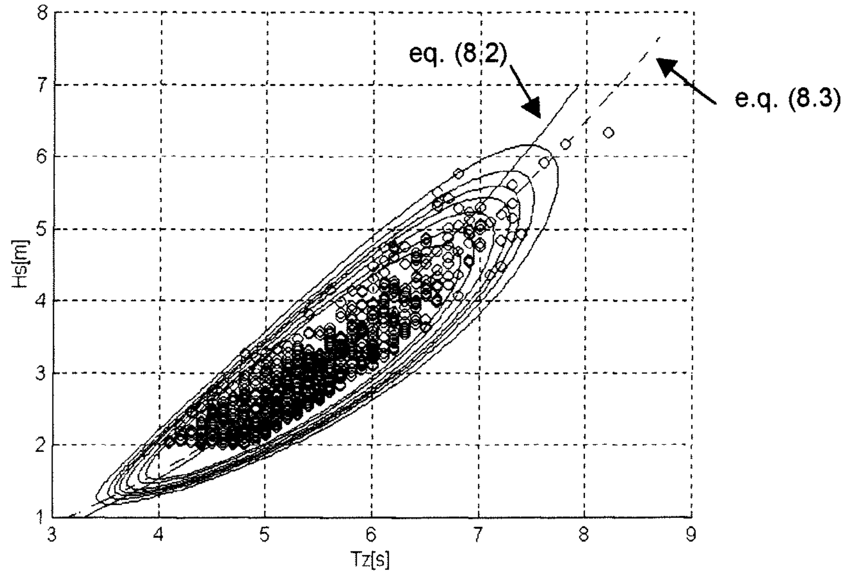


Fig 8.50 Contour plot of the bivariate Log-normal distribution. (H_s>2.00 m; observations of wind waves)

In section 3.4.2, it has been shown that the model relates the significant wave height and the zero-up-crossing period by the following complex formula

$$(8.2) \quad g(H_s) = E(T_z \mid H_s) = \exp\left[\lambda_{T_z} + \rho\left(\frac{\delta_{H_s}}{\delta_{T_z}}\right)\left(\log H_s - \lambda_{H_s}\right)\right]$$

On basis of physical consideration, one could suggest to describe the relation with a parabolic function similar to the formula of the deep water wave steepness (e.q. 3.10). In that case, the relation follows as

$$(8.3) \quad g(H_s) = a(H_s)^{1/2}$$

in which

a = constant.

By comparing both curves, one gets an impression of the relation between the physical proces and the purely statistical model. In figure 8.50, the parabolic curve $s_{50\%}$ representing the value of the wave steepness that is not reached by 50% of the waves (The wave steepness is described by a Weibull distribution. See bivariate model 4. An example of the curve $s_{50\%}$ is given on page 3-13) is plotted together with the exponential function (e.q. 8.2). As can be seen from the figure, for lower values of H_s and T_z the lines are almost identical. However, in the extreme part of the plot the difference between the curves of the functions increases.

Model 3: The model based on a marginal distribution for H_s and a conditional function for T_z
In this bivariate model, the marginal distribution of H_s is modelled by a Gumbel distribution and the conditional distribution of T_z is described by a Log-normal distribution.

The parameters of the conditional distribution are defined as functions of Hs. Empirical regression functions are used to describe these relationships. In the left part of figure 8.51, the estimation of the scale parameter is shown. As can be noticed from the figure, it is far from easy to obtain a reasonable fit. The linear function seems to be the best option. The value of the scale parameter decreases with an increasing wave height. This agrees with the fact that in general the range of periods narrows at the higher values of H_s.

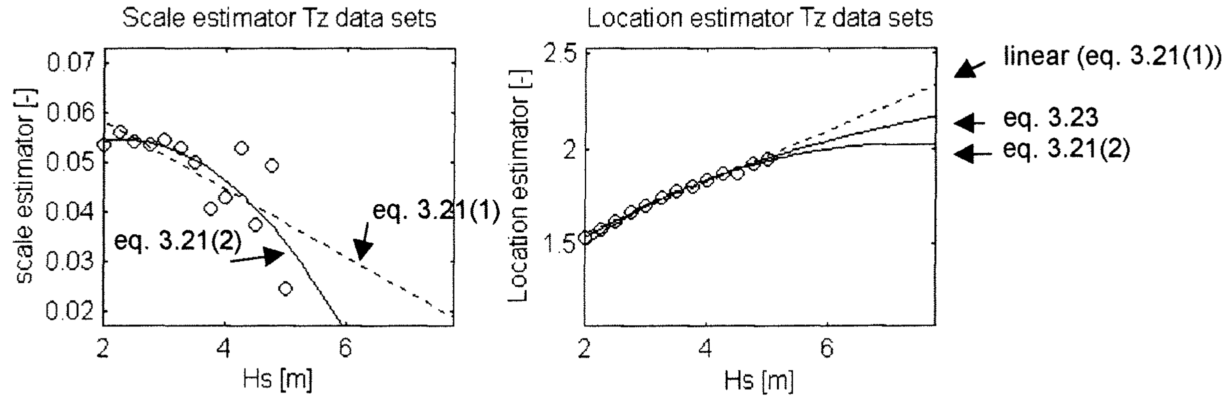


Fig 8.51 Estimators of the scale and the location parameter of the conditional distribution function. The estimators correspond to chosen classes of H_s (see section 3.4.). Used distributions: Gumbel distribution (H_s) and Log-normal distribution (T_p). Shown regression lines: linear (dashed line; eq. 3.21(1)) and parabolic (solid line; eq. 3.21 (2)). For the location parameter, the parabolic function defined by eq. 3.23 is also shown.

In the right part of the figure, the estimation procedure for the location parameter is illustrated. In the figure, three regression functions are shown: the linear and the parabolic function of eq. (3.21), which are fully empirical, and the parabolic function that is similar with the formula of the deep water wave steepness (eq. 3.23). On basis of visual inspection, the conclusion can be drawn that all functions fit the data equally well. On basis of physical grounds, however, it is suggested to prefer the latter parabolic function (eq. 3.23).

In figure 8.52 the bivariate model is compared with model 4, which is based on the marginal distributions of the significant wave height and the wave steepness. As can be seen, for model 3 the width of the conditional probability function (pdf) of T_z decreases with increasing H_s. For model 4 the width of the conditional pdf is constant. The two models become similar when in the case of model 3 the scale parameter of the conditional pdf of T_z is constant and the location parameter is described by the parabolic function of eq. 3.23.

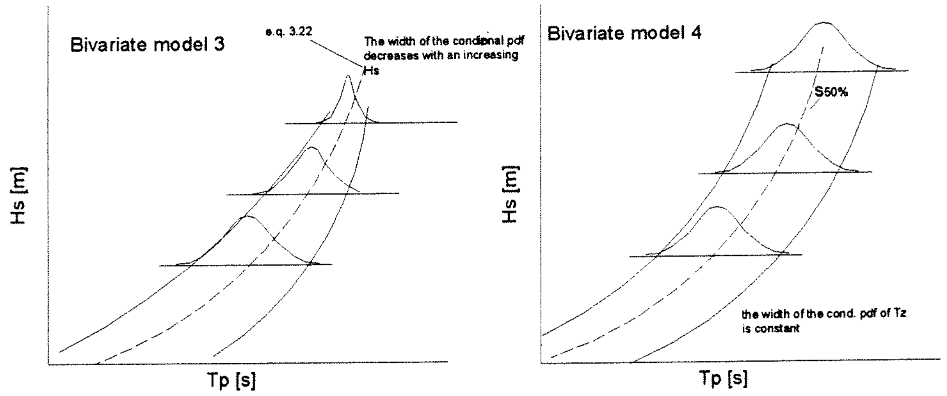

Bivariate model 3
e.q. 3.22
The width of the condional pdf decreases with an increasing Hs
Hs [m]
Tp [s]
Bivariate model 4
S50%
the width of the cond. pdf of Tz is constant
Hs [m]
Tp [s]

Fig 8.52 Comparison between model 3 and model 4. For both models, the conditional probability density function of T_Z for a number of significant wave height levels has been given

In figure 8.53, the contour plot of bivariate model 3 is shown.

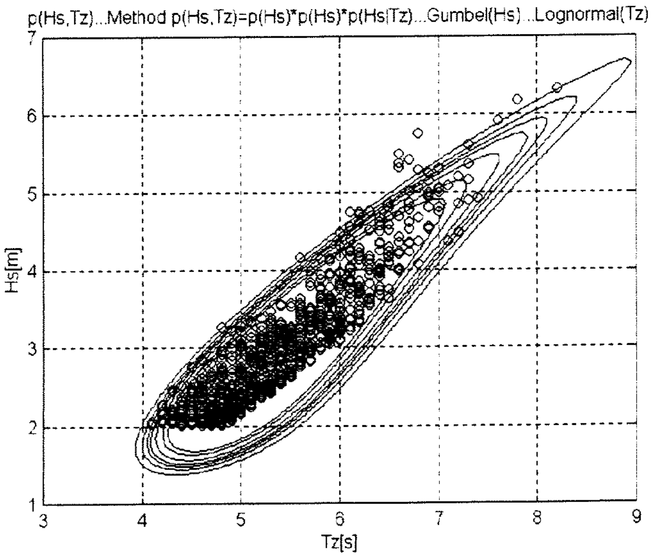


Fig. 8.53 Contour plot of bivariate model 3 consisting of a Gumbel (H_S) and a Log-normal (T_Z) distribution (Regression lines for parameters of conditional function both linear). (H_s>2.00 m; wind waves)

Model 4: the model based on the marginal distribution of the significant wave height and the wave steepness

In figure 8.54 the observations of H_s are plotted against the observations of s. According to the figure, the correlation between the data points is weak. In figure 8.55 the Weibull distribution is fitted to the observations of s. In figure 8.56, the contour plot of the bivariate model is shown consisting of a Gumbel distribution for H_s and a weibull distribution for s.

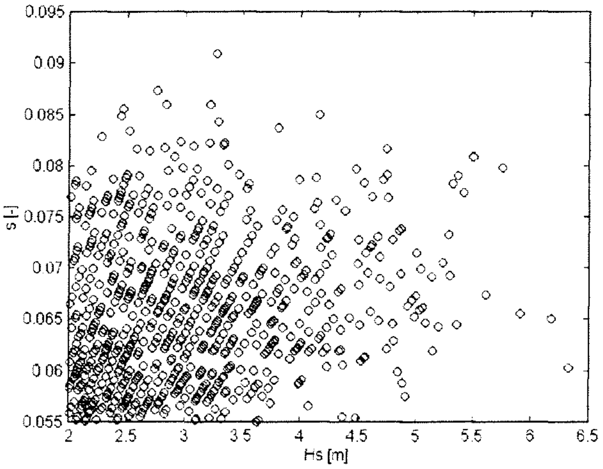


Fig 8.54 Observations of H_s plotted against s

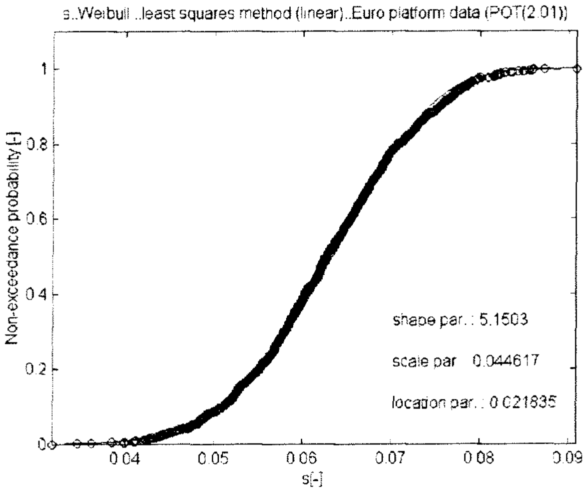


Fig 8.55 Weibull distribution fitted to observations of s

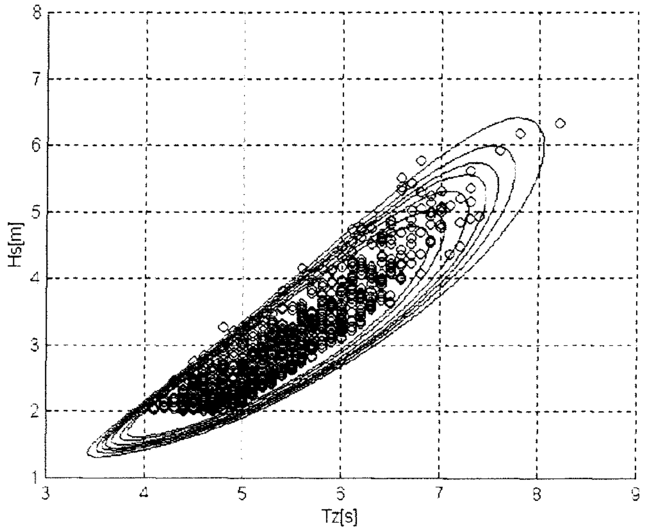


Fig 8.56 Contour plot of bivariate model 4 consisting of Gumbel cdf for H_s and Weibull cdf for T_z (H_s>2.00 m)

Model 5: The bivariate model with given marginals

The bivariate model tested is composed form a Gumbel distribution for H_s and Log-normal distribution for T_z. The contour plot of the model is shown 8.57.

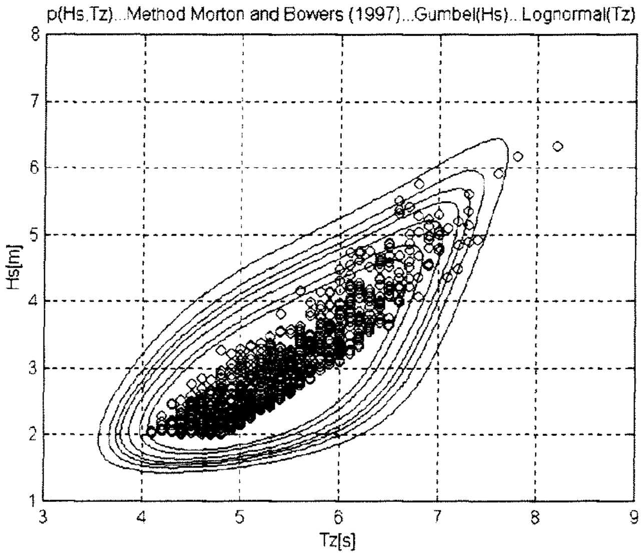


Fig 8.57 Contour plot of bivariate model 5 consisting of Gumbel cdf for H_s and Log-normal distribution for T_z

Data set 2: The observations above H_s = 4.50 m (59 points)

The fit of the models is illustrated in the figures below.

- Fig. 8.58 : Contour plot of the bivariate Log-normal distribution.
(Shown quantile line : (1) MAX (H_s) + MAX (T_z))

- Fig. 8.59 : Contour plot of model 3: Weibull (H_s) + Log-normal (T_z).
(Shown quantile lines : (1) Lin LS (H_s) + MAX (T_z), (2) Non lin LS (H_s) + MAX (T_z))

- Fig. 8.60 : The wave steepness observations plotted against the significant wave heights

- Fig. 8.61 : The Weibull distribution fitted to the wave steepness data

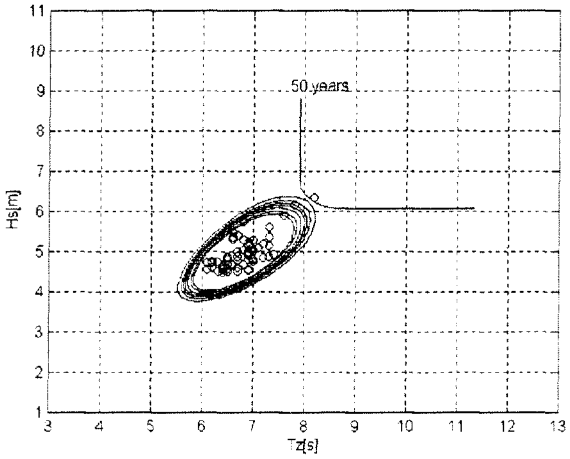


Fig 8.58

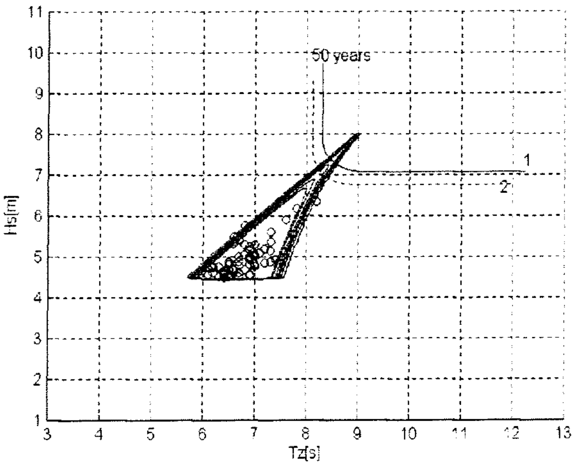


Fig 8.59

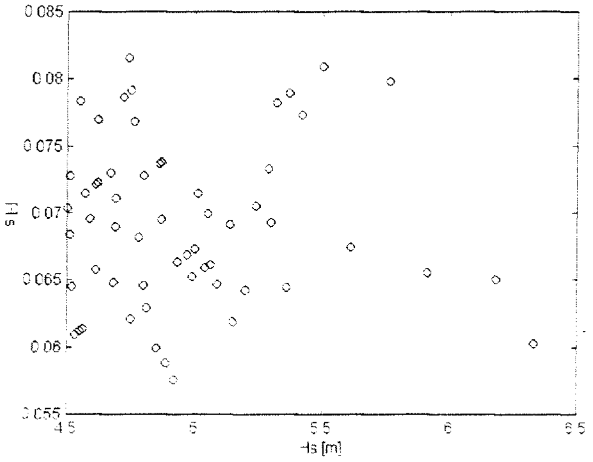


Fig 8.60

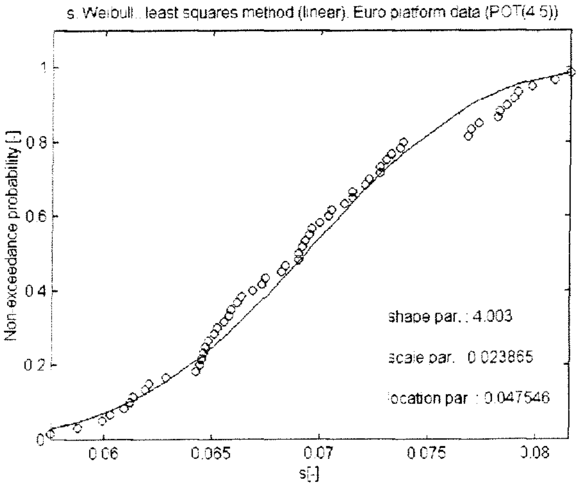


Fig 8.61

- Fig. 8.62 : Contour plot of model 4: Weibull (H_s) + Weibull (s).
  (Shown quantile lines : (1) Lin LS (H_s) + MAX (T_z), (2) Non lin LS (H_s) + MAX (T_z))

- Fig. 8.63 : Contour plot of model 5: Weibull (H_s) + Log-normal (T_z).
  (Shown quantile lines : (1) Lin LS (H_s) + MAX (T_z), (2) Non lin LS (H_s) + MAX (T_z))

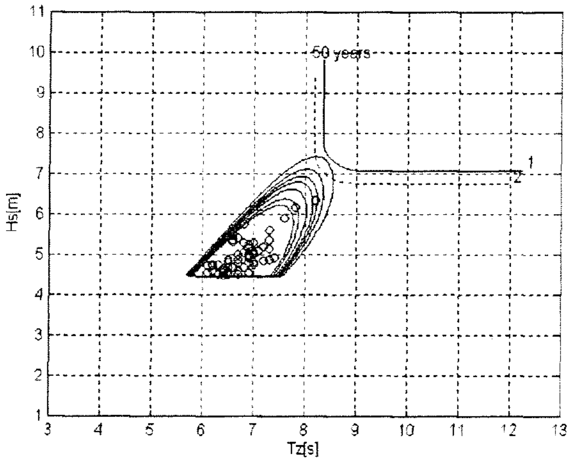


Fig 8.62

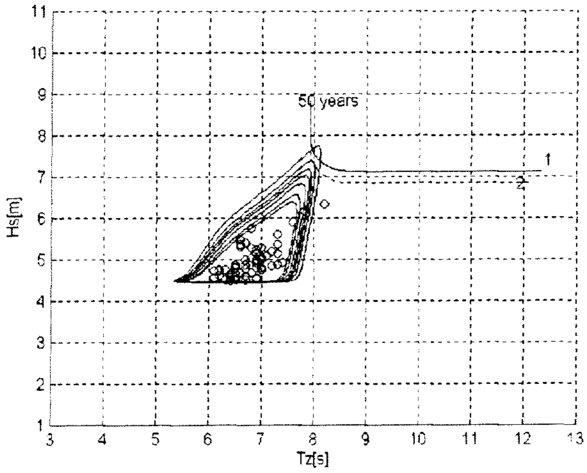


Fig 8.63

Data set 2: The observations above H_s = 5.00 m (22 points)

The fit of the models is illustrated in the figures below.

- Fig 8.64 : Contour plot of the bivariate Log-normal distribution
  (Shown quantile lines : (1) MAX (H_s) + MAX (T_z))
- Fig. 8.65 : The wave steepness observations plotted against the significant wave height observations
- Fig 8.66 : The Weibull distribution fitted the wave steepness data
- Fig. 8.67 : Contour plot of model 4: Weibull (H_s) + Weibull (s).
  (Shown quantile lines : (1) MAX (H_s) + Lin LS (s), (2) Lin LS (H_s)+ Lin LS (s),
  (3) Non LS (H_s) +Lin LS (s))

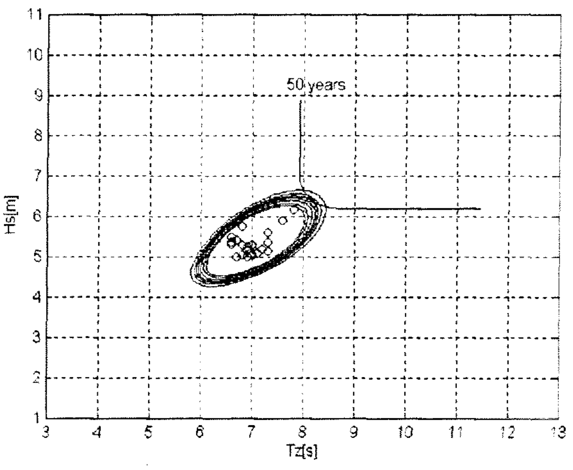


Fig 8.64

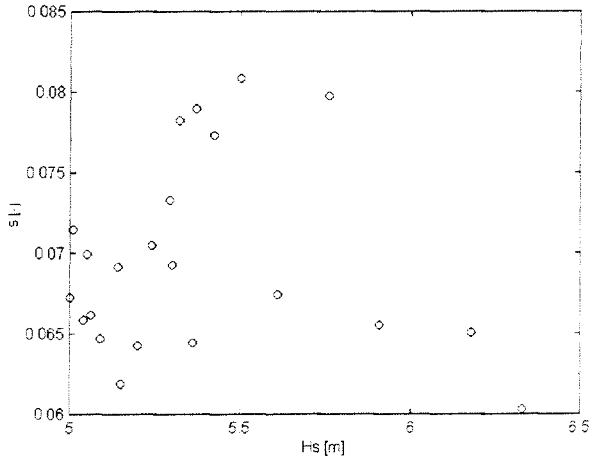


Fig 8.65

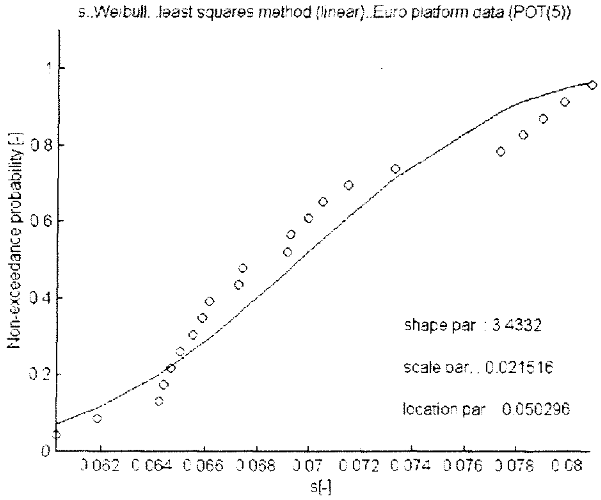


Fig 8.66

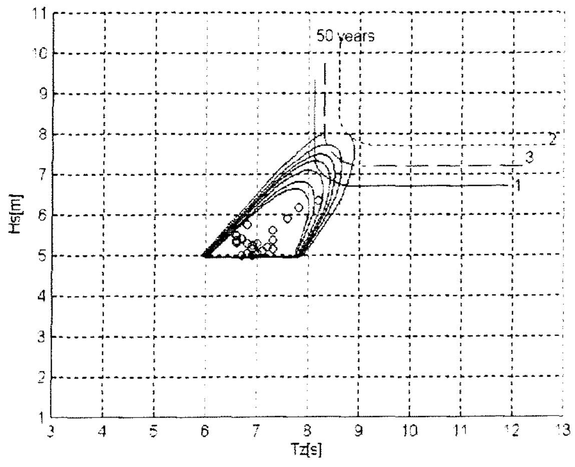


Fig 8.67

- Fig 8.68 : Contour plot of model 5: Weibull (H_s) + Weibull (T_z).
(Shown quantile lines : (1) MAX (H_s) + MAX (T_z), (2) LIN LS (H_s) + MAX (T_z), (3) Non Lin LS (H_s) + MAX (T_z))

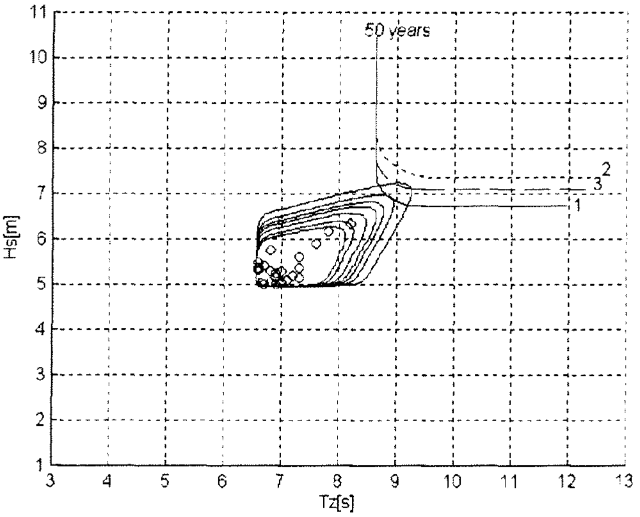


Fig 8.68

Goodness of fit

Table 8.15 shows the computed chi-square values.

Table 8.15 Chi-square values

Treshold level [m]	Model	Marginal distribution for H_s		Marginal distribution for T_p		Chi-square value
	Nr	Distribution	Parameter estimation methods	Distribution	Parameter estimation methods.	
2.00	1	Log-normal	MOM	Log-normal	MAX	231.48
2.00	3	Gumbel	MOM	Log-normal	MAX	156.12
2.00	3	Gumbel	MAX	Log-normal	MAX	182.25
2.00	3	Gumbel	Lin LS	Log-normal	MAX	261
2.00	4	Gumbel	MOM	Weibull (wavesteepness)	Lin LS	225.37
2.00	4	Gumbel	MAX	Weibull	MAX	245.98
2.00	4	Gumbel	Lin LS	Weibull	MAX	223.06
2.00	5	Gumbel	MOM	Weibull	MAX	177.34
2.00	5	Gumbel	MAX	Weibull	MAX	195.29
2.00	5	Gumbel	Lin LS	Weibull	MAX	178.31
4.50	1	Log-normal	MAX	Log-normal	MAX	44.17
4.50	3	Weibull	Lin LS	Log-normal	MAX	67.79
4.50	3	Weibull	Non lin LS	Log-normal	MAX	77.02
4.50	4	Weibull	Lin LS	Weibull (wavesteepness)	Lin LS	12.67
4.50	4	Weibull	Non lin LS	Weibull	MAX	12.59
4.50	5	Weibull	Lin LS	Log-normal	MAX	100.20
4.50	5	Weibull	Nn lin LS	Log-normal	MAX	60.84
5.00	1	Log-normal	MAX	Log-normal	MAX	17.33
5.00	4	Weibull	MAX	Weibull	MAX	11.07
5.00	4	Weibull	Lin LS	Weibull (wavesteepness)	Lin LS	9.79
5.00	4	Weibull	Non lin LS	Weibull	MAX	10.62
5.00	5	Weibull	MAX	Weibull	MAX	8.63
5.00	5	Weibull	Lin LS	Weibull	MAX	10.56
5.00	5	Weibull	Non lin LS	Weibull	MAX	9.56

Return values

The values of (H_s(50),T_p(50)) for s=7% are shown in table 8.16.

It must be stressed that the shown return values are under assumption that the Euro platform is located in deep water. However, this assumption is doubtful, especially for the extreme significant wave heights and periods.

Table 8.16 (H_s(50),T_p(50)) corresponding to s=7%

Treshold level [m]	Model	Marginal distribution for H_s		Marginal distribution for T_p		(H_s(50),T_p(50)) [m,s]
	Nr	Distribution	Parameter estimation methods	Distribution	Parameter estimation methods.	
4.50	1	Log-normal	MAX	Log-normal	MAX	(6.82,7.90)
4.50	3	Weibull	Lin LS	Log-normal	MAX	(7.50,8.30)
4.50	4	Weibull	Non lin LS	Weibull (wave steepness)	Lin LS	(7.72,8.40)
4.50	5	Weibull	Non lin LS	Log-normal	MAX	(6.99,8.00)
5.00	1	Log-normal	MAX	Log-normal	MAX	(6.80,7.90)
5.00	4	Weibull	Lin LS	Weibull (wave steepness)	Lin LS	(7.35,8.30)
5.00	5	Weibull	MAX	Weibull	MAX	(8.09,8.60)

Discussion

- To obtain an accurate fit of the models to the extreme sea states, only the extreme data points should be involved in the analysis. When a low treshold level is used for the selection of data, the estimated models tend to overpredict the extremes.
- The bivariate Log-normal distribution underpredicts the extreme sea states. This follows directly from the marginal analysis
- The quantile lines of the models again agree with the marginal return values
- The assumption that the wave steepness is independent of the significant wave height is acceptable for the used data.
- Model 4 appears to be a good alternative for the bivariate statistics of H_s and T. Mainly when the significant wave height is modelled by the Weibull distribution (in that case the probability of sea states below the treshold level is equal to zero, at least when λ_{Hs}=treshold level-0.1 m) the model fits very well the data.
- Model 3, the model with a conditional distribution for T_z, appears to be only applicable when a sufficient number of observations is available.

8.5 References

[1] Johnson, M.E., *Multivariate statistical simulation.* New York: John Wiley, 1987

[2] Mardia, K.V., *Families of bivariate distributions.* London: Griffin, 1970

[3] Mathiesen, M. et al., *"Intercomparison of extremal wave analysis methods using numerically simulated data".* New orleans: Proc. WAVES '93 conf., july 1993, pp 963-977

[4] Metcalfe, A. V., *Statistics in Civil Engineering.* New York: John Wiley & Sons Inc, 1997

[5] Morton, I.D., Bowers, J., *"Extreme value analysis in a multivariate offshore environment".* Applied Ocean Research, 1997, v 18, pp. 303-317

9 Conclusions and recommendations

In this report, probability distributions for the long-term, significant wave height and wave period have been compared. The marginals as well as the joint statistics of the two variables have been analyzed. The study has been focussed on extreme, deep-water wave fields. For this, sets of data measured at Karwar (India) and at the Euro platform located in the North Sea have been used. It must be stressed that in this study it is assumed that the wave field near the Euro platform is a deep-water wave field. However, this assumption is doubtful, especially for the extreme values of H_s and T_z.

In the following part, conclusions and recommendations are given for each phase of the presented wave analysis.

Data selection

- In the Peak over Treshold method the choice of the treshold strongly determines the estimated return values. In each case study it is found that a lower treshold gives more peak wave data and an increase in the estimated return value. This corresponds to results of earlier case studies (Mathiesen et al. 1993, Maes et al. (1994)).
- For an extremal analysis, only the extreme observations of H_s and T should be taken into account. If other types of waves such as swell are included, it is seen that in such case the distributions tend to overpredict the extreme observations.

Marginal distributions

- With regard to the significant wave height, the Gumbel and Weibull distribution provided the best fits to the data. The Log-normal tends to underpredict the extreme observations, which was earlier found by Fang and Hogben (1982) and Ochi (1978). The Frechet distribution provided a very poor fit. It must be noted that these results strongly depend on the data sets used.
- With regard to the wave period (the zero-up-crossing period or the spectral peak period), the Log-normal distribution gave the closest fit to the data. This also followed from studies made by Burrows et al (1986), Haver (1985), and Mathisen et al. (1990).
- The distribution of the deepwater wave steepness has been described with the Gumbel, Normal and Weibull distribution. The advantage of the Weibull distribution is that it contains three parameters. Due to the inclusion of the shape parameter, the distribution can be fitted very close to the steepness data.

Bivariate distributions

- The bivariate Log-normal distribution tend to underpredict the upper sea sates. This directly follows from the poor fit of the marginal Log-normal distribution. The disadvantage of this model is obvious: only the Log-normal can be used for the marginal description of H_s and T.
- The Fang and Hogben distribution has previously been proposed as a modified version of bivariate Log-normal distribution. In both case studies, however, the model was almost identical to the "normal" bivariate Log-normal distribution. No improvements were found in the fit to the extreme sea states.
- The conditional distribution approach, model 3, performs very well when a large number of observations are involved in the fittings procedure. However, when a small number of observations is available, (when a relative high treshold level is used), this model is less suitable. A Linear regression function appears to be best the best alternative for the description of the parameters of the conditional distribution as a function of H_s.
- Model 4, the model based on transforming the joint distribution of the significant wave height and the (deepwater) wave steepness, provided good results. It must be noted, however, that the assumption is being made that H_s and s are independent. The observations of H_s and s that have been analyzed showed no correlation.
- Model 5, the bivariate model proposed by Morton and Bowers (1997) is relative complicated. The model is not fully understood by the author. Since the model does not provide a much better fit to the data than the other models, there seems to be no reason to use this model instead of the other, much simpler, models.

- In general, it is seen that the fit of the bivariate models to wave data is accurate when there is a high correlation between the two wave variables. In the case of the North Sea data, the (linear) correlation coefficient between the observed values of H_s and T_z was 0.85, and consequently the models fitted the data very well, especially when a high treshold level was used.

Parameter estimation methods

- The major differences found in this study are due to the different methods (treshold levels) of data selection, i.e. which peak wave data should be included in the statistical analysis. The differences between the results obtained through the various fitting techniques was relative small. Trends similar with previous studies (Mathisen et al. (1993)) were found: in general, the least squares methods provide a higher prediction of return values than the maximum likelihood method and the method of moments.

Goodness of fit

- The data set of the south-west monsoon at Karwar and the sets of the Euro platform were suspected to be inhomogeneous. The empirical cdf's of the data showed some irregularities, which indicate the inhomogeneity of the data. In the case of the monsoon data, this might be explained by the fact that the set contains observations of both wind waves and swell. In the case of the North Sea data, the set might probably contain storms of different directions.
  Due to the inhomogeneity of the data, it was hard to obtain a reasonable fit of the probability functions
- The fits of the marginal distributions to the data have mainly been judged visually. Furthermore, the linear correlation coefficient has been used just as two empirical rejection criteria, i.e. the DOL and REC criteria (see section 5.2). The selection of the best fitting distribution on basis of the correlation coefficient did agree with the visual judgement.
  The two criteria of Goda gave generally poor results. Only distributions were rejected which very clearly did not fit the data.
- The fit of the bivariate models was assessed on basis of the two dimensional histogram (scatter diagram) of the two variables. A visual judgement was made by comparing the scatter diagram of the observations to scatter diagrams computed with the theoretical models (section 5.3.1). Furthermore, a two-dimensional chi-square test was applied. Both tests perform well when a large amount of data points is available. However, often the number of data points is too small to compose a 2-dimenional histogram. As far as known to the author, no better goodness of fit criteria is present in literature. Therefore, it might be useful to study this topic in future.

References

[1] Burrows,R., Salih, B. A., *"Statistical Modelling of long-term wave climates".* Proc. 20th Int. conf. coastal. eng., 1986, pp. 42-56

[2] Fang, Z.S., and Hogben, N., *"Analysis and prediction of long-term probability distributions of wave height and periods".* London: Technical report, National Maritime Institute, 1982 (*)

[3] Ochi, M. K., *"Wave statistics for the design of ships and structures".* New York: Proc. SNAME Conf., 1978 (*)

[4] Haver,S. *"Wave climate off northern Norway".* Applied ocean research, 1985,v 7, p 85-92

[3] Maes,M. A.,and Gu,G.Z.,*"Techniques used to determine extreme wave heights from the NESS Data set"*. Gaithersburg: Proc. of the conf. on extreme value theory and applications, volume 2, 1994, pp. 435-444

[4] Mathisen,J., and Bitner-Gregersen, E., *"Joint distributions for significant wave height and wave zero-up-crossing period".* Applied Ocean Research, 1990, Vol. 12, no. 2, pp 93-103

[5] Mathiesen, M. et al., *"Intercomparison of extreme wave analysis: a comparitive study".* New orleans: Proc. WAVES '93 conf., july 1993, pp 963-977

[6] Morton, I.D., Bowers, J., *"Extreme value analysis in a multivariate offshore environment".* Applied Ocean Research, 1997, v 18, pp. 303-317

Appendix 1

This appendix provides a description of the computer program that has been written for the declustering of wave observations. The description is specified to the wave data set that consists of observations, which come from the Euro platform located in the North Sea. (See chapter 8).

To decluster the observations, first the time interval between successive storm events has to be chosen. With respect to the wave field of the North Sea, earlier studies used a time interval varying from 18 h (Mathiesen (1993)) to 30 h (Morton and Bowers(1997)). Here 24 h is taken as minimum time between storm events.

The declustering of the data is started by dividing the data set in parts of 24 hours. For each part, the maximum wave height is determined. The wave period that is selected, corresponds to the maximum significant wave height in the time interval. This is illustrated in figure 1.

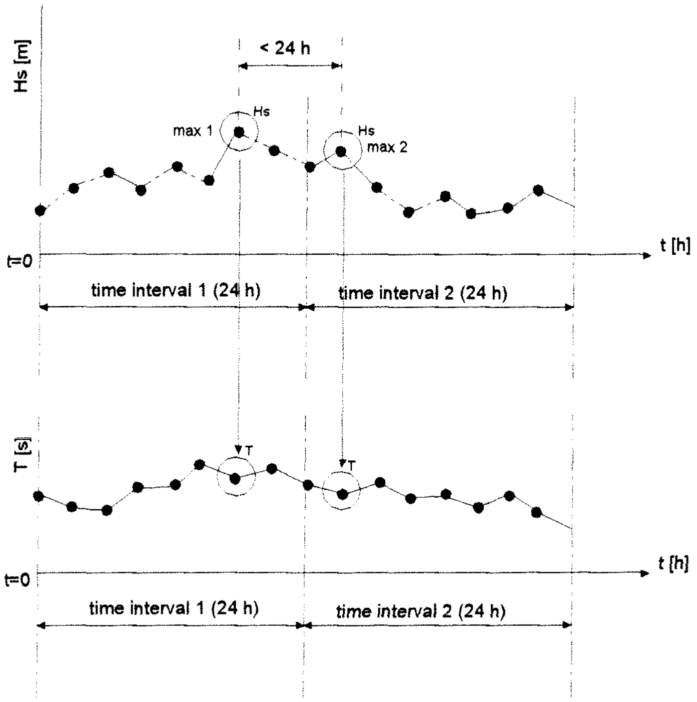
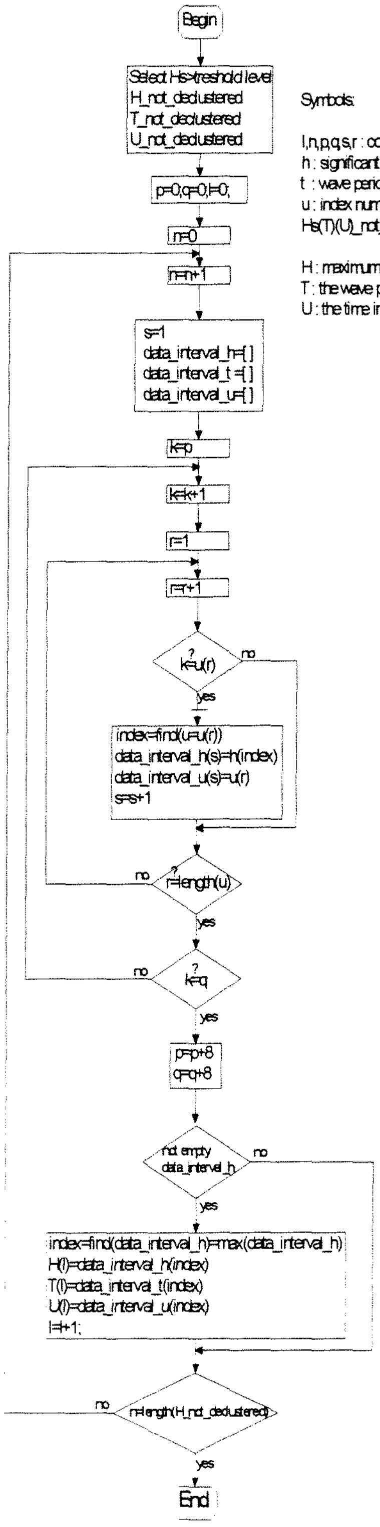


Fig 1. Declustering of data

Figure 2 and 3 show the flows charts of the computer program that has been written. The first chart flow covers the above procedure.

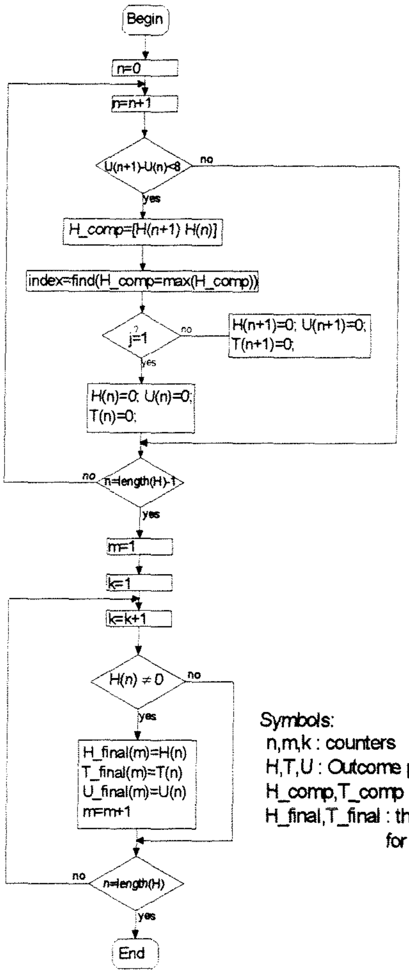
The next procedure, presented in the second flow chart, is the declustering of the selected maximum wave heights. As shown in figure 1, it is possible that the time between the selected maximum significant wave heights of two successive intervals is less than 24 hours. In such case, the program censors the lowest of the two maxima.



Symbols:

l,n,p,q,s,r : counters
h : significant wave height
t : wave period (Tz or Tp, depending on the used sample)
u : index number point of time measurements (u=1: first measurement)
Hs(T)(U)_not_declustered : temporary array containing the measurements (h/t/u)
of the considered time interval
H : maximum value of Hs of the considered time interval
T : the wave period corresponding to the maximum value of Hs
U : the time index corresponding to the maximum value of Hs

Fig 2. flow chart 1



Symbols:
n,m,k : counters
H,T,U : Outcome previous flow chart
H_comp,T_comp : temporary arrays
H_final,T_final : the extreme observations which are used
for the extreme wave analysis

Fig 3. flow chart 2

Appendix 2

In this appendix some additional information is given about the bivariate Log-normal distribution.

The aim of this appendix is to show some typical features of this distribution. The bivariate Log-normal distribution is given by

(1)
$$f(x,y) = \frac{0.5}{xy\pi\delta_x\delta_y\sqrt{1-\rho^2}} * \\ *\exp\left\{-\frac{0.5}{1-\rho^2}\left[\frac{(\log x-\lambda_x)^2}{\delta_x{}^2} - \frac{2\rho(\log x-\lambda_x)(\log y-\lambda_y)}{\delta_x\delta_y} + \frac{(\log y-\lambda_y)^2}{\delta_y{}^2}\right]\right\}$$

in which x represents the wave period and y represents the significant wave height. For this distribution the contour lines with equal probability density are ellipses:

(2)
$$\frac{(\log x-\lambda_x)^2}{\delta_x{}^2} - \frac{2\rho(\log x-\lambda_x)(\log y-\lambda_y)}{\delta_x\delta_y} + \frac{(\log y-\lambda_y)^2}{\delta_y{}^2} = C$$

In this equation C is some constant value. Tangents of these ellipses are defined as

(3)
$$y' = \frac{\left(\frac{\partial f}{\partial x}\right)}{\left(\frac{\partial f}{\partial x}\right)}$$

Vertical tangents are defined as

(4)
$$\frac{\partial f}{dy} = 0$$

which follow as

(5)
$$\log(y) - \lambda_y = \rho\frac{\delta_y}{\delta_x}[\log(x) - \lambda_x]$$

This equation can be simplified to

(6)
$$y = \exp\left[\lambda_y + \rho\frac{\delta_y}{\delta_x}(\log(x) - \lambda_x)\right]$$

Simultaneously the equation for horizontal tangents can be derived. This gives

(7)
$$x = \exp\left[\lambda_x + \rho\frac{\delta_x}{\delta_y}(\log(y) - \lambda_y)\right]$$

Equation (6) and (7) are equal to the mean value of the conditional distribution functions f(y|x) and f(y|x), respectively. The function f(x|y) is defined as:

(8)
$$f(x\,|\,y) = \frac{f(x,y)}{f(y)} = \frac{1}{x\delta_x\sqrt{2\pi}(1-\rho)}\exp\left(-\frac{1}{2}\left\{\frac{x-\left[\lambda_x+\rho\left(\frac{\delta_x}{\delta_y}\right)(\log y-\lambda_y)\right]}{\delta_x\sqrt{1-\rho^2}}\right\}^2\right)$$

Thus

(9)
$$E(x\,|\,y) = \exp\left[\lambda_x+\rho\left(\frac{\delta_y}{\delta_x}\right)(\log y-\lambda_x)\right]$$

Equation (9) is called *the regression of x on y.* In figure 1 and 2 the regression functions defined by equation 6 and 7 (or 9) are shown, together with some contour lines of the bivariate Log-normal distribution function. The line, which represents equation (6), is indicated with 1. Equation (7) is presented by line 2.

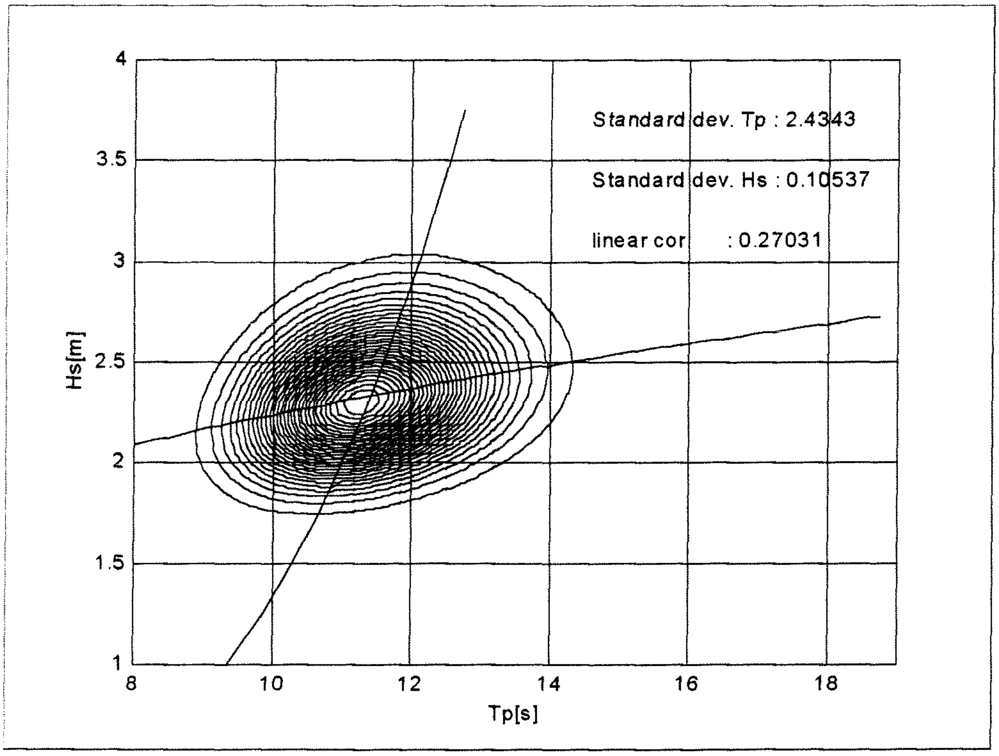


Fig 1 An example of a contour plot of a bivariate Lognormal distribution (1)

Observing equation (6) and (7), it can be seen that the parameters ρ, δ_x and δ_y determine the steepness of the regression lines. When the linear correlation value increases, both lines become steeper. That means that the ellipses become narrower. This agrees with the fact that when the spread of the data points decreases, the value of the correlation coefficient increases.

When the standard deviation of the wave period decreases, line 1 becomes steeper. The slope of line 2 decreases. The complete distribution turns to the left and the slope of the ellipses become steeper. Apparently this bivariate distribution function adjusts to the shape of the marginal distributions.

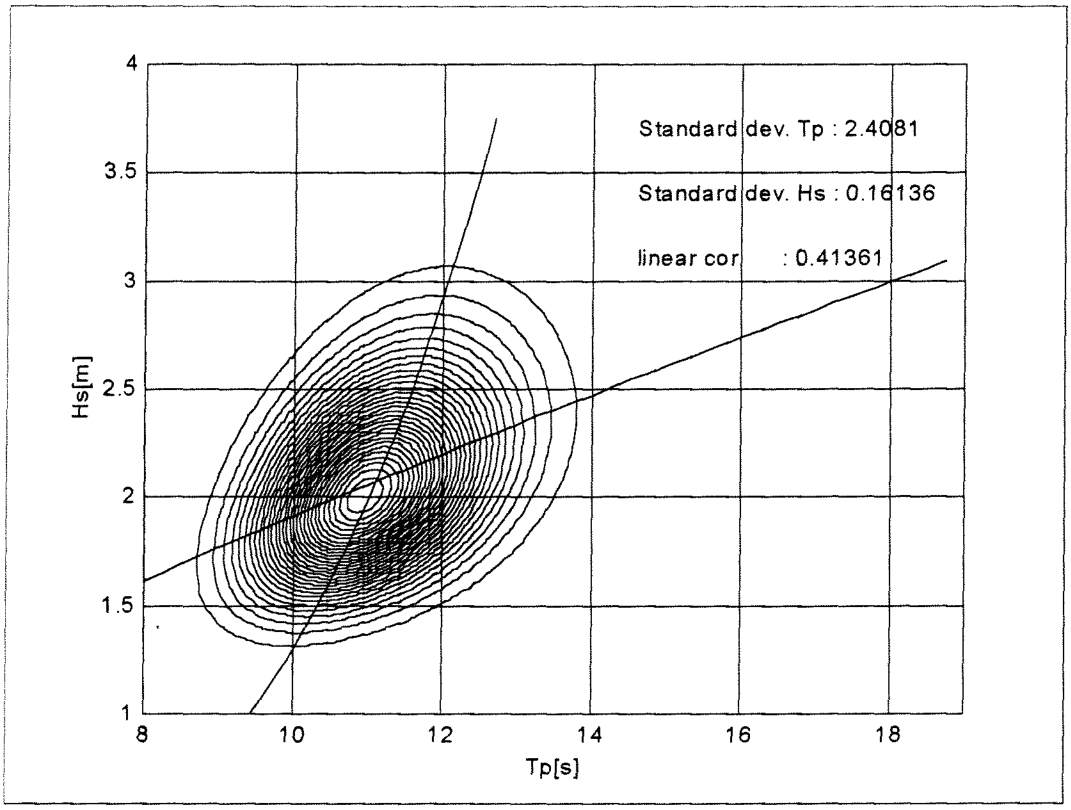


Fig 2 An example of a contour plot of a bivariate Lognormal distribution (2)

Appendix 3

In this appendix a detailed description of the bivariate model of H_s and T that has been developed by Morton and Bowers (1997) is given.

1 Organization appendix

The organization of this appendix is as follows: section 2 provides a description of the transformation of the variables. The choice of the joint treshold is dealt with in section 3. The modelling of the dependence model is described in section 4, and the construction of the joint probability density function is shown in section 5.

2 Transforming the variables

Having determined marginal distributions, the next stage is to consider the nature of the joint distribution. As a preparatory step, it is recommended (Coles and Tawn (1994)) that the marginal extremes are transformed such that their cumulative distributions become unit Frechet. This transformation both scales the variables and also places a greater weight on the more extreme observations, which helps to distinguish them. They appear to be a non-homogenous scattered collection of points in space, while the less extremes observations appear homogenous and are thus collapsed downwards the axes.

The unit Frechet transformation is undertaken by identifying a function Z(X) such that Z has a cumulative distribution

(1) $$P(Z < z) = \begin{cases} 0 \Rightarrow z \leq 0 \\ \exp(-z^{-1}) \Rightarrow z > 0 \end{cases}$$

In the article of Morton and Bowers, the Generalized Pareto distribution (GPD) has been chosen as marginal cdf. Here the marginal distribution functions of chapter 5 are selected.

The required transformations for these distribution functions are listed in table 1.

Tabel1 Transformation formula unit Frechet space for the marginal cdf's

	Z_j
Exponential	$\dfrac{1}{\log\left[1-\exp\left(-\left(\dfrac{x-\lambda}{\delta}\right)\right)\right]}$
Gumbel	$\exp\left[-\dfrac{(-x+\lambda)}{\delta}\right]$
Weibull	$\dfrac{1}{\log\left[1-\exp\left(-\left(\dfrac{x-\lambda}{\delta}\right)^{\beta}\right)\right]}$
Frechet	$\left(\dfrac{x-\lambda}{\delta}\right)^{\beta}$
Log-normal	$\dfrac{1}{\log\left[\dfrac{1}{\sigma\sqrt{2\pi}}\int\dfrac{1}{x}\exp\left(-\dfrac{1}{2}\left[\dfrac{\log(x)-\mu}{\sigma}\right]^{2}\right)dx\right]}$

Having transformed the marginal variables, a second transformation is recommended (Coles and Tawn(1994)) in order to combine these variables in such a way that their joint distribution may be better appreciated. Pseudo-polar co-ordinates are constructed from unit Frechet variables. The radial components correspond to the combined magnitude of the extreme event while the angular components capture the dependencies between the variables. In the bivariate case, the pseudo-polar co-ordinates are defined as

(2) $$r = \frac{(Z_1 + Z_2)}{n}$$

$$w = \frac{Z_1}{nr}$$

3 Choice of joint treshold

The transformations described by equations (1)–(2) are designed to help distinguish the characteristics of the marginal variables and inter-relationships between those variables. The inter-relationships are then modelled by the dependence structure, which reflects the behaviour of the joint extremes; this proces is summarised in block C of fig 3.11.

Morton and Bowers described a graphical method of choosing an appropriate joint treshold. In this method, the angular co-ordinate w is be plotted against the radial co-ordinate, r for various values of r_{min}. It is suggested that independence between the two pseudo-polar co-ordinates is satisfied when the variance increases appreciably and at this point the joint treshold, u_r, is chosen.

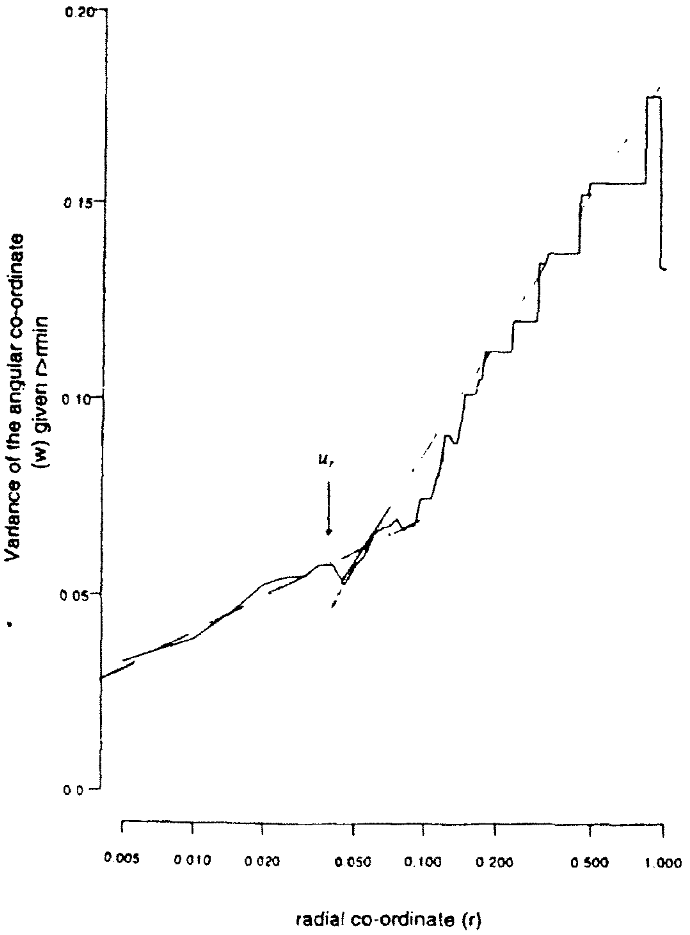


Fig 2 The variance of w given r (Taken from Morton and Bowers(1997))

An application of this procedure is shown in figure 2. In this plot there is a significant change in the variance of when r=0.033 as indicated by an arrow. Therefore this value was selected as the joint treshold. The joint treshold for both data sets is thus expressed in r-values.

4 Modelling the dependence structure

Various models of the dependence between extreme variables have been suggested: logistic, bilogistic and Dirichlet. However, it appears that the choice of dependence model is not usually critical to the accuracy of the final model: previous studies (Coles and Tawn (1994); Anderson and Nadarajah (1993); Cavanie (1993)) have found that each of the dependence models appear to describe the distribution of bivariate data equally well.

Hence in the present study the simplest model was adopted, the logistic:

(3) $$V(z) = \left(\left(\frac{1}{z_1} \right)^{\varphi} + \left(\frac{1}{z_2} \right) \right)^{\frac{1}{\varphi}}$$

The parameter φ, the dependence measure, is determined by a consideration of the distribution of w. The distribution of $Z=(Z_1,Z_2)$ is modelled (Coles and Tawn (1994); Anderson and Nadarajah (1993)) as a point proces, where $P_n = \{n^{-1}Z_i;\ i=1,2,\ldots,n\}$, n being the number of observations. The limiting proces of P_n can be described by an intensity measure Λ where

(4) $$\Lambda(dr * dw) = \frac{dr}{r^2} dH(w)$$

and H is the dependence function, which satisfies

(5) $$\int_{s_p} w_j dH(w) = 1$$

The measure density, h, of the dependence function, H, in the case of a logistic model is

(6) $$h(w) = (\varphi - 1)(w(1-w))^{\varphi - 2}(w^{\varphi} + (1-w)^{\varphi}))^{\frac{1}{\varphi} - 2}$$

In this model $\varphi=1$ corresponds to independence between the two variables and $\varphi=\infty$ implies perfect dependence. The dependence measure has a corresponding negative log-likelihood function that can be used to determine a suitable estimate of φ

(7) $$-l(\varphi; w) = -n\log(\varphi - 1) - (\varphi - 2)\sum_{i=1}^{n}\log(w_i(1-w_i)) - \left(\frac{1}{\varphi} - 2\right)\sum_{i=1}^{n}\log(w_i^{\varphi} + (1-w_i)^{\varphi})$$

5 Constructing the joint probability density function

Having established suitable marginal distributions and the dependence structure, a joint pdf may be constructed. The details of the proces, in the bivariate case, may be described in terms of

(8) $$\frac{\partial^2}{\partial x_1 \partial x_2} P(X_1 > x_1, X_2 > x_2)$$

where

(9) $$P(X_1 > x_1, X_2 > x_2) = \exp(-V(z))$$

Therefore,

(10) $$\frac{\partial^2}{\partial x_1 \partial x_2} P(X_1 > x_1, X_2 > x_2) = \left(V_1(z)V_2(z) - V_{12}(z)\right)\frac{\partial z_1}{\partial x_1}\frac{\partial z_2}{\partial x_2}\exp(-V(z))$$

where

(11) $$V(z) = \left(\left(\frac{1}{z_1}\right)^{\varphi} + \left(\frac{1}{z_2}\right)^{\varphi}\right)^{\frac{1}{\varphi}}$$

is the logistic model and

(12) $$V_1(z) = \frac{\partial V}{\partial z_1} = (-z_1^{-\varphi-1})\left(\left(\frac{1}{z_1}\right)^{\varphi} + \left(\frac{1}{z_2}\right)^{\varphi}\right)^{\frac{1}{\varphi}-1}$$

(13) $$V_2(z) = \frac{\partial V}{\partial z_2} = (-z_2^{-\varphi-1})\left(\left(\frac{1}{z_1}\right)^{\varphi} + \left(\frac{1}{z_2}\right)^{\varphi}\right)^{\frac{1}{\varphi}-1}$$

(14) $$V_{12}(z) = \frac{\partial^2 V}{\partial z_1 \partial z_2} = (-z_1^{-\varphi-1})(-z_2^{-\varphi-1})(1-\varphi)\left(\left(\frac{1}{z_1}\right)^{\varphi} + \left(\frac{1}{z_2}\right)^{\varphi}\right)^{\frac{1}{\varphi}-2}$$

For each of the candidate marginal distribution functions, except for the Log-normal distribution, the formula of the two first order derivatives of Z, $\frac{\partial Z_j}{\partial x_j}$ (j=1,2) are listed in table 2.

Tabel2 First order derivates of Z

	$\frac{\partial Z_j}{\partial x}$
Exponential	$\frac{-1}{\left[\log\left(1-\exp\left(-\left(\frac{x-\lambda}{\delta}\right)\right)\right)^2\delta\right]} * \frac{\exp\left[-\left(\frac{x-\lambda}{\delta}\right)\right]}{\left[1-\exp\left(-\left(\frac{x-\lambda}{\delta}\right)\right)\right]}$
Gumbel	$\frac{1}{\delta}\exp\left[-\frac{(x-\lambda)}{\delta}\right]$
Weibull	$\frac{1}{\log\left[1-\exp\left(-\left(\frac{x-\lambda}{\delta}\right)^\beta\right)\right]^2}\left(\frac{x-\lambda}{\delta}\right)^\beta\frac{\beta}{x-\lambda}\frac{\exp\left[-\left(\frac{x-\lambda}{\delta}\right)^\beta\right]}{\left[1-\exp\left(-\left(\frac{x-\lambda}{\delta}\right)^\beta\right)\right]}$
Frechet	$\left(\frac{x-\lambda}{\delta}\right)^\beta\frac{\beta}{(x-\lambda)}$

For the cumulative distribution function of the Log-normal distribution there exists no analytical expression. Only via numerical integration techniques this function can be approximated. Therefore, the first order derivative of Z for a Log-normal distribution also has to be determined numerically. In this case the derivative is calculated with

$$\frac{\partial Z_j}{\partial x} \approx \frac{Z_{j+1}-Z_{j-1}}{2\Delta x} \tag{15}$$

In equation (15) Δx represents a step size. Here a step size of 0.05 m for the significant wave height and a step size of 0.05 s for the wave period are taken.

The resultant pdf describes the distribution of the joint extremes. When there is a reasonably high dependence between variables and only the extremes are of interest, this pdf is sufficient (Morton and Bowers (1997)).

Appendix 4

In this appendix the derivation of the slope (A) and intercept (B) of the (linear) empirical regression function is given.

The linear least squares method is based on minimization of the distance

(1) $$\sum_{i=1}^{n}(y_i-(B+Ax_i))^2$$

or

(2) $$\chi^2=\min_{\theta\to\hat{\theta}}\sum_{i=1}^{n}(y_i-(A+Bx_i))^2$$

The minimum of (1) is obtained by equating to zero the partial derivatives of (2) with respect to the parameters:

(3) $$\frac{\partial\chi^2}{\partial A}=0 \qquad \frac{\partial\chi^2}{\partial B}=0$$

This gives

(4) $$\frac{\partial\chi^2}{\partial A}=\sum_{i=1}^{n}-2y_i+2nA+\sum_{i=1}^{n}2Bx_i=0$$

$$\frac{\partial\chi^2}{\partial B}=\sum_{i=1}^{n}-2y_ix_i+\sum_{i=1}^{n}2x_iA+\sum_{i=1}^{n}2Bx_i^2=0$$

A and B solved gives:

(5) $$B=\frac{\sum_{i=0}^{n-1}x_iy_i-\sum_{i=0}^{n-1}\left(\sum_{i=0}^{n-1}\frac{1}{n}y_i\right)x_i}{\sum_{i=0}^{n-1}x_i^2-\sum_{i=0}^{n-1}\left(\sum_{i=0}^{n-1}\frac{1}{n}x_i\right)x_i} \qquad A=\sum_{i=0}^{n-1}\frac{1}{n}y_i-\sum_{i=1}^{n}\frac{B}{n}x_i$$

Appendix 5

In this appendix the derivation of the maximum likelihood estimators is presented for each selected distribution function.

The maximum likelihood function of the Exponential distribution
The loglikelihood function of a random sample {x$_i$},i=1,2...n from this distribution is

(1) $$\log L = -n\log(\delta) - \frac{\sum_{i=1}^{n} x_i}{\delta} + \frac{n\lambda}{\delta}$$

Maximum likelihood estimating equations are obtained by finding the root of the partial derivatives of log *L* with respect to the parameters. Since $\lambda \le \min(x_i)$, λ is maximum when

(2) $$\hat{\lambda} = x_1 \qquad (x_1 = \min(x_i))$$

The partial derivative $\partial L / \partial\delta$ is:

(3) $$\frac{\partial \log L}{\partial\delta} = -\frac{n}{\delta} + \frac{\sum_{i=1}^{n} x_i}{\delta^2} - \frac{n\lambda}{\delta^2}$$

Setting this expression to zero gives the following simultaneous equation for the maximum likelihood estimator (MLE) $\hat{\delta}$:

(4) $$\hat{\delta} = \bar{x} - x_1$$

The maximum likelihood function of the maximum Gumbel distribution
The loglikelihood function of a random sample {x$_i$},i=1,2...n from this distribution is

(5) $$\log L = -n\log\delta - \sum_{i=1}^{n}\frac{x_i-\lambda}{\delta} - \sum_{i=1}^{n}\exp\left[-\left(\frac{x_i-\lambda}{\delta}\right)\right]$$

Again maximum likelihood estimating equations are obtained by finding the root of the partial derivatives of log *L* with respect to the parameters. Parameter estimates are then obtained as the simultaneous solutions of these equations.

The partial derivatives of log *L* are:

(6) $$\frac{\partial \log L}{\partial\lambda} = \frac{n}{\delta} - \sum_{i=1}^{n}\frac{1}{\delta}\exp\left[-\left(\frac{x_i-\lambda}{\delta}\right)\right]$$

(7) $$\frac{\partial \log L}{\partial\delta} = -\frac{n}{\delta} + \sum_{i=1}^{n}\frac{x_i-\lambda}{\delta^2} - \sum_{i=1}^{n}\frac{x_i-\lambda}{\delta^2}\exp\left[-\left(\frac{x_i-\lambda}{\delta}\right)\right]$$

Setting these expressions to zero gives the following simultaneous equations for the maximum likelihood estimators (MLEs) $\hat{\lambda}$ and $\hat{\delta}$ (Tucker, 1991):

(8) $$\hat{\lambda} = -\hat{\delta}\log(n^{-1}\sum_{i=1}^{n}\exp\left[-\frac{x_i}{\hat{\delta}}\right]$$

(9) $$\hat{\delta} = n^{-1}\sum_{i=1}^{n} x_i - \frac{\sum_{i=1}^{n} x_i \exp\left[-\frac{x_i}{\hat{\delta}}\right]}{\sum_{i=1}^{n}\exp\left[-\frac{x_i}{\hat{\delta}}\right]}$$

The maximum likelihood function of the 3-parameter minimum Weibull distribution
The loglikelihood function of a random sample $\{x_i\}$, i=1,2...n from this distribution is

(10) $$\log L = n\log\left[\left(\frac{\beta}{\delta}\right)\right] + (\beta-1)\left\{\sum_{i=1}^{n}\log\left(\frac{x_i-\lambda}{\delta}\right)\right\} - \sum_{i=1}^{n}\left(\frac{x_i-\lambda}{\delta}\right)^{\beta}$$

The estimating equations follow as

$$\frac{\partial \log L}{\partial \lambda} = \frac{\delta}{\theta}\sum_{i=1}^{n}(x_i-\lambda)^{\beta-1} - (\beta-1)\sum_{i=1}^{n}(x_i-\lambda)^{-1} = 0$$

(11) $$\frac{\partial \log L}{\partial \beta} = \frac{n}{\beta} + \sum_{i=1}^{n}\log(x_i-\lambda) - \frac{1}{\theta}\sum_{i=1}^{n}(x_i-\lambda)^{\beta}\log(x_i-\lambda) = 0$$

$$\frac{\partial \log L}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n}(x_i-\lambda)^{\beta} = 0$$

where, in order to simplify the derivatives, the scale parameter δ has been replaced by

(12) $$\theta = \delta^{\beta}.$$

The three equations do not yield explicit solutions for the estimates. However as shown by Cohen(1965), θ can be estimated from the last two equations to give

(13) $$\left[\frac{\sum_{i=1}^{n}(x_i-\hat{\lambda})^{\hat{\beta}}\log(x_i-\hat{\lambda})}{\sum_{i=1}^{n}(x_i-\hat{\lambda})^{\hat{\beta}}} - \frac{1}{\hat{\delta}}\right] - \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\lambda}) = 0.$$

From the last equation of (11), we have

(14) $$\hat{\theta} = \sum_{i=1}^{n} (x_i - \lambda)^{\hat{\beta}}$$

When λ is known, equation (13) can be solved iteratively for $\hat{\beta}$. When λ is unknown, the three parameter distribution is converted to the two parameter distribution by preselecting the location parameter by inspection of data in a sample. Usually the parameter is set at the lowest value –0.1 of the data set. By employing a trial and error procedure the best required estimation will be reached.

With λ_1 fixed, (13) is solved for δ_1, and θ_1 follows from (14). $(\partial \log L / \partial\lambda)_1$ can be calculated by substituting λ_1, δ_1, and θ_1 into the first equation of (11). If $(\partial \log L / \partial\lambda)_1$=0, then $\hat{\lambda} = \lambda_1$, $\hat{\delta} = \delta_1$, $\hat{\theta} = \theta_1$, and the estimation proces is completed. Other, the cycle of computations with a new approximation λ_2 will be repeated until a value for λ is found for which eq (11) is approximately equal to zero. For this procedure one can define a tolerance interval. When the outcome of eq. (11) falls within this interval the iteration procedure is ended. The procedure is illustrated by fig. 1.

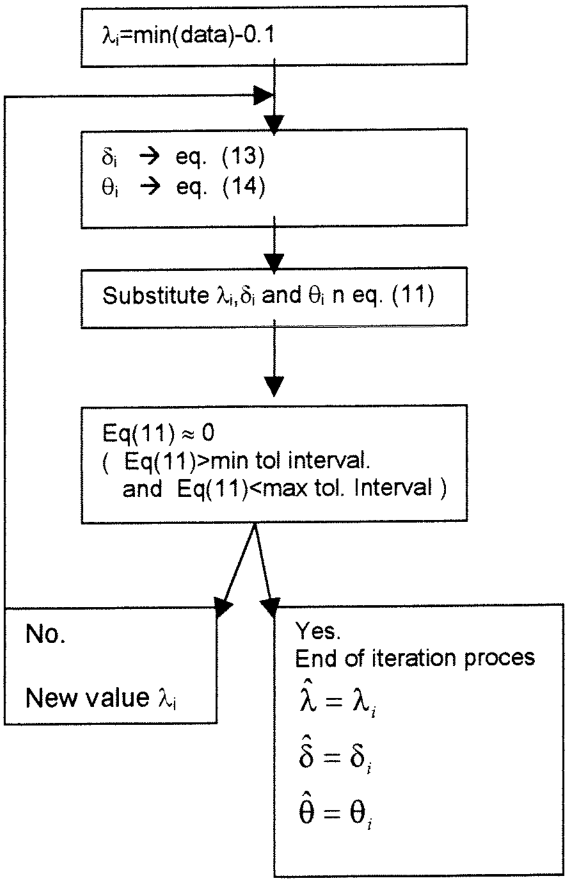

λ_i=min(data)-0.1
δ_i → eq. (13)
θ_i → eq. (14)
Substitute λ_i,δ_i and θ_i n eq. (11)
Eq(11) ≈ 0
(Eq(11)>min tol interval.
and Eq(11)<max tol. Interval)
No.
New value λ_i
Yes.
End of iteration proces
$\hat{\lambda} = \lambda_i$
$\hat{\delta} = \delta_i$
$\hat{\theta} = \theta_i$

Fig 1.Maximum likelihood estimation procedure for the 3-par. Weibull distribution

The maximum likelihood function of the 2-parameter maximum Frechet distribution
Estimation of the two-parameter Frechet distribution can be reduced to the estimation of the Gumbel distribution. Therefore, the following transformation is used

(15) $$Y = \log(x - \lambda)$$

The Frechet distribution of X is transformed into the Gumbel distribution of Y with the following relation between the parameters

(16) $$\lambda_G = \log(\delta_F)$$

(17) $$\delta_G = \frac{1}{\beta_F}$$

The maximum likelihood function of the 3-parameter maximum Frechet distribution
The loglikelihood function of a random sample {x$_i$}, i=1,2...n from this distribution is

(18) $$\log L = n\log\left(\frac{\beta}{\delta}\right) + (\beta+1)\left\{\sum_{i=}^{n}\log\left(\frac{\delta}{x_i-\lambda}\right)\right\} - \sum_{i=1}^{n}\left(\frac{\delta}{x_i-\lambda}\right)^{\beta}$$

The estimating equations follow as

$$\frac{\partial \log L}{\partial \delta} = \frac{n}{\beta} + (\beta+1)\frac{n}{\delta} - \sum_{i=1}^{n}\left(\frac{\delta}{(x_i-\lambda)}\right)^{\beta}\frac{\beta}{\delta} = 0$$

$$\frac{\partial \log L}{\partial \lambda} = (\beta+1)\sum_{i=1}^{n}\frac{1}{(x_i-\lambda)} - \sum_{i=1}^{n}\left(\frac{\delta}{(x_i-\lambda)}\right)^{\beta}\frac{\beta}{(x_i-\lambda)}$$

(19)

$$\frac{\partial \log L}{\partial \beta} =$$

$$= -\frac{n}{\delta} + n\log(\delta) + \sum_{i=1}^{n} -\log(x_i-\lambda) +$$

$$+\sum_{i=1}^{n} -\log(x_i-1) - \sum_{i=1}^{n}\left(\frac{\delta}{(x_i-\lambda)}\right)^{\beta}\log\left(\frac{\delta}{(x_i-\lambda)}\right) = 0$$

The above set of equations has been determined by using the mathematical spreadsheet program Mathcad. In spite of the Weibull distribution, no simplified formulas have been found in literature. With the above set of equations, the maximum likelihood estimators of the three distribution parameters can be calculated iteratively. In this case, direct maximization of the (log) likelihood function will probably be much simpler.

The maximum likelihood function of the Log-normal distribution
The loglikelihood function of a random sample {x$_i$}, i=1,2...n from this distribution is

(20) $$\log L = -n\log(\sqrt{2\pi\delta}) - \sum_{i=1}^{n}\log x_i + \frac{\sum_{i=1}^{n}\left[-(\log x_i)^2 + 2\lambda\log x_i\right]}{2\delta^2} - \frac{n\lambda^2}{2\delta^2}$$

The estimating equations follow as

$$\frac{\partial\log L}{\partial\lambda} = \frac{2\sum_{i=1}^{n}\log x_i}{2\delta^2} - \frac{2n\mu}{2\delta^2}$$

(21) $$\frac{\partial\log L}{\partial\delta} = -\frac{n}{\delta} - \frac{\sum_{i=1}^{n}\left[-(\log(x_i))^2 + 2\lambda\log x_i\right]}{\delta^3} + \frac{n\lambda^2}{\delta^3}$$

Setting these expressions to zero gives the following simultaneous equations for the maximum likelihood estimators (MLEs) $\hat{\mu}$ and $\hat{\sigma}$:

(22) $$\hat{\lambda} = \bar{Z}$$

$$\hat{\delta} = \left[n^{-1}\sum_{j=1}^{n}(Z_j - \bar{Z})^2\right]^{1/2}$$

with

$$Z_i = \log(x_i)$$

Appendix 6

In this appendix an analytical derivation of a confidence interval is given which includes statistical uncertainty. The presented confidence bands are specified to the Gumbel distribution. For other distribution functions the calculation procedure is similar.

Consider the Gumbel distribution,

(1) $$F(x) = \exp\left(\exp\left[-\frac{(x-\lambda)}{\delta}\right]\right)$$

The value of a return value of this function can be determined after transformation of the cumulative distribution function:

(2) $$\hat{x}_p = \hat{\lambda} - \hat{\delta}\log(-\log(F(x))$$

or

(3) $$\hat{x}_p = \hat{\lambda} + \hat{\delta}u$$

with

(4) $$u = -\log(-\log(F(x))$$

In eq. (2) and (3) $\hat{x}_p, \hat{\lambda}$ and $\hat{\delta}$ are mean values. With eq. (3) the variance of $\hat{x}_p$ can be derived from the variance of $\hat{\lambda}$ and $\hat{\delta}$:

(5) $$\begin{aligned}&\mathrm{var}(\hat{x}_p) = \\ &= E((\hat{x}_p - \mu(\hat{x}_p))^2) \\ &= E((\hat{\lambda} + \hat{\delta}u - \mu(\hat{\lambda}) - \mu(\hat{\delta})u)^2) \\ &= E((\hat{\lambda} - \mu(\hat{\lambda}))^2 + 2u(\hat{\lambda} - \mu(\hat{\lambda}))(\hat{\delta} - \mu(\hat{\delta})) + u^2(\hat{\delta} - \mu(\hat{\delta}))^2) \\ &= \mathrm{var}(\hat{\lambda}) + 2u(\mathrm{cov}(\hat{\lambda}, \hat{\delta})) + u^2\,\mathrm{var}(\hat{\delta})\end{aligned}$$

The estimators of the Gumbel distributions are assumed to be Gaussian distributed. The variance of these normal distributions is computed numerically by using the bootstrap procedure. (See appendix [4.3]). For the covariance between $\hat{\lambda}$ and $\hat{\delta}$, the covariance value between the two sets of bootstrap data, i.e. the sets of data with 500 estimators of $\hat{\lambda}$ and $\hat{\delta}$, are used.

When the variance of a return value is known, the corresponding confidence intervals can be calculated. Considering a 95% confidence interval, the expression follows as (Groeneboom et al. (1995))

(6) $$\left\{x_p; E(\hat{x}_p) - 1.96\sqrt{\mathrm{var}(\hat{x}_p)} < E(\hat{x}_p) < E(\hat{x}_p) + 1.96\sqrt{\mathrm{var}(\hat{x}_p)}\right\}$$

Appendix 7

In this appendix the theoretical scatterdiagrams of the first case studie are presented. They belong to the bivariate functions fitted to the Karwar data above the treshold level H_s=1.95 m. For each bivariate model only one theoretical scatterdiagram is presented (thus for only one parameter estimation method : the method of moments). It must be noted that due to round of errors, the total number of observations of the theoretical scatterdiagrams do no always conform to the total number of observed observations. Therefore the theoretical diagrams must be seen as a global indication of the fit of the functions to the data.

Figure 1 to 4:

- Treshold level : **1.95 m**
- Number of observations : 167 points
- Fig. 1 : Theoretical scatterdiagram of the Bivariate Log-normal distribution
- Fig. 2 : Theoretical scatterdiagram of model 3: Gumbel (H_s) + Log-normal (T_z).
- Fig. 3 : Theoretical scatterdiagram of model 3: Weibull (H_s) + Log-normal (T_z).
- Fig. 4 : Theoretical scatterdiagram of model 4: Gumbel (H_s) + Gumbel (s).

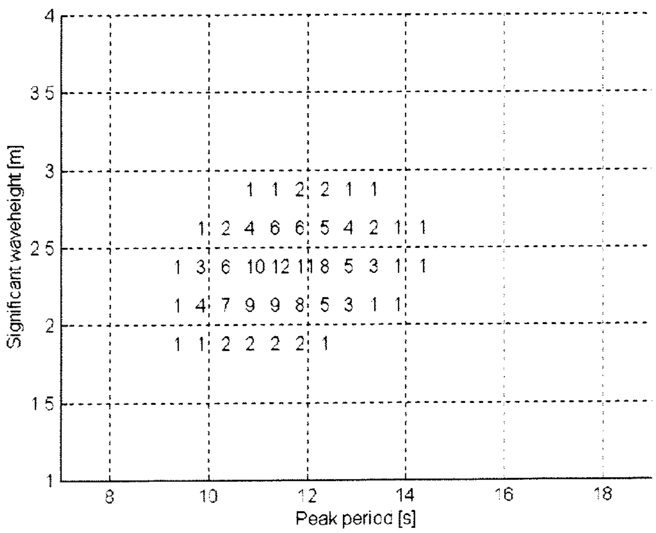


Fig 1

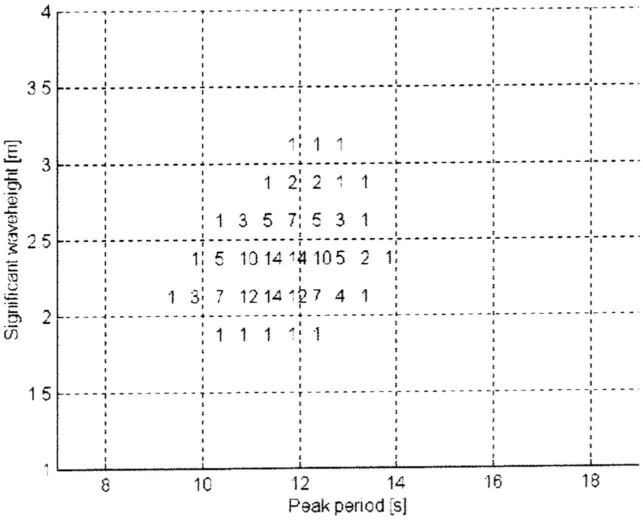


Fig 2

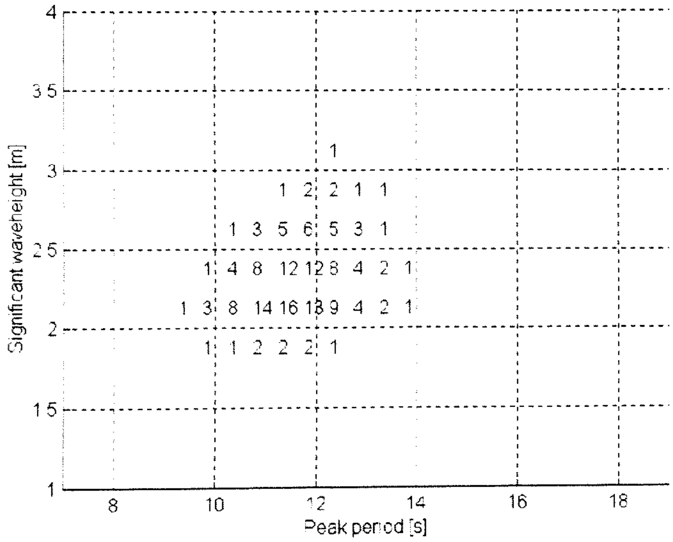


Fig 3

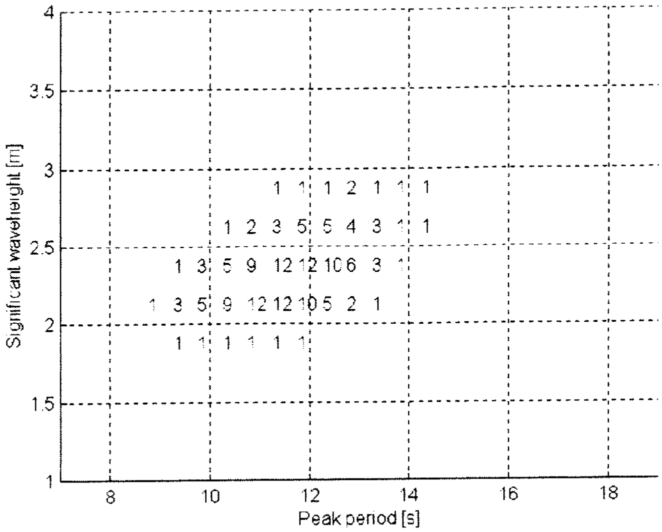


Fig 4

Fiure 5:

- Treshold level : **1.95 m**
- Number of observations : 167 points
- Theoretical scatterdiagram of model 5 : Weibull (H_s) + Log-normal (T_p).

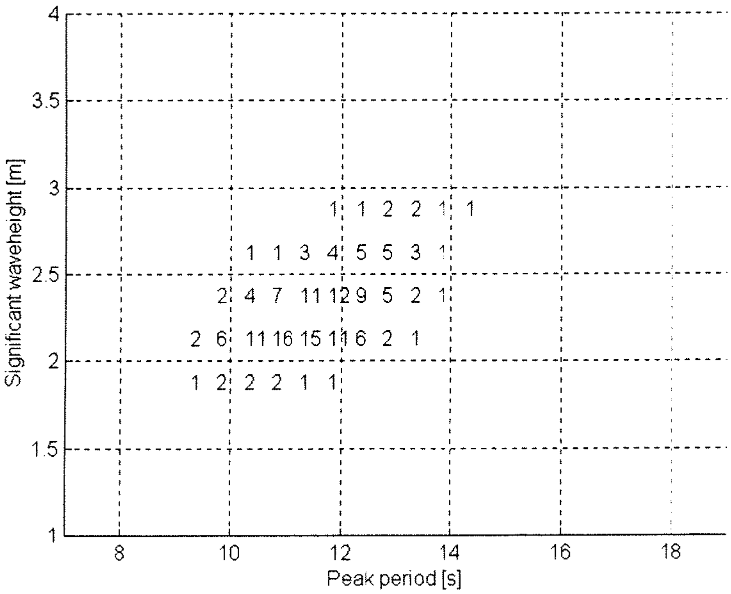


Fig 5