Statistical Modelling of Severe Wind Gust

X.G. Lin

Risk Modelling Project, Minerals and Geohazards Division, Geoscience Australia, Canberra, Australia
E-mail: xunguo.lin@ga.gov.au

Abstract: The modelling of severe wind gust is a fundamental part of any wind hazard assessment. Statistical modelling approaches are usually used to describe the probability of occurrences of extreme winds. The classical extreme value theory based on the generalized extreme value (GEV) distribution and the more recently developed peaks over threshold (POT) and the use of generalized Pareto distribution (GPD) are investigated. Practical issues such as the threshold selection and model validation are also discussed. Severe wind gust models were developed for a number of locations in Perth. Estimates of 50 and 100 year return period events are compared with estimates published by the Standards Australia for Perth (Region A), which are based on a single site. For coast sites, the predicted wind gust speeds are similar to those published by the Standards, while the inland sites have smaller wind gust speeds.

Keywords: *Wind gust modelling; Extreme value theory; Generalized Extreme Value distribution; Generalized Pareto distribution; Peaks over threshold; Return periods*

1. INTRODUCTION

Extreme wind is one of the major natural hazards experienced in Perth. These extreme winds are generally produced by cold fronts and not from cyclones. To predict the extreme wind with a given return period statistical analysis of wind records was used.

Extreme value theory is a statistical technique for describing the unusual (extreme) rather than the usual events. This work quantifies the hazard components in a risk assessment of the Perth metropolitan area to winds that result in damage to buildings and infrastructures. These are the extreme winds of interest.

2. EXTREME VALUE ANALYSIS

The classical extreme value theory is based on the analysis of the largest (or smallest) value in an epoch. In wind engineering, an epoch is assumed to be a calendar year. The maximum yearly (3 seconds) wind gust is used in the classical extreme value analysis.

In the past twenty years, a new body of extreme value theory has been developed and is referred to as ‘peaks over threshold’ (POT) modelling. This theory allows for the use of all available data exceeding a sufficiently high threshold.

A review of methods for calculating extreme wind speeds using extreme value techniques can be found in Palutikof et al. (1999). An introduction of extreme value theory can be found in Embrechts et al. (2001) and Colse (2001).

2.1 Generalized Extreme Value Distribution

The classical extreme value theory is based on three asymptotic extreme value distributions identified by Fisher and Tippett (1928). The Generalized Extreme Value (GEV) distribution introduced by Jenkinson (1955) combines the three distributions into a single mathematical form with the cumulative distribution function (CDF):

$$H(x;\xi,\sigma,\mu)=\begin{cases} e^{-\left(1-\xi\frac{x-\mu}{\sigma}\right)^{1/\xi}} & \xi\neq 0 \\ e^{-e^{-(x-\mu)/\sigma}} & \xi=0 \end{cases} \quad (1)$$

where ξ, σ and μ are the shape, scale and location parameter, respectively, and x is the maximum of an epoch.

When $\xi = 0$, it is the Type I GEV or so called Gumbel distribution; when $\xi < 0$, the GEV is called the Type II (or Frechet) distribution, which has a right long tail; when $\xi > 0$, it is the Type III GEV (a form of the Weibull distribution) and has a short tail. Type III GEV has a theoretical upper bound ($\mu+\sigma/\xi$), that may be useful for estimates of extreme values (such as wind gust). Many scientists believe that due to physical and meteorological limitations, there is an upper bound to the maximum wind gust.

2.2 Peaks Over Threshold and Generalized Pareto Distribution

A major criticism of traditional extreme value theory is that it only considers a single maximum within each epoch. This approach ignores other extreme events that may have occurred in each epoch. An alternative approach, often referred to as the peaks over threshold (POT) approach, is to consider all values greater than a given threshold value. Given a threshold u, the distribution of excess values of x over u is defined by:

$$F_u(y) = \Pr\{X - u \le x \mid X > u\} = \frac{F(x) - F(u)}{1 - F(u)} \tag{2}$$

which represents the probability that the value of x exceeds u by at most an amount y, where $y = x - u$. Balkema and de Haan (1974) and Pickands (1975) show that for a sufficiently high threshold, u, the distribution function of the excess, $F_u(y)$, converges to the generalized Pareto distribution (GPD) which has a CDF given by:

$$G(x;\xi,\sigma,\mu) = \begin{cases} 1 - \left(1 - \xi \dfrac{x-\mu}{\sigma}\right)^{1/\xi} & \xi \neq 0 \\ 1 - e^{-(x-\mu)/\sigma} & \xi = 0. \end{cases} \tag{3}$$

When $\xi = 0$, the GPD corresponds to an exponential distribution (medium-size tail); when $\xi < 0$, it takes the form of the ordinary Pareto distribution (long tailed); when $\xi > 0$, it is known as a Pareto II type distribution (short tailed), which is also upper bounded by ($\mu + \sigma/\xi$).

The parameters of a GPD can be estimated with various methods such as the maximum likelihood (ML) method (see Davison, 1984) and probability weighted moments (PWM) method (see Hosking and Wallis, 1987).

An important property of the GPD is that if $\xi > -1$, then the conditional mean exceedance (CME) over a threshold, u, is a linear function of u:

$$E(X - u \mid X > u) = \frac{\sigma - \xi u}{1 + \xi}. \tag{4}$$

The linearity of the CME plot can thus be used as an indicator of the appropriateness of the GPD model.

The shape and scale parameters of GPD may also be estimated using this property. Define the following sample mean excess (SME) function as,

$$e_n(u) = \frac{\sum_{i=1}^{n} (X_i - u)^+}{\sum_{i=1}^{n} 1_{\{X_i > u\}}} \tag{5}$$

with respect to the threshold u, where the '+' sign ensures only the positive results of $(X_i - u)$ will be taken into account. That is, the SME is the sum of the excesses over the threshold u divided by the number of data points which exceed u. The SME is an empirical estimate of the CME and ξ and σ of the GPD can be determined by the slope and intercept of the SME plot, using the following equations.

$$\text{Slope} = \frac{-\xi}{1+\xi} \tag{6}$$

$$\text{Intercept} = \frac{\sigma}{1+\xi}. \tag{7}$$

2.3 Original Data Fitting

From (2) and the fact that $F_u(y)$ converges to $G(x)$ when u is large, we have the following equation

$$G(x-u) = \frac{F(x) - F_n(u)}{1 - F_n(u)} \tag{8}$$

which can be re-arranged as

$$F(x) = (1 - F_n(u))G(x-u) + F_n(u). \tag{9}$$

Given a threshold u, the $F_n(u)$ can be estimated using $(n - N_u)/n$, where n is the sample size and N_u is the number of exceedances. In the case of $\xi \neq 0$, (9) can be simplified to

$$F(x) = 1 - \frac{N_u}{n}\left(1 - \xi \frac{x-u}{\sigma}\right)^{1/\xi} \tag{10}$$

where $x > u$.

It can be seen that (10) is also a GPD with parameters (ξ, σ', μ') where

$$\sigma' = \sigma(1 - F_n(u))^{-\xi} \tag{11}$$

$$\mu' = u + \sigma'\left((1 - F_n(u))^{\xi} - 1\right)/\xi \tag{12}$$

and ξ and σ are the fitted GPD parameters to $x - u$, where $x > u$, using the POT method.

2.4 Return Period

When the threshold is chosen sufficiently large, it is assumed that the number of exceedances N_u (where u is the threshold) has an approximate Poisson distribution with parameter λ (the rate of exceedances per year, also called the crossing rate). Hence λT is the number of exceedances in T years. Let λ_U be the number of events exceeding a very high level U. That is,

$$\lambda_U = \lambda T \cdot \Pr\{X > U\} = \lambda T(1 - F(U)). \tag{13}$$

Assume U_T is the event with the largest value in T years, and by definition $\lambda_{U_T} = 1$ (i.e. it only happens once in T years).

Now

$$\lambda_{U_T} = \lambda T(1 - F(U_T)) = 1 \tag{14}$$

so

$$F(U_T) = 1 - \frac{1}{\lambda T} \tag{15}$$

or

$$U_T = F^{-1}\left(1 - \frac{1}{\lambda T}\right) \tag{16}$$

where $F^{-1}(\cdot)$ is the inverse of the CDF of the GPD (or the GEV). For GPD, the crossing rate λ can be estimated by N_u/T_{data}, where T_{data} is the number of years for which data has been recorded. For GEV, the crossing rate λ has the value 1 if the yearly maximum is used, or 12 if the monthly maximum is used.

2.5 Quantile Estimation

The quantile estimate for GEV can be obtained by inverting (1) and using (16):

$$U_T = \begin{cases} \mu + \dfrac{\sigma}{\xi}\left\{1 - \left[-\ln\left(1 - \dfrac{1}{\lambda T}\right)\right]^{\xi}\right\} & \xi \neq 0 \\ \mu - \sigma \ln\left[-\ln\left(1 - \dfrac{1}{\lambda T}\right)\right] & \xi = 0. \end{cases} \tag{17}$$

For GPD, the quantile estimate is obtained by inverting (3):

$$U_T = \begin{cases} \mu' + \dfrac{\sigma'}{\xi}\left[1 - \dfrac{1}{(\lambda T)^{\xi}}\right] & \xi \neq 0 \\ \mu' + \sigma' \ln(\lambda T) & \xi = 0 \end{cases} \tag{18}$$

where σ' and μ' are given in (11) and (12).

2.6 Threshold Selection

The threshold should be set high enough so that the exceedances approximate Poisson arrival rate. Otherwise the distribution of selected extremes may not converge to the GPD. However, the threshold must be low enough to ensure there is enough data points left for satisfactory determination of the GPD parameters.

There are techniques available that may assist in the selection of an optimal threshold. One approach is to use the SME plots defined in (5). If the data is from a GPD, its CME would be linear with respect to u (see (4)). Hence, an appropriate threshold can be chosen by selecting the lowest value above which the SME graph is approximately a straight line.

2.7 Model Validation

An important question when considering any numerical model of a complex physical process is the reliability and appropriateness of the chosen model. Graphical methods are commonly used in model validation. Two popular methods involve plotting CDF and quantile-quantile plot (QQ-plot).

A CDF plot of the fitted GPD and the empirical distribution derived from the data will show the degree of GPD fitting towards the data.

The QQ-plot is a convenient visual diagnostic tool to examine whether a sample comes from a specific distribution. Specifically, the quantiles of an empirical distribution are plotted against the quantiles of a hypothesized distribution, in our case, the fitted GPD. If the sample (data) comes from the GPD, the QQ-plot will approximate a straight line.

Confidence intervals of the distribution parameters and return levels are also an indicator of the quality and appropriateness of the selected model. In general, smaller confidence intervals indicate a better model fit compared to the larger ones. A method of estimating confidence intervals using profile likelihood techniques has been described in Coles (2001).

2.8 Procedure For POT Method

The following procedure summarises the POT method step by step:

- Set up a threshold u assisted by SME plots;
- Fit the GPD parameters ξ, σ, μ using the data exceeding u (i.e. $x - u$);
- Verify the model using the CDF plot and the QQ-plot of the exceedances GPD and the empirical distribution (of exceedances);
- Calculate the overall fitted tail GPD parameters (ξ, σ', μ') for all data using (11) and (12);
- Verify the overall model using CDF plot of the overall fitted tail GPD and the empirical distribution (of all data);
- Estimate quantiles using (18) with the overall fitted parameters (ξ, σ', μ').

3. PERTH SEVERE WIND MODELLING

Six years of wind gust data from an automatic weather station (AWS) at Gooseberry Hill in Perth was used to demonstrate the techniques.

The GEV and GPD analyses were done by using WAFO (2000). WAFO is a MATLAB toolbox for analysis of random waves and loads, which has a sub-toolbox of extreme value analysis.

3.1 The GEV approach

The yearly maxima data contained only 5 values from each of the 5 completed record years. It is not appropriate to fit the GEV model to this limited data and consequently the GEV model is fitted to the maximum monthly values. The fitted GEV model uses parameters estimated using the methods of maximum likelihood (ML) and probability weighted moments (PWM) are plotted against the recorded monthly maximum wind gusts in Figure 1.

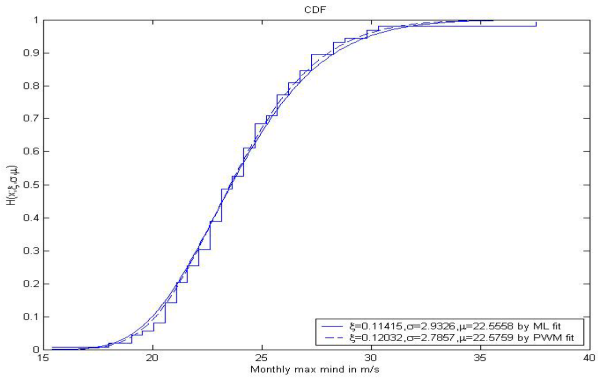

Figure 1. Gooseberry Hill – CDF plot of monthly maximum gust (steps) and fitted GEVs using ML and PWM.

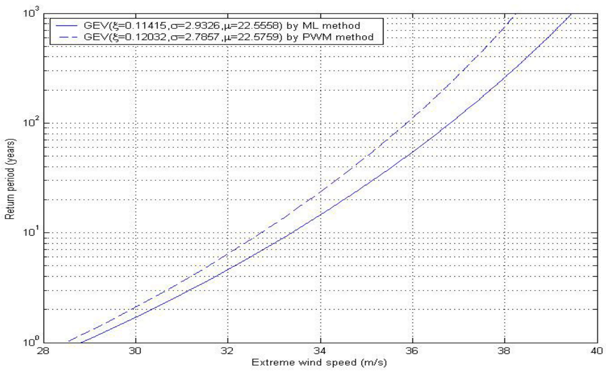

Figure 2. Gooseberry Hill – Return periods and levels.

The predicted maximum wind gusts corresponding to various return periods have been plotted in Figure 2. There is a slight difference between the GEV model fitted by different parameter estimating methods. In Figure 3, the estimated 50 and 100 year return levels are drawn with the monthly maximum gust. The predicted return levels can be compared visually with the historical records.

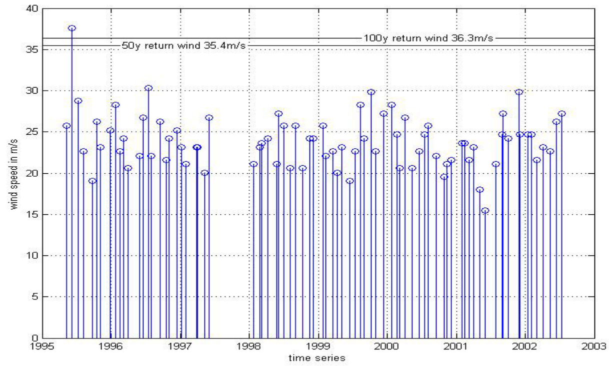

Figure 3. Gooseberry Hill – Time series plot of monthly maximum gust with estimated 50 and 100 year return levels indicated.

3.2 The POT approach

Before attempting the POT analysis an appropriate threshold value was determined using the SME plot defined by (5) (Figure 4). The SME is approximately constant (Figure 4) when $u \geq$ 20m/s. Consequently, a threshold of 20 m/s has been chosen. There are 325 daily maximum wind gusts that exceed 20 m/s and these were used to estimate the GPD parameters. This is approximately 4 times as many data points compared with the GEV's case.

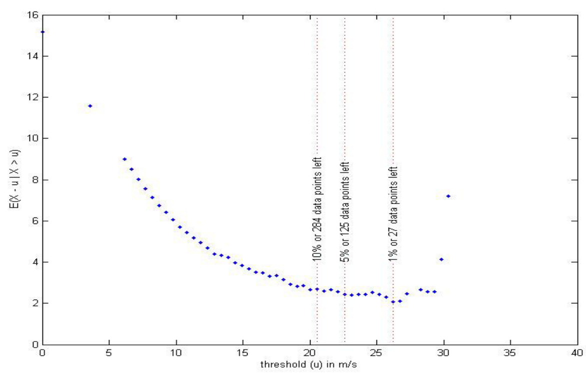

Figure 4. Gooseberry Hill – SME plot versus threshold.

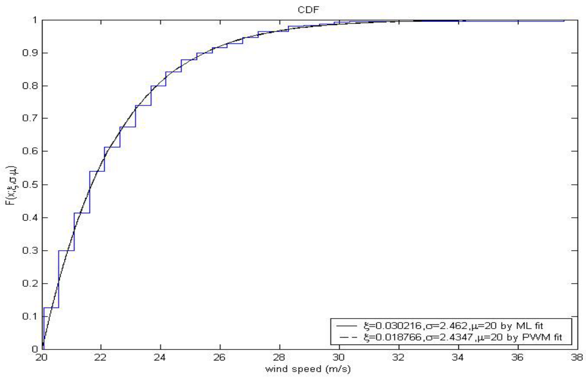

Figure 5. Gooseberry Hill – CDF plot of exceedances (steps) versus fitted GPD using exceedances with *u*=20m/s.

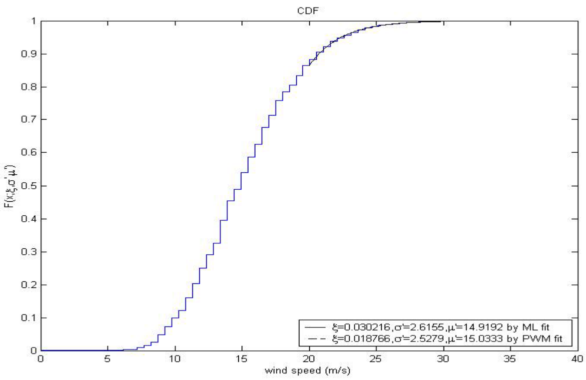

Figure 6. Gooseberry Hill – CDF plot of all daily gust versus overall fitted tail GPD model.

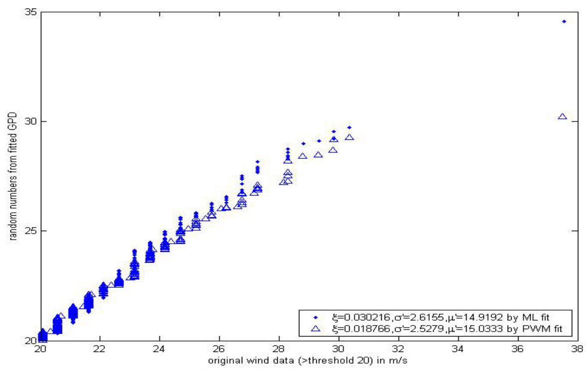

Figure 7. Gooseberry Hill – QQ-plot of exceedances and the fitted GPD with u = 20 m/s.

The CDF plots of exceedances and the fitted GPD using exceedances are shown in Figure 5. The whole record and the overall fitted tail (larger than 20 m/s) GPD are shown in Figure 6. Both figures show a very good fit using both methods (ML and PWM). The QQ-plot also suggests that both models fit the available data with the quantiles generally plotting as a straight line except one point (Figure 7). The return level plot is shown in Figure 8 together with the empirical return levels (the dots) which suggested a good fit from the GPD models. The time series of daily gust is plotted in Figure 9 with estimated 50 and 100 year level gusts shown.

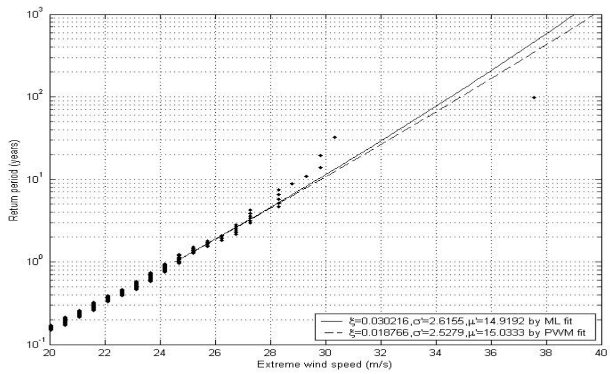

Figure 8. Gooseberry Hill – Return periods versus return levels.

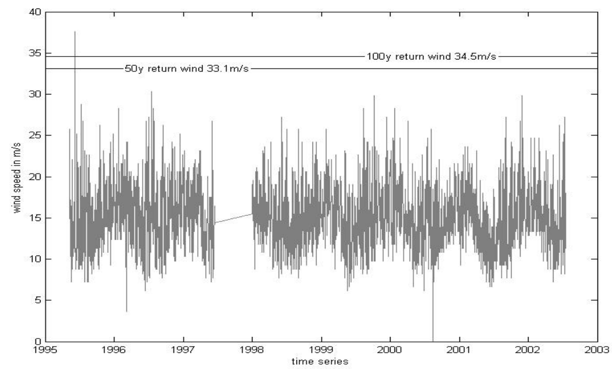

Figure 9. Gooseberry Hill – Time series plot with 50 and 100 year return level gusts indicated.

3.3 Results for Other AWS sites

The modelling of severe wind gust at up to six other AWS locations has also been carried out by fitting the yearly (or monthly) maximum gust by the GEV and the daily gust by the GPD. The initial estimates are shown in Table 1. It can be seen that the return levels estimated from the GEV are slightly larger than the ones from the GPD. Most of the shape parameters fitted were positive in both the GEV and GPD cases, which indicate that they are from Type III distributions. This result is consistent with a study of extreme wind modelling in the US (Lechner et al., 1992).

The return wind levels for Region A (including Perth) have recently been published by Standards Australia (AS/NZS 1170.2, 2002). These are based on a single site, the Perth Airport and may not be representative of the Perth area. These are also included in Table 1. It can be seen that all the inland sites have a return wind speed smaller than those recommended in the Standard. However, all the three coast sites (Ocean Reef, Rottnest Is and Swanbourne) have return wind speeds close to the Standard with one exception at Swanbourne, which is slightly larger. The data from Swanbourne may need further investigation because the GEV model fitted using the yearly maximum gust was different to the ones fitted by the monthly maximum gust and their shape parameters have different signs.

Table 1. Estimated 50 and 100 year return levels (rounded to the nearest 1 m/s) for the AWS sites in Perth.

AWS site	Data period[1]	By GEV		By GPD	
		50y	100y	50y	100y
Airport	44-02	35m/s	36m/s	33m/s	35m/s
Gooseberry	95-02	35m/s	36m/s	33m/s	35m/s
Jandakot	89-02	35m/s	36m/s	33m/s	34m/s
Ocean Reef	86-02	39m/s	40m/s	38m/s	39m/s
Pearce	67-02	36m/s	38m/s	33m/s	35m/s
Rottnest Is	84-02	39m/s	40m/s	38m/s	39m/s
Swanbourne	86-02	40m/s	41m/s	39m/s	41m/s
Wind Standard		**39**m/s	**41**m/s	**39**m/s	**41**m/s

4. CONCLUSIONS

Extreme value analysis has been carried out on daily wind gust data recorded in Perth using the GEV and GPD analyses. Results show that at coastal sites the predicted wind gust speeds are similar to these published by Standards Australia for Perth (Region A). However, all the inland sites studied have smaller wind gust speeds than the ones published by Standards Australia.

The future damages to building and infrastructures caused by those predicted return period winds can be estimated using a wind damage model that is calibrated by the historical storm damage.

5. ACKNOWLEDGEMENTS

The author wishes to thank Joe Courtney (from Australian Bureau of Meteorology, Perth Office) for his kind collaboration and providing meteorological data.

[1] Some data periods were not continuous.

6. REFERENCES

AS/NZS 1170.2, *Structural design actions, Part 2: Wind actions*, Australian/New Zealand Standard, page 13, 2002.

Balkema, A.A. and L. de Haan, Residual lifetime at great age, *Annals of Probability*, 2, 792–804, 1974.

Colse, S., *An introduction to statistical modelling of extreme values*, Springer, London, 2001.

Davison, A.C., Modelling excesses over high thresholds, with an application, *In Statistical Extremes and Applications*, (ed. J. Tiago de Oliviera), Reidel, Dordrecht, 461–482, 1984.

Embrechts, P., C. Kluppelberg and T. Mikosch, *Modelling extreme events for insurance and finance*, Springer, New York, third printing, 2001.

Fisher, R.A. and L.H.C. Tippett, Limiting forms of the frequency distribution of the largest or smallest member of a sample, Proceedings of the Cambridge Philosophical Society, 24, 180–190, 1928.

Hosking, J.R.M. and J.R. Wallis, Parameter and quantile estimation for the generalized Pareto distribution, *Technometrics*, 29(3), 339–349, 1987.

Jenkinson, A.F., The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quarterly Journal of the Royal Meteorological Society*, 81, 158–171, 1955.

Lechner, J.A., S.D. Leigh and E. Simiu, Recent approaches to extreme value estimation with application to wind speeds, Part 1: the Pickands method, *Journal of Wind Engineering and Industrial Aerodynamics*, (combined vols) 41–44, 509–519, 1992.

Palutikof, J.P., B.B. Brabson, D.H. Lister and S.T. Adcock, A review of methods to calculate extreme wind speeds, *Meteorological Applications*, 6(1), 119–132, 1999.

Pickands, J., Statistical inference using extreme order statistics, *Annals of Statistics*, 3, 119–131, 1975.

WAFO – *A MATLAB toolbox for analysis of random waves and loads,* Lund University, Sweden, homepage http://www.maths.lth.se/matstat/wafo/, 2000.