Estimation of extreme values from sampled time series

A. Naess

Centre for Ships and Ocean Structures
& Department of Mathematical Sciences
Norwegian University of Science and Technology
Trondheim, Norway

O. Gaidai

Centre for Ships and Ocean Structures
Norwegian University of Science and Technology
Trondheim, Norway

Abstract

The paper focuses on the development of a method for extreme value estimation based on sampled time series. It is limited to the case when the extreme values asymptotically follow the Gumbel distribution. The method is designed to account for statistical dependence between the data points in a rational way. This avoids the problem of declustering of data to ensure independence, which is a common problem for the peaks-over-threshold method. The goal has been to establish an accurate method for prediction of e.g. extreme wind speeds based on recorded data. The method will be demonstrated by application to both synthetic and real data. From a practical point of view, it seems to perform better than the POT and Gumbel methods, and it is applicable to nonstationary time series.

Keywords: Extreme value estimation, Sampled time series, Approximation by conditioning, Mean exceedance rate, Monte Carlo simulation.

1 Introduction

Extreme value statistics, even in applications, have very often been based on asymptotic results. This is done either by assuming that the epochal extremes, for example yearly extreme wind speeds, are distributed according to the generalized (asymptotic) extreme value distribution with unknown parameters to be estimated on the basis of the observed data. Or it is assumed that the exceedances above high thresholds follow a generalized (asymptotic) Pareto distribution with parameters to be estimated from the data, see [1]. The major problem with both of these approaches is that the asymptotic extreme value theory itself cannot be used in practice to decide to what extent it is applicable for the observed data. Hence, the assumption that an asymptotic extreme value distribution is the appropriate distribution for the observed data is based more or less on faith or convenience.

In an effort to ameliorate this situation, we have developed an approach to this problem that is less restrictive and more flexible than the ones based on asymptotic theory. In particular, it has the capability to capture subasymptotic behaviour of the data, which seems to be of some importance for accurate prediction. However, the proposed approach in its present form has one limitation, it is restricted to cases where the Gumbel distribution is the appropriate asymptotic extreme value distribution.

2 Cascade of Conditioning Approximations

Consider a stochastic process $Z(t)$, which has been observed over a time interval, $(0, T)$ say. Assume that values $X_1, \ldots, X_N$, which have been derived from the observed process, are allocated to the discrete times $t_1, \ldots, t_N$ in $(0, T)$. This could be simply the observed values of $Z(t)$ at each t_j, $j = 1, \ldots, N$, or it could be average values or peak values over smaller time intervals centered at the t_j's. Our goal in this paper is to accurately determine the distribution function of the extreme value $M_N = \max\{X_j\,; j = 1, \ldots, N\}$. Specifically, we want to estimate $P(\eta) = \text{Prob}(M_N \leq \eta)$ accurately for large values of η. An underlying premise for the development in this paper is that a rational approach to the study of the extreme values of the sampled time series is to consider exceedances of the individual random variables X_j above given thresholds, as in classical extreme value theory. The alternative approach of considering the exceedances by upcrossing of given thresholds by a continuous stochastic process has been developed in [2, 3]. The approach taken in this paper would seem to be the appropriate way to deal with the recorded data time series of, for example, the daily largest wind speeds observed at a given location.

In the following we outline possible approaches for practical implementation of a cascade of approximations based on conditioning, where the first one is a Markov-like approximation in the sense that it is a one-step memory approximation. This approximation concept is described in [4, 5].

From the definition of $P(\eta)$ it follows that

$$\begin{aligned} P(\eta) &= \text{Prob}\{X_1 \le \eta, \ldots, X_N \le \eta\} \\ &= \text{Prob}\{X_N \le \eta | X_1 \le \eta, \ldots, X_{N-1} \le \eta\}\text{Prob}\{X_1 \le \eta, \ldots, X_{N-1} \le \eta\} \\ &= \prod_{j=2}^{N} \text{Prob}\{X_j \le \eta | X_1 \le \eta, \ldots, X_{j-1}\} \cdot P(X_1 \le \eta) \end{aligned} \tag{1}$$

In general, the variables X_j are statistically dependent. Hence, instead of assuming that all the X_j are statistically independent, which leads to the classical approximation

$$P(\eta) \approx \prod_{j=1}^{N} P(X_j \le \eta), \tag{2}$$

the following Markov-like, or one-step memory, assumption will to a certain extent account for dependence between the X_j,

$$\text{Prob}\{X_j \le \eta | X_1 \le \eta, \ldots, X_{j-1} \le \eta\} \approx \text{Prob}\{X_j \le \eta | X_{j-1} \le \eta\}, \tag{3}$$

for $2 \le j \le N$. This can be extended to

$$\text{Prob}\{X_j \le \eta | X_1 \le \eta, \ldots, X_{j-1} \le \eta\} \approx \text{Prob}\{X_j \le \eta | X_{j-2} \le \eta, X_{j-1} \le \eta\}, \tag{4}$$

for $3 \le j \le N$, and so on.

Eqs. (3) and (4) represent refinements of the independence assumption. One would expect that such approximations would be increasingly more able to capture statistical dependence between neighboring data in the time series. As will be seen in the examples in the following section, $P(\eta)$ computed using Eq. (4) is quite close to the value obtained using Eq. (3). This indicates that in practice, Eq. (3) is often able to capture the effect of statistical dependence in e.g. wind speed data with good accuracy. However, there is no noticeable increase of numerical effort by using Eq. (4), or its further refinements by including three or more preceding peaks.

Combining Eq. (1) with Eq. (3), the following relation is obtained

$$P(\eta) \approx \frac{\prod_{j=2}^{N} p_{2j}(\eta)}{\prod_{j=2}^{N-1} p_{1j}(\eta)} \tag{5}$$

where we have introduced the notation $p_{kj}(\eta) = \text{Prob}\{X_{j-k+1} \le \eta, \ldots, X_j \le \eta\}$ for $j \ge k$.

It is of interest to compare the values for $P(\eta)$ obtained by using Eq. 5 as compared to Eq. 2. Now, Eq. 2 can be rewritten in the form

$$P(\eta) \approx \prod_{j=1}^{N} \Big(1 - \alpha_{1j}(\eta)\Big), \tag{6}$$

where

$$\alpha_{1j}(\eta) = \text{Prob}\{X_j > \eta\} = 1 - p_{1j}(\eta). \tag{7}$$

Then

$$P(\eta) \approx P_1(\eta) = \exp\Big(-\sum_{j=1}^{N} \alpha_{1j}(\eta)\Big) \tag{8}$$

Alternatively, Eq. 5 gives

$$P(\eta) \approx \prod_{j=2}^{N} \Big(1 - \alpha_{2j}(\eta)\Big)\, p_{1N}(\eta), \tag{9}$$

where $\alpha_{kj}(\eta) = 1 - p_{kj}(\eta)/p_{k-1,j-1}(\eta)$, for $j \geq k \geq 2$. That is

$$\alpha_{kj}(\eta) = \text{Prob}\{X_j > \eta \,|\, X_{j-k+1} \leq \eta, \ldots, X_{j-1} \leq \eta\} \tag{10}$$

denotes the exceedance probability conditional on $k-1$ previous non-exceedances. From Eq. 9 it is obtained that,

$$P(\eta) \approx P_2(\eta) = \exp\Big(-\sum_{j=2}^{N} \alpha_{2j}(\eta) - \alpha_{1N}(\eta)\Big), \tag{11}$$

since $p_{1N}(\eta) \approx \exp(-\alpha_{1N}(\eta))$.

Conditioning on two previous observations X_{j-2}, X_{j-1} preceding X_j gives

$$P(\eta) \approx P_3(\eta) = \exp\Big(-\sum_{j=3}^{N} \alpha_{3j}(\eta) - \alpha_{2N}(\eta) - \alpha_{1N}(\eta)\Big), \tag{12}$$

while conditioning on three prior observations leads to the equation

$$P(\eta) \approx P_4(\eta) = \exp\Big(-\sum_{j=4}^{N} \alpha_{4j}(\eta) - \alpha_{3N}(\eta) - \alpha_{2N}(\eta) - \alpha_{1N}(\eta)\Big), \tag{13}$$

and so on. Therefore, extreme value prediction by the conditioning approach described above reduces to estimation of (combinations) of the $\alpha_{kj}(\eta)$ functions. For most practical applications $N >> 1$, so that $(k \geq 2)$

$$P_k(\eta) \approx \exp\Big(-\sum_{j=k}^{N} \alpha_{kj}(\eta)\Big). \tag{14}$$

Going back to Eq. (8), and the definition of $\alpha_{1j}(\eta)$, it follows that $\sum_{j=1}^{N} \alpha_{1j}(\eta)$ is equal to the expected number of exceedances of the threshold η during the time interval $(0, T)$. Eq. (8) therefore expresses the approximation that the stream of exceedance events constitute a (non-stationary) Poisson process. This opens for an understanding of Eq. (11) and subsequent approximations by interpreting the expressions $\sum_{j=k}^{N} \alpha_{kj}(\eta) + \alpha_{k-1,N}(\eta) + \ldots + \alpha_{1N}(\eta) \approx$

$\sum_{j=k}^{N} \alpha_{kj}(\eta)$ as the expected effective number of (independent) exceedances provided by conditioning on $k-1$ previous observations.

3 Empirical Estimation of the Mean Exceedance Rates

It is expedient to introduce the concept of (conditional) average exceedance rates (AER) as follows,

$$\bar{\varepsilon}_k(\eta) = \frac{1}{N-k+1} \sum_{j=k}^{N} \alpha_{kj}(\eta)\,,\; k = 1, 2, \ldots \tag{15}$$

In practice there are typically two scenarios for the underlying process $Z(t)$. Either we may consider it to be a stationary process, or, in fact, even an ergodic process. The alternative is to view $Z(t)$ as a process that depends on certain parameters whose variation in time may be modelled as an ergodic process in its own right. For each set of values of the parameters, the premise is that $Z(t)$ can be modelled as an ergodic process. This would be the scenario that can be used to model long-term statistics [6].

For both these scenarios, the empirical estimation of the conditional AER $\bar{\varepsilon}_k(\eta)$ proceeds in a completely analogous way by counting the total number of favourable incidents, that is, exceedances combined with the requisite number of preceding non-exceedances, for the total data time series and then finally dividing by $N-k+1 \approx N$. This can be shown for the long-term situation by using a similar analysis as in [6].

A few more details on the numerical estimation of $\bar{\varepsilon}_k(\eta)$ for $k \geq 2$ are useful. We start by introducing the following random functions,

$$A_{kj}(\eta) = \mathbf{1}\{X_j > \eta, X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}\,,\; j = k, \ldots, N,\; k = 2, 3, \ldots \tag{16}$$

and

$$B_{kj}(\eta) = \mathbf{1}\{X_{j-1} \leq \eta, \ldots, X_{j-k+1} \leq \eta\}\,,\; j = k, \ldots, N,\; k = 2, \ldots\,, \tag{17}$$

where $\mathbf{1}\{\mathcal{A}\}$ denotes the indicator function of some event $\mathcal{A}$. Then

$$\alpha_{kj}(\eta) = \frac{\mathrm{E}[A_{kj}(\eta)]}{\mathrm{E}[B_{jk}(\eta)]}\,,\; j = k, \ldots, N,\; k = 2, \ldots\,, \tag{18}$$

where $\mathrm{E}[\cdot]$ denotes the expectation operator. Assuming an ergodic process, then obviously $\overline{\varepsilon}_k(\eta) = \alpha_{kk}(\eta) = \ldots = \alpha_{kN}(\eta)$, and it may be assumed that for the time series at hand

$$\overline{\varepsilon}_k(\eta) = \lim_{N\to\infty} \frac{\sum_{j=k}^{N} A_{kj}(\eta)}{\sum_{j=k}^{N} B_{jk}(\eta)}\,. \tag{19}$$

Clearly, $\lim_{\eta\to\infty} \sum_{j=k}^{N} B_{jk}(\eta) = N-k+1 \approx N$. Hence, $\lim_{\eta\to\infty} \tilde{\varepsilon}_k(\eta)/\overline{\varepsilon}_k(\eta) = 1$, where

$$\tilde{\varepsilon}_k(\eta) = \lim_{N\to\infty} \frac{\sum_{j=k}^{N} A_{kj}(\eta)}{N-k+1}\,. \tag{20}$$

In the following we shall use $\tilde{\varepsilon}_k(\eta)$ instead of $\overline{\varepsilon}_k(\eta)$ for $k \geq 2$. The advantage of using the modified conditional AER function $\tilde{\varepsilon}_k(\eta)$ for $k \geq 2$ is that it is easier to use for non-stationary or long-term statistics than $\overline{\varepsilon}_k(\eta)$. Since our focus is on the values of the AER at the extreme levels, we may use any function that provides correct estimates of the AER function at the extreme levels.

For both stationary and non-stationary time series, the sample estimate of $\tilde{\varepsilon}_k(\eta)$ would be,

$$\hat{\varepsilon}_k(\eta) = \frac{1}{R} \sum_{r=1}^{R} \hat{\varepsilon}_k^{(r)}(\eta)\,, \tag{21}$$

where R is the number of realizations (samples), and

$$\hat{\varepsilon}_k^{(r)}(\eta) = \frac{\sum_{j=k}^{N} A_{kj}^{(r)}(\eta)}{N-k+1}\,, \tag{22}$$

where the index (r) refers to realization no. r.

It is of interest to note what events are actually counted for the calculation of the various $\hat{\varepsilon}_k(\eta)$, $k \geq 2$. Let us start with $\hat{\varepsilon}_2(\eta)$. It follows from the definition of $\tilde{\varepsilon}_2(\eta)$ that $\tilde{\varepsilon}_2(\eta)\,(N-1)$ can be interpreted as the expected number of exceedances above the level η satisfying the condition that an exceedance is counted only if it is immediately preceded by a non-exceedance. A reinterpretation of this is that $\hat{\varepsilon}_2(\eta)\,(N-1)$ equals the average number of clumps of exceedances above η for the realizations considered, where a clump of exceedances is defined as a maximum number of consecutive exceedances above η. In general, $\hat{\varepsilon}_k(\eta)\,(N-1)$ then equals the average number of clumps of exceedances above η separated by at least $k-1$ non-exceedances.

Now, let us look at the problem of estimating a confidence interval for $\tilde{\varepsilon}_k(\eta)$. The sample standard deviation $\hat{s}_k(\eta)$ can be estimated by the standard formula,

$$\hat{s}_k(\eta)^2 = \frac{1}{R-1}\sum_{r=1}^{R}\Big(\hat{\varepsilon}_k^{(r)}(\eta) - \hat{\varepsilon}_k(\eta)\Big)^2 . \tag{23}$$

Assuming that realizations are independent, for a suitable number R, e.g. $R \geq 20$, Eq. (23) leads to a good approximation of the 95 % confidence interval $\mathrm{CI} = \big(\mathrm{CI}^-(\eta), \mathrm{CI}^+(\eta)\big)$ for the value $\tilde{\varepsilon}_k(\eta)$, where

$$\mathrm{CI}^{\pm}(\eta) = \hat{\varepsilon}_k(\eta) \pm 1.96\,\hat{s}_k(\eta)/\sqrt{R}\,. \tag{24}$$

The approach to extreme value prediction presented in this paper derives from an assumption about the sampled time series to be used as a basis for prediction. This assumption derives from an underlying premise concerning the relevant asymptotic extreme value distribution, which is assumed here to be of Gumbel type. The implication of this assumption on the possible sub-asymptotic functional forms of $\tilde{\varepsilon}_k(\eta)$ cannot easily be decided. However, using the asymptotic form as a guide, it is assumed that the behaviour of the mean exceedance rate in the tail is dominated by a function of the form $\exp\{-a(\eta-b)^c\}$ ($\eta \geq \eta_1 \geq b$) where a, b and c are suitable constants, and η_1 is an appropriately chosen tail level. Hence, it will be assumed that

$$\tilde{\varepsilon}_k(\eta) \approx q_k(\eta)\,\exp\{-a_k(\eta-b_k)^{c_k}\}\,,\;\; \eta \geq \eta_1, \tag{25}$$

where the function $q_k(\eta)$ is slowly varying compared with the exponential function $\exp\{-a_k(\eta - b_k)^{c_k}\}$ and a_k, b_k, and c_k are suitable constants, that in general will be dependent on k. Note that the value $c_k = 1$ corresponds to the asymptotic case.

From Eq. (25) it follows that,

$$-\log\Big|\log\left(\tilde{\varepsilon}_k(\eta)/q_k(\eta)\right)\Big| \approx -c_k \log(\eta - b_k) - \log(a_k). \tag{26}$$

Therefore, under the assumptions made, a plot of $-\log\Big|\log\left(\tilde{\varepsilon}_k(\eta)/q_k(\eta)\right)\Big|$ versus $\log(\eta - b_k)$ will exhibit an almost perfectly linear tail behaviour.

It is realized that if the function $q_k(\eta)$ could be replaced by a constant value, q_k say, we would immediately be in a position to apply a linear extrapolation strategy. In general, $q_k(\eta)$ is not constant, but its variation in the tail region is usually sufficiently slow to allow for its replacement by a constant. A similar approach was successfully used by the authors for mean up-crossing rate estimation for extreme value analysis of the response processes related to different dynamic systems, cf. [2, 3]. Details and examples are discussed later in this paper.

Since the linearity in the plotting procedure described above is dependent on an appropriate choice of the parameters (b_k, q_k), it is important to discuss this issue in some detail.

First, we cut from consideration the very tail of the data, where uncertainty is high. As a practical procedure we suggest to neglect data points, where the relative confidence band width is greater than some constant δ, that is,

$$\frac{1.96\hat{s}_k(\xi)/\sqrt{R}}{\tilde{\varepsilon}_k(\eta)} > \delta \tag{27}$$

where the value chosen for δ is dependent on the actual 'roughness' of the data tail, but its value would typically be in the interval $(0.5, 1)$. Next, we come to the estimation of the predicted response level and its 95% confidence interval.

First, the tail marker η_1 is identified from visual inspection of the log plot $(\eta, \ln \tilde{\varepsilon}_k(\eta))$. The value chosen for η_1 corresponds to the beginning of regular tail behaviour in a sense to be discussed below. Next, initial estimates for b and q are found by the procedure to linearize the tail on the transformed scale. Occasionally, some care should be exercised if an optimization algorithm is used to determine b and q. It may happen that the optimization is actually almost ill defined in practice in the sense that the algorithm may have difficulties locating an optimal (b, q) pair. It has been observed that in such cases the optimal value of b may for example seem to be at a very large negative value. It can be shown that a good approximation would then often be obtained by putting $b = 0$ and $c = 1$. This can be partly explained by noting that in the special case that $c = 1$, the is no unique optimal pair of parameters (b, q). In fact, it is recognized that an infinity of (b, q)-pairs exist that result in exactly the same AER. The initial value of the parameters a and c would generally be determined from the initial 'optimal' straight line $-cx - \log a$ $(x = \log(\eta - b))$ approximating the data tail on the transformed plot (26).

Instead of doing the optimization directly on the loglog-log plot, which is appealing in the sense that it only involves two parameters, a more robust optimization may in fact be obtained by doing it on the log plot even if the optimization must be carried out with respect to all four parameters a, b, c, q. Our experience has been that the Levenberg-Marquardt least squares optimization method is well suited for the task [7]. The mean square error function to be minimized is written as

$$\sum_{j=1}^{N} w_j \left| \log \hat{\nu}^+(\eta_j) - \log q + a(\eta_j - b)^c \right|^2 , \tag{28}$$

where $w_j = \Big(\log CI^+(\eta_j) - \log CI^-(\eta_j) \Big)^{-2}$ denotes a weight factor that puts more emphasis on the more reliable data points. The choice of weight factor is, of course, to some extent arbitrary, and if it is considered more appropriate to put a stronger emphasis on the larger data, this can be simply achieved by replacing the exponent -2 by, for example, -1 in the definition of w_j.

The practical approach currently adopted is to get a first idea of the values of the parameters a, b, c, q by having a look at the loglog-log plot. These values may then be used as starting values for the Levenberg-Marquardt algorithm. For estimation of the confidence interval for the predicted return value provided by the optimal curve, the empirical confidence band is reanchored to the optimal curve. The range of fitted curves that stay within the reanchored confidence band will determine an optimized confidence interval of the predicted return value. As a final point, it has been observed that the predicted return value is not very sensitive to the choice of η_1.

To offer a comparison of the predictions obtained by the method proposed in this paper with those obtained by other methods, we shall use the predictions given by the two methods that seem to be most favored by practitioners, the Gumbel method and the peaks-over-threshold method.

4 The Gumbel Method

The Gumbel method is based on recording epochal extreme values and fitting these values to a corresponding Gumbel distribution. By assuming that the recorded extreme value data are Gumbel distributed, then representing the obtained data set of extreme values as a Gumbel probability plot should ideally result in a straight line. In practice, one cannot expect this to happen, but on the premise that the data follow a Gumbel distribution, a straight line can be fitted to the data. Due to its simplicity, a popular method for fitting this straight line is the method of moments. That is, writing the Gumbel distribution of the extreme value M_N as

$$\text{Prob}(M_N \leq \eta) = \exp\Big\{-\exp\Big(-a(\eta - b)\Big)\Big\}, \tag{29}$$

it is known that the parameters a and b are related to the mean value m_M and standard deviation σ_M of $M(T)$ as follows: $b = m_M - 0.57722\,a^{-1}$ and $a = 1.28255/\sigma_M$ [8]. The estimates of m_M and σ_M obtained from the available sample therefore provides estimates of a and b, which leads to the fitted Gumbel distribution by the moment method.

Typically, a specified fractile value of the fitted Gumbel distribution is then extracted and used in a design consideration. To be specific, let us assume that the requested fractile value is the $100(1-\alpha)\%$ fractile, where α is usually a small number, for example $\alpha = 0.1$. To quantify the uncertainty associated with the obtained $100(1-\alpha)\%$ fractile value based on a sample of size $\tilde{N}$, the 95% confidence interval of this value is often used. A good estimate of this confidence interval can be obtained by using a parametric bootstrapping method [9, 10]. In our context, this simply means that the initial sample of $\tilde{N}$ extreme values is assumed to have been generated from an underlying Gumbel distribution, whose parameters are, of course, unknown. If this Gumbel distribution had been known, it could have been used to generate a large number of (independent) samples of size $\tilde{N}$. For each sample, a new Gumbel distribution would be fitted and the corresponding $100(1-\alpha)\%$ fractile value identified. If the number of samples had been large enough, then a very accurate estimate of the 95% confidence interval on the $100(1-\alpha)\%$ fractile value based on a sample of size $\tilde{N}$ could be found. Since the true parameter values of the underlying Gumbel distribution are unknown, they are replaced by the estimated values obtained from the initial sample. This fitted Gumbel distribution is then used as described above to provide an approximate 95% confidence interval. Note that the assumption that the initial $\tilde{N}$ extreme values are actually generated with good approximation from a Gumbel distribution cannot easily be verified in general, which is a drawback of this method. Compared with the POT method, the Gumbel method would also seem to use much less of the information available in the data. This may explain why the POT method has become increasingly popular over the past years, but the Gumbel method is still widely used in practice.

5 The Peaks-Over-Threshold Method

The Generalized Pareto Distribution

The POT method is based on what is called the generalized Pareto (GP) distribution (defined below) in the following manner: It has been shown [11] that asymptotically, the excess values above a high level will follow a GP distribution if and only if the parent distribution belongs to the domain of attraction of one of the extreme value distributions. The assumption of a Poisson process model for the exceedance times combined with GP distributed excesses can be shown to lead to the generalized extreme value (GEV) distribution for corresponding extremes, see below. The expression for the GP distribution is

$$G(y) = G(y; a, c) = \text{Prob}(Y \leq y) = 1 - \left(1 + c\,\frac{y}{a}\right)_{+}^{-1/c} . \tag{30}$$

Here $a > 0$ is a scale parameter and c ($-\infty < c < \infty$) determines the shape of the distribution. $(z)_{+} = \max(0, z)$.

The asymptotic result referred to above implies that Eq. (30) can be used to represent the conditional cumulative distribution function of the excess $Y = X - u$ of the observed variate X over the threshold u, given that $X > u$ for u sufficiently large [11]. The cases $c > 0$, $c = 0$ and $c < 0$ correspond to Fréchet (Type II), Gumbel (Type I), and reverse Weibull (Type III) domains of attraction, respectively, cf. section below.

For $c = 0$, which corresponds to the Gumbel extreme value distribution, the expression between the parentheses in Eq. (30) is understood in a limiting sense as,

$$G(y) = G(y; a, 0) = \exp(-y/a)\,. \tag{31}$$

Return Periods

The return period R of a given wind speed, in years, is defined as the inverse of the probability that the specified wind speed will be exceeded in any one year. If λ denotes the mean exceedance rate of the threshold u per year (i.e., the average number of data points above the threshold u per year), then the return period R of the value of X corresponding to the level $x_R = u + y$ is given by the relation

$$R = \frac{1}{\lambda\,\mathrm{Prob}(X > x_R)} = \frac{1}{\lambda\,\mathrm{Prob}(Y > y)} \tag{32}$$

Hence, it follows that

$$\mathrm{Prob}(Y \le y) = 1 - 1/(\lambda R)\,. \tag{33}$$

Invoking equation (1) for $c \neq 0$ leads to the result

$$x_R = u - a[1 - (\lambda R)^c]/c\,. \tag{34}$$

Similarly, for $c = 0$, it is found that,

$$x_R = u + a\ln(\lambda R)\,, \tag{35}$$

where u is the threshold used in the estimation of c and a.

The de Haan Estimators

Let n denote the total number of data points, while the number of observations above the threshold value u is denoted by k. The threshold u then represents the $(k+1)$-th highest data point(s). An estimate for λ is $\hat{\lambda} = k/n_{yrs}$, where

n_{yrs} denotes the length of the record in years. The highest, second highest, ... , k-th highest, $(k+1)$-th highest variates are denoted by X^*_n, X^*_{n-1}, ...,$X^*_{n-(k+1)}$, $X^*_{n-k} = u$, respectively.

The parameter estimators proposed by [12] are based on the following two quantities,

$$H_{k,n} = \frac{1}{k}\sum_{i=0}^{k-1}\{\ln(X^*_{n-i}) - \ln(X^*_{n-k})\}\,, \tag{36}$$

and

$$H^{(2)}_{k,n} = \frac{1}{k}\sum_{i=0}^{k-1}\{\ln(X^*_{n-i}) - \ln(X^*_{n-k})\}^2\,. \tag{37}$$

Estimators for a and c are then given by the relations

$$\hat{a} = \rho\, X^*_{n-k}\, H_{k,n} = \rho\, u\, H_{k,n}\,, \tag{38}$$

and

$$\hat{c} = H_{k,n} + 1 - \frac{1}{2}\left\{1 - \frac{(H_{k,n})^2}{H^{(2)}_{k,n}}\right\}^{-1}\,, \tag{39}$$

where $\rho = 1$ if $\hat{c} \geq 0$, while $\rho = 1 - \hat{c}$ if $\hat{c} < 0$.

Subject to general conditions on the underlying probability law, [12] showed that $\hat{a} \to a$ and $\hat{c} \to c$ as $n \to \infty$ (in probability).

The Hill estimators [13] are closely related to the de Haan estimators. Their application to problems similar to the ones discussed in this paper were investigated by [14]. The conclusion was that the Hill estimators lead to results that are quite similar to those provided by the de Haan estimators. Since the Hill estimators require considerably higher numerical efforts than the de Haan estimators, we have therefore chosen to exclude the Hill estimators from the present study. Moment estimators also represent an alternative to the estimation of distribution parameters, and apparently they may provide more accurate results than obtained by using de Haan estimators [15]. In the present paper we have chosen to apply de Haan estimators as these are commonly known and extensively used. However, it may be noted that questions have been asked about the applicability of these estimators for limited sets of data [16].

6 Synthetic data

In this section we illustrate the philosophy of tail linearization and also the 95% CI estimation. We consider 20 years of synthetic wind speed data, amounting to 2000 data points, which is not much for detailed statistics. However, this case may represent a real situation when nothing but a limited data sample is available. In this case it is crucial to provide extreme value estimates utilizing all data available. As we shall see, the tail extrapolation technique is significantly better than asymptotic methods such as POT or Gumbel.

The extreme value statistics will first be analyzed by application to synthetic data [15], for which the exact extreme values can be calculated. In particular, it is assumed that the underlying (normalized) stochastic process $Z(t)$ is stationary and Gaussian with mean value zero and standard deviation equal to one. It is also assumed that the mean zero up-crossing rate $\nu^+(0)$ is such that the product $\nu^+(0)T = 10^3$ where $T = 1$ year, which seems to be typical for the wind speed process. Using the Poisson assumption, the distribution of the yearly extreme value of $Z(t)$ is then calculated by the formula

$$F^{1\mathrm{yr}}(\eta) = \exp\left\{-\nu^+(\eta)T\right\} = \exp\left\{-10^3 \exp\left(-\frac{\eta^2}{2}\right)\right\}, \tag{40}$$

where $T = 1$ year and $\nu^+(\eta)$ is the mean up-crossing rate per year, η is the scaled wind speed. The 100-year return period value $\eta^{100\mathrm{yr}}$ is then calculated from the relation $F^{1\mathrm{yr}}(\eta^{100\mathrm{yr}}) = 1 - 1/100$, which gives $\eta^{100\mathrm{yr}} = 4.80$.

The Monte Carlo simulated data to be used for the synthetic example are generated based on the observation that the peak events extracted from measurements of the wind speed process, are usually separated by 3-4 days. This is done to obtain approximately independent data, cf. [17]. In accordance with this, peak event data are generated from the extreme value distribution

$$F^{3\mathrm{d}}(\eta) = \exp\left\{-q \exp\left(-\frac{\eta^2}{2}\right)\right\}\,, \tag{41}$$

where $q = \nu^+(0)T = 10$, which corresponds to approximately $T = 3.65$ days.

Since the data points (i.e. $T = 3.65$ days maxima) are independent, $\overline{\varepsilon}_k(\eta)$ is independent of k. Therefore we put $k = 1$. Fig. 1 presents a plot of the average exceedance rate (AER) $\hat{\varepsilon}_1$ for 20 years of simulated data. Since we have 100 discrete time steps in one year, the data amounts to 2000 data points. The transformed plot (26) is a loglog-log plot $-\log|\log\hat{\varepsilon}_1(\eta) - d|$ versus $\log(\eta - b)$, where $d = -\log q$. The parameters q and b were found by optimization. From the simulated data the predicted 100 year return level is $\tilde{\eta}^{100\mathrm{yr}} = 4.58$, while the exact value $\eta^{100\mathrm{yr}} = 4.80$, see Fig. 1. Estimation of the 95% confidence interval gives $(4.44,\ 5.02)$. The transformed plot is shown in Fig. 2. Fig. 3 presents the analytical solution for $\overline{\varepsilon}_1$ on the transformed scale (26).

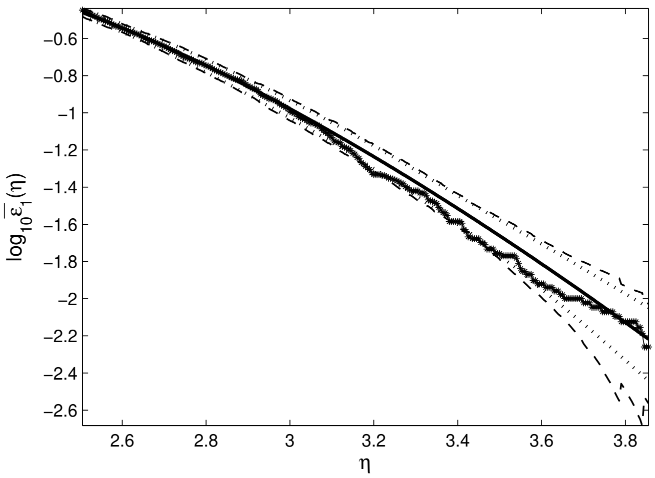

Fig. 1. Synthetic data AER $\hat{\varepsilon}_1$, Monte Carlo simulation (*); optimized curve fit (—); empirical 95% confidence band (- -); optimized confidence band (· · ··). $\eta_1 = 2.5$

Fig. 4 presents POT predictions for different threshold numbers based on de

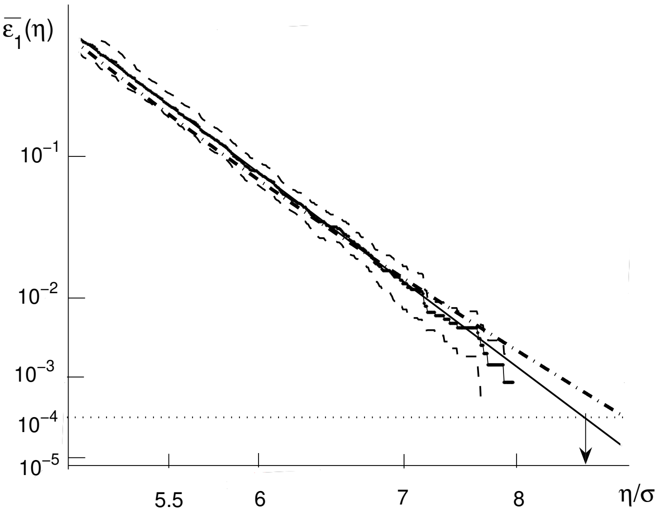

Fig. 2. Synthetic data AER $\hat{\varepsilon}_1$. Transformed plot; q=0.513; b=2.22; η_1 = 2.5. $\sigma = 0.52$. Monte Carlo simulation (*); optimized curve fit (—); exact result (−·)

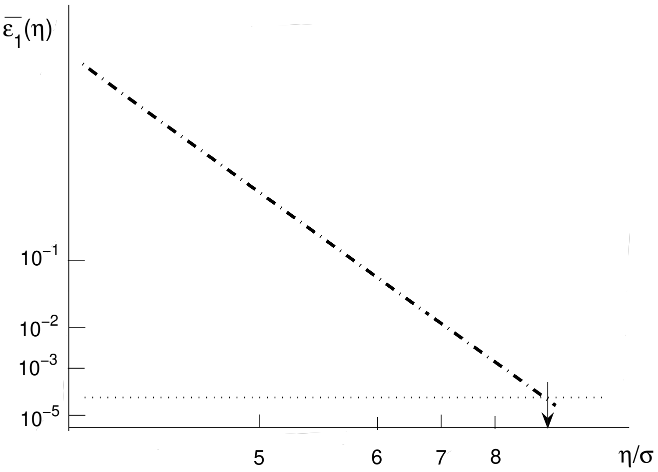

Fig. 3. Exact AER $\bar{\varepsilon}_1$. Transformed plot q=0.72; b=1.96; η_1 = 2.18. E = 2.34, $\sigma = 0.52$. Exact 100 year level = 4.80.

Haan estimators. In Fig. 5 is shown the parametrically bootstrapped PDF of the POT prediction for threshold number $n = 140$ based on de Haan estimators. The predicted value is 4.56, while the 95% confidence interval is (4.04, 5.00). The same data set as in Figs. 3 and 1 was used. It is seen that

POT tends to underestimate the exact value, which is 4.80. A clear advantage of the AER method compared to POT is seen when comparing 95% confidence intervals.

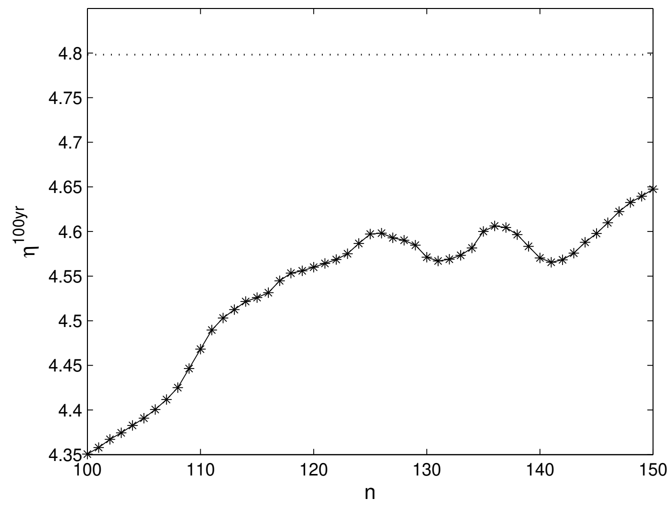

Fig. 4. The point estimate $\tilde{\eta}^{100\mathrm{yr}}$ of the 100-year return period based on 25 years synthetic data as a function of the number n of data points above threshold. Return level estimate $= 4.56$ at $n = 140$.

In order to get more insight into AER and POT statistics, 10 independent 20 yr MC simulations where done. Table 1 compares predicted values and confidence intervals. It must be mentioned that when analyzing a lot of data samples, POT as implemented here, generally underestimates the 100 yr return level value. It is seen that AER extrapolation gives consistently more accurate results than POT.

7 Measured wind speed data

In this section we analyze wind speed data, measured at Ørland Airport, Norway. Extreme wind speed prediction is an important issue for design of

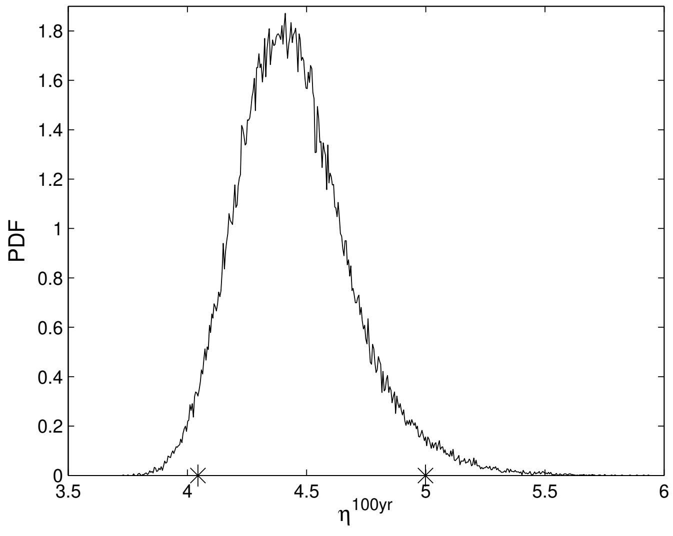

Fig. 5. Bootstrapped PDF of $\tilde{\eta}^{100\mathrm{yr}}$ by de Haan's estimator for 100000 bootstrap samples. $n = 140$. Stars indicate 95% confidence interval, CI = (4.02, 5.01).

structures exposed to the weather variations. Significant efforts have been devoted to the problem of predicting extreme wind speeds on the basis of measured data by various authors over several decades, see e.g. [18, 19, 20, 21] for extensive references to previous work.

Hourly maximum gust wind was recorded during 7 years 1998-2004. The objective is to estimate 100 year return wind speed. Along with hourly 7 years data, 20 years 1985-2004 daily and yearly wind speed maxima were available from Ørland station. Yearly maxima gives an opportunity to compare Gumbel predictions based on 20 yearly maxima with the results obtained by the extrapolation technique based on 7 years of hourly data and 20 years of daily maxima. Variation in the wind speed caused by seasonal and daily variations in the wind climate during the year makes the wind speed a non-stationary process on the scale of months. Moreover due to global climate change yearly statistics may vary on the scale of years. The latter is, however, a slow process and for the purpose of long-term prediction we assume here that within a time

Table 1

Return level estimates and 95% CI comparison

No.	AER $\hat{\eta}^{100}$	AER CI	POT $\hat{\eta}^{100}$	POT CI
1	4.71	(4.44, 5.02)	4.56	(4.02, 5.01)
2	4.97	(4.75, 5.21)	4.23	(3.93, 4.45)
3	4.71	(4.53, 4.91)	4.94	(4.22, 5.92)
4	4.89	(4.57, 5.22)	4.39	(4.09, 4.66)
5	4.92	(4.62, 5.27)	4.22	(3.87, 4.59)
6	5.05	(4.75, 5.37)	4.54	(4.11, 5.86)
7	4.98	(4.67, 5.37)	5.47	(4.31, 7.16)
8	4.55	(4.35, 4.73)	4.55	(4.12, 4.90)
9	4.86	(4.64, 5.08)	4.73	(4.14, 5.25)
10	4.84	(4.55, 5.15)	5.09	(4.26, 6.13)
Average	4.85	(4.60, 5.00)	4.66	(4.11, 5.39)

span of 100 years a quasi-stationary model of the wind speeds applies.

Fig. 6 highlights the cascade of conditional AER estimates $\tilde{\varepsilon}_1 \ldots \tilde{\varepsilon}_5$ for the case of 7 years of hourly data. Here $\tilde{\varepsilon}_5$ is considered to represent the final converged results. By 'converged' we mean that $\tilde{\varepsilon}_4 \approx \tilde{\varepsilon}_5$ in the tail, so that there is no need to consider conditioning of an even higher order than 5. Fig. 6 reveals a rather significant statistical dependence between consecutive data, which is clearly reflected in the effect of conditioning on previous data values. It is also interesting to observe that this effect is almost fully captured already by $\tilde{\varepsilon}_2$, that is, by conditioning only on the value of the previous data point.

Subsequent conditioning on more than one previous value does not lead to substantial changes in AER values.

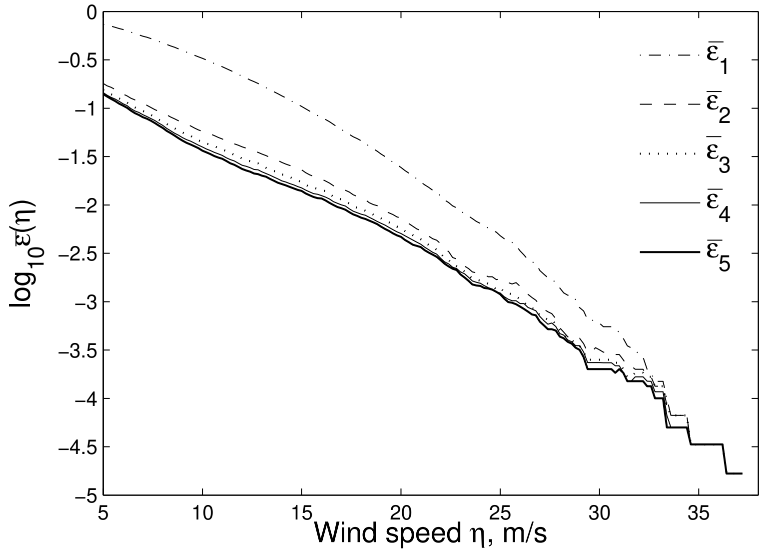

Fig. 6. Ørland wind speed statistics, 7 years hourly data. Comparison between AER estimates for different degrees of conditioning on the decimal log scale.

In Fig. 7 is shown the results of parametric estimation of the return value and its 95% CI for 7 years of hourly maxima. The corresponding transformed plot is shown in Fig. 8.

It is clearly seen that linear extrapolation fits well to the transformed data. The predicted 100 year return speed is $\eta^{100\text{yr}} = 44.45$ m/s with 95% confidence interval $(41.65, 44.56)$. We realize that $R = 7$ years of data may not be enough to guarantee Eq. (24), since we required $R \geq 20$. Nevertheless, even with $R = 7$, Eq. (24) is probably a reasonably good engineering estimate of the confidence interval.

Fig. 9 presents a Gumbel plot based on the 7 yearly extremes extracted from the 7 years of hourly data. The Gumbel prediction $\eta^{100\text{yr}} = 38.81$ m/s, with a parametric bootstrapped 95% confidence interval equal to (35.46, 42.39) m/s.

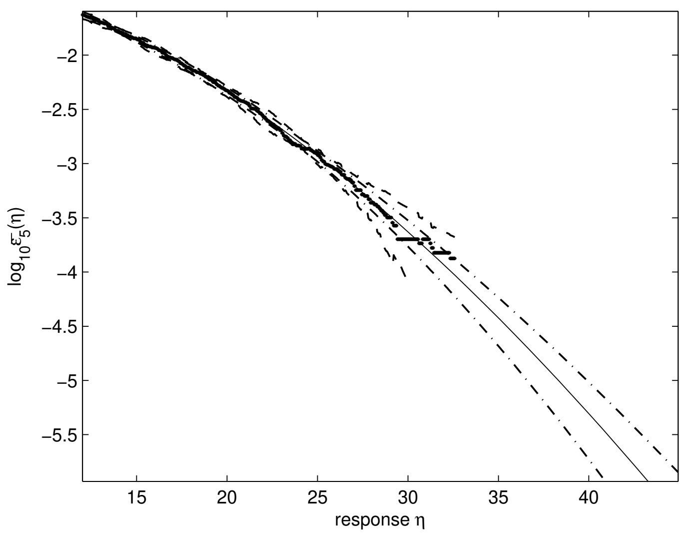

Fig. 7. Ørland wind speed statistics, 7 years hourly data. $\hat{\varepsilon}_5(\eta)$ (*); Optimized curve fit (—); Empirical 95% confidence band (- -); Optimized confidence band ($\cdot\cdot\cdot\cdot$); Predicted 100 year return level estimate = 44.45 m/s. Predicted 95% CI = (41.04, 48.30) m/s. η_1 = 15 m/s.

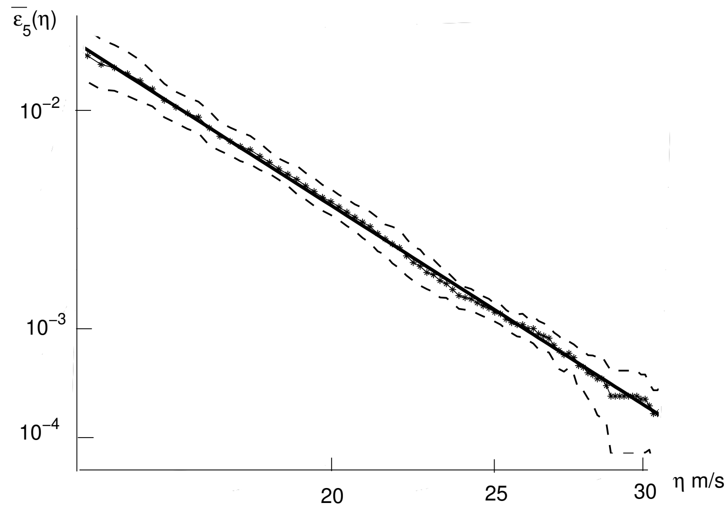

Fig. 8. Ørland wind speed statistics, 7 years hourly data. Transformed plot of $\hat{\varepsilon}_5(\eta)$, q=0.036; b=8.84; η_1 = 15 m/s.

Fig. 10 shows the bootstrap PDF based on 100000 samples.

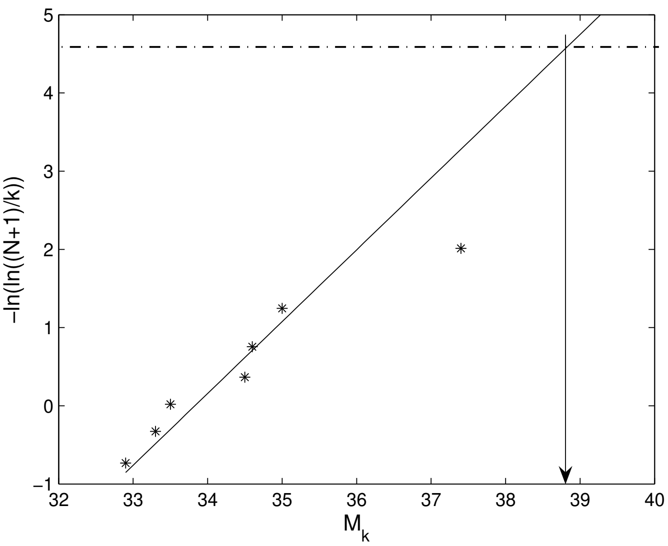

Fig. 9. Ørland wind speed statistics, 7 years of hourly data. Gumbel plot. Return level prediction = 38.81 m/s, CI = (35.46, 42.39) m/s.

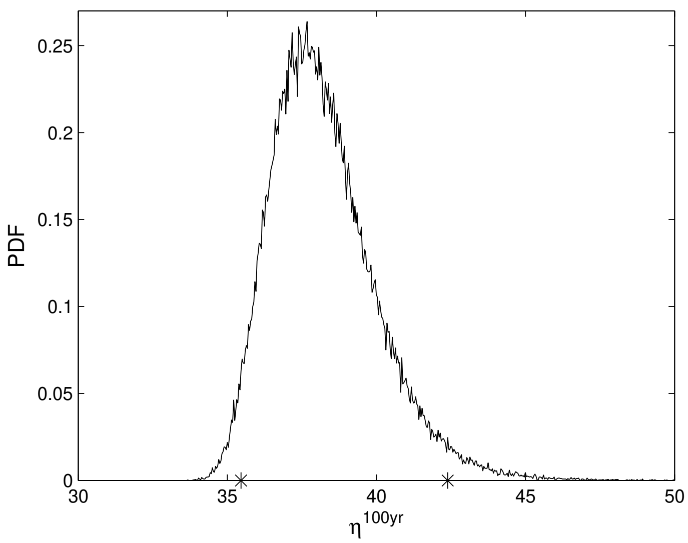

Fig. 10. 7 years hourly data. Bootstrapped PDF of $\eta^{100\mathrm{yr}}$ from Gumbel distribution for 100000 bootstrap samples. Stars indicate 95% confidence interval (35.46, 42.39) m/s.

Fig. 11 presents POT predictions for different threshold numbers based on de Haan estimators for the case of 7 years of hourly data. In Fig. 12 is shown the parametrically bootstrapped PDF of the POT prediction for threshold number $n = 50$ based on de Haan estimators for 7 years of hourly data. The POT prediction is $\eta^{100\mathrm{yr}} = 38.50$ m/s, while the bootstrapped 95% confidence interval is found to be (34.26, 45.84) m/s based on 100000 generated samples.

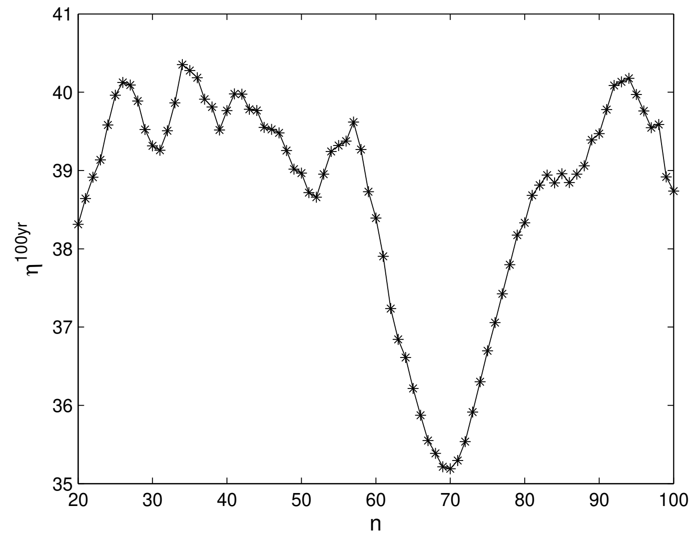

Fig. 11. 7 years hourly data. The point estimate $\eta^{100\mathrm{yr}}$ of the 100-year return level based on 7 years hourly data as a function of the number n of data points above threshold. Return level estimate = 38.97 m/s, $n = 50$.

Fig. 13 highlights the cascade of conditional AER estimates $\tilde{\varepsilon}_1 \ldots \tilde{\varepsilon}_4$ for the case of 20 years of daily maximum wind speeds. Here $\tilde{\varepsilon}_4$ was taken as the final converged results. By comparing with Fig. 6, it is seen that the effect of conditioning is weaker for the present case than for the previous one. This is to be expected since we are now dealing with daily data, as opposed to hourly data. Considering the substantial difference in time separation between the data points in the two sets of data, it is to be expected that the statistical dependence is weaker in the present case due to the much longer time separation.

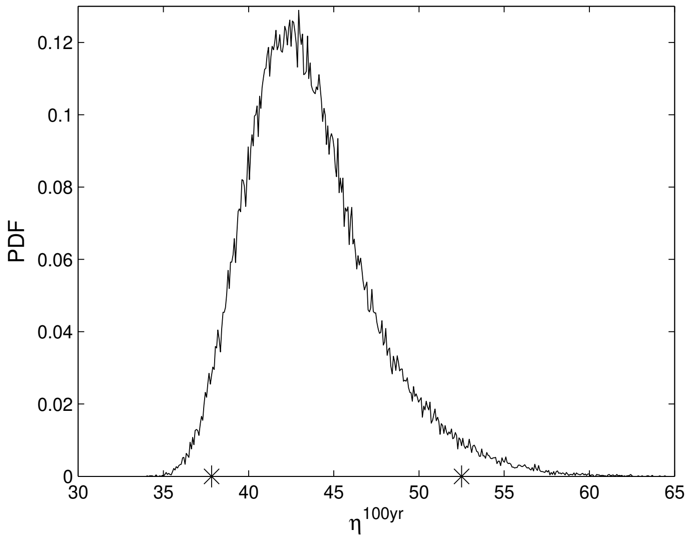

Fig. 12. 7 years hourly data. Bootstrapped PDF of $\eta^{100\text{yr}}$ by de Haan's estimator for 100000 bootstrap samples. $n = 50$. Stars indicate 95% confidence interval (37.83, 52.50).

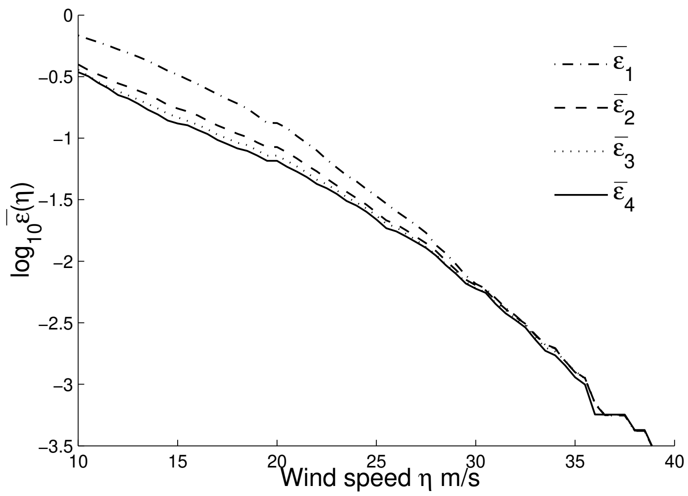

Fig. 13. Ørland wind speed statistics, 20 years daily data. Comparison between AER estimates for different degrees of conditioning on the decimal log scale.

In Fig. 14 is shown the results of parametric estimation of the return value and its 95% CI for 20 years of daily maxima. The corresponding transformed plot is presented in Fig. 15.

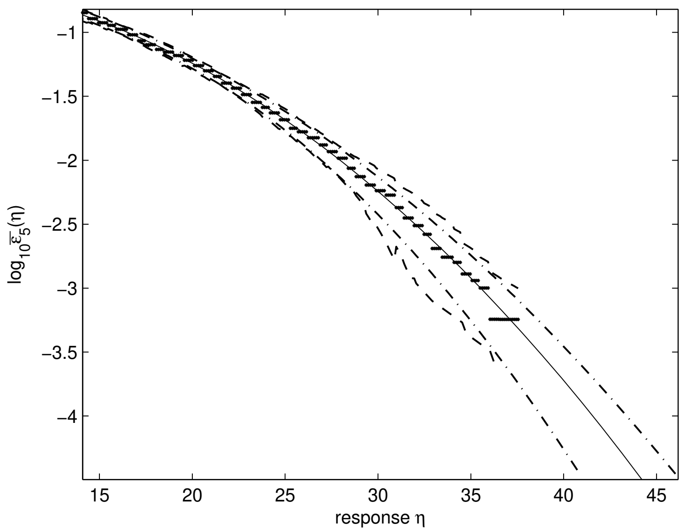

Fig. 14. Ørland wind speed statistics, 20 years of daily maximum data. $\hat{\varepsilon}_4(\eta)$ (*); Optimized curve fit (—); Empirical 95% confidence band (- -); Optimized confidence band (· · ··); Predicted 100 year return level estimate = 44.51 m/s. Predicted 95% CI = (40.41, 48.03) m/s. $\eta_1 = 16$ m/s.

Fig. 16 presents the Gumbel plot based on the 20 yearly extreme values derived from 20 years of daily maxima. The Gumbel prediction $\eta^{100\text{yr}} = 47.10$ m/s, with parametric bootstrapped 95% confidence interval (40.90, 54.33) m/s. 100000 bootstrap samples were generated. Fig. 17 shows the bootstrap PDF.

Fig. 18 presents POT predictions for different threshold numbers based on de Haan estimators, 20 years daily data. A conspicuous feature of this plot is the large variability in the prediction depending on the number of data. Fig. 19 shows the parametrically bootstrapped PDF of the POT prediction for threshold number $n = 140$ based on de Haan estimators, 20 years of daily maximum data. The POT prediction is $\eta^{100\text{yr}} = 42.22$ m/s, and the bootstrapped 95% confidence interval is (36.68, 45.83) m/s. 100000 samples were generated.

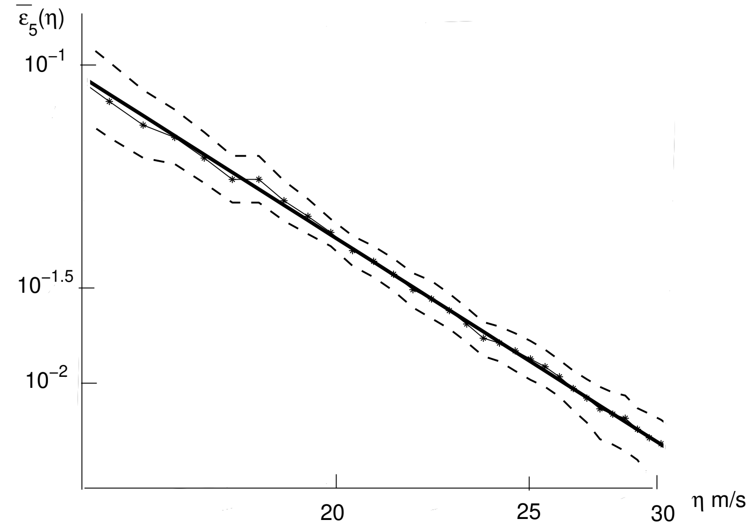

Fig. 15. Ørland wind speed statistics, 20 years of daily maximum data. Transformed plot of $\hat{\varepsilon}_4(\eta)$, q=0.159; b=10.38; η_1 = 16 m/s.

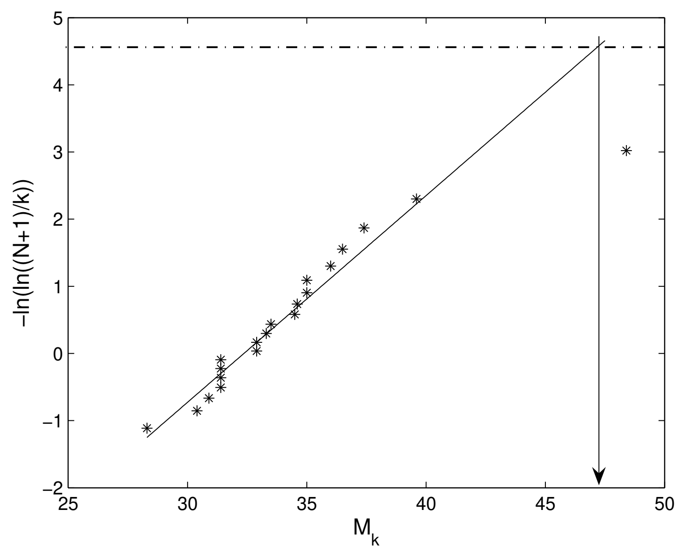

Fig. 16. Ørland wind speed statistics. 20 years daily data, Gumbel plot. 100 year return level prediction = 47.10 m/s, CI = (40.90, 54.33) m/s.

When comparing the results from 20 years of data with 7 years of data, it is important to note that the wind climate seems to have been milder during

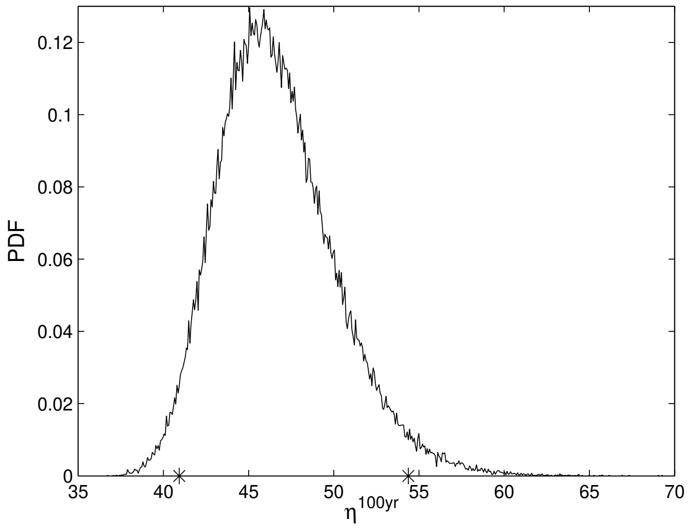

Fig. 17. 20 years daily data. Bootstrapped PDF of $\eta^{100\mathrm{yr}}$ from Gumbel distribution for 100000 bootstrap samples. Stars indicate 95% confidence interval (40.90, 54.33) m/s.

the last 7 years when compared with the whole 20 year period. However, it is interesting to see that the 100 year return level estimates based on the AER method are less sensitive to this than the Gumbel and POT methods, which may be explained by the fact that the AER method is less dependent on the largest values. A general comment is that the POT method based on de Haan estimators tends to underestimate the target value also for the measured data. Another clear advantage of the AER method over the POT method is seen when comparing the predicted 95% confidence intervals.

8 Concluding remarks

This paper studies a new method for extreme return level estimation for sampled time series. The method is based on the introduction of a conditional average exceedance rate, which allows dependence in the time series to be

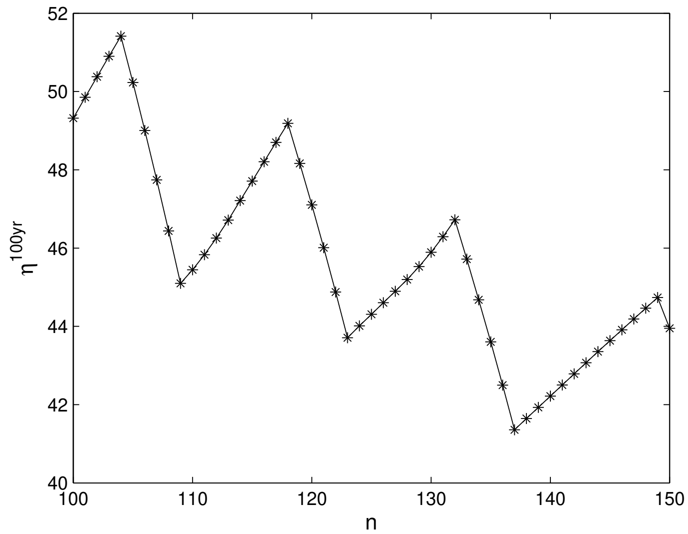

Fig. 18. 20 years daily data. The point estimate $\eta^{100\mathrm{yr}}$ of the 100-year return period based on 20 years daily data as a function of the number n of data points above threshold. Level estimate 42.22 m/s, $n = 140$.

properly accounted for. Declustering of the data is therefore avoided, and all the data are used in the analysis. Significantly, the proposed method also aims at capturing the subasymptotic form of the extreme value distribution.

Results for wind speeds, both synthetic and measured, are used to illustrate the method. Its validation is done by comparison with exact results (when available), or other widely used methods for extreme value statistics, such as the Gumbel and peaks-over-threshold (POT) methods. Comparison of the various predictions indicate that the proposed method provides more accurate results than the Gumbel and POT methods.

The proposed method applies to nonstationary time series, but it cannot directly predict the effect of climate change in the form of long-term trends in the average exceedance rate extending beyond the data. This must be incorporated into the analysis by explicit modelling techniques.

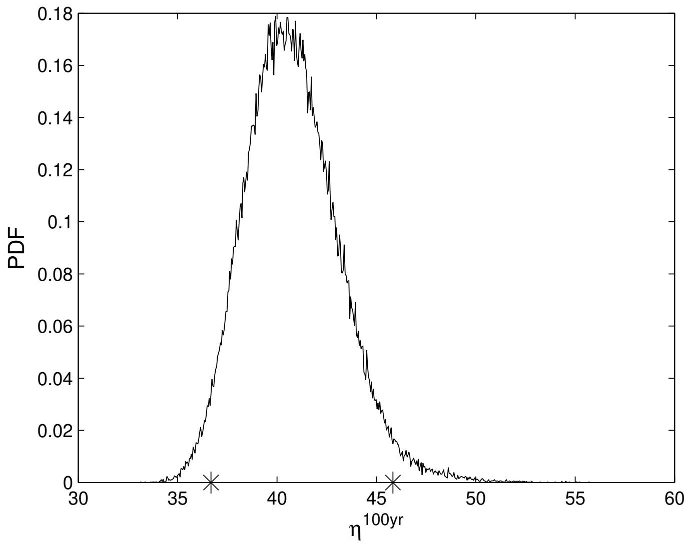

Fig. 19. 20 years daily data. Bootstrapped PDF of $\eta^{100\text{yr}}$ by de Haan's estimator for 100000 bootstrap samples. $n = 140$. Stars indicate the 95% confidence interval (36.68, 45.83) m/s.

9 Acknowledgements

This work was supported by the Research Council of Norway (NFR) through the Centre for Ships and Ocean Structures (CeSOS) at the Norwegian University of Science and Technology. The authors are grateful to Dr. Knut Harstveit of The Norwegian Meteorological Institute in Oslo for providing wind speed data.

References

[1] S. Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer-Verlag, London, 2001.

[2] A. Naess and O. Gaidai. Monte Carlo methods for estimating the extreme response of dynamical systems. *Journal of Engineering Mechanics,*

ASCE, 134(8):, 2008.

[3] A. Naess, O. Gaidai, and S. Haver. Efficient estimation of extreme response of drag dominated offshore structures by Monte Carlo simulation. *Ocean Engineering*, 34(16):2188–2197, 2007.

[4] A. Naess. The joint crossing frequency of stochastic processes and its application to wave theory. *Applied Ocean Research*, 7(1):35–50, 1985.

[5] A. Naess. Approximate first-passage and extremes of narrow-band Gaussian and non-Gaussian random vibrations. *Journal of Sound and Vibration*, 138(3):365–380, 1990.

[6] A. Naess. On the long-term statistics of extremes. *Applied Ocean Research*, 6(4):227–228, 1984.

[7] P. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, London, 1981.

[8] K. V. Bury. *Statistical Models in Applied Sciences*. John Wiley & Sons, Inc., New York, 1975.

[9] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

[10] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, London, 1997.

[11] J. Pickands. Statistical inference using order statistics. *Annals of Statistics*, 3:119–131, 1975.

[12] L. de Haan. Extreme value statistics. In J. Galambos, J. A. Lechner, and E. Simiu, editors, *Extreme Value Theory and Applications*. Kluwer Academic Publishers, Dordrecht, 1994.

[13] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels. *Statistics of Extremes*. John Wiley & Sons, Ltd., Chichester, UK, 2004.

[14] A. Naess and P. H. Clausen. Statistical extrapolation and the peaks over threshold method. omae–99–6422. In *Proceedings 18th International Conference on Offshore Mechanics and Arctic Engineering*. New York, ASME, 1999.

[15] A. Naess and P. H. Clausen. Combination of peaks-over-threshold and bootstrapping methods for extreme value prediction. *Structural Safety*, 23:315–330, 2001.

[16] J. Galambos and N. Macri. Classical extreme value model and prediction of extreme winds. *Journal of Structural Engineering, ASCE*, 125(7):792–794, 1999.

[17] A. Naess. Statistical extrapolation of extreme value data based on the peak over threshold method. *Journal of Offshore Mechanics and Arctic Engineering*, 120:91–96, 1998.

[18] N. J. Cook. Towards better estimation of extreme winds. *Journal of Wind Engineering and Industrial Aerodynamics*, 9(3):295–323, 1982.

[19] A. Naess. Estimation of long return period design values for wind speeds. *Journal of Engineering Mechanics, ASCE*, 124(3):252–259, 1998.

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

[21] O. Perrin, H. Rootzen, and R. Taesler. A discussion of statistical methods used to estimate extreme wind speeds. *Theoretical and Applied Climatology*, 85(3-4):203–215, 2006.