ELSEVIER

Journal of Wind Engineering
and Industrial Aerodynamics 59 (1996) 1–22

JOURNAL OF wind engineering AND industrial aerodynamics

Gumbel re-visited – a new look at extreme value statistics applied to wind speeds

R.I. Harris

College of Aeronautics, Cranfield University, Cranfield, Bedford, UK

Received 15 November 1994; revised 25 September 1995; accepted 6 November 1995

Abstract

This paper re-examines the traditional Gumbel extreme value method for analysing annual maximum windspeeds or similar data, with a view to improving the process. The advantages of transforming the data before applying an extreme value analysis are explained, and it is noted that the traditional Gumbel plot should really have its axes interchanged. An improved set of plotting positions based on the mean values of the order statistics are derived, together with a means of obtaining the standard deviation of each position. This enables a fitting procedure using weighted least squares to be adopted, which gives results similar to the traditional Lieblein BLUE process, but with the advantages that it does not require tabulated coefficients, is available for any number of data up to at least 50, and provides a quantitative measure of goodness of fit. A computer program which automatically implements the method is provided.

1. Introduction

For the extraction of extreme wind speeds, two methods are in current use. Where time series of meteorological data are available which are sufficiently continuous to enable the identification of individual storms, then the method of independent storms [1], is the usual first choice. In cases where the data is limited to annual largest values, or similar, then the usual choice is the Gumbel–Lieblein BLUE method [2]. This latter method requires the application of tables of coefficients, which are available only for sample sizes up to 24, and which must be stored as part of any computer program. For sample sizes greater than 24, the Gumbel–Lieblein procedure can still be applied. The method has been described by Cook [1] and involves selecting a subset of size n ($n \leqslant 24$) from the original data. The selection is made by ranking the data as usual, and then selecting the values which have probability ordinates closest to those $(m/(n+1)$ for

SSDI 0167-6105(95)00029-1

$m = 1, 2, \ldots, n$) which would be obtained from ranking a set of n data. The n selected data are then fitted using the Lieblein BLUE technique appropriate for a sample of size n. In practice the value of n is usually fixed at say $n = 10$, and the procedure becomes programmable, if not particularly convenient.

The work described in this paper was originally undertaken as part of an effort to provide a suite of computer programs for processing wind engineering data. One of the requirements was a program for extracting extreme values, and this led to the search for a method which would be easily programmable, would not require stored tables of coefficients, and could be applied uniformly to any size of data sample. These requirements led the writer to a re-examination from first principles of the treatment of extreme value data, and to the work reported in this paper.

In all that follows, it will be assumed that continuous records are not available and so a Gumbel rather than a Cook type of analysis is being applied to a set of e.g. annual extreme wind data. It will also be assumed for the purposes of discussion that this extreme data is hourly mean wind data, although the discussion applies equally to mean data over other periods, including so-called gust data which is really a mean over a period of about 3 s.

2. The Gumbel method

It is undoubtedly true that many wind engineers use the Gumbel method as a procedure for analysing data, without being familiar with its underlying theoretical origins. Both the theory and practice were described originally in the book by Gumbel [3], but this also contains a wealth of other information from which it is not easy to extract the material required.

One of the earliest descriptions of the Gumbel analysis procedure applied to wind data was given by Shellard [4] and may be summarised as follows. Given a set of N data values which represent, say the annual largest values of the hourly mean wind speed at a particular site, the values are ranked in ascending order of size, i.e. the smallest is accorded rank $m = 1$, and the largest rank $m = N$. Each of the windspeed values is plotted versus the variable $y = \ln\{-\ln[m/(N+1)]\}$ derived from its rank, and a straight line is fitted to the plot, from which the slope is obtained and also the intercept on the windspeed axis for $y = 0$. The values of $m/(N+1)$ are identified with the probability of the corresponding wind speed not being exceeded in one year. The fitted straight line is then usually used for extrapolation to determine a value of windspeed for which the probability of not being exceeded in one year is 0.98 (i.e. $m/(N+1) = 0.98$). This is the same as the value with a probability of 0.02 of being equalled or exceeded in one year – the so-called once-in-fifty-year windspeed. This value forms the basis of wind loading codes in the UK, but occasionally for design purposes some other value may be used which also has to be derived by extrapolation on the Gumbel plot.

The analysis procedure just outlined is arithmetically simple and is robust, in the sense that it assumes no additional knowledge of the data being analysed. It does not even require that the extreme data conforms to the expected Fisher–Tippett type 1 or double-exponential distribution. If the data does not conform to this distribution, then

as long as the method is graphically based, this is revealed by the fact that the plot of y versus windspeed does not give a straight line. Indeed, some modern methods which seek to identify to which of the possible asymptotic distributions a set of extreme data conforms, are based on examining the curvature on the standard plot, see Ref. [5]. Unfortunately, in practice, curvature can also arise not because of a failure to fit the type 1 asymptote, but because the data comes from a "mixed climate" or because, as will be demonstrated later in this paper, the data have not been plotted in the optimum form for the sample size available.

The Lieblein BLUE method gives values for the intercept and slope directly from arithmetical operations on the extreme wind data. There is, therefore, a temptation not to bother to draw the traditional Gumbel plot. For the reasons just given, this temptation should be resisted.

As already stated, the Gumbel method makes no use of any other information which may be available about the data. If such information is available, and it can be utilised, then it is reasonable to expect that an improvement in the results can thereby be obtained. The remainder of this paper is concerned with the search for these improvements.

For these purposes, it will be assumed that it is known that the wind data is drawn from a parent distribution of the exponential type. This means that the probability distribution can be represented by the form

$$F(x) = 1 - \exp[-h(x)],$$

where $h(x)$ is a positive function of x which increases with x faster than $\ln(x)$. That is so that

$$\lim_{x\to\infty}\left(\frac{\ln(x)}{h(x)}\right) = 0.$$

There is a considerable body of experimental data which suggests that not only hourly mean speeds, but also wind speeds averaged over shorter periods, conform to a Weibull parent distribution of the form

$$F(x) = 1 - \exp(-x^k), \qquad k > 0,$$

which is a distribution of the exponential type in the sense just defined.

The other assumption which will be made is that the extreme value data comes from a "simple" wind climate. At most sites, wind speeds arise from a variety of physically distinct meteorological mechanisms, such as depressions, thunderstorms, tropical cyclones, etc. For a wind climate to be "simple" in the sense used here, means that, at least for large wind speeds, one of these mechanisms is completely dominant – that is all the extreme data is produced by one, and only one, of the possible physical mechanisms. For sites in temperate latitudes, such as the UK, this is a very reasonable assumption, the dominant mechanism being depressions.

3. Effects of correlation

The original Gumbel treatment of extreme values dealt only with extremes of independent identically distributed samples. Each largest value was assumed to be the largest taken from M such samples, all of which had the same parent probability distribution $F(x)$. The central result of the Gumbel treatment is that under these conditions the largest value has a probability distribution given by $[F(x)]^M$. Thus, in the case of hourly mean speeds, if the values of each hourly mean were statistically independent of those for neighbouring hours, then it would be expected that the annual largest value of the hourly mean wind speed would have a probability distribution given by $[F(x)]^{8766}$, where $F(x)$ is the Weibull form and 8766 is the number of hours in a year. This is not observed in practice, because values of the hourly mean for adjacent hours are not independent.

More recent studies (see Ref. [5] for more details) have shown that provided the data is statistically stationary, and hence has a finite correlation time, τ, then the distribution of the largest value from M samples, with parent distribution $F(x)$, is given by $[F(x)]^n$, where $n \ll M$ and, if T is the record length, then $n \sim T/\tau$. For the case of the annual largest mean wind speed, direct comparison with observations has shown that $n \sim 100$. Thus, for correlated wind data, the original Gumbel form is still valid, but the value of n remains to be determined. As noted elsewhere [6] it is this uncertainty in the value of n which makes a direct theoretical attack on the problem difficult.

4. The asymptotic distribution of extreme values

The central result of the Gumbel treatment is that for parent distributions of the exponential type, the distribution of the largest value conforms asymptotically to the Fisher–Tippett type 1 or double exponential form. That is

$$\Psi(x) = \exp[-\exp(-y)], \tag{4.1}$$

where

$$y = \alpha(x - \check{x}) = \alpha x - \Pi. \tag{4.2}$$

$\check{x}$ is thus the mode of the extreme value distribution and α a measure of the dispersion. The dimensions of α are the inverse of those of x or $\check{x}$, and hence Π is a dimensionless quantity, called by Gumbel the "characteristic product". Gumbel refers to $\check{x}$ as the "characteristic largest value".

The Fisher–Tippett type 1 (FT-1) form is an asymptotic version of the exact result, i.e. $[F(x)]^n$, which becomes increasingly accurate as the value of n increases. Thus if the asymptotic result is applied in practice, for finite n, then some systematic error is to be expected in addition to any random errors due to the "fit" of measurements. In applying extreme value methods, engineers must be concerned with minimising this systematic error between the asymptotic form which they use and the exact result (which is not accessible owing to the uncertainty in n). Almost always having fitted the asymptotic

form they then use it to extrapolate to find future risks, and consequently they also need to be aware of the effect of the accuracy of the fit of the asymptote on the extrapolation process.

In most of the derivations of the FT-1 asymptote, it is not easy to see at what point approximations are made and thus it is difficult to estimate the errors involved. A notable exception is a derivation by Cramer [7]. This starts from the assumption of a parent distribution of the exponential type. That is

$$F(x) = 1 - \exp[-h(x)]. \tag{4.3}$$

Thus the exact distribution of the largest value is

$$\Phi(1,n) = [F(x)]^n = \{1 - \exp[-h(x)]\}^n. \tag{4.4}$$

The characteristic largest value, $\check{x}$, is defined by the relation

$$1 - F(\check{x}) = 1/n = \exp[-h(\check{x})]. \tag{4.5}$$

Thus

$$\Phi(1,n) = \left(1 - \frac{\exp[h(\check{x}) - h(x)]}{n}\right)^n. \tag{4.6}$$

Now since $h(x)$ is a positive and increasing function of x, as x increases beyond $\check{x}$, $\exp[h(\check{x}) - h(x)] \ll 1$, and for wind data $n \sim 100$. A well-known result for the exponential function, originally due to Cauchy [8] is

$$\lim_{n\to\infty}\left(1 - \frac{\epsilon}{n}\right)^n = \exp(-\varepsilon) \tag{4.7}$$

and the convergence of this expression to the limit is such that if ϵ is small and n is large, the exponential limit is a good approximation to the l.h.s., the leading error term being $\epsilon^2/2n$. Thus, applying this result to (4.6) gives

$$\Phi(1,n) \simeq \exp\{-\exp[h(\check{x}) - h(x)]\}. \tag{4.8}$$

Note that apart from the use of the Cauchy formula, (4.8) is exact – it contains no other approximations. Taking logarithms twice gives

$$-\ln\{-\ln[\Phi(1,n)]\} = h(x) - h(\check{x}). \tag{4.9}$$

The right-hand side of (4.9) is now expanded in a Taylor series about $x = \check{x}$ and gives

$$-\ln\{-\ln[\Phi(1,n)]\} = (x - \check{x})\frac{h'(\check{x})}{1!} + (x - \check{x})^2\frac{h''(\check{x})}{2!}) + \ldots . \tag{4.10}$$

The familiar Fisher–Tippet type 1 asymptotic formula is obtained by truncating the Taylor series at the first term, and thus

$$y = -\ln\{-\ln[\Psi(x)]\} \simeq \alpha(x - \check{x}), \tag{4.11}$$

where $\Psi(x)$ is used to denote the asymptotic form of $\Phi(1,n)$ and

$$\alpha = h'(\check{x}) = n(\mathrm{d}F/\mathrm{d}x)_{x=\check{x}}. \tag{4.12}$$

Consequently, the errors involved in the asymptotic formula are twofold. Firstly, the error in the use of the Cauchy formula for the exponential, which, as already stated, should be small, and secondly the effect of truncating the Taylor series. In this case the errors involved can be estimated from the remaining terms in the Taylor series.

In the original Gumbel method [3] or the Gumbel–Lieblein BLUE method [2], an estimate of y is plotted versus x and the best linear fit is obtained, from which the parameters α and x are estimated. Thus a straight line is being fitted to what, from (4.9), should in general, be a curve $h(x)$. The fitted line is the tangent to this curve at $x = \check{x}$, and thus as one moves away from the neighbourhood of $\check{x}$ an error will be introduced which may be positive or negative, depending on whether the curve is concave upwards or downwards, i.e. on the sign of the first term neglected in the Taylor series. If, as is usual, the fitted line is then used for extrapolation, to estimate values of y for values of x beyond the available data, then this curvature has an obvious bearing on the accuracy of the extrapolation. For instance, if the parent distribution is a Rayleigh distribution so that $h(x) = x^2$, then the first error term is positive, so that in a conventional Gumbel plot (with wind speed values on the vertical axis) the curve is concave downwards, and thus for a specified probability of occurrence, extrapolation *overestimates* wind speeds – facts already demonstrated by Cook [1].

The remedy for this situation is obtained by noting that if $h(x) \equiv x$ (i.e. the parent distribution is exponential, at least in the right-hand tail), then the Taylor expansion terminates after the first term, and thus the tangent fitted at $x = \check{x}$ corresponds to the exact formula (4.9). Thus in the general case if estimates of y are plotted not versus x, but versus $h(x)$, then not only will the same experimental data be a better fit to the line, *but the accuracy of extrapolation along the line is guaranteed.* For a Weibull parent, $h(x) = x^k$. For most UK sites observations have shown that the value of k lies between 1.8 and 2.2, and thus in the situation postulated at the beginning of this paper where parent data is not available, plotting y estimates versus $q = x^2$ as recommended by Cook [1], is a reasonable compromise.

5. Plotting axes – the wind and probability variates

It is the normal convention when plotting one quantity as a function of another, to plot the dependent variable (supposedly contaminated by error) on the vertical or y-axis and the independent variable (supposedly error-free) on the horizontal or x-axis. Thus, in fitting a line to the data, the regression line of y on x is calculated (which, of course, is different from the regression line of x on y). In the usual Gumbel analysis of extreme wind speeds, experimentally observed wind speeds (or, following the previous section, values which are functions of the observed wind speeds) are plotted versus probabilities derived from ordering the data. These probability values are known exactly for a specified number of samples. The natural tendency, therefore, is to regard the wind speed variable as being the dependent (i.e. error prone) variable and the probability

variable as being the independent (i.e. error-free) variable. Thus, in a conventional Gumbel plot, one finds wind speeds on the vertical axis and the reduced probability variate on the horizontal axis, in spite of the universal use of the letter y to denote this variate. The book by Castillo [5] is the first publication this writer has seen in which this convention is not followed.

The variation in the observed wind speeds derives from two sources. Firstly, even if the wind speeds could be observed with perfect accuracy, there would be the natural sampling variation, because the measurements are of a quantity varying according to the laws of probability – indeed finding out information about the nature of the probability distribution is the reason for plotting the observations. Secondly, there is variability introduced by the observation process – i.e. the instrumentation used. The observations available can thus be regarded either as information about the idealised probability of wind speeds, but contaminated by (random) instrumentation errors, or as exact information about the probability distribution of measured values, the latter being (if the errors are additive and independent) a convolution of the perfect wind speed distribution and the distribution of instrument errors. In practice, since only measured values are ever going to be available, it seems more sensible to adopt the second view, and in this sense, therefore, the observations are exact. (*Note.* In this paragraph it has been assumed that the quantisation error involved in turning the observations into numerical data is negligible – it should be in a properly designed instrumentation system.)

In order to assess the variability of the probability variate it is necessary to understand its derivation.

This depends on two results:

(i) If a variable x has a probability density $f(x)$ then a new variable z related to x by the transformation $z = g(x)$, will have a probability density $h(z)$ given by

$$h(z)\,\mathrm{d}z = f(x)\left|\frac{\mathrm{d}x}{\mathrm{d}g}\right|\mathrm{d}z. \tag{5.1}$$

(ii) If a variable x has probability density $f(x)$ and probability distribution $F(x)$ (so that $f(x) = \mathrm{d}F/\mathrm{d}x$) and if N independent samples from this distribution are ranked from $m = 1$ for the smallest to $m = N$ for largest, then the exact probability density for the value of rank m is given by [3]

$$\Phi_m(x)\,\mathrm{d}x = \frac{N!}{(N-m)!(m-1)!}[F(x)]^{m-1}[1-F(x)]^{N-m}f(x)\,\mathrm{d}x. \tag{5.2}$$

Following (i) above, the transformation $z = F(x)$ is introduced which gives

$$h_m(z)\,\mathrm{d}z = \frac{N!}{(N-m)!(m-1)!}z^{m-1}(1-z)^{N-m}\,\mathrm{d}z, \tag{5.3}$$

since in this case $|\mathrm{d}x/\mathrm{d}g| = 1/f(x)$.

$h_m(z)$ is still the probability density for the value of rank m, but now written in terms of the new variable z. Note that by definition $F(x)$ is a single-valued increasing function of x, so there is a one-to-one relationship between a given value of x and the corresponding value of z. Whatever the range of the variable x, this range is transformed into $0 \leqslant z \leqslant 1$.

The *mean* value of a variable z which has a probability density $h_m(z)$ is given by

$$\bar{z} = \int_0^1 z h_m(z)\,\mathrm{d}z \tag{5.4}$$

$$= \frac{N!}{(N-m)!(m-1)!} \int_0^1 z^m (1-z)^{N-m}\,\mathrm{d}z \tag{5.5}$$

$$= \frac{m}{N+1}, \tag{5.6}$$

a result which follows from the definition of the beta function [9].

Thus the plotting position used for the sample of rank m is the mean value of the parent probability distribution $F(x)$ associated with that sample value.

The value is a mean in the following sense. Suppose that there are not one but L sets, each of N independent samples all taken from the same parent distribution, $F(x)$, which is a known function. Then for a particular choice of m, if each set is ranked in the usual way, there will be L individual mth ranked values of x. If $F(x)$ is a known function, then each of these values can be translated into a corresponding value of $z = F(x)$. The result just devised shows that the mean value of z taken over the ensemble of L values tends, for L large, to the value $m/(N+1)$. For one particular set of N samples, therefore, the plotting position $m/(N+1)$ is not exact, but is subject to sampling error, and therefore for this reason the probability variate should properly be regarded as the dependent (error-prone) variable and therefore plotted along the vertical or y-axis – as the designation y for this variate suggests, and the target for fitting thus becomes the regression line of y on the measured variable.

Note that nowhere in this discussion of the probability variate has it been assumed that the sample under consideration consists of extreme values. Indeed, the use of sample ordering is one standard technique which can be used to fit observations to any known or postulated probability distribution. The other frequently used alternative methods are the fitting of probability density to the observed histogram of the sample, and the direct fitting of the distribution to the empirically observed distribution or ogive [10]. For samples consisting of extreme values, the distribution being fitted is the FT-1 asymptote, and the "peaky" nature of the probability density makes the histogram technique inappropriate, while the ogive method produces larger errors at the extremes of the distribution and thus again is unsuitable, which is why the standard Gumbel analysis uses the method based on order statistics.

6. Improved probability plotting positions

As noted earlier, the standard Gumbel method is robust because it makes no assumptions about the data being plotted. If it is known that the data should conform to the FT-1 asymptote, it is possible to improve the technique. Since the standard plot is aimed at fitting a straight line to the data, it is *not* the probability ordinate that is plotted,

but the reduced variate $y = -\ln\{-\ln[\Psi(x)]\}$. Thus in a standard Gumbel plot, for a sample of given rank, the probability variate is first approximated by its mean value, and then a non-linear transformation is applied before plotting. The result of taking a mean and applying a non-linear transformation depends on the order in which these operations are applied. If a non-linearly transformed value is to be plotted and fitted, then it is more correct to apply the non-linear transformation first, and then to use the mean value of the transformed variable as the plotting position. Thus, it is the mean value of $-\ln[-\ln(z)]$, and not the mean value of z, which should be required for each of the samples.

At this point it is convenient to reverse the usual Gumbel data ordering procedure, and introduce the ranking $\nu = 1$ for the largest sample down to $\nu = N$ for the smallest. In terms of this new system, the probability density for the νth ranked value is, in terms of z,

$$h_\nu(z) = \frac{N!}{(\nu-1)!(N-\nu)!} z^{N-\nu}(1-z)^{\nu-1}, \tag{6.1}$$

i.e. Eq. (5.3) with $N - m$ now equal to $\nu - 1$.

Thus the plotting position for the νth ranked value is given by

$$\bar{y}_\nu = \frac{N!}{(\nu-1)!(N-\nu)!} \int_0^1 -\ln[-\ln(z)] z^{N-\nu}(1-z)^{\nu-1}\,dz. \tag{6.2}$$

The values of this integral for general N, ν can be obtained analytically by introducing the substitution $z = \exp(-t)$, from which

$$\bar{y}_\nu = \frac{N!}{(\nu-1)!(N-\nu)!} \int_0^\infty [-\ln(t)] \exp[-(N-\nu+1)t][1-\exp(-t)]^{\nu-1}\,dt. \tag{6.3}$$

The following result is given by Gradshteyn and Ryzhik [11],

$$\int_0^\infty \exp(-\mu x)\ln(x)\,dx = -\frac{1}{\mu}[C + \ln(\mu)], \qquad \mathrm{Re}\,\mu > 0, \tag{6.4}$$

where C is Euler's constant ($= 0.577215\ldots$).

Applying this result to the integrals (6.3) gives

$$\bar{y}_1 = C + \ln(N), \tag{6.5}$$

$$\bar{y}_2 = C + \frac{N\ln(N-1) - (N-1)\ln(N)}{1!}, \tag{6.6}$$

$$\bar{y}_3 = C + \frac{1}{2!}[N(N-1)\ln(N-2) - 2N(N-2)\ln(N-1) + (N-1)(N-2)\ln(N)], \tag{6.7}$$

$$\overline{y}_4 = C + \frac{1}{3!}[N(N-1)(N-2)\ln(N-3) - 3N(N-1)(N-3)\ln(N-2) + 3N(N-2)(N-3)\ln(N-1) - (N-1)(N-2)(N-3)\ln(N)], \quad (6.8)$$

which establishes the form of the results for larger values of ν. The results (6.5) and (6.6) were given by Gumbel [3], but no further use was made of them for data analysis, possibly because at the time of publication of the book (1958), there were no feasible means of routinely performing the necessary computations.

For values of N in the range usually required for practical analysis, say $10 \leqslant N \leqslant 50$, the analytical formulae (6.5)–(6.8), etc., are not very suitable for computing the required plotting positions, because the alternating signs inevitably lead to computation of relatively small values by differencing several large ones, with consequent loss of precision.

However for practical values of N and $\nu \geqslant 2$, the integrand in (6.2) is a smooth well-behaved function and is zero at both the limits. Thus the most suitable and economical technique for evaluating $\overline{y}_\nu$ for a given N, is to use the analytical result for $\overline{y}_1$ and $\overline{y}_2$, and to compute the remaining values by numerical integration. The method used is discussed in Appendix A.

7. Deviation of the probability plotting positions

The original Gumbel method used classical least squares to fit a straight line to the plotted probability ordinates and thus to estimate the slope and intercept parameters. The classical least squares method presumes that the errors associated with any of the plotted ordinates are of similar magnitude. It is thus not really suitable for Gumbel plots, because for extreme value data, the scatter or standard deviation associated with the plotted probability ordinates, $\overline{y}_\nu$, is not uniform, but varies systematically, being largest for the largest values, i.e. $\overline{y}_1$, $\overline{y}_2$. Consequently classical least squares gives too much weight to the deviations of the largest extreme values. One of the reasons for the introduction of the Lieblein BLUE method [2] was to correct this defect.

Therefore a necessary step before the improved plotting positions introduced in the previous section can be used, is the calculation of the standard deviation associated with each position. If this is denoted by σ_ν, then

$$\sigma_\nu = [\overline{y_\nu^2} - (\overline{y}_\nu)^2]^{1/2}, \quad (7.1)$$

so that in order to calculate σ_ν for each ν, the mean-square value of each plotting position, $\overline{y_\nu^2}$ is required. The derivation follows that for $\overline{y}_\nu$ given in Section 6, and the result is

$$\overline{y_\nu^2} = \frac{N!}{(\nu-1)!(N-\nu)!} \int_0^1 \{-\ln[-\ln(z)]\}^2 z^{N-\nu}(1-z)^{\nu-1}\,dz. \quad (7.2)$$

Again the substitution $z = \exp(-t)$ is introduced, from which

$$\overline{y_\nu^2} = \frac{N!}{(\nu-1)!(N-\nu)!}\int_0^\infty [\ln(t)]^2 \exp[-(N-\nu+1)t][1-\exp(-t)]^{\nu-1}\,dt. \tag{7.3}$$

Gradshteyn and Ryzhik [11] give the result

$$\int_0^\infty [\ln(x)]^2 \exp(-\mu x)\,dx = \frac{1}{\mu}\{\tfrac{1}{6}\pi^2 + [C+\ln(\mu)]^2\}, \qquad \mathrm{Re}\,\mu > 0, \tag{7.4}$$

where, as previously, C is Euler's constant.

Applying this result to the integrals (7.3) gives

$$\overline{y_1^2} = \tfrac{1}{6}\pi^2 + [C+\ln(N)]^2, \tag{7.5}$$

$$\overline{y_2^2} = \tfrac{1}{6}\pi^2 + C^2 + 2C[N\ln(N-1) - (N-1)\ln(N)] + N[\ln(N-1)]^2 - (N-1)[\ln(N)]^2. \tag{7.6}$$

Further results can be derived analytically, but the algebra rapidly becomes impossibly tedious, and, as for $\overline{y_\nu}$, the resulting expressions are unsuitable for accurate computation. Substituting for $\overline{y_\nu^2}$ and $(\overline{y_\nu})^2$ in (7.1) leads to

$$\sigma_1 = \pi/\sqrt{6}, \tag{7.7}$$

$$\sigma_2 = \{\tfrac{1}{6}\pi^2 - N(N-1)[\ln(1-1/N)]^2\}^{1/2} \tag{7.8}$$

$$\simeq [\tfrac{1}{6}\pi^2 - 1 + O(N^{-2})]^{1/2}. \tag{7.9}$$

For calculation of $\overline{y_\nu^2}$ and hence σ_ν for practical values of N, i.e. $10 \leqslant N \leqslant 50$, it is again necessary to revert to numerical methods for calculating $\overline{y_\nu^2}$ for $\nu \geqslant 3$. The method used is closely related to that used for $\overline{y}_\nu$ and details are given in Appendix A.

As an example, the plotting positions and corresponding standard deviations have been calculated for the case $N = 21$, and are show in Table 1. For comparison, the values of the orthodox Gumbel plotting positions are given in the second column.

From this table it can be seen that the Gumbel plotting positions obtained from $-\ln\{\ln[(N-\nu+1)/(N+1)]\}$ differ quite markedly from the values based on $\overline{y_\nu}$, especially at either end of the ranked values. Secondly, the computed values of σ_ν emphasise the much larger statistical variability associated with the largest values (those with small values of ν), and thus confirm that classical least squares is not a suitable fitting technique.

8. Fitting a straight line to the data

Given a set of N pairs of values of q_v (= windspeed^2) and $\overline{y}_\nu$, the requirement is to find the straight line $y = \alpha q - \Pi$ which is the best fit to the data. As has been

Table 1

Order (ν)	Gumbel (y)	$\bar{y}_\nu$	σ_ν
1	3.0679	3.6217	1.2825
2	2.3506	2.5971	0.8032
3	1.9200	2.0715	0.6288
4	1.6061	1.7113	0.5334
5	1.3555	1.4332	0.4714
6	1.1443	1.2037	0.4273
7	0.9597	1.0058	0.3939
8	0.7941	0.8300	0.3676
9	0.6423	0.6700	0.3463
10	0.5007	0.5215	0.3288
11	0.3665	0.3815	0.3142
12	0.2377	0.2473	0.3020
13	0.1123	0.1168	0.2918
14	−0.0115	−0.0119	0.2834
15	−0.1355	−0.1409	0.2767
16	−0.2618	−0.2727	0.2718
17	−0.3931	−0.4103	0.2691
18	−0.5334	−0.5587	0.2692
19	−0.6894	−0.7262	0.2739
20	−0.8746	−0.9315	0.2879
21	−1.1285	−1.2378	0.3319

demonstrated, the classical least squares technique is unsuitable because it presumes that the deviations of all points from the line are statistically identical. Instead, a weighted least squares technique can be used. Weights w_ν are introduced to compensate for the systematic change in deviation, and the problem then becomes that of finding values α and Π which minimise S^2, where

$$S^2 = \sum_{\nu=1}^{N} w_\nu (\bar{y}_\nu - \alpha q_\nu + \Pi)^2. \tag{8.1}$$

As usual the partial derivatives $\partial S^2/\partial\alpha$ and $\partial S^2/\partial\Pi$ are found, and equated to zero, and the resulting pair of simultaneous linear equations are then solved for α and Π. The resulting expressions are considerably simplified if the condition

$$\sum_{\nu=1}^{N} w_\nu = 1 \tag{8.2}$$

is imposed on the weights, w_ν, which can be done without loss of generality. With this condition, the resulting expressions for α and Π are

$$\alpha = \frac{\sum_1^N w_\nu \bar{y}_\nu q_\nu - (\sum_1^N w_\nu \bar{y}_\nu)(\sum_1^N w_\nu q_\nu)}{\sum_1^N w_\nu q_\nu^2 - (\sum_1^N w_\nu q_\nu)^2}, \tag{8.3}$$

$$\Pi = \alpha \sum_1^N w_\nu q_\nu - \sum_1^N w_\nu \bar{y}_\nu. \tag{8.4}$$

The choice of the weights, w_ν, becomes obvious, if it is noted that the variance of the quantity

$$(\bar{y}_\nu - \alpha q_\nu + \Pi)^2/\sigma_\nu^2 \tag{8.5}$$

is constant, independent of the order ν, and thus the weights, w_ν, must be chosen to be proportional to $1/\sigma_\nu^2$. Hence, any weight, w_ν, is obtained from the expression

$$w_\nu = \frac{1/\sigma_\nu^2}{\sum_{\nu=1}^{N} 1/\sigma_\nu^2} \tag{8.6}$$

to preserve the condition that the weights sum to unity. Since for the proposed new plotting positions, the variances, σ_ν^2, are all available, the required weights can easily be calculated from (8.6).

It is useful to have a quantitative measure of the goodness of fit of the line to the observations. In classical least squares, the appropriate measure is the residual standard deviation. In this case, because the standard deviation of individual observations is not uniform, a modified form of residual standard deviation is required. The appropriate form is

$$\sigma_r = S\sqrt{N/(N-2)}. \tag{8.7}$$

Here the factor $\sqrt{N/(N-2)}$ is required to obtain an unbiased estimate of σ_r, since two parameters (α and Π) have already been calculated from the observations.

9. Application to wind data

Recently Cook [12] has, on behalf of the Building Research Establishment, obtained from the UK Meteorological Office an extreme value analysis of the data from 13 UK observation stations for the years 1970–90 inclusive. Accordingly, it was decided to select one of these 13 stations as a test case, and after consultation with the Meteorological Office, the Honington Station was chosen as the least likely to be affected by variation in the roughness of the fetch for different wind directions.

The Meteorological Office (M.O.) supplied the values for the annual maximum hourly mean wind speed (in knots) for this station for the 21 years 1970–90. Unfortunately, it is M.O. standard practice to supply such data rounded to the nearest knot, in spite of the fact that each hourly mean is compounded from automatically recorded consecutive one minute means which would justify a precision of 0.1 knot. (Data to this higher precision is available only at considerable extra cost.) This means that the data used does to some extent violate the assumptions made with regard to quantisation error in Section 5, and it also leads to spurious "ties" in the ranking process.

However, since only this form of data was readily available, it was input into the computer program given in Appendix B, and the resulting output is shown in Table 2. Since tabulated coefficients for the Lieblein BLUE method are available for a data set of 21 points, this method was also applied to the values of q_ν derived from the data, and the results of this calculation are also given:

Table 2

ν	q_ν	$\bar{y}_\nu$	w_ν
1	610.5	3.6217	0.00331
2	424.0	2.5971	0.00844
3	382.6	2.0715	0.01378
4	382.6	1.7113	0.01915
5	382.6	1.4332	0.02451
6	362.8	1.2037	0.02984
7	343.4	1.0058	0.03511
8	324.6	0.8300	0.04031
9	306.3	0.6700	0.04541
10	288.6	0.5215	0.05038
11	288.6	0.3815	0.05517
12	288.6	0.2473	0.05973
13	288.6	0.1168	0.06399
14	271.3	−0.0119	0.06784
15	271.3	−0.1409	0.07115
16	254.7	−0.2727	0.07372
17	254.7	−0.4103	0.07523
18	238.5	−0.5587	0.07517
19	207.8	−0.7262	0.07262
20	207.8	−0.9315	0.06571
21	193.2	−1.2378	0.04944

(i) The resulting extreme value parameters are: mode of 271.6 $(\mathrm{m/s})^2$, $\alpha = 0.01437$ $(\mathrm{m/s})^{-2}$, characteristic product of 3.903 and weighted residual standard deviation of 0.1686.

(ii) Application of the Gumbel–Lieblein BLUE method to the same set of q_ν data gives the following results: mode of 272.9 $(\mathrm{m/s})^2$, $\alpha = 0.01463$ $(\mathrm{m/s})^{-2}$, characteristic product of 3.993.

These results are plotted in Fig. 1. As can be seen, the new method gives a result which is virtually identical to that obtained by the use of the Gumbel–Lieblein method.

10. Concluding remarks

A new method for analysing annual or similar extreme values has been devised which appears to give results virtually identical to the Gumbel–Lieblein method. In contrast to the latter method, the new method is fully automated, requires no look-up of tabulated coefficients, and can be used for any number of data points. (The program listed in Appendix B has been tested satisfactorily for up to 50 data points.)

Appendix A

Evaluation of the integrals for the mean and mean square of the probability ordinates

From (6.2) in the main text the integral to be evaluated for $\bar{y}_\nu$ is

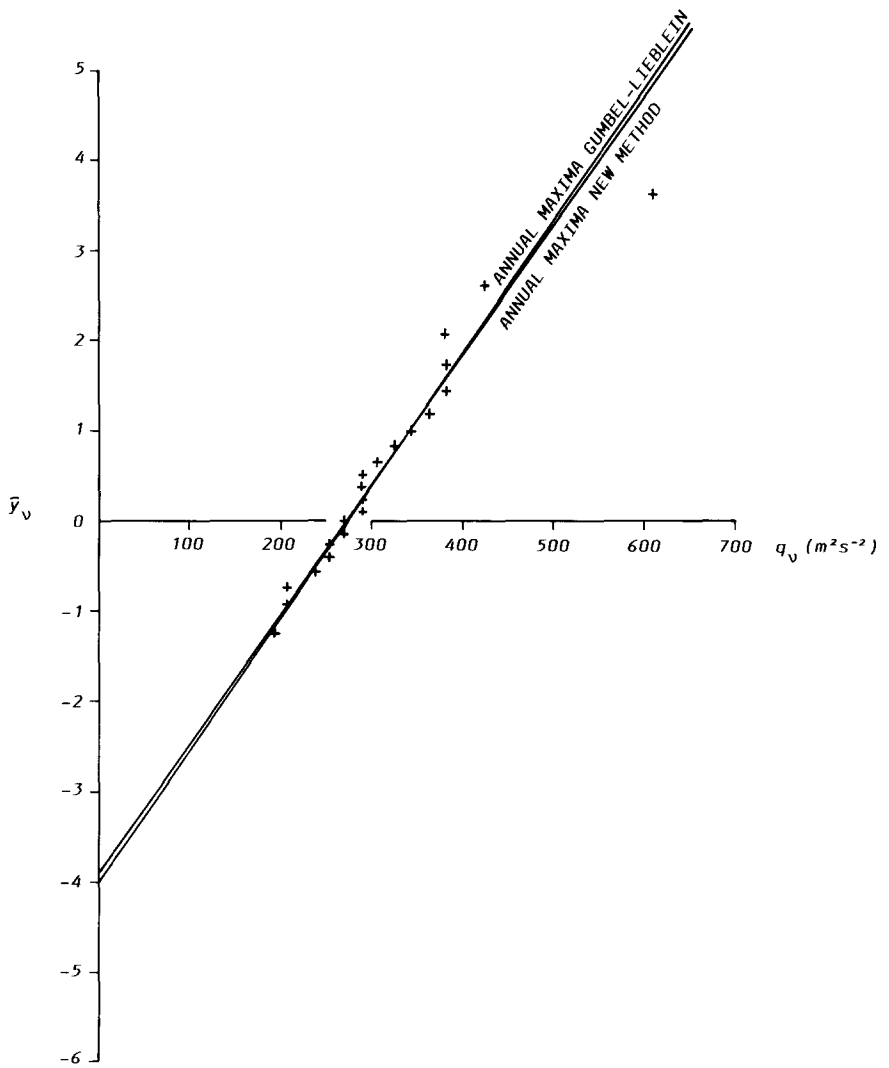

Fig. 1. Extreme value analysis of the Honington data.

$$\bar{y}_\nu = \frac{N!}{(\nu-1)!(N-\nu)!}\int_0^1 -\ln[-\ln(z)]z^{N-\nu}(1-z)^{\nu-1}\,dz. \tag{A.1}$$

For $\nu \geqslant 2$, this integrand is finite everywhere, but for N large it changes very rapidly near both limits. Consequently, rather than attempt to evaluate it numerically in this form, the same substitution as used for the analytical treatment, namely $z = \exp(-t)$, is introduced giving

$$\overline{y}_\nu = \frac{N!}{(\nu-1)!(N-\nu)!}\int_0^\infty [-\ln(t)]\exp[-(N-\nu+1)t][1-\exp(-t)]^{\nu-1}\,dt. \tag{A.2}$$

Integrating once by parts gives for $1 \leqslant \nu \leqslant N-1$

$$\overline{y}_{\nu+1} = \overline{y}_\nu - \frac{N!}{\nu!(N-\nu)!}\int_0^\infty \frac{\exp[-(N-\nu)t]}{t}[1-\exp(-t)]^\nu\,dt. \tag{A.3}$$

The integrand in (A.3) is (for the specified range of ν) zero at both limits and converges rapidly to zero for modest values of t. For numerical evaluation this renders it suitable for application of the Romberg method. This is a standard technique for numerical integration described in many texts (see, e.g., Ref. [13]). Basically it uses repeated trapezium rule integration with halving of the interval and Richardson extrapolation, which also allows for automatic termination either when a prescribed level of accuracy has been achieved (10^{-6} in this implementation) or when the extrapolation begins to be contaminated by the build-up of rounding error. (Not an issue in this case, as double-length arithmetic can be used without an unacceptable time penalty.) The infinite upper limit is replaced by a finite one, which is either the value of t for which the integrand first becomes less than the machine accuracy (10^{-25}) or 1.6, whichever is larger.

This procedure is used for given a value of N, to evaluate the integral in (A.3) for $2 \leqslant \nu \leqslant N-1$, and with the aid of the recurrence relation (A.3) and the analytical results for y_1 and y_2, the remaining values $\overline{y_3}, \ldots, \overline{y_N}$ can be evaluated.

From (7.2) in the main text, the integral for $\overline{y_\nu^2}$ is

$$\overline{y_\nu^2} = \frac{N!}{(\nu-1)!(N-\nu)!}\int_0^1 \{-\ln[-\ln(z)]\}^2 z^{N-\nu}(1-z)^{\nu-1}\,dz. \tag{A.4}$$

Again, as for the analytical procedure, the substitution $z = \exp(-t)$ is introduced, giving

$$\overline{y}_\nu^2 = \frac{N!}{(\nu-1)!(N-\nu)!}\int_0^\infty [\ln(t)]^2\exp[-(N-\nu+1)t][1-\exp(-t)]^{\nu-1}\,dt. \tag{A.5}$$

Again for $2 \leqslant \nu \leqslant N-1$, integrating by parts once gives

$$\overline{y}_{\nu+1}^2 = \overline{y}_\nu^2 - \frac{N!}{\nu!(N-\nu)!}\int_0^\infty \frac{2[-\ln(t)]}{t}\exp[-(N-\nu)t][1-\exp(-t)]^\nu\,dt. \tag{A.6}$$

Consider the integrand in (A.6). For the range of ν specified, it is zero at both limits, and by virtue of the presence of the term $\ln(t)$, also has a zero at $t = 1$. Therefore numerical evaluation by the Romberg method is carried out by dividing the range of integration

into two portions: (i) 0 to 1; (ii) 1 to ∞. Each of these is evaluated separately by the Romberg technique. Range (i) being finite, causes no problem, whilst for (ii) the value of the upper limit is rendered finite by the same means as the integrand involved in the evaluation of $\overline{y}_\nu$.

Given numerical values of the integral for $2 \leqslant \nu \leqslant N-1$, the analytical results for $\overline{y}_1^2$ and $\overline{y}_2^2$, together with the recurrence relation (A.6) are then used to evaluate $\overline{y}_3^2, \ldots, \overline{y}_N^2$.

Appendix A

Computer implementation of the method

A computer program has been written to implement the method described in the paper, and is presented in the following.

```
100 REM Program for analysis of largest values by Modified Gumbel method
110 DEFDBL A-H,K-Z
120 DIM WINDSPEED(50),MEAN(50),VARIANCE(50),WEIGHT(50),Q(50),S(30),T(30)

130 REM Read in number of points to be analysed
140 CLS:PRINT "Enter total number of data values"
150 INPUT IN
160 REM Read in and display extreme value data
170 FOR I=1 TO IN
180 INPUT WINDSPEED(I)
190 CLS
200 PRINT "Total number of data values is",IN
210 FOR J=1 TO I
220 PRINT J,WINDSPEED(J)
230 NEXT J
240 NEXT I
250 PRINT "Data input complete. Check! Press any key to continue"
260 A$=INKEY$:IF A$="" THEN GOTO 270

270 REM End of data input
280 REM Convert data to metres/sec and square
290 FOR I=1 TO IN
300 Q(I)=((WINDSPEED(I)*.51477#)^2#)
310 NEXT I
320 REM End of data conversion

330 REM Sort data into descending order
340 FOR J=2 TO IN
350 TEMP=Q(J)
360 FOR I=J-1 TO 1 STEP -1
370 IF Q(I)>TEMP THEN GOTO 410
380 Q(I+1)=Q(I)
390 NEXT I
400 I=0
410 Q(I+1)=TEMP
```

```
420 NEXT J
430 REM End of sort

440 REM Evaluate analytical results for largest and next-largest points
450 N=CDBL(IN)
460 MEAN(1)=.577215664901533#+LOG(N)
470 MEAN(2)=.577215664901533#+N*LOG(N-1)-(N-1)*LOG(N)
480 VARIANCE(1)=9.869604401089359#/6#
490 VARIANCE(2)=VARIANCE(1)-N*(N-1)*(LOG(1#-1#/N)^2#)
500 REM End of Analytical Results
510 REM Begin main cycle to calculate recurrence integrals for means
520 REM and meansquares
530 DEF FNMEAN(N,NU,X)=EXP(-X*(N-NU))*((1#-EXP(-X))^NU)/X
540 DEF FNMEANSQUARE(N,NU,X)=2#*(-LOG(X))*EXP(-X*(N-NU))*((1#-EXP(-X))^NU)/X
550 FOR INU=3 TO N
560 NU=CDBL(INU-1)
570 REM Use subroutine for the Log of the Gamma Function to calculate the
580 REM Pascal triangle ratios of factorials outside the integral
590 XX=N+1#:GOSUB 2000
600 PASCAL=GL
610 XX=NU+1#:GOSUB 2000
620 PASCAL=PASCAL-GL
630 XX=N-NU+1#:GOSUB 2000
640 PASCAL=EXP(PASCAL-GL)
650 REM End of calculation;required ratio in PASCAL
660 REM Begin numerical integration for recurrence integral for mean
670 LOWERLIMIT=0#
680 UPPERLIMIT=55#
690 S(0)=0#:REM Initial Trapezium Rule estimate from end points
700 REM Set effective finite upper limit based on smallness of integrand
710 IF FNMEAN(N,NU,UPPERLIMIT)<1D-25 AND UPPERLIMIT>1.6#
THEN UPPERLIMIT=UPPERLIMIT-.1#: GOTO 710
720 SELECT$="MEAN":GOSUB 1500
730 MEAN=RESULT*PASCAL
740 REM End of integration for recurrence integral for mean
750 REM Begin first integration for meansquare over range 0 to 1
760 LOWERLIMIT=0#
770 UPPERLIMIT=1#
780 S(0)=0#:REM Initial Trapezium Rule estimate from end points
790 SELECT$="MEANSQUARE":GOSUB 1500
800 MEANSQUARE=RESULT
810 REM End of first integration for meansquare
820 REM Begin second integration for meansquare over range 1 to infinity
830 LOWERLIMIT=1#
840 UPPERLIMIT=55#
850 S(0)=0#:REM Initial Trapezium Rule estimate from end points
860 REM Set effective finite upper limit based on smallness of integrand
870 IF ABS(FNMEANSQUARE(N,NU,UPPERLIMIT))<1D-25 AND UPPERLIMIT>1.6#
THEN UPPERLIMIT=UPPERLIMIT-.1#: GOTO 870
880 SELECT$="MEANSQUARE":GOSUB 1500
890 MEANSQUARE=(MEANSQUARE+RESULT)*PASCAL
900 REM End of integration for meansquare
910 REM Obtain values of mean and variance for current value of NU, using
```

```
920 REM recurrence integrals just obtained, and values for previous NU
930 MEAN(INU)=MEAN(INU-1)-MEAN
940 VARIANCE(INU)=VARIANCE(INU-1)+(MEAN(INU-1)^2#)-MEANSQUARE-(MEAN(INU)^2#)
950 NEXT INU
960 REM End of calculation of means and variances

970 REM Begin calculation from variances of weights for weighted least squares
980 WEIGHT=0#
990 FOR I=1 TO IN
1000 WEIGHT(I)=1#/VARIANCE(I)
1010 WEIGHT=WEIGHT+WEIGHT(I)
1020 NEXT I
1030 FOR I=1 TO IN
1040 WEIGHT(I)=WEIGHT(I)/WEIGHT
1050 LPRINT USING "#############";I;:LPRINT USING "############.#";Q(I);
:LPRINT USING "########.####";MEAN(I);:LPRINT USING "#######.#####";WEIGHT(I)
1060 NEXT I
1070 REM End of calculation of weights
1080 REM Begin weighted least square fit to data to find ALPHA and PIRATIO
1090 WYQ=0#:WY=0#:WQ=0#:WQQ=0#
1100 FOR I=1 TO IN
1110 WYQ=WYQ+WEIGHT(I)*MEAN(I)*Q(I)
1120 WY=WY+WEIGHT(I)*MEAN(I)
1130 WQ=WQ+WEIGHT(I)*Q(I)
1140 WQQ=WQQ+WEIGHT(I)*Q(I)*Q(I)
1150 NEXT I
1160 ALPHA=(WYQ-WY*WQ)/(WQQ-WQ*WQ)
1170 PIRATIO=ALPHA*WQ-WY
1180 MODE=PIRATIO/ALPHA
1190 LPRINT:LPRINT:LPRINT:LPRINT "EXTREME VALUE PARAMETERS ARE:"
1200 LPRINT "MODE";:LPRINT USING "######.#";MODE;
1210 LPRINT "     ALPHA";:LPRINT USING "####.#####";ALPHA;
1220 LPRINT "     CHARACTERISTIC PRODUCT";:LPRINT USING "####.######";PIRATIO
1230 REM End of calculation of weighted least square parameters
1240 REM Begin calculation of weighted residual standard deviation
1250 SIGMAR=0#
1260 FOR I=1 TO IN
1270 SIGMAR=SIGMAR+WEIGHT(I)*((MEAN(I)-ALPHA*Q(I)+PIRATIO)^2#)
1280 NEXT I
1290 SIGMAR=SQR(SIGMAR*(N/(N-2#)))
1300 LPRINT:LPRINT"WEIGHTED RESIDUAL STANDARD DEVIATION";
:LPRINT USING"####.######";SIGMAR
1310 END

1500 REM Subroutine for Romberg Integration
1510 J=0:REM Generation counter
1520 K=-1:REM Interval Halving Counter
1530 ACCURACY=.000001
1540 H=UPPERLIMIT-LOWERLIMIT
1550 REM Up-date cycle by evaluating trapezium estimate at half intervals
1560 K=K+1
1570 M=2^K
1580 SUM=0#:ARG=-.5*H+LOWERLIMIT
```

```
1590 FOR I=1 TO M
1600 ARG=ARG+H
1610 IF SELECT$="MEAN" THEN SUM=SUM+FNMEAN(N,NU,ARG)
1620 IF SELECT$="MEANSQUARE" THEN SUM=SUM+FNMEANSQUARE(N,NU,ARG)
1630 NEXT I
1640 T(0)=.5*(S(0)+SUM*H)
1650 H=.5*H
1660 REM Trapezium up-date cycle complete
1670 REM Apply Romberg generation rule
1680 J=J+1
1690 FOR I=1 TO J
1700 X=4^I
1710 T(I)=(X*T(I-1)-S(I-1))/(X-1)
1720 NEXT I
1730 EPSILON=ABS((T(J-1)-S(J-1))):IF ABS(T(J-1))=0# THEN GOTO 1740
ELSE EPSILON=EPSILON/ABS(T(J))
1740 IF EPSILON<ACCURACY THEN GOTO 1790
1750 FOR I=0 TO J
1760 S(I)=T(I)
1770 NEXT I
1780 GOTO 1550
1790 RESULT=T(J)
1800 RETURN
1810 REM Romberg integration cycle completed;Result in RESULT
2000 REM This routine computes the value of LN(GAMMA(XX#)) as GL#
2010 REM IF GAMMA(XX#) is negative then the logarithm of its modulus is given
2020 IF XX#<0# THEN VV#=1#-XX# ELSE VV#=XX#
2030 FF#=41.62443691643907#
2040 HH#=VV#/(VV#+1#)
2050 FF#=FF#-51.22424102237477#*HH#
2060 HH#=(VV#-1#)*HH#/(VV#+2#)
2070 FF#=FF#+11.33875581348898#*HH#
2080 HH#=(VV#-2#)*HH#/(VV#+3#)
2090 FF#=FF#-.747732687772388#*HH#
2100 HH#=(VV#-3#)*HH#/(VV#+4#)
2110 FF#=FF#+.008782877493061#*HH#
2120 HH#=(VV#-4#)*HH#/(VV#+5#)
2130 FF#=FF#-.000001899030264#*HH#
2140 HH#=(VV#-5#)*HH#/(VV#+6#)
2150 FF#=FF#+.000000001946335#*HH#
2160 HH#=(VV#-6#)*HH#/(VV#+7#)
2170 FF#=FF#-.000000000199345#*HH#
2180 HH#=(VV#-7#)*HH#/(VV#+8#)
2190 FF#=FF#+.000000000008433#*HH#
2200 HH#=(VV#-8#)*HH#/(VV#+9#)
2210 FF#=FF#+.000000000001486#*HH#
2220 HH#=(VV#-9#)*HH#/(VV#+10#)
2230 FF#=FF#-.000000000000806#*HH#
2240 HH#=(VV#-10#)*HH#/(VV#+11#)
2250 FF#=FF#+.000000000000293#*HH#
2260 HH#=(VV#-11#)*HH#/(VV#+12#)
2270 FF#=FF#-.000000000000102#*HH#
2280 HH#=(VV#-12#)*HH#/(VV#+13#)
```

```
2290 FF#=FF#+.000000000000037#*HH#
2300 HH#=(VV#-13#)*HH#/(VV#+14#)
2310 FF#=FF#-.000000000000014#*HH#
2320 HH#=(VV#-14#)*HH#/(VV#+15#)
2330 FF#=FF#+.000000000000006#*HH#
2340 HH#=LOG(2.506628274631001#)+(VV#+.5#)*LOG(VV#+5.5#)+(-VV#-5.5#)
2350 FF#=HH#+LOG(FF#)-LOG(VV#)
2360 IF XX#>=0# THEN GL#=FF#:GOTO 2390
2370 PI#=3.141592653589793#
2380 GL#=PI#/(FF#*ABS(SIN(XX#*PI#)))
2390 RETURN
```

The language chosen is MICROSOFT GW-BASIC, on the basis that this can easily be run on most PCs. (More advanced versions of BASIC are downwards compatible.) The source code can be converted to another language such as FORTRAN reasonably painlessly, and compilers for GW-BASIC are also readily available.

The program has a number of sections bounded by comment statements.

The first section requests the user to enter the total number of data points (windspeeds) to be analysed, followed by the numerical values of the windspeeds. They are displayed on entry so that a visual check can be made.

The second section converts the source data from knots (as supplied by the UK Meteorological Office) to metres per second, and then squares the result in accordance with the discussion in Section 4.

In the third section the data are sorted into descending order of magnitude, using an adaptation of the FORTRAN ascending order sort routine given in Ref. [13].

In the fourth section the analytical results for the mean and variance of the largest and next-to-largest windspeeds are evaluated.

The fifth section contains the main computation by numerical integration, for the means and variances of the probability values for $3 \leqslant \nu \leqslant N$ and uses two subroutines, one of which provides the Romberg numerical integration, and the other the evaluation of the logarithm of the gamma function. This is used for the evaluation of the ${}_N C_\nu$ terms outside the integrals. The Log Gamma subroutine is taken from the book by Luke [14].

The sixth and final section of the main program uses the variances previously calculated to obtain the weights for the weighted least squares process and then evaluates the mode, measure of dispersion (α) and characteristic product, i.e. the slope and intercept of the weighted least squares straight line fitted to the data, and also the weighted residual standard deviation.

A compiled version of this program running on the writer's 486/25 PC took only a few seconds from the end of data entry to compute and print the results.

References

[1] N.J. Cook, Towards better estimation of extreme winds, J. Wind Eng. Ind. Aerodyn. 9 (1982) 295–323.

[2] J. Lieblein, Efficient methods of extreme-value methodology, report NBSIR 74-602, National Bureau of Standards, Washington (1974).

[3] E.J. Gumbel, Statistics of extremes (Columbia Univ. Press, New York, 1958).

[4] H.C. Shellard, The estimation of design wind speed, in: Proc. 1st Int. Conf. on Wind effects on buildings and structures, HMSO, London, 1965.
[5] E. Castillo, Extreme value theory in engineering (Academic Press, New York, 1988).
[6] R.I. Harris, An improved method for the prediction of extreme values of wind effects on simple buildings and structures, J. Wind Eng. Ind. Aerodyn. 9 (1982) 343–380.
[7] H. Cramer, in: Statistics of extremes, ed. E.J. Gumbel (Columbia Univ. Press, New York, 1958) p. 170.
[8] A.L. Cauchy, Cours d'analyse I, p. 167.
[9] W. Magnus and F. Oberhettinger, Special functions of mathematical physics (Chelsea, New York, 1949).
[10] E.S. Keeping, Introduction to statistical inference (Van Nostrand, Princeton, 1962).
[11] I.S. Gradshteyn and I.M. Ryzhik, Tables of integrals, series and products (Academic Press, New York, 1965).
[12] N.J. Cook, private communication.
[13] W.H. Press et al., Numerical recipes (Cambridge Univ. Press, Cambridge, 1989).
[14] Y.L. Luke, The special functions and their approximations, Vol. II (Academic Press, New York, 1969).