NOVEL EXTREME VALUE ESTIMATION PROCEDURES: APPLICATION TO EXTREME WIND DATA

John Gross (National Institute of Standards and Technology)
Alan Heckert (National Institute of Standards and Technology)
James Lechner (National Institute of Standards and Technology)
Emil Simiu (National Institute of Standards and Technology and
The Johns Hopkins University)

1. Introduction

The past two decades have seen the development of a large body of extreme value theory based on the application of the Generalized Pareto Distribution (GPD) to the excess of the extreme variate over a fixed threshold. For sufficiently large values of the extreme variates, the GPD with tail length parameters $c > 0$ and $c < 0$ is equivalent, respectively, to the Type II (Fréchet) and Type III (reverse Weibull) distribution of the largest values. The Type I (Gumbel) distribution is equivalent to the limit of the GPD as $c \rightarrow 0$. Owing to these equivalences, the GPD can be used to model extreme data obtained by either the 'peaks over threshold' approach or the epochal approach.

The overall purpose of our investigation is to assess and use the potential of GPD/extreme value theory for improving our knowledge of extreme wind speed behavior. In particular we are interested in examining the issue of the extreme distribution tail length.

J. Galambos et al. (eds.), Extreme Value Theory and Applications, 139–158.

This issue has a fairly long history. The 1972 building code requirements for minimum design loads [1] were based on the assumption that extreme winds in regions not prone to hurricanes are described by a Fréchet distribution with tail length parameter $\gamma = 9$. This assumption was examined in Ref. [11] which concluded that extreme wind speeds are better modeled by the Gumbel distribution than by the longer-tailed Fréchet distribution. However, even estimates based on the Gumbel model lead to failure probability estimates that appear to be unrealistically high [6]. This also appears to be the case for safety margins for wind loads ("load factors") [16]. The question was therefore examined whether the extreme wind distribution is in fact shorter-tailed than the Gumbel distribution [14,17]. Ref. [14], in which this question was answered affirmatively, compared the goodness of fit of the Gumbel and the regular Weibull distribution, which has a shorter but infinite upper tail. For theoretical reasons recalled earlier, the comparison should be made between the Gumbel and the **reverse** Weibull distribution, which has a limited upper tail.

We seek answers to the following questions: (1) how can we use existing probability and statistics knowledge to improve estimates of extreme wind speeds, given commonly available lengths of record (i.e., about 3 to 50 years)? (2) are wind speeds better modeled by the reverse Weibull than by the Gumbel distribution? and, if so, (3) are data sets of moderate size (say, 50 years) sufficient for estimating with acceptable confidence the length of the finite distribution tail?

The first phase of this investigation was devoted to preliminary research on the relative performance of various existing estimation methods, and on determining whether extreme wind speed data are better fitted by the reverse Weibull than by the Gumbel distribution. Based on the results of the first phase we concluded that a second phase was warranted. In this paper we review the research conducted in the first phase and describe objectives and approaches for the second, forthcoming phase.

Details on Monte Carlo simulations and analyses of largest yearly data sets are provided in Sections 2 and 3, respectively. Section 4 outlines the approach and objectives proposed for the second phase of our investigation. Section 5 presents our conclusions. An Appendix reviews the Conditional Mean Exceedance (CME), Pickands and de Haan-Dekkers-Einmahl (de Haan) methods for estimating GPD parameters, and provides expressions for estimating variates with specified mean recurrence intervals.

2. Monte Carlo Simulations

2.1 Populations, Data Samples and Numerical Results

Two sets of Monte Carlo simulations, denoted as Set I and Set II, were run and analyzed. Set I [9] was based on the assumptions that the population mean and standard deviation are $E(X)$ = 22.35 m/s (50 mph) and $s(X)$ = 2.79 m/s (6.25 mph), respectively (these values are typical of samples of annual maximum speeds in extratropical storms [12]), and that the population distributions are: (1) Gumbel (i.e., c=0); (2) reverse Weibull with γ=2 (i.e., c=-0.5 [18]; the justification for this assumption is discussed in Section 3), and (3) the normal distribution. For each of the three population distributions, 250 samples of size N=10,000 were generated. From these samples, 250 samples of size n=1000, 250 samples of size n=250, and 250 samples of size n=50 were generated by taking from the 10,000 data the largest 1000, 250 and 50 data, respectively. Our purpose in generating large samples (N=10,000) and analyzing data exceeding relatively high thresholds was to gain insight into the possible dependence of an estimator's performance on the magnitude of the threshold, that is, on how closely the sample being analyzed conforms to the asymptotic assumption inherent in GPD theory.

The estimate of the mean crossing rate for the sample of size n=50 is λ=50/10,000=0.005/year; for n=250 and n=1000, the estimates are λ=0.025/year and λ=0.1/year, respectively. These estimates are reasonable as long as the population data are interpreted as annual maximum wind speeds, which is consistent with the assumed values of $E(X)$ and $s(X)$. (In Ref. [9] the mean crossing rates were arbitrarily assumed to be 1.25/year, 6.25/year and 25/year for n = 50, 250, and 1000, respectively. For any given probability of occurrence of a variate, the respective nominal mean recurrence intervals, R, in years were therefore correspondingly smaller in Ref. [9] than in this paper, see Eq. 8 in the Appendix).

For each of the three crossing rates the CME, Pickands and de Haan estimation methods were applied to obtain estimates of the GPD parameters and, based on these parameters, estimates of extreme wind speeds for mean recurrence intervals of R=250, 5000, and 50,000 years were computed. A measure of the performance of the estimators is the root-mean-square-error which is the square root of the sum of the variance and the square of the bias (the difference between the estimated mean and the population mean). The means (M), standard deviations (SD) and root-mean-square-errors (RMSE) of the estimates based on the sets of simulated 250 samples are shown in Tables 1, 2 and 3 for the Gumbel, reverse Weibull and normal distributions.

Set II [7] was based on the assumptions that the population mean and standard deviation were $E(X)$ = 12.96 m/s (29 mph) and $s(X)$ = 2.91 m/s (6.5 mph), respectively. Further, it was assumed that the population distributions are (1) Gumbel (i.e., c=0); reverse Weibull distribution with γ=-1/c=3.64; and normal. The assumed population parameters were estimated from two sets of maximum daily data recorded over periods of about 25 years at Boise, Idaho and Toledo, Ohio. Those sets were censored below, that is, only speeds larger than 8.05 m/s (18 mph) were used in the estimation of $E(X)$, $s(X)$ and γ. This limit was chosen because it was judged that lower wind speeds (e.g., morning breezes) are likely not to belong to the same meteorological class as extreme winds and, if included in the analysis, may vitiate the probabilistic description of the extremes. The resulting censored samples included about 750 data (i.e., roughly 30 data per year of record). The mean crossing rate was therefore assumed to be λ=30/year. For each population distribution, 500 samples of size 750 (25 years of data) and 1200 (40 years of data) were generated. For various crossing rates (λ = 20, 15, 10, 5, 2 and 1 per year), the CME, Pickands, and de Haan estimation methods were applied to obtain estimates of the GPD parameters and, based on these parameters, estimates of extreme wind speeds for mean recurrence intervals of 50, 500 and 5000 years were computed. The means, standard deviations and root-mean-square-errors based on the 500 simulated samples are shown in Tables 4, 5, and 6 for the Gumbel, reverse Weibull and normal distributions based on 25-year (750 data) samples.

2.2 Comparison of Estimation Methods

We compare the relative efficiency of the CME, Pickands and de Haan estimation methods. For both Set I and Set II it is apparent from Tables 1 through 6 that, with insignificant exceptions, the Pickands estimator was outperformed by the CME and de Haan estimators. We note that the results given in Tables 1-6 for the Pickands estimator were based on the NIST implementation (see Appendix). Similar or worse results were obtained for the original Pickands estimator. For example, for 25-year simulated records taken from a Gumbel population for λ=10/yr and R=50 yrs, 500 yrs and 5000

λ	Method	a		c		R = 250 yrs			R = 5000 yrs			R = 50,000 yrs		
		M	SD	M	SD	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)
0.100	CME	5.26	0.24	-0.06	0.03	32.85	0.27	0.38	38.15	0.91	1.75	41.65	1.65	3.44
	DEH	5.03	0.22	-0.02	0.03	33.05	0.28	0.29	39.35	0.99	1.03	44.03	1.87	1.98
	PIC	5.11	0.29	-0.03	0.06	33.02	0.43	0.44	39.10	1.76	1.85	43.05	2.83	3.26
0.025	CME	4.96	0.52	-0.02	0.08	33.08	0.31	0.31	39.41	1.31	1.33	43.73	2.50	2.66
	DEH	4.94	0.44	-0.01	0.06	33.08	0.30	0.31	39.47	1.09	1.11	44.13	2.20	2.27
	PIC	4.93	0.60	-0.01	0.12	33.08	0.34	0.34	39.60	1.99	1.99	43.25	3.23	3.51
0.005	CME	5.03	1.06	-0.05	0.17	33.09	0.34	0.34	39.31	1.16	1.21	43.20	2.60	2.99
	DEH	5.02	0.95	-0.05	0.15	33.09	0.34	0.34	39.32	1.14	1.19	43.49	2.57	2.82
	PIC	4.89	1.16	-0.02	0.25	33.07	0.34	0.34	39.41	1.87	1.87	42.59	3.13	3.76
Population				0.00		33.12			39.65			44.66		

Table 1 - Threshold Procedure / Gumbel Distribution / 10,000 Years of Simulated Data

λ	Method	a		c		R = 250 yrs			R = 5000 yrs			R = 50,000 yrs		
		M	SD	M	SD	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)
0.100	CME	2.27	0.10	-0.53	0.03	27.31	0.03	0.02	27.58	0.04	0.05	27.63	0.04	0.05
	DEH	2.28	0.12	-0.54	0.05	27.30	0.03	0.05	27.58	0.07	0.08	27.62	0.08	0.09
	PIC	2.27	0.18	-0.52	0.29	27.30	0.04	0.04	27.58	0.08	0.09	27.62	0.10	0.11
0.025	CME	1.08	0.09	-0.51	0.06	27.31	0.03	0.03	27.60	0.03	0.04	27.66	0.04	0.04
	DEH	1.09	0.11	-0.52	0.09	27.31	0.03	0.03	27.60	0.05	0.05	27.66	0.07	0.07
	PIC	1.08	0.12	-0.51	0.11	27.31	0.03	0.03	27.61	0.08	0.08	27.67	0.12	0.12
0.005	CME	0.49	0.09	-0.52	0.14	27.31	0.03	0.03	27.60	0.03	0.03	27.66	0.04	0.04
	DEH	0.49	0.11	-0.54	0.21	27.31	0.03	0.03	27.60	0.03	0.04	27.66	0.06	0.06
	PIC	0.39	0.23	0.26	1.59	27.30	0.04	0.04	25.65	10.52	2.19	27.63	0.19	0.19
Population				-0.50		27.31			27.61			27.67		

Table 2 - Threshold Procedure / Reverse Weibull Distribution / 10,000 Years of Simulated Data

λ	Method	a		c		R = 250 yrs			R = 5000 yrs			R = 50,000 yrs		
		M	SD	M	SD	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)
0.100	CME	1.53	0.06	-0.17	0.03	25.64	0.06	0.06	26.59	0.17	0.22	27.07	0.25	0.50
	DEH	1.51	0.08	-0.14	0.04	25.70	0.06	0.08	26.76	0.20	0.20	27.34	0.33	0.36
	PIC	1.56	0.14	-0.17	0.29	25.64	0.08	0.08	26.54	0.25	0.33	26.95	0.39	0.64
0.025	CME	1.17	0.11	-0.11	0.08	25.67	0.06	0.06	26.79	0.22	0.22	27.46	0.47	0.47
	DEH	1.17	0.11	-0.10	0.07	25.67	0.06	0.06	26.79	0.20	0.20	27.48	0.42	0.42
	PIC	0.56	0.64	2.28	2.16	25.45	20.28	5.05	26.67	0.25	0.28	27.18	0.50	0.59
0.005	CME	1.00	0.22	-0.22	0.17	25.64	0.06	0.08	26.76	0.20	0.20	27.43	0.59	0.59
	DEH	1.00	0.20	-0.11	0.15	25.64	0.06	0.08	26.76	0.20	0.20	27.46	0.47	0.47
	PIC	0.36	0.59	2.27	1.85	25.59	0.08	0.11	28.43	27.23	5.83	27.29	2.46	2.48
Population						25.64			26.76			27.48		

Table 3 - Threshold Procedure / Normal Distribution / 10,000 Years of Simulated Data

λ	Method	a		c		R = 50 yrs			R = 500 yrs			R = 5000 yrs		
		M	SD	M	SD	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)
20	CME	6.60	0.67	-0.08	0.07	27.34	2.06	2.24	31.31	3.94	4.48	34.92	6.49	7.50
	DEH	7.17	0.52	-0.17	0.06	24.46	1.68	4.12	26.44	2.47	7.42	27.84	3.21	11.28
	PIC	7.70	0.51	-0.23	0.06	23.49	1.58	4.99	24.89	2.30	8.86	25.78	2.93	13.20
15	CME	6.10	0.70	-0.06	0.08	27.71	2.21	2.28	32.20	4.54	4.71	36.58	8.08	8.35
	DEH	6.34	0.51	-0.10	0.06	26.29	1.99	2.77	29.50	3.27	5.12	32.21	4.69	7.98
	PIC	6.75	0.55	-0.16	0.08	25.21	2.32	3.80	27.65	3.79	6.92	29.58	5.40	10.56
10	CME	5.73	0.79	-0.04	0.09	27.92	2.33	2.35	32.88	5.19	5.22	38.04	10.09	10.11
	DEH	5.81	0.55	-0.06	0.07	27.38	2.30	2.45	31.61	4.11	4.50	35.59	6.40	7.09
	PIC	6.02	0.63	-0.09	0.11	26.98	3.31	3.53	31.06	6.55	6.97	35.13	11.37	11.90
5	CME	5.50	1.07	-0.04	0.13	28.00	2.45	2.47	33.40	6.22	6.22	39.68	14.00	14.04
	DEH	5.47	0.74	-0.04	0.10	27.96	2.71	2.72	33.00	5.41	5.43	38.26	9.36	9.37
	PIC	5.56	0.85	-0.05	0.16	28.31	4.68	4.68	34.67	12.39	12.45	43.41	29.91	30.28
2	CME	5.48	1.61	-0.05	0.20	27.91	2.49	2.51	33.74	7.35	7.35	41.93	19.85	20.12
	DEH	5.42	1.23	-0.06	0.17	27.92	2.82	2.83	33.17	6.18	6.19	39.15	11.73	11.74
	PIC	5.32	1.35	-0.05	0.25	28.71	5.78	5.80	38.36	26.40	26.86	50.65	47.68	49.17
1	CME	5.75	2.17	-0.09	0.27	27.88	2.61	2.63	33.82	8.31	8.31	43.63	26.09	26.56
	DEH	5.58	1.69	-0.08	0.24	28.08	3.04	3.04	33.69	7.41	7.41	40.84	15.71	15.86
	PIC	5.18	1.95	-0.02	0.38	29.60	6.69	6.83	46.55	47.02	48.82	61.20	65.07	68.86
Population						28.22			33.44			38.66		

Table 4 - Threshold Procedure / Gumbel Distribution / 25 Years of Simulated Data

λ	Method	a		c		R = 50 yrs			R = 500 yrs			R = 5000 yrs		
		M	SD	M	SD	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)
20	CME	8.77	0.56	-0.42	0.05	20.52	0.40	0.67	20.86	0.49	0.94	20.99	0.54	1.20
	DEH	9.80	0.66	-0.59	0.08	19.05	0.59	2.09	19.16	0.65	2.58	19.19	0.67	2.95
	PIC	9.62	0.58	-0.54	0.05	19.42	0.42	1.69	19.56	0.47	2.15	19.60	0.49	2.52
15	CME	7.08	0.52	-0.37	0.05	20.75	0.43	0.53	21.19	0.55	0.72	21.39	0.63	0.93
	DEH	7.49	0.53	-0.45	0.08	20.03	0.69	1.24	20.30	0.81	1.59	20.40	0.88	1.88
	PIC	7.58	0.56	-0.46	0.07	20.02	0.59	1.19	20.28	0.72	1.56	20.37	0.78	1.86
10	CME	5.69	0.50	-0.34	0.06	20.90	0.45	0.47	21.44	0.61	0.65	21.70	0.72	0.81
	DEH	5.85	0.49	-0.38	0.08	20.60	0.76	0.88	21.05	0.97	1.15	21.25	1.11	1.38
	PIC	5.90	0.58	-0.38	0.10	20.58	0.82	0.95	21.04	1.14	1.29	21.26	1.36	1.59
5	CME	4.35	0.59	-0.32	0.09	20.98	0.47	0.48	21.59	0.71	0.71	21.93	0.90	0.91
	DEH	4.39	0.53	-0.33	0.11	20.94	0.81	0.82	21.56	1.18	1.18	21.91	1.47	1.48
	PIC	4.40	0.64	-0.33	0.14	20.97	1.05	1.06	21.68	1.82	1.82	22.16	2.70	2.70
2	CME	3.25	0.74	-0.31	0.16	21.00	0.50	0.50	21.72	0.89	0.90	22.20	1.35	1.35
	DEH	3.30	0.69	-0.34	0.19	20.96	0.72	0.73	21.64	1.18	1.18	22.07	1.61	1.61
	PIC	3.17	0.82	-0.27	0.44	21.08	1.14	1.14	22.14	2.89	2.93	23.31	6.61	6.73
1	CME	2.75	0.86	-0.33	0.24	21.00	0.51	0.51	21.77	1.04	1.05	22.37	1.83	1.86
	DEH	2.75	0.82	-0.34	0.28	21.03	0.70	0.71	21.83	1.30	1.31	22.43	2.01	2.04
	PIC	2.41	1.09	-0.02	0.91	25.34	33.14	33.42	23.43	6.58	6.81	28.05	27.41	28.05
Population				-0.50		21.05			21.72			22.07		

Table 5 - Threshold Procedure / Reverse Weibull Distribution / 25 Years of Simulated Data

λ	Method	a		c		R = 50 yrs			R = 500 yrs			R = 5000 yrs		
		M	SD	M	SD	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)
20	CME	7.79	0.59	-0.33	0.05	21.30	0.63	1.17	21.93	0.85	2.30	22.24	1.01	3.49
	DEH	9.06	0.65	-0.54	0.08	19.06	0.68	3.30	19.20	0.76	4.91	19.25	0.79	6.39
	PIC	8.99	0.56	-0.49	0.05	19.61	0.45	2.71	19.81	0.51	4.28	19.87	0.54	5.74
15	CME	6.24	0.55	-0.27	0.06	21.68	0.71	0.93	22.56	1.05	1.83	23.07	1.33	2.85
	DEH	6.82	0.51	-0.39	0.08	20.33	0.83	2.12	20.74	1.02	3.47	20.93	1.15	4.80
	PIC	7.04	0.53	-0.40	0.07	20.33	0.74	2.09	20.71	0.94	3.48	20.88	1.08	4.83
10	CME	5.08	0.55	-0.22	0.07	21.94	0.78	0.85	23.09	1.30	1.62	23.84	1.82	2.52
	DEH	5.34	0.47	-0.29	0.08	21.23	0.88	1.38	21.99	1.21	2.40	22.42	1.46	3.49
	PIC	5.55	0.55	-0.32	0.09	21.05	0.98	1.58	21.72	1.41	2.73	22.10	1.77	3.91
5	CME	4.04	0.62	-0.18	0.11	22.13	0.88	0.89	23.56	1.67	1.74	24.66	2.62	2.78
	DEH	4.14	0.51	-0.21	0.11	21.88	1.07	1.14	23.10	1.77	2.02	23.97	2.53	3.00
	PIC	4.23	0.63	-0.24	0.14	21.82	1.42	1.49	23.11	2.65	2.82	24.15	4.18	4.42
2	CME	3.39	0.83	-0.17	0.17	22.18	0.94	0.95	23.83	2.04	2.06	25.34	3.76	3.77
	DEH	3.40	0.66	-0.19	0.16	22.14	1.16	1.17	23.71	2.29	2.32	25.06	3.79	3.82
	PIC	3.35	0.85	-0.18	0.24	22.38	1.99	2.00	24.95	6.20	6.27	28.93	17.07	17.39
1	CME	3.20	1.09	-0.20	0.25	22.16	0.95	0.96	23.86	2.24	2.25	25.67	4.77	4.77
	DEH	3.15	0.87	-0.21	0.24	22.20	1.19	1.19	23.87	2.46	2.47	25.44	4.33	4.33
	PIC	2.97	1.15	-0.11	0.55	25.01	26.12	26.26	27.31	16.21	34.08	33.16	31.85	25.58
Population						22.29			24.06			25.59		

Table 6 - Threshold Procedure / Normal Distribution / 25 Years of Simulated Data

yrs, the original Pickands method yielded root-mean-square-errors of 7.37, 25.58 and 41.47, respectively, whereas the NIST implementation of the Pickands method yielded 3.53, 6.97 and 11.90.

From the results for Set II, we observe that the de Haan estimator outperforms the CME estimator for small values of $\boldsymbol{\lambda}$ (low crossing rates) and large mean recurrence intervals, $\boldsymbol{R}$. Similar results were obtained from the analyses of 40-year (1200 data) sets. The results we noted appear to be consistent with those obtained for Set I (Tables 1, 2 and 3).

2.3 Optimal Crossing Rates

A high threshold reduces the bias since it conforms best with the asymptotic assumption on which the GPD distribution is based; however, because it results in a smaller number of data, it increases the sampling error. An optimal exceedance rate $\boldsymbol{\lambda}$ exists for which the root-mean-square-error is a minimum. From Tables 4, 5 and 6 it is seen that the optimal $\boldsymbol{\lambda}$ depends on the population distribution and the mean recurrence interval. Results shown in Tables 1 through 3 are consistent with this observation.

Tables 4 through 6 suggest that, with no significant error, an approximately optimal threshold may be assumed to correspond to a mean exceedance rate of 5/yr to 15/yr. The 40-year sets yielded similar results.

2.4 Comparison of Threshold and Epochal Procedures

Next we compare the epochal procedure, traditionally used in the estimation of extreme wind speeds, with the threshold procedure. Epochal extreme wind data were generated by taking the largest variate in each successive set of 30 (maximum crossing rate) data from the parent populations with mean $\boldsymbol{E(X)}$ = 12.96 m/s (29 mph) and standard deviation $\boldsymbol{s(X)}$ = 2.91 m/s (6.5 mph). This resulted in 500 sets of 25 yearly maxima and 500 sets of 40 yearly maxima [7]. The CME, Pickands, de Haan and Probability Plot Correlation Coefficient (PPCC) [13] estimation methods were applied to the data. Results are given in Table 7.

For the epochal approach, a comparison among estimation methods, based on the root-mean-square-error, showed that the CME method performed best in a majority of the cases.

Method	a		c		R = 50 yrs			R = 500 yrs			R = 5000 yrs		
	M (m/s)	SD (m/s)	M (m/s)	SD (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)	M (m/s)	SD (m/s)	RMSE (m/s)
Gumbel Distribution													
CME	11.39	4.19	-0.34	0.32	27.54	2.35	2.45	31.38	6.28	6.61	35.79	16.02	16.27
DEH	17.57	7.92	-1.08	0.73	24.25	2.52	4.70	24.97	3.62	9.21	25.37	4.61	14.06
PIC	11.60	3.02	-0.52	0.25	25.44	2.48	3.73	27.13	5.51	8.38	28.58	13.87	17.15
PPCC					29.12	3.79	3.89	39.99	29.54	30.25	77.74	285.83	288.48
Population					28.22			33.44			38.66		
Reverse Weibull Distribution													
CME	8.03	3.03	-0.61	0.35	21.91	0.98	1.05	22.72	1.80	2.24	23.20	2.65	3.56
DEH	14.00	10.35	-1.74	1.34	20.33	1.15	2.28	20.46	1.48	3.88	20.53	1.69	5.33
PIC	7.08	1.83	-0.70	0.26	20.78	1.09	1.86	21.19	1.60	3.29	21.38	2.13	4.72
PPCC					22.36	1.16	1.17	24.44	3.00	3.03	26.76	6.54	6.65
Population					22.29			24.06			25.59		
Normal Distribution													
CME	8.39	2.95	-0.84	0.33	20.82	0.47	0.52	21.08	0.70	0.94	21.17	0.86	1.25
DEH	15.01	8.81	-2.21	1.29	19.65	0.66	1.55	19.69	0.71	2.15	19.69	0.73	2.49
PIC	6.72	1.73	-0.87	0.24	19.92	0.63	1.30	20.07	0.80	1.82	20.13	1.00	2.19
PPCC					20.96	0.52	0.53	21.77	1.00	1.01	22.34	1.64	1.66
Population					21.05			21.72			22.07		

Table 7 - Epochal Procedure / 25 Years of Simulated Data

Results for the epochal procedure using the CME estimation method were then compared with results for the threshold procedure using a crossing rate of 10/yr. It was found that for the Gumbel and reverse Weibull distributions, the threshold procedure produced better estimates. For the normal distribution, the epochal method performed better. This may be due to the fact that, unlike the Gumbel and the reverse Weibull distribution, the normal distribution is not an extreme value distribution. Consequently, data taken from a Gumbel or reverse Weibull population that exceed a specified threshold (with λ=10/yr, say) have an extreme value distribution while similarly obtained data from a normal population do not. On the other hand, data consisting of *maximum yearly values* taken from a normal population *may* be characterized more closely by an extreme value distribution.

3. Analyses of Largest Observed Maximum Yearly Data

We summarize results of analyses performed on sets of about 20 to 45 yearly maximum wind speeds recorded at various U.S. sites [8]. The CME method was used to estimate the tail length parameter *c* of the corresponding GPDs. For more than two-thirds of 95 data samples at stations not affected by hurricanes, the estimated *c* values were between approximately -0.35 and -0.80. CME graphs for 32 selected stations are shown in Fig. 1 [8]. The choice c=-0.5 in the Monte Carlo simulations of Section 2 was based on these results.

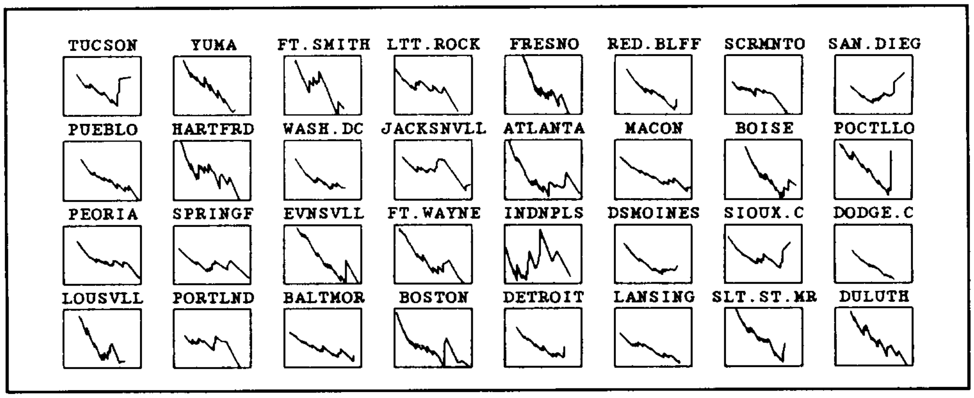

Fig. 1 - Cumulative Mean Exceedance (CME) Plots for Extreme Wind Speeds at 32 Stations

Reference [8] also reports simulations from Gumbel, reverse Weibull and Fréchet distributions intended to provide a framework for interpreting the results of the observed maximum yearly data. The simulation results tended to support the view that most of the stations are in fact best fitted by reverse Weibull distributions. This is also true of analyses of recorded extreme wind speed data reported in Ref. [19].

4. Approach and Objectives of Current Research

The purpose of the second phase is to verify and refine the work performed so far by analyzing a relatively large number of sets of daily maximum fastest mile wind speeds. To this end, we assembled 48 complete sets of daily maximum fastest mile wind speeds recorded over 15 to 26 years. For each of these sets the data were reduced to a common elevation using the procedure of Ref. [12]. For a few of the sets a small number (less than one percent) of the daily data were missing. We filled in the missing data by estimating their values from records of wind speeds measured on the respective days at 3-hour intervals. In most cases the estimated wind speeds being filled in were relatively small — 13.4 m/s (30 mph) or less — so that errors in their estimation could not be expected to alter significantly the estimates of the extremes.

In the second phase of our investigation we intend to address the following topics:

(a) From the sets of daily data, it is necessary to extract sets of uncorrelated or weakly correlated wind speeds. The latter must then be analyzed to determine their best fitting distributions. Monte Carlo simulations similar to those reviewed in Section 3 will then be performed from populations described by these best-fitting distributions.

(b) The sets of uncorrelated or weakly correlated data will contain a number of speeds irrelevant to the estimation of the extremes. These speeds (e.g., morning breezes) are generated by meteorological processes that differ from those that generate the extreme winds. We plan to investigate the effect of eliminating all data below a reasonably selected threshold — say 7 m/s (16 mph). We will also investigate the effect of the choice of the censoring threshold.

(c) In spite of the elimination of irrelevant data, the resulting data sets will still be meteorologically inhomogeneous: we are unable in practice — at least for the time being — to separate data associated with thunderstorms from data associated with large-scale extratropical storms. However, we can create separate sets of data for distinct periods of the year, e.g., for each season, or even each month. The probability that a given wind speed will not be exceeded during N years is then estimated as the probability that it will not be exceeded by wind speeds in any of the segregated categories. Under the reasonable assumption of independence, that probability will be equal to the product of the probabilities estimated for each of those categories. The predominance (or rarity) of thunderstorms among high winds occurring during certain periods of the year will be reflected by the probability distributions fitted to the corresponding segregated data sets. Errors inherent in the mixing of thunderstorms and non-thunderstorm winds in the unsegregated data sets would thus be reduced, if not altogether eliminated. We plan to compare results based on data segregated by seasons or months with results obtained from the unsegregated data sets.

(d) An important practical problem in extreme wind climatology is the estimation of extreme wind speeds from short-term records (e.g., three-year records). Epochal estimates based on maximum monthly winds have shown that 3-year data sets can yield respectable estimates of extreme speeds [15]. However, these estimates raise the question of the effect of seasonality — at most stations summer winds have considerably lower means than winter winds. The significance of this effect on the reliability of estimates based on monthly extremes has not been studied so far. We plan to obtain estimates based on, say, 3-year records by using the 'peaks over threshold' approach, which reduces seasonality effects, since one would expect most low-season winds to be discarded via the thresholding procedure. We plan to compare estimates based on the 'peaks over threshold' procedure applied to short (e.g., 3-year) data sets with estimates based on the entire (15-year to 26-year) period of record.

5. Summary and Conclusions

In this paper we reviewed results obtained in a first phase of an investigation on the application of the GPD-based approach to the estimation of extreme winds. The results were the following: (1) for data exceeding a sufficiently high threshold, as well as for epochal data, the Cumulative Mean Exceedance (CME) and the de Haan-Dekkers-Einmahl (de Haan) estimation methods performed significantly better than the Pickands method; (2) for the threshold approach, the performance of the estimates was optimal for mean crossing rates of about 10 to 15 per year; (3) for the epochal approach, the CME method performed better than the Probability Plot Correlation Coefficient (PPCC) method; (4) for data from Gumbel and reverse Weibull distributions the threshold approach performed better than the epochal approach; however, the reverse was true for data from a normal distribution; (5) most sets of observed largest yearly extratropical wind speed data (not including tornado winds) were better fitted by reverse Weibull distributions than by Gumbel distributions.

We also outlined our approach and objectives for a second phase of this research. This approach uses statistical analyses of data taken from sets of maximum daily fastest mile speeds and Monte Carlo simulations based on probabilistic models derived from those analyses. The objectives are to rank the performances of the CME and de Haan procedures, account for seasonality and correlation effects, verify whether most extreme speed data sets are in fact better fitted by reverse Weibull than by Gumbel distributions, confirm or refine earlier estimates of optimal exceedance rates for 'peaks over threshold' procedures, confirm or refine earlier conclusions on the relative performance of 'peaks over threshold' and epochal procedures, and estimate confidence levels for 'peaks over threshold' estimates of extreme winds based on short records.

Acknowledgement

E. Simiu acknowledges with thanks partial support from the National Science Foundation (Grant MSM-9013116 to the Department of Civil Engineering, The Johns Hopkins University) and the helpful interaction on this project with R.B. Corotis of The Johns Hopkins University.

References

[1] American National Standard A58.1, Building Code Requirements for Minimum Design Loads for Buildings and Other Structures, American National Standards Institute, New York, 1972.

[2] Castillo, E., Extreme Value Theory in Engineering, Academic Press, New York, 1988.

[3] Davisson, A.C., and Smith, R.L., Models of exceedances over high thresholds," Journal of the Royal Statistical Society, B52 (1990), 339-442.

[4] Dekkers, A.L.M. and de Haan, L., On the estimation of the extreme-value index and large quantile estimation, Annals of Statistics, 17 (1989), 1795-1932.

[5] Dekkers, A.L.M., Einmahl, J.H.J., and de Haan, L., a moment estimator for the index of an extreme-value distribution, Annals of Statistics, 17 (1989), 1833-1855.

[6] Ellingwood, B. et al., Development of a probability-based load criterion for American National Standard A58, NBS Special Publication 577, National Bureau of Standards, Washington, D.C., 1980.

[7] Gross, J.L., Heckert, N.A., Lechner, J.A. and Simiu, E., Modeling of extreme loading by 'peaks over threshold' methods, Dynamic Response and Progressive Failure of Special Structures, ASCE, New York, in press.

[8] Lechner, J.A., Leigh, S.D. and Simiu, E., Recent approaches to extreme value estimation with application to extreme wind speeds, J. Wind Eng. Ind. Aerod., 41-44 (1992), 509-519.

[9] Lechner, J.A., Simiu, E. and Heckert, N.A., Assessment of 'peaks over threshold' methods for estimating extreme value distribution tails, Structural Safety (in press).

[10] Pickands, J., Statistical inference using order statistics, Annals of Statistics, 3 (1975), 119-131.

[11] Simiu, E., Biétry J., and Filliben, J.J., Sampling errors in estimation of extreme wind speeds, J. Struct. Div., ASCE, 104 (1978), 491-501.

[12] Simiu, E., Changery, M. and Filliben, J.J., Extreme Wind Speeds at 129 Stations in the Contiguous United States, NBS Building Science Series 18, National Bureau of Standards, Washington, D.C., 1979.

[13] Simiu, E. and Filliben, J.J., Probability distributions of extreme wind speeds, J. Struct. Div., 102 (1976), 1861-1877.

[14] Simiu, E. and Filliben, J.J., Weibull distributions and extreme wind speeds, J. Struct. Div., 106 (1980) 2365-2374; errata in J. Struct. Div., 107 (1981), 716-717.

[15] Simiu, E., Filliben, J.J. and Shaver, J.R., Short-term records and extreme wind speeds, J. Struct. Div. 108 (1092), 1467-1484.

[16] Simiu, E., and Scanlan, R. H., Wind Effects on Structures, second edition, Wiley, New York, 1986.

[17] Simiu, E., Shaver, J.R. and Filliben, J.J., Wind speed distributions and reliability estimates, J. Struct. Div., 107 (1981) 1003-1007; errata in J. Struct. Div., 107 (1981), 2052.

[18] Smith, R. L., Extreme value theory. In Handbook of Applicable Mathematics, Supplement (eds. Ledermann W. et al., Wiley, New York, 1989, pp. 437-472.

[19] Walshaw, D., Getting the most from your extreme wind data: a step by step guide, NIST Journal of Research (in press).

Appendix

Generalized Pareto Distribution

The expression for the Generalized Pareto Distribution (GPD) is:

$$G(y) = Prob[Y<y] = 1 - \left[1 + (\frac{cy}{a})\right]^{-1/c} \qquad a>0,\ 1+(\frac{cy}{a})>0 \quad \ldots (1)$$

Eq. 1 can be used to represent the conditional cumulative distribution of the excess $\boldsymbol{Y = X - u}$ of the variate $\boldsymbol{X}$ over the threshold $\boldsymbol{u}$, given $\boldsymbol{X > u}$ for $\boldsymbol{u}$ sufficiently large [10].

Conditional Mean Exceedance (CME), Pickands and de Haan Methods

1. The CME is the expectation of the amount by which a value exceeds a threshold $\boldsymbol{u}$, conditional on that threshold being attained. If the exceedance data are fitted by the GPD model and $\boldsymbol{c < 1}$, $\boldsymbol{u > 0}$, and $\boldsymbol{a+uc > 0}$, then the CME plot (i.e., CME vs. $\boldsymbol{u}$) should follow a line with intercept $\boldsymbol{a/(1-c)}$ and slope $\boldsymbol{c/(1-c)}$ [3]. The linearity of the CME plot can thus be used as an indicator of the appropriateness of the GPD model, and both $\boldsymbol{c}$ and $\boldsymbol{a}$ can be estimated from the CME plot.

2. Following Pickands' [10] notation, let $X_{(1)} \geq \ldots \geq X_{(n)}$ denote the order statistics (ordered sample values) of a sample of size $\boldsymbol{n}$. For $s = 1, 2, \ldots, [n/4]$

([ ] denoting largest integer part of), one computes $F_s(x)$, the empirical estimate of the exceedance CDF

$$F_s(x) = Prob(X - X_{(4s)} < x \mid X > X_{(4s)}) \ldots\ldots\ldots\ldots \quad (2)$$

and $G_s(x)$, the Generalized Pareto distribution, with a and c estimated by

$$\hat{c} = \frac{\log\{(X_{(s)} - X_{(2s)})/(X_{(2s)} - X_{(4s)})\}}{\log(2)} \ldots\ldots\ldots\ldots \quad (3)$$

$$\hat{a} = \frac{\hat{c}(X_{(2s)} - X_{(4s)})}{2^{\hat{c}} - 1} \ldots\ldots\ldots\ldots \quad (4)$$

One takes for Pickands estimators of c and a those values which minimize (for $1 < s < [n/4]$) the maximum distance between the empirical exceedance CDF and the GPD model. Pickands' method can be shown to be consistent.

Following a critique of an earlier implementation of the Pickands method [2, 10] an alternative implementation was developed [8] which entailed the following steps: (1) choose as threshold u an order statistic of the sample; (2) compute the empirical exceedance CDF for the data above u; (3) nonlinear least-squares fit the GPD model for the parameters c and a; (4) plot the resulting c estimates against u for each order statistic.

3. Following Dekkers, Einmahl and de Haan, [4, 5] consider an integer-valued function of n, $k(n)$, such that, as $n \to \infty$, $k(n) \to \infty$ and $k(n)/n \to 0$ (e.g., $k(n) = [\sqrt{n}]$. Compute the quantities

$$M_n^{(1)} = \frac{1}{k(n)} \sum_{i=1}^{k(n)} \{\log X_{(i)} - \log X_{(k(n))}\} \ldots\ldots\ldots\ldots \quad (5)$$

$$M_n^{(2)} = \frac{1}{k(n)} \sum_{i=1}^{k(n)} \{\log X_{(i)} - \log X_{(k(n))}\}^2 \ldots\ldots\ldots\ldots \quad (6)$$

The estimator of c is then

$$\hat{c} = M_n^{(1)} + 1 - \frac{1}{2\{1 - (M_n^{(1)})^2/(M_n^{(2)})\}} \ldots\ldots\ldots\ldots \quad (7)$$

The estimator of a is obtained as the CME value for $X_{(k(n))}$ times $1 - \hat{c}$, where $\hat{c}$ is given by Eq. 7.

Estimation of Variates with Specified Mean Recurrence Intervals

The estimates of the wind speeds, x_R, corresponding to the mean recurrence intervals, R (in years), are of interest. Let λ denote the mean crossing rate of the threshold, u, per year (i.e., the average number per year of data points above u). We have

$$Prob[Y<y] = 1 - 1/(\lambda R) \qquad (8)$$

$$1 - [1+(cy/a)]^{-1/c} = 1 - 1/(\lambda R) \qquad (9)$$

$$y = -a[1-(\lambda R)^c]/c \qquad (10)$$

The value being sought is

$$x_R = y + u \qquad (11)$$

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

John Gross, Building and Fire Research Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899.

Alan Heckert, Computing and Applied Mathematics Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899.

James Lechner, Computing and Applied Mathematics Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899.

Emil Simiu, Building and Fire Research Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899.