On the distributions used and the corresponding fitting techniques

Abstarct

This is a supplement to the report for the geometrical entities and techniques that better fit to the study of wind and wave forecasts for selected areas in Greece and provides detailed information on the distributions tested for the wind and wave data under study as well as on the fitting tests adopted. The distributions are described by means of their probability density functions and the corresponding cumulative functions. On the other hand, the fitting tests adopted are discussed focusing on the optimization criteria and the significance level at which they are tested. The results obtained are presented for an indicative test case in which the prevalence of the adopted Weibull distribution is proved. Similar outputs have been recorded for all the other sites under study.

The Distributions Tested

For each area under study the probability density functions that optimally fit the three data sets obtained (observations, MARINA project's results, operational forecasts) are defined. The distributions used in order to select the one that optimally fits the data are the following:

Normal

Probability density function

$$f(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)}{\sigma\sqrt{2\pi}}$$

Cumulative density function

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

σ: scale parameter (σ>0)

μ: location parameter

Weibull

Probability density function

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)$$

Cumulative density function

$$F(x) = 1 - \exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)$$

α: shape parameter (a>0)

β: scale parameter (β>0)

γ: location parameter (γ=0 : gives the two-parameter Weibull distribution)

Lognormal

Probability Density Function

$$f(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)^2\right)}{(x-\gamma)\,\sigma\sqrt{2\pi}}$$

Cumulative distribution

$$F(x) = \Phi\left(\frac{\ln(x-\gamma)-\mu}{\sigma}\right)$$

μ: shape parameter (μ>0)

σ: scale parameter (σ>0)

γ: location parameter (γ=0 : gives the two-parameter Lognormal distribution)

Generalized Extreme Value distribution

Probability density function

$$f(x) = \begin{cases} \frac{1}{\sigma} \exp\left(-(1+k\,z)^{-1/k}\right)(1+k\,z)^{-1-1/k} & k \neq 0 \\ \frac{1}{\sigma} \exp(-z - \exp(-z)) & k = 0 \end{cases}$$

Cumulative density function

$$F(x) = \begin{cases} \exp\left(-(1+k\,z)^{-1/k}\right) & k \neq 0 \\ \exp(-\exp(-z)) & k = 0 \end{cases}$$

where $z \equiv \frac{x-\mu}{\sigma}$

k: shape parameter

σ: scale parameter (σ>0)

μ: location parameter

Exponential

Probability density function

$$f(x) = \lambda \exp(-\lambda(x-\gamma))$$

Cumulative density function

$$F(x) = 1 - \exp(-\lambda(x-\gamma))$$

λ: inverse scale parameter (λ>0)

γ: continuous location parameter (γ=0 : one-parameter exponential distribution)

Log-Logistic

Probability density function

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{-2}$$

Cumulative density function

$$F(x) = \left(1+\left(\frac{\beta}{x-\gamma}\right)^{\alpha}\right)^{-1}$$

a: shape parameter (a>0)

β: scale parameter (β>0)

γ: location parameter (γ=0 : gives the two-parameter Lognormal distribution)

Gamma

Probability density function

$$f(x) = \frac{(x-\gamma)^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\exp(-(x-\gamma)/\beta)$$

Cumulative density function

$$F(x) = \frac{\Gamma_{(x-\gamma)/\beta}(\alpha)}{\Gamma(\alpha)}$$

a: shape parameter (a>0)

β: scale parameter (β>0)

γ: location parameter (γ=0 : gives the two-parameter Gamma distribution)

The aforementioned distributions have been checked at several levels of statistical significance by utilizing different distribution fitting tests.

Distribution Fitting Tests

The goodness of fit (GOF) tests check how well the distribution selected fits to the data under study. The tests selected for this analysis are the Kolmogorov-Smirnov, the Anderson –Darling and Chi-Squared. A ranking according to each test assist towards the

selection of the optimal distribution. In the following paragraphs a short description about the fitting tests is presented.

Kolmogorov-Smirnov Test

This test is utilized to check whether the chosen hypothesized (theoretical) cumulative distribution and the empirical cumulative distribution function of the sample are close. More precisely, it is based on the maximum difference between the empirical cumulative distribution function and the hypothesized one:

$$\mathrm{D}=\max_{1\le i\le n}\left(F(x_i)-\frac{i-1}{n},\frac{i}{n}-F(x_i)\right)$$

where the empirical CDF of a random sample $x_1, \dots, x_n$ is given by

$$Fn(x)=\frac{1}{n}[Number\ of\ observations \le x]$$

The two hypotheses made for this test are:

H_0 : The data to follow the chosen theoretical distribution

H_A : The data not to follow the chosen theoretical distribution

A check is performed regarding the null hypothesis H_0 in order to reject it or accept it at a chosen significance level specified through fixed critical values (a). If D > a then the null hypothesis is rejected at the chosen significance level a.

P-Value

The P-value, in contrast to fixed values, is calculated based on the test statistic, and denotes the threshold value of the significance level in the sense that the null hypothesis (H_o) will be accepted for all values of less than the P-value. For example, if P=0.025, the null hypothesis will be accepted at all significance levels less than P and rejected at higher levels, including 0.05 and 0.1.

The P-value can be useful in particular when the null hypothesis is rejected at all predefined significance levels and we need to know at which level it could be accepted.

Anderson-Darling

This test gives more emphasis more on the tails of the distributions. The Anderson-Darling statistic (A^2) is defined as:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F(X_i) + \ln(1 - F(X_{n-1+1}))]$$

and it is applied in this study using the same hypothesis tests with the Kolmogorov-Smirnov test. The critical values as calculated by the approximation formula are the same for all distributions and depend only on the sample size.

Chi-Squared

This test is applied to binned data, so the value of the test statistic depends on how the data are binned. In order to calculate the number of bins (k) the following empirical formula based on the sample size (N) is used:

$$K = 1 + \log_2 N$$

The data can be grouped into intervals of equal probability or equal width. The first approach is generally more preferable since it handles peaked data much better. Each bin should contain at least 5 or more data points, so certain adjacent bins sometimes need to be joined together for this condition to be satisfied.

The Chi-Squared statistic is defined as:

$$X^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency for bin i and E_i is the expected frequency for bin I calculated by:

$$E_i = F(X_2) - F(X_1)$$

where F is the CDF of the probability distribution being tested and X_1, X_2 are the limits for the i-bin.

The two hypotheses made for this test are:

H_0: The data to follow the chosen theoretical distribution

H_A: The data not to follow the chosen theoretical distribution

The hypothesis regarding the distributional form is rejected at the chosen significance level (a) if the test statistic is greater than the critical value defined as

$$X^2_{1-a,k-1}$$

This is the Chi-Squared inverse CDF with k-1 degrees of freedom and a significance level of a.

The fixed values of (0.01, 0.05 etc.) are generally used to evaluate the null hypothesis (H_0) at various significance levels. A value of 0.05 is typically used for most applications; however, a lower value may be applied.

The P-values based on the Chi-Squared test statistics (χ^2) are calculated for each fitted distribution.

For assessing the "goodness of fit" the following probability plots are used.

Probability Difference Graph

The probability difference graph is a plot of the difference between the empirical CDF and the theoretical CDF:

$$Diff(x) = F_n(x) - F(x)$$

This graph can be used to determine how well the theoretical distribution fits to the observed data and compare the goodness of fit of several fitted distributions. It is displayed as a continuous curve or a scatterplot for continuous distributions and a collection of vertical lines (at each integer x) for discrete distributions:

P-P plots

The probability-probability (P-P) plot is a graph of the empirical CDF values plotted against the theoretical CDF values. It is used to determine how well a specific distribution fits to the observed data. This plot will be approximately linear if the specified theoretical distribution is the correct model.

Q-Q plot

The quantile-quantile (Q-Q) plot is a graph of the input (observed) data values plotted against the theoretical (fitted) distribution quantiles. Both axes of this graph are in units of the input data set.

The Q-Q graphs are produced by plotting the observed data values X_i (i = 1, ... , n) against the X-axis, and the following values against the Y-axis:

$$F^{-1}(F_n(x_i) - \frac{0.5}{n})$$

Where:

$F^{-1}(x)$: inverse cumulative distribution function (ICDF);

$F_n(x)$: empirical CDF;

n: sample size.

The Q-Q plot will be approximately linear if the specified theoretical distribution is the correct model.

A Test Case

A test case was selected for the island of Kerkyra (Corfu) which is located in the Ionian Sea (Figure 1). The data used is the 10m wind speed as obtained from the atmospheric model SKIRON operated for the MARINA Platform project. The previous discussed distributions are tested through the 3 goodness of fitting tests resulting to the ranking Table 1.

Figure 1: The location of the test case site at Kerkyra Island

Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
	Statistic	Rank	Statistic	Rank	Statistic	Rank
Weibull	0,01684	1	5,97	1	76,955	1
Weibull (3P)	0,01723	2	7,3504	2	86,696	2
Gen. Extreme Value	0,02926	3	26,743	3	205,42	3
Lognormal (3P)	0,03208	4	34,398	5	277,42	5
Gamma (3P)	0,03344	5	27,006	4	222,71	4
Log-Logistic (3P)	0,03888	6	59,669	6	537,37	6
Gamma	0,04115	7	76,639	7	555,29	7
Normal	0,05595	8	123,82	8	841,01	8
Lognormal	0,07884	9	189,95	10	1520,7	9
Log-Logistic	0,08515	10	189,51	9	1636,2	10
Exponential (2P)	0,16379	11	946,13	11	4332,4	11
Exponential	0,16455	12	955,51	12	4375,0	12

Table 1: Ranking table according to three fitting tests (Kolmogorov-Smirnov, Anderson-Darling, Chi-Squered)

According to all fitting tests Weibull is the best distribution to describe 10m wind speed for the specific location. On the other hand Weibull (3P), GEV, Log-normal (3P) and Gamma(3P) reveal a quite similar behavior to the first ranked one (Weibull) as it can be seen from the following probability plots.

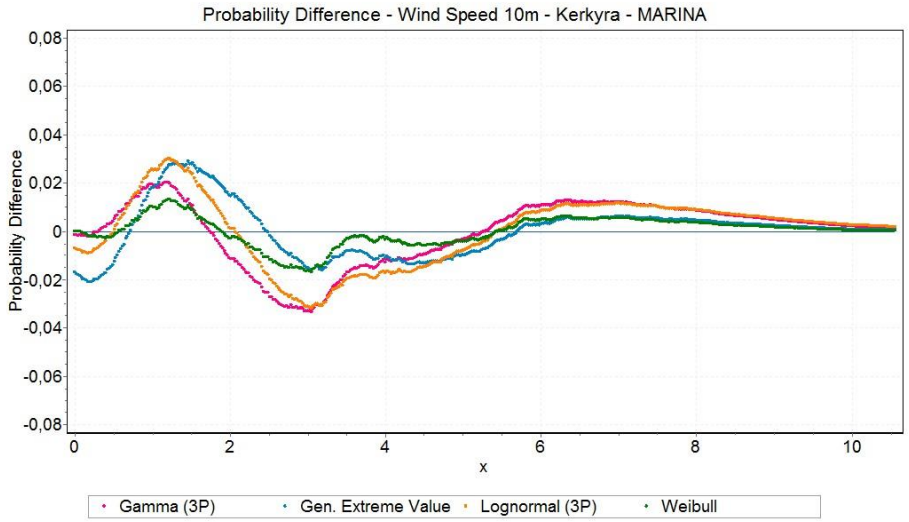

Figure 2: Probability difference plot for wind speed at 10m (SKIRON-MARINA)

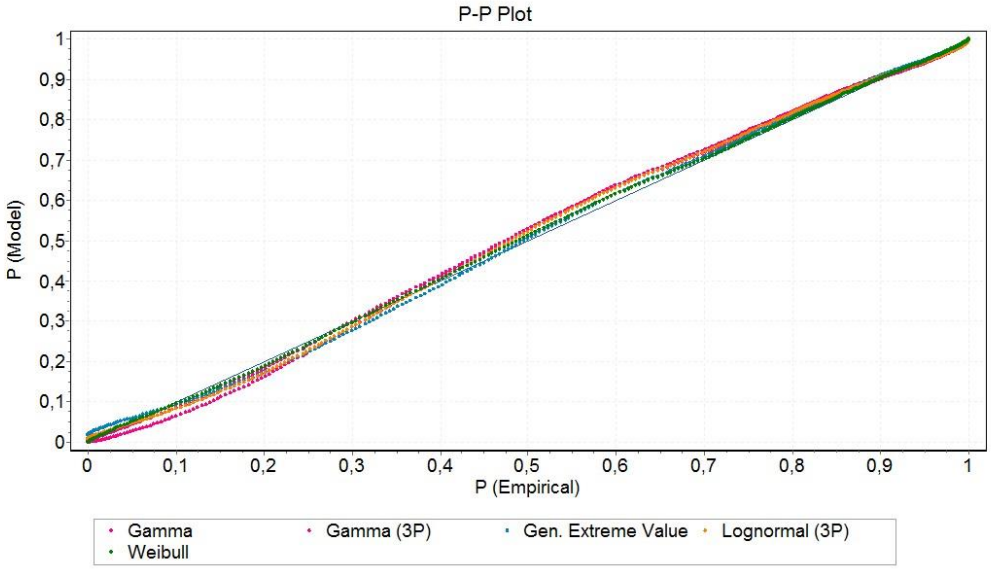

Figure 3: P-P plot for wind speed at 10m (SKIRON-MARINA)

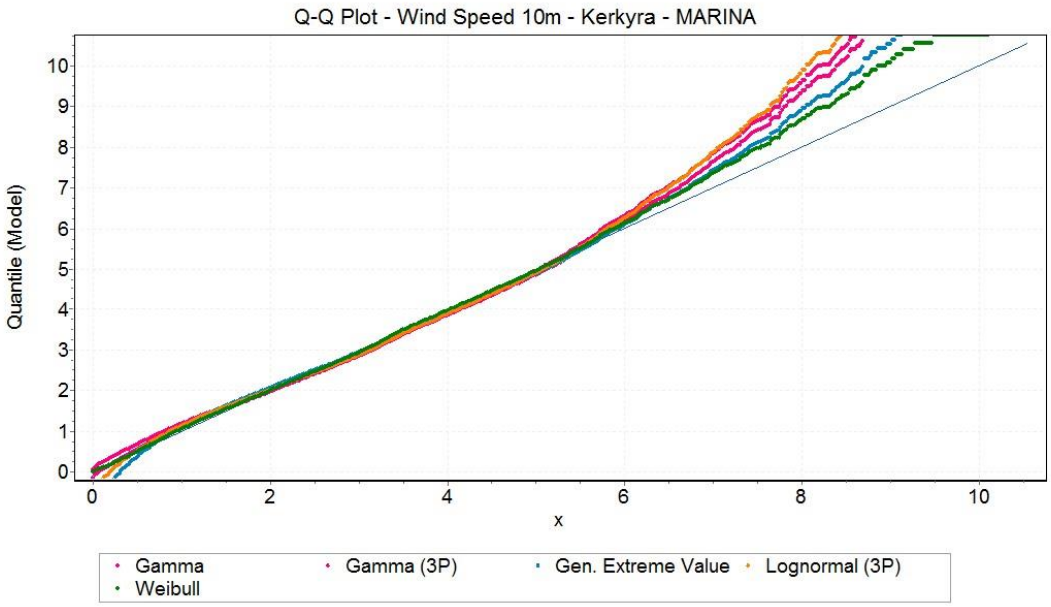

Figure 4: Q-Q plot for wind speed at 10m (SKIRON-MARINA)

All the tests (for this sample case as well as for the other examined) were performed using the mathematical software MATLAB and its Toolboxes.

References

1. Amari S-I, 1985. Differential Geometrical Methods in Statistics, Springer Lecture, Notes in Statistics 28, Springer-Verlag, Berlin.
2. Amari S-I., Nagaoka H., 2000. Methods of Information Geometry, American Mathematical Society, Oxford University Press, Oxford.
3. Arwini K., Dodson C.T.J., 2007: Alpha-geometry of the Weibull manifold. Second Basic Science Conference, Tripoli, Libya.
4. D'Agostino R. B., Stephens M.A., 1986: Goodness-of-fit Techniques, New York: Marcel Dekker.
5. Kalnay E., 2002: Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, 341.
6. Muraleedharan G., Rao A.D., Kurup P.G., Unnikrishnan N., Mourani S., 2007: Modified Weibull distribution for maximum and significant wave height simulation and prediction, Coastal Engineering 54, 630–638.
7. Papoulis A., 1991. Probability, Random Variables and Stochastic Processes 3rd edition, McGraw-Hill, New York.