*Environmental Sciences Proceedings*

MDPI

*Proceedings*

# Assessment of the Couple between the Historical Sample and the Theoretical Probability Distributions for Maximum flow Values Based on a Fuzzy Methodology †

**Matthaios Saridakis, Mike Spiliotis *, Panagiotis Angelidis and Basil Papadopoulos**

Section of Hydraulic Engineering, Department of Civil Engineering, Democritus University of Thrace, Kimmeria Campus, 67100 Xanthi, Greece; msarida@civil.duth.gr (M.S.); pangelid@civil.duth.gr (P.A.); papadob@civil.duth.gr (B.P.)

* Correspondence: mspiliot@civil.duth.gr

† Presented at the 4thEWaS International Conference: Valuing the Water, Carbon, Ecological Footprints of Human Activities, Online, 24–27 June 2020.

Published: 13 August 2020

**Abstract:** In this article, an adjustment of the extreme theoretical probability distributions upon the sample is proposed, based on the conventional fuzzy linear regression model of Tanaka [1], where all the data must be included within the produced fuzzy band. This is achieved by using the quintile approach, which relates the observed return period with the theoretical cumulative probability. A new contribution of this work is the use of the fuzzified maximum likelihood, as a measure of goodness of fit. The model is applied for real data from the Strymonas River, regarding the annual maximum flow, and finally, useful conclusions are made.

**Keywords:** observed probabilities; maximum likelihood; fuzzy regression; extension principle; Strymonas River

## 1. Introduction

Fuzzy logic has proven to be a particularly useful tool in the hands of engineers, and its use in recent decades has been widespread in hydrology and hydraulics. Fuzzy linear regression provides a functional fuzzy relationship between dependent and independent variables, where uncertainty manifests itself in the coefficients of the independent variables.

Next, the fuzzy linear regression of Tanaka [1] is used. In general, the fuzzy regression analysis gives a fuzzy functional relationship between the dependent and independent variables [2,3]. In contrast to the statistical regression, the fuzzy regression model of Tanaka [1] has no error term, while the uncertainty is incorporated into the model with the use of fuzzy numbers [4,5]. The data of the fuzzy regression can be either fuzzy or crisp. Usually, the data are rather crisp numbers (observed data), and thus, the uncertainty arises from the used fuzzy model, that is, the fuzzy coefficients. The inclusion property of the produced fuzzy band, that is, the requirement that all the data must be included within the produced fuzzy band, creates the constraints.

Papadopoulos et al. and Spiliotis et al. [5,6] proposed a fuzzy hybrid frequency factor-based method, with the use of fuzzy regression, in order to improve the couple between the theoretical and the observed probability distributions. These articles deal with either annual cumulative streamflow or precipitation. In this work, the annual maximum flow is studied, and furthermore, the fuzzified likelihood is used as an additional measure of suitability.

## 2. Basic Notations and the Proposed Methodology

### *2.1. Fundamentals of Fuzzy Sets and Logic*

A fuzzy set A on a universe set X is a mapping $A: X \to [0,1]$, assigning to each element $x \in X$ a degree of membership $0 \le A(x) \le 1$. The membership function $A(x)$ can be also presented as $\mu_A(x)$.

If $A$ is a fuzzy set, and let any number $a \in [0,1]$ by the $\alpha$-cut, $A[\alpha]$ and the strong $\alpha$-cut, $A[\alpha]^+$ ,the crisp sets are defined, respectively as:

$$A[\alpha] = \{x \in X : A(x) \ge \alpha\} \tag{1}$$

$$A[\alpha]^+ = \{x \in X : A(x) > \alpha\} \text{ (strong } \alpha\text{-cut)}, \tag{2}$$

The 0-cut can be defined as follows:

$$A[0]^+ = \{x \in X : A(x) > 0\}, \tag{3}$$

In order to have a closed interval containing the boundaries, Hanss [7] proposed the phrase worst-case interval *W*, which is the union of the 0-strongcut and the boundaries. It is worth noting that, by using the $\alpha$-cut concept, we can move from the fuzzy sets to the conventional crisp mathematical methodologies.

A special kind of fuzzy sets is the fuzzy numbers. In this work, fuzzy symmetric triangular numbers are used, which are special kinds of fuzzy numbers. The fuzzy symmetric triangular numbers have the following membership function:

$$\mu_A(x) = \begin{cases} 1 - \dfrac{|x-a|}{w}, & if \quad a - w \le x \le a + w \\ 0, & otherwise \end{cases}, \quad w > 0 \tag{4}$$

in which a is the center and *w* the spread of the fuzzy number.

The operation of the usual crisp functions, if the inputs are fuzzy sets, can be extended based on the extension principle. In most cases, it is preferable to use α-cuts in the fuzzy analysis [8]. If *g* is a continuous function in the extension principle, the use of α-cuts can be made by determining the α-cuts of the function *f*, as follows [9,10]. Then, based on the min intersection, it holds:

$$\begin{cases} g^L\left(x, \tilde{A}_1, \tilde{A}_2\right)_\alpha = \min\left\{g(y_1, y_2) \,|x \, is \, given, y_1 \in A_1[\alpha], y_2 \in A_2[\alpha]\right\}, \\ g^R\left(x, \tilde{A}_1, \tilde{A}_2\right)_\alpha = \max\left\{g(y_1, y_2) \,|x \, is \, given, y_1 \in A_1[\alpha], y_2 \in A_2[\alpha]\right\}. \end{cases} \tag{5}$$

From the theorem of global existence for maxima and minima of functions with many variables, it is known that, if the domain of a real function is closed and bounded and the real function is continuous, then the function will have its absolute minimum and maximum values at some points in the domain [8,11]. Based on this theorem, it is evident that the $\alpha$-cut for any real continuous function with real variables in this domain can be determined, given that the inputs are fuzzy triangular numbers [8].

### *2.2. Fuzzy Linear Regression*

The fuzzy linear regression model proposed by Tanaka [1] has the following form:

$$\tilde{Y}_j = \tilde{A}_0 + \tilde{A}_1 x_{1j} + ... \tilde{A}_i x_{ij} + .... + \tilde{A}_n x_{nj} \; with \; j = 1, \ldots, m, \; i = 1, \ldots, n, \tag{6}$$

where $n$ is the number of independent variables $x_{ij}$ (here only $n$ =1 which is related with the observed return period), $M$ is the number of data, $\tilde{Y}_j$ is the fuzzy predicted value of the dependent variable considering the $j$th data (here, the maximum annual flow).

According to Tanaka [1], all the data must be included within the produced fuzzy band (inclusion principle). Based on the extension principle and the concept of $\alpha$-cuts, the inclusion principle is equivalent to (by using fuzzy symmetrical triangular numbers as coefficients):

$$\sum_{i=0}^{n} a_i x_{ij} - (1-h)\sum_{i=0}^{n} c_i \left|x_{ij}\right| \le y_j \le \sum_{i=0}^{n} a_i x_{ij} + (1-h)\sum_{i=0}^{n} c_i \left|x_{ij}\right|,\ j = 1,...,M\ , \tag{7}$$

where $a_i$ is the center and $c_i$ the spread of the fuzzy number which represents the fuzzy coefficient.

Finally, the sum of the produced semi-widths for the produced dependent variable for all the data is proposed as an objective function:

$$J = min\left\{ Mc_0 + \sum_{j=1}^{M}\sum_{i=1}^{n} c_i \left|x_{ij}\right| \right\}, \tag{8}$$

In other words, the problem of fuzzy linear regression is reduced to a linear programming problem [1,4].

### *2.3. Observed Probabilities*

Let a historical sample. The rank order method involves ordering the data from the largest hydrological value to the smallest hydrological value, assigning a rank of 1 to the largest value and a rank of $N$ to the smallest value. An empirical distribution is used to compute the plotting position probabilities as follows [12]:

$$P\left(Q \ge q\right) = \frac{m}{N+1}, \tag{9}$$

Therefore, the cumulative probability or non-exceedance probability can be determined as follows:

$$P\left(Q < q\right) = 1 - \frac{m}{N+1}, \tag{10}$$

The concept of the observed probability is critical, since the suitability of the used theoretical probability function is based on the comparison between the observed and the theoretical probability values.

### *2.4. Generalized Extreme Value (GEV) Distribution*

The probability density function of the GEV distribution is of the form [13]:

$$f(x) = \frac{1}{\alpha}[1 - k(\frac{x-u}{\alpha})]^{1/k-1} e^{-[1-k(\frac{x-u}{\alpha})]^{1/k}}\ , \tag{11}$$

The GEV distribution function is as follows [14]:

$$F(x) = \exp\{-[1 - k(\frac{x-u}{\alpha})]^{1/k}\}, \tag{12}$$

The distribution function of $x$ given by Equation (12) can be written in the inverse form [13]:

$$x = u + \frac{\alpha}{k}[1 - (-\log F)^k], \tag{13}$$

By substituting $F = 1 - 1/T$ where $T$ is the return period, the T-year quantile estimate of the annual maximum flow, $\hat{q}_T$, is obtained as follows:

$$\hat{q}_T = \hat{u} + \frac{\hat{\alpha}}{\hat{k}}[1-\{-\log(1-\frac{1}{T})\}^{\hat{k}}], \tag{14}$$

$C_s$is the skewness coefficient, which can be estimated as:

$$C_S \simeq \frac{a'}{s^3}, \tag{15}$$

Furthermore, $a'$ is the asymmetry of a sample, of which the unbiased estimation is:

$$a' = \frac{N}{(N-1)(N-2)}\sum_{j=1}^{N}(x_j - \overline{x})^3, \tag{16}$$

where $N$ is the magnitude of the historical sample. Approximate relationships between the value of $k$ and the skewness coefficient $C_S$, obtained through regression analysis, are given in pages 225, 226 and 227 (7.1.12, 7.1.13 and 7.1.14) in [13]. This approach was adopted by the authors.

In this article, the term $[1-\{-\log(1-\frac{1}{T})\}^{\hat{k}}]$ is considered as a crisp number, while the terms $\hat{u}$ and $\frac{\hat{\alpha}}{\hat{k}}$ are considered as fuzzy numbers. Therefore, the following fuzzy regression model is modulated:

$$\begin{aligned} &\tilde{q}_{Tj} = \tilde{a}_0 + \tilde{a}_1 x_j \\ &where: x_j = [1-\{-\log(1-\frac{1}{T_j})\}^{\hat{k}}], \end{aligned} \tag{17}$$

in which $\tilde{q}_{Tj}$, $\tilde{a}_0$, $\tilde{a}_1$ are the fuzzified dependent variable, the constant term and the coefficient of the considered independent variable. It should be clarified that the independent variables, as well as all the observed values, are crisp numbers. The method does not require any transformation of the crisp data to fuzzy data.

### *2.5. The Extreme Value Type I EV1(2) Distribution*

The probability density function of the EV1(2) distribution is given by the following equation [13]:

$$f(x) = \frac{1}{\alpha}\exp[-(\frac{x-\beta}{\alpha}) - e^{-(\frac{x-\beta}{\alpha})}], \tag{18}$$

The variable $x$ takes values in the range $-\infty < x < \infty$. The distribution function of x is given by the following equation:

$$F(x) = \exp[-e^{-(\frac{x-\beta}{\alpha})}], \tag{19}$$

The EV1(2) distribution is a special case of the GEV distribution, discussed in Section 2.4, in which the shape parameter $k$ is equal to zero. The distribution function of EV1(2) (Equation (14)) can be obtained in the inverse form as follows [13]:

$$x = \beta - \alpha \cdot \log(-\log F), \tag{20}$$

The T-year quantile is calculated by substituting $F = 1-(1/T)$, where $T$ is the return period, to obtainthe following equation:

$$x_T = \hat{\beta} - \alpha \cdot \log[-\log(1-1/T)], \tag{21}$$

More specifically, the term $\log[-\log(1-1/T)]$ is considered as a crisp number, while the terms $\hat{\beta}$ and $\hat{\alpha}$ are considered as fuzzy numbers. Inthe same way, by using the fuzzy linear regression, the following fuzzy relation is determined, with the aim of fuzzy linear regression model:

$$\begin{aligned} &\tilde{q}_{T_j} = \tilde{a}_0 + \tilde{a}_1 x_j \\ &where: x_j = -\log[-\log(1-1/T_j)]' \end{aligned} \quad (22)$$

*2.6. Measures of Suitability*

Since a fuzzy approach is dealt, the usual statistical test cannot be used to check the suitability of the proposed model. The first model of suitability, which can be used, is the produced uncertainty of the produced fuzzy band, *J* (Equation (8)):

$$Mc_0 + c_1 \sum_{j=1}^{M} |x_j|, \quad (23)$$

where $M, c_0, c_1$ represent the number of data, the semi-width of the constant term and the semi-width of the independent variable coefficient, respectively.

In this article, the use of the likelihood is used as a measure of suitability (and not to determine the parameters). Since the parameters are fuzzy, then the fuzzified likelihood is constructed by determining the borders of the $\alpha$-cut for the likelihood, for several representative values of the $\alpha$-cut, as follows:

$$\begin{cases} g^L\left(x_j, \tilde{a}_0, \tilde{a}_1\right)_\alpha = \min\left\{\sum_{j=1}^{M} \ln f(y_1, y_2) \middle| x_j \; is \; given \, (crisp), y_1 \in a_0[\alpha], y_2 \in a_1[\alpha]\right\}, \\ g^R\left(x_j, \tilde{a}_0, \tilde{a}_1,\right)_\alpha = \max\left\{\sum_{j=1}^{M} \ln f(y_1, y_2) \middle| x_j \; is \; given \, (crisp), y_1 \in a_0[\alpha], y_2 \in a_1[\alpha]\right\}, \end{cases} \quad (24)$$

To summarize the proposed methodology, the following steps are adopted: (1) modulating the independent and dependent variables based on the Weibull 1939 empirical distribution and the examined theoretical probability distribution. (2) Applying fuzzy regression; (3) the magnitude of the produced fuzzy band and the fuzzfied likelihoods are used as suitability measures.

## 3. Case Study

The case under investigation is a northern region (Marino Pole) of the Strymonas River (Figure 1). The Strymonas (Struma) River is one of the largest rivers in theBalkan Peninsula in terms of length, since it crosses the Bulgarian and Greek borders. It rises in the Vitosha Mountain in Bulgaria, and has its outlet in the Aegean Sea. Its drainage area is 17,330 km², of which 10,797 km² is in Bulgaria, 6295 km² is in Greece, and the rest is in North Macedonia. In Greece, it is the main waterway feeding and exiting from Lake Kerkini, a significant center for migratory wildfowl. The river's length is 415 km (of which 290 km is in Bulgaria), making it the country's fifth-longest and one of the longest rivers that run solely in the interior of the Balkans. The location of the analysis is the Marino Pole, which is located in Bulgaria.

**Figure 1.** The Basin of Strymonas River in Bulgaria and the location of the case study.

Firstly, the GEV theoretical probability distribution is examined. Since the fuzzy regression enables the determination of two parameters, the third parameter is estimated at the beginning of the method, based on the moment method. Hence, based on the sample, the skewness coefficient is $C_S = 0.406632$ and finally, $\hat{k} = 0.157612$. Subsequently, the fuzzy linear regression of Tanaka with the worst-case interval W is applied, which leads to the following fuzzy curve (Figure 2):

$$\tilde{q}_{T_j} = (145.7, 27.8) + (1128.6, 43.3) \cdot [1 - \{-\log(1 - \frac{1}{T_j})\}^{\hat{k}}] \quad (25)$$

Based on the solution, the sum of the produced semi-widths *(J)* is:

$$J = 971.7549 \ {}^{m^3}\!/_{s} \quad (26)$$

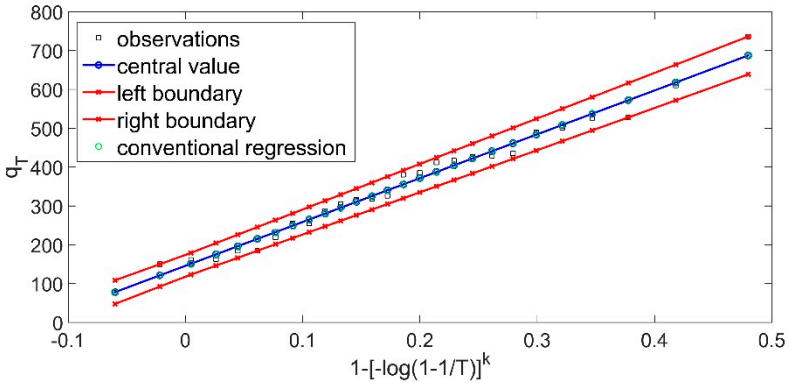


**Figure 2.** Graphical representation of the fuzzy linear regression for the Generalized Extreme Value (GEV) distribution.

Another interesting point of view is that, as can be seen from Figure 2, all the (crisp)observations are included within the produced fuzzy band a property, which is not satisfied according to the conventional methodologies.

Secondly, the extreme value type I [*EV1(2)*] theoretical distribution is examined. The fuzzy linear regression was applied (Tanaka et al. 1987) based on the aforementioned quantile approach, which leads to the following fuzzy curve:

$$\tilde{q}_{T_j} = (158.62, 27.01) + (323.18, 20.1) \cdot \{-\log[-\log(1 - 1/T_j)]\} \quad (27)$$

while the sum of the produced semi-widths is:

$$J = 1060.4 \ {}^{m^3}\!/_{s} \quad (28)$$

Both fuzzy curves are depicted in Figures 2 and 3. Based on the magnitude of the produced fuzzy band (that is, the value of the objective function), the GEV theoretical probability distribution is preferred. An interesting point is that every fuzzy linear regression problem based on the Tanaka methodology has a solution. The critical point is the magnitude of the produced fuzzy band. From Figures 2 and 3, it is evident that the fuzzy band can be characterized as functional. Another important point of view is that all the data must be included within the produced fuzzy band, and hence all the empirical probabilities are within the produced fuzzy band.

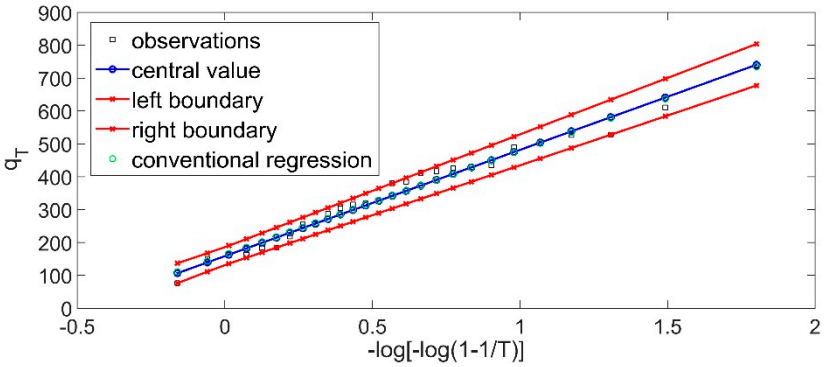


**Figure 3.** Graphical representation of the fuzzy linear regression for the extreme value type I [EV1(2)] distribution.

As mentionedbefore, the second measure of suitability is the produced fuzzy likelihood $g$, which in its crisp expression, is a continuous function. Based on the extension principle (Equation (24)), the fuzzified likelihood is built by determining several $\alpha$-cuts. The fuzzified likelihoods for both examined theoretical probability distributions are presented in Figure 4. The two likelihoods have the shape of fuzzy numbers. Even if the comparison between two fuzzy numbers is an ill constructed problem, by applying several widely used measures, we conclude that the fuzzy likelihood of *GEV* is greater than the fuzzy likelihood of *EV1(2)* and hence the *GEV* is preferred according to this criterion. We highlight that the fuzzified likelihood is determined based on the fuzzy regression, or in other words, the fuzzified likelihood is used as a suitability measure, and not in order to determine the parameters of the probability distribution.

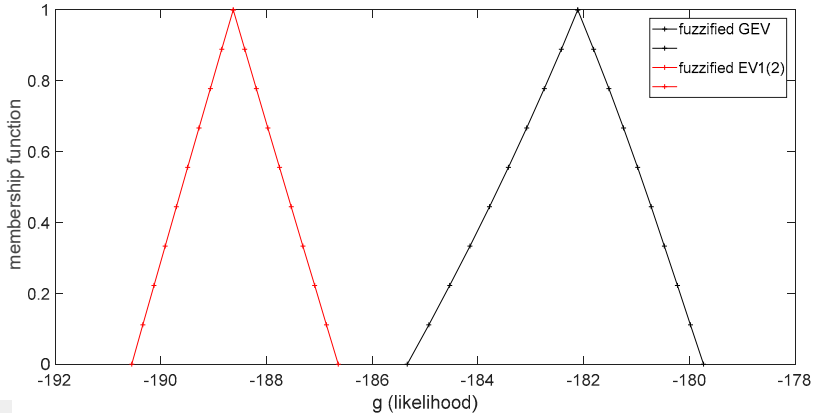


**Figure 4.** The fuzzified likelihoods for *GEV* and *EV1(2)* theoretical probability distributions.

## 4. Conclusions

In order to achieve the couple between the examined extreme theoretical probability distribution and the sample, a hybrid fuzzy regression based approach is developed. The fuzzy regression model is formulated according to the quantile approach, which relates the observed return period with the theoretical cumulative probability. The problem of fuzzy linear regression for crisp data and fuzzy triangular numbers as coefficients concludes to an equivalent linear programming problem.

Two theoretical probability distributions were applied to study the annual maximum river in Strymonas River: the GEV and the EV1(2). To evaluate the proposed approach, two measures of suitability are proposed. The first one is the magnitude of the produced fuzzy bands, which is the objective function regarding the equivalent optimization problem. The second measure is the fuzzified likelihood, which is built according to the extension principle and the solution of the fuzzy regression. It is concluded that, in both cases, a functional uncertainty appears; therefore, the use of the fuzzified theoretical probability function is successful and can be further utilized. Secondly, by comparing both objective functions and the fuzzified likelihood, the use of the fuzzified GEV must

be preferred. Hence, the proposed measure of suitability can be used, in order to select the fuzzified theoretical probability function.

**Author Contributions:** All authors have read and agree to the published version of the manuscript. All authors contributed equally.

**Funding**: This research received no external funding.

**Conflicts of Interest**: The authors declare no conflict of interest.

## References

1. Tanaka, H. Fuzzy data analysis by possibilistic linear models. *Fuzzy Sets Syst.***1987**, *24*, 363–375.
2. Tzimopoulos, C.H.; Papadopoulos, K.; Papadopoulos, B. Fuzzyregressionwithapplications in hydrology. *Int. J. Eng. Innov. Technol. (IJEIT)* **2016**, *5*, 69–75.
3. Tsakiris, G.; Tigkas, D.; Spiliotis, M. Assessment of interconnection between two adjacent watersheds using deterministic and fuzzy approaches. *Eur. Water* **2006**, *15*, 15–22.
4. Spiliotis, M.; Hrissanthou, V. Fuzzy and crisp regression analysis between sediment transport rates and stream discharge in the case of two basins in northeastern Greece. In *Conventional and Fuzzy Regression: Theory and Engineering Applications*, 1st ed.; Hrissanthou, V., Spiliotis, M., Eds.; Nova Science Publishers: New York, NY, USA, 2018; pp. 1–49.
5. Papadopoulos, C.H.; Spiliotis, M.; Angelidis, P.; Papadopoulos, B. A hybrid fuzzy frequency factor based methodology for analyzing the hydrological drought. *J. Desal. Water Treat.***2019**, *167*, 385–397, doi:10.5004/dwt.2019.24549.
6. Spiliotis, M.; Angelidis, P.; Papadopoulos, B. A hybrid probabilistic bi-sector fuzzy regression based methodology for normal distributed hydrological variable. *Evolv. Syst.* **2019**, 1–14, doi:10.1007/s12530-019-09284-7.
7. Hanss, M. *Applied Fuzzy Arithmetic, an Introduction with Engineering Applications*; Springer: Berlin, Germany, 2005; p. 256.
8. Tsakiris, G.; Spiliotis, M. Uncertainty in the analysis of urban water supply and distribution systems. *J. Hydroinformatics* **2017**, *19*, 823–837, doi:10.2166/hydro.2017.134.
9. Buckley, J.; Eslami, E. *An Introduction to Fuzzy Logic and Fuzzy Sets*; Springer: Berlin, Germany, 2002; p. 285.
10. Buckley, J.; Eslami, E.; Feuring, T. Solvingfuzzyequations. *Fuzzy Math. Econ. Eng.* **2002**,*91*,19–46.
11. Marsden, J.; Tromba, A. *Vector Calculus*, 5th ed.; W.H. Freeman and Company: New York, NY, USA, 2003.
12. Weibull, W. *A Statistical Theory of the Strength of Materials*; Generalstabens Litografiska Anstalts Förlag: Stockholm, Sweden, 1939.
13. Ramachandra Rao, A.; Khaled Hamed, H. *Flood Frequency Analysis*; CRC-Press: Boca Raton, FL, USA, 1999.
14. Jenkinson, A.F. The frequency Distribution of the Annual Maximum (or Minimum) Values of Meteorological Elements. *Q. J. R. Meteorol. Soc.***1955**, *87*, 158–171.