A Hybrid Fuzzy Regression-Based Methodology for Normal Distribution (Case Study: Cumulative Annual Precipitation)

M. Spiliotis, P. Angelidis, B. Papadopoulos

A Hybrid Fuzzy Regression-Based Methodology for Normal Distribution (case study: Cumulative Annual Precipitation)

Spiliotis M. [1][0000-0001-9051-028X] Angelidis P.[2][0000-0002-0871-4484] and Papadopoulos B.[2][0000-0001-6519-895X]

[1] Dep. of Civil Engineering, Democritus University of Thrace
Kimmeria Campus, 67100 Xanthi, m.spiliotis@gmail.com
[2] Dep. of Civil Engineering, Democritus University of Thrace
Kimmeria Campus, 67100 Xanthi

Abstract. An advantage of the probabilistic approach is the exploitation of the observed probability values in order to test the goodness-of-fit for the examined theoretical probability distribution function (pdf). Since in fact, the interest of the engineers is to determine the hydrological variable which corresponds to a selected return period, a fuzzy linear relation between the standardized normal variable Z and the examined hydrologic random variable is achieved in condition that the hydrological variable is normally distributed. In this work, for the first time, the implementation of the fuzzy linear regression of Tanaka is proposed, to achieve a fuzzy relation between the standardized variable Z and the annual cumulative precipitation. Thus, all the historical data are included in the produced fuzzy band. The proposed innovative methodology provides the opportunity to achieve simultaneously a fuzzy assessment of the mean value and the standard deviation based on the solution of the fuzzy linear regression. The suitability test of the examined theoretical pdf is founded on the comparison of the spread of the fuzzy band and the distance between the achieved central values of the mean value and the standard deviation with the unbiased statistical estimation of the same variables.

Keywords: Fuzzy linear regression, fuzzy sets, empirical probability function, normal distribution, cumulative annual precipitation.

1 Introduction

During the hydraulic design and management the first step is the assessment of the hydrological variables. Unfortunately, these cannot be treated with crisp values. There is plenty of academic works and engineering experience about the probabilistic implementation of the hydrological variables. However, the couple between the statistic and theoretical probability distributions is not without problems.

Thus, in this article, for first time, a hybrid model is proposed to improve the matching between the sample and the normal probability distribution function. This hybrid approach treats the couple between the normal distribution density function and the observed probabilities enhanced by the using of fuzzy linear regression.

An advantage of the probabilistic approach as a choice to treat the uncertainty is the exploitation of the cumulative empirical (observed) probability distribution in order to test the goodness-of-fit for a theoretical probability distribution with respect to the historical sample. In this article, a hybrid methodology is developed so that, both the concepts of the observed probability and the flexibility of the fuzzy approach can be used.

More specifically, the main idea is to express the probabilistic parameters as fuzzy numbers and thus, the flexibility of the fuzzy arithmetic and other concepts of the fuzzy sets can be exploited.

Mainly, in contrast with the traditional approaches, the proposed hybrid, fuzzy enhanced methodology produces a fuzzy relation which can be seen as a fuzzy band, which contains all the observed data, that is, all the pairs between the observed hydrological values and the observed probability. This fuzzy relation (it can be seen as a fuzzy band) is modulated based on the fuzzy linear regression model of Tanaka (1987). Therefore, the thresholds of fuzziness are modulated with respect to the observed data, by following a constraining optimization problem, and not by following an a- priori information as in several cases of the fuzzy optimization problems.

Fuzzy linear regression may be a useful tool to express functional relationships between variables, especially when the available data are not sufficient (Ganoulis 1994). For instance, Kitsikoudis et al. (2016) employed a fuzzy regression to produce a lower and an upper limit for the critical dimensionless shear stress. Thus, we avoid the ambiguity of selecting a threshold for the initiation of motion, and hence, the model provides a smoother transition to the state of general movement.

More generally, for studying complex physical phenomena such as the interconnection between adjacent watersheds (Tsakiris et al. 2006), the rainfall-runoff process without all the involved parameters, the use of fuzzy models should be investigated. In contrast to the statistical regression, fuzzy regression analysis has no error term, while the uncertainty is incorporated in the model by means of fuzzy numbers (Spiliotis & Bellos 2016; Papadopoulos & Sirpi 1999).

In brief, the proposed methodology can be divided into three steps: modulation of the data (an analysis based on the conventional statistical approach), application of the fuzzy linear regression (which concludes to a constrained optimization problem) and finally evaluation of the solution achieved where the assumption of the considered *pdf* is tested. The proposed methodology concludes to a fuzzy relation between the probability and the hydrological variables simultaneously with the fuzzy assessment of the mean value and the standard deviation. In this work, the methodology is developed for normal distributed annual cumulative precipitation.

2   Basic Notions

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. Based on the Weibull empirical distribution (which is widely used in the Greek regions) to compute the plotting position probabilities, the cumulative exceedance probability can be calculated as follows (e.g. Chow et al. 1988):

$$Prob(P \geq p) = \frac{m}{N+1} \tag{1}$$

where m is the rank of the value in a list ordered by descending magnitude (Chow et al. 1988). As P many hydrological random variables can be considered. In this work we focus on the *annual cumulative precipitation*.

It is evident that the cumulative probability of non-exceedance can be calculated as follows based on the exceedance probability (given that that P is a continuous random variable and according to the common practice in hydrology):

$$Prob(P \leq p) = 1 - \frac{m}{N+1} \tag{2}$$

The Return period of an event, T, of a given magnitude may be viewed as the average recurrence length of time (usually years) between events equaling or exceeding a specified magnitude (Chow et al. 1998):

$$Prob(P \leq p) = 1 - \frac{1}{T} \tag{3}$$

In case of the normal distribution, based on the calculated unbiased estimation of the mean and the standard deviation with respect to the available historical sample, $\bar{p}, s,$ correspondingly, it holds:

$$Prob(P \leq p_T) = \left(1 - \frac{1}{T}\right), \tag{4}$$

$$\left.\begin{array}{l} Prob(P \leq p_T) = \left(1 - \frac{1}{T}\right) = \Phi(Z_T) \\[2ex] Z_T = \frac{p_T - \bar{p}}{\sigma} \end{array}\right\} \Rightarrow \frac{p_T - \bar{p}}{\sigma} = Z_T \Rightarrow p_T = \bar{p} + \sigma Z_T \tag{5}$$

where Z is the standardized normal variable and p_T is the magnitude of the event having a return period T.

A used technique in hydrology is to investigate a linear relation between the standardized normal variable Z and the hydrological variable (here the cumulative annual

precipitation), instead of the probability plot. Similarly with the probability plot, the calculation of the normal variable Z is based on the empirical probability of the sample which corresponds to each pair between the standardized normal variable and the examined hydrological variable:

$$p_T = a_0 + a_1 Z_T \tag{6}$$

It is desirable, that the coefficients α_0 and α_1 to be close to the unbiased estimation of the mean value and the standard deviation of the sample, $\bar{p}$ and s respectively, as it occurs in the case of the crisp formulation. In the past, the graphical method with specified paper plots were used in order to rather manually identify the proper curve and the corresponding parameters. Also the conventional regression can be used to determine the parameters α_0 and α_1.

The main proposal of this article is the application of the fuzzy linear regression to determine the parameters α_0 and α_1. That is,

$$\tilde{p}_T = \tilde{a}_0 + \tilde{a}_1 Z_T \tag{7}$$

An interesting point is that additionally, a fuzzy approach for the mean value and the standard deviation can be achieved. The proposed methodology is graphically presented in Fig.1 and it will be presented bellow more analytically. As aforementioned, the produced fuzzy band will contain all the observed data and this is one significant advantage of the proposed methodology.

3   Proposed methodology

The aim of the proposed methodology is to provide a fuzzy relation between the standardized normal variable Z and the examined hydrological variable simultaneously with a fuzzy estimation of the mean value and the standard deviation. The proposed methodology is based on the observed probabilities of the sample and thereafter the fuzzy regression is used. However, the proposed method holds is the hypothesis of normal distributed random variable holds. Next the main three steps of the proposed methodology are presented.

3.1 Modulating the independent and dependent variables

Firstly, the data are ordered from the largest hydrological value to the smallest hydrological value, assigning a rank. Then, based on the Weibull 1939 empirical distribution, the cumulative non-exceedance probability is calculated. Subsequently, a hypothesis must be made about the theoretical probability distribution. Here, the normal distribution is examined. Hence, the standardized normal variable Z can be determined for each pair of data based on the empirical probabilities.

As aforementioned the independent variable is the Z_T term. To determine the Z_T term the following two assumptions must be made:

$$Z_T = \Phi^{-1}\left(1 - \frac{m}{N+1}\right)$$

(8)

It should be clarified that the standard statistical tests of fitness can be used only to predispose the probability distribution. The proposed methodology does not use the standard statistical tests of fitness. Furthermore, the evaluation of the solution achieved is based on the fuzzy solution and not on any conventional statistical test.
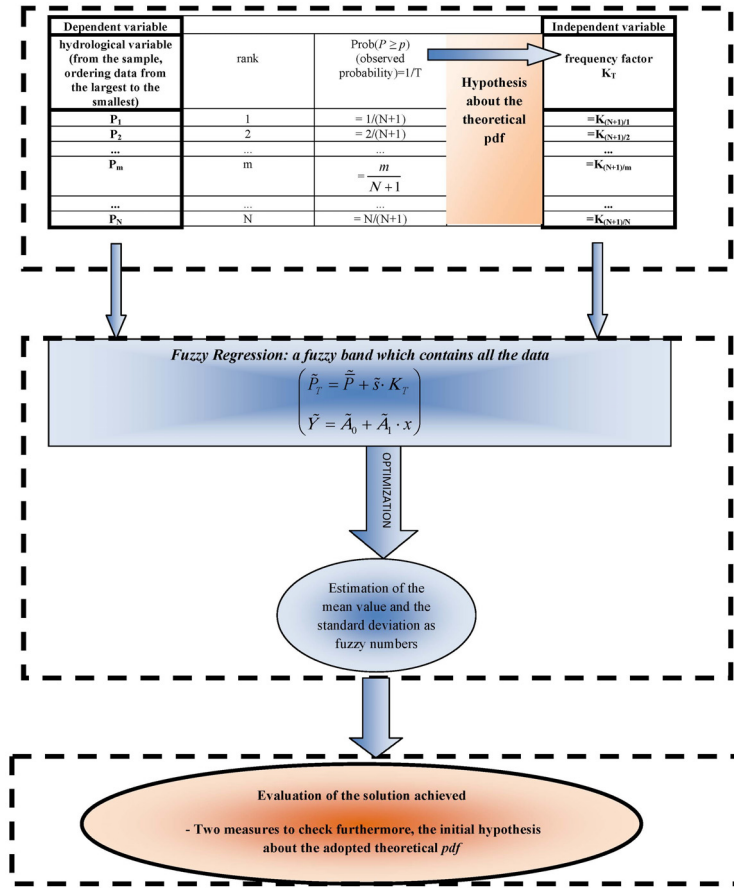

Fig.1. Schematic representation of the proposed methodology.

3.2 Applying Fuzzy Regression

Fuzzy linear regression model proposed by Tanaka (1987) has the following form:

$$\tilde{Y}_j = \tilde{A}_0 + \tilde{A}_1 x_{1j} + ... \tilde{A}_i x_{ij} + ... + \tilde{A}_N x_{Nj}$$

$$\text{with } j = 1, ..., M, \ i = 1, ..., N \tag{9}$$

where N is the number of independent variables, M is the number of data, and $\tilde{A}_i$ = $(a_i, c_i)_L$ are symmetric fuzzy triangular numbers selected as coefficients (Fig. 2) (Spiliotis et al. 2015; Papadopoulos & Sirpi 2004):

In case of normally distributed annual cumulative precipitation, the following fuzzy linear regression model is examined:

$$\tilde{p}_{Tj} = \tilde{\tilde{p}} + \tilde{s} \cdot Z_{T,j}$$

$$\left(\tilde{Y}_j = \tilde{A}_0 + \tilde{A}_1 \cdot x_{1j} \right) \tag{10}$$

where $\tilde{p}_{T,j}$ is the fuzzy assessment of the hydrological magnitude of the event (dependent variable) having a return period T for the jth set of data.

With respect to Eq. 10, the mean and the standard deviation $\tilde{\tilde{p}}, \tilde{s}$ respectively, are the fuzzy coefficients (fuzzy symmetrical triangular numbers in this application). The fuzzy coefficients are determined based on the Tanaka fuzzy linear regression method. The standardized normal variable is the independent variable, which takes only crisp values as well as the dependent variable which is the annual cumulative precipitation.

Thus, the proposed methodology can produce a fuzzy estimation of the mean value and the standard deviation. This fuzzy estimation is based on the fuzzy regression and hence, it is founded with the respect to the data. However, we highlight that the fuzzy estimation of the mean value and the standard deviation is related with the assumption of the normal distribution.

In this point it should be referred, that many problems of fuzzy logic, as the fuzzy arithmetic, the fuzzy linear regression e.t.c. are formulated with the aim of *a-cuts*. The *α-cuts* could be characterized as the bridge between the fuzzy and the crisp (conventional) sets. The *α–cut* set of the fuzzy number A (with $0 < \alpha \leq 1$) is defined as follows:

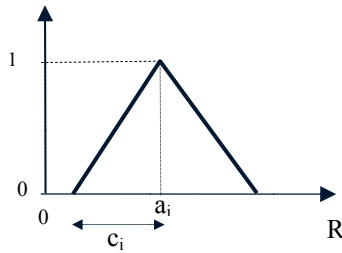$$[A]_\alpha = \left\{ x \middle| \mu_A(x) \geq \alpha \right\} \tag{11}$$

Fig. 2. Fuzzy triangular symmetrical number.

Note that the *a–cut* set is a crisp set determined from the fuzzy set according to a selected value of the membership function and, alternatively, a fuzzy set can be practically decomposed to a significant number of *a-cut sets*. In case of *a = 0*, the above definition (Eq. 14) can be modified without the equality in order to describe the *zero-cut* (Kitsikoudis et al. 2016; Spiliotis & Bellos 2016; Buckley & Eslami 2002).

The concept of inclusion can be defined with respect to the *a-cut*. The inclusion of a fuzzy set A to the fuzzy set B with the associated degree $0 \leq h \leq 1$ (it is a critical point, that the inclusion property is based on a selected *h-cut* between all the other *α-cuts*) is defined as follows:

$$[A]_h \subseteq [B]_h \tag{12}$$

The constraints of the problems are modulated based on the concept of inclusion. Thus, all the data (which are crisp numbers) must be included in the produced fuzzy band:

$$p_j \in \left[p_{h,j}^L, \ p_{h,j}^R \right] \tag{13}$$

By taking into account the fuzzy arithmetic, for a selected level *h*, the inclusion constraints in case the decision variables are selected to be symmetrical triangular numbers, are equivalent to:

$$
\begin{cases}
Inclusion\ Constra\,int\,s \\
\left(\bar{p} + s \cdot Z_{T,j} \right) - \left(1 - h\right)\left(w_{\bar{p}} + w_s \left| Z_{T,j} \right| \right) = p_{h,j}^L \leq p_j \\
\left(\bar{p} + s \cdot Z_{T,j} \right) + \left(1 - h\right)\left(w_{\bar{p}} + w_s \left| Z_{T,j} \right| \right) = p_{h,j}^R \geq p_j \\
w_{\bar{p}}, w_s \geq 0
\end{cases}
\tag{14}
$$

Since the fuzzy regression leads to a constrained optimization problem, the assessment of the fitness is based on the produced fuzzy band. The smaller the fuzzy band, the more proper the fuzzy model becomes. Hence, Tanaka (1987) suggested the minimization of the sum of the produced fuzzy semi-spreads for all the data (Spiliotis and Bellos, 2016, Kitsikoudis et al. 2016; Papadopoulos & Sirpi 1999):

$$min\ J \left(= \sum_{j=1}^{M} w_{pj} = \left\{ M \cdot w_{\bar{p}} + w_s \sum_{j=1}^{M} \left| Z_{T,j} \right| \right\} \right) \ (objective\ function) \tag{15}$$

The above measure is the sum of all the produced semi-spreads (from the fuzzy band) for all the observed data.

Therefore, in brief, based on the Tanaka methodology, the parameters of the fuzzy coefficients, that is, the centers and the widths of the fuzzy coefficients, are determined by solving a constrained optimization problem (Eqs. 14 and 15). In the examined problem, these fuzzy coefficients can be viewed as an estimation of the mean

value and the standard deviation. The objective function is selected to be the total semi-spread of the fuzzy outputs.

3.3 Evaluation of the solution achieved

In fact, every fuzzy regression problem based on Tanaka formulation will have a solution, and hence, the magnitude of the produced fuzzy band could be a criterion about how successful is the proposed fuzzy linear transformation. A large semi-spread, J, indicates that the axis of the standardized normal variable Z must be changed and hence, other probability distribution could be examined (Spiliotis and Papadopoulos, 2017).

Another measure to test the suitability of the examined probability distribution is proposed, F which is the Euclidean distance between the central values of the mean value and the standard deviation with the unbiased (usual statistical) estimation of the same variables, $\hat{\mu}, \hat{\sigma}$:

$$F = \sqrt{\left(\bar{p} - \hat{\mu}\right)^2 + \left(s - \hat{\sigma}\right)^2} \tag{16}$$

4 Application in case of annual rainfall time series

The meteorological station of Aldeia Nova de São Bento belongs to the National Information System for the Water Resources (SNIRH) of Portugal. The station is located inside the area of Guadiana between the Guadiana River and the Portugal – Spain border, as shown in Figure 3 (Angelidis et al., 2012). The data cover a large time span, from 1931 up to 2007. The proposed methodology was applied to express the annual cumulative rainfall with respect to the for normal distributed rainfall.

Indeed, by using the statistical test of Kolmogorov-Smirnov (K-S) with an interval confidence $(1-\alpha)=0.80$, the result indicates that the normal distribution can be used to express the annual rainfall in Aldeia Nova de São Bento. As aforementioned the statistical test is not used directly in the proposed methodology, but it simply indicates a proper algebraic transformation.
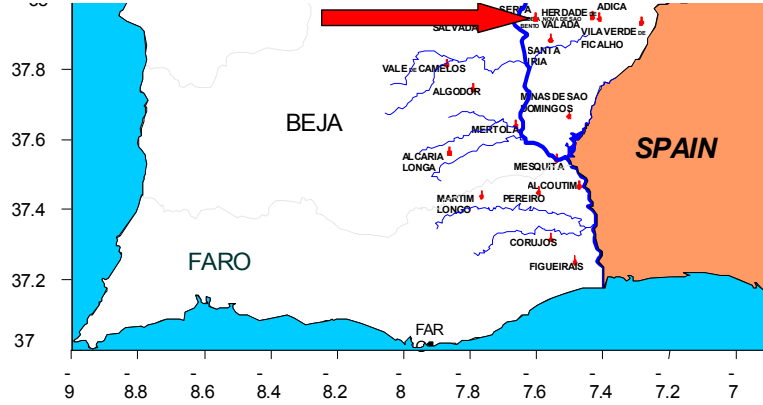
Fig. 3. Location of the Aldeia Nova de São Bento rain station.

Fuzzy regression is applied for a selected $h = 0$ whilst fuzzy symmetrical triangular numbers were selected as fuzzy coefficients. Finally, based on the above assumptions, the problem of fuzzy regression leads to a linear programming problem. All in all, the following fuzzy curve between the standard normal random variable Z and the annual rainfall is produced through the constrained optimization problem according to Eqs. 14 and 15 (Fig. 4):

$$\tilde{P}_T = \left(161.96,\, 3.58\right)_L \cdot Z_T + \left(551.50, 39.95\right)_L \tag{17}$$

where $\tilde{P}_T$ is a fuzzy number which represents the estimated cumulative annual runoff (dependent variable). Hence, by adopting the normal distribution, according to Eq.17, we achieve simultaneously a fuzzy assessment of the mean value $\left(551.50, 39.95\right)_L$ and the standard deviation $\left(161.96,\, 3.58\right)_L$. The first term in the bracket expresses the central value (Fig. 2) and the second term, the semi-width of each fuzzy symmetrical triangular number.

The objective function of the fuzzy regression problem which expresses the total uncertainty is equal to:

$$J_N = 75 \cdot 39.95 + 3.58 \sum_{j=1}^{75} \left| Z_{T,j} \right| = 3{,}202.8 \tag{18}$$

In addition, from the fuzzy curve (Eq. 17), it is easy to see that the central values of the fuzzy coefficients, which can be seen as a fuzzy approximation of the mean value and the standard deviation, are very close to their unbiased estimation based on the historical sample $\left(\hat{\mu} = 545.50, \hat{\sigma} = 159.91\right)$. Indeed, the Euclidean distance between the central values of the mean value and the standard deviation with the unbiased (usual statistical) estimation of the same variables, $\hat{\mu}, \hat{\sigma}$ has a rather small value:

$$F = \sqrt{(545.50 - 551.50)^2 + (161.96 - 159.91)^2} = 6.34\,mm \tag{19}$$

The values of both the objective function (Eq. 18) and the proposed measure F (Eq. 19) indicate that the fuzzified normal distribution can be used to express the annual cumulative rainfall in Aldeia Nova de São Bento. Therefore, from the application, it is evident that the objective function and the proposed measure of suitability, F, provide us the opportunity to test the suitability of the initial hypothesis about the use of the normal distribution as the theoretical probability distribution.
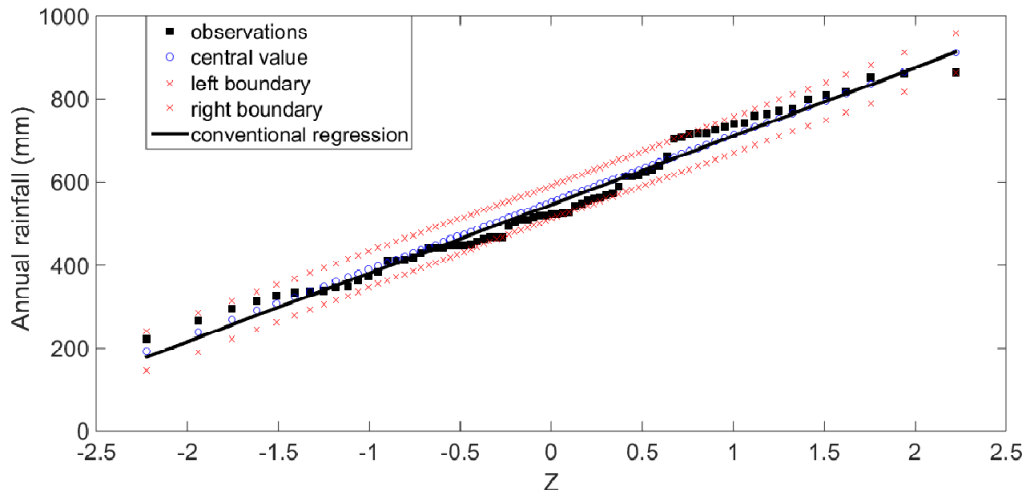

Fig.4. Observed data, fuzzy and conventional (crisp) regression between the standardize normal variable Z (based on the observed cumulative non-exceedance probabilities) and the annual rainfall of the sample.

Another interesting perspective to discuss is the proposed methodology in relation with the statistic test of Kolmogorov–Smirnov. The statistic test of Kolmogorov–Smirnov involves the absolute values of the distance between the observed probability and the probability value from the adopted theoretical cumulative distribution function for all the sample. By applying the proposed methodology, all the data, that is, each pair which includes the value of the hydrological value and the corresponding non-exceedance empirical probability is included in the produced fuzzy band at least to some degree (Fig. 4).

An interesting point is that the proposed fuzzy estimation of the mean value and the standard deviation is more founded to fuzzy logic and without subjective choices which must be made by the user, compared with Sfiris et al., 2014 approach. According to Sfiris et al., 2014, the mean value and the standard deviation are estimated as fuzzy numbers starting from the conventional probabilistic confidence interval. However, a confidence interval must be selected from the user, in order to move from the probabilistic approach to the fuzzy approach. In addition, the transition from the asymptotic to non asymptotic membership function was done by selecting a suitable

function. According to the new methodology, the mean value and the standard deviation are estimated as fuzzy numbers based on the observed data by following a fuzzy regression approach without specific considerations from the user. Furthermore, two measures to test the validity of the method are proposed.

It should be clarified also that the proposed methodology is based on the probabilistic model with crisp data whilst the fuzziness arises from the matching between the theoretical probability distribution function and the observed probabilities. The method should not be confused with either methods which are based on fuzzy data (fuzzy sample, e.g. Viertl, 2011) or methods with starting from a probabilistic approach conclude to a fuzzy approach (e.g. Sfiris and Papadopoulos, 2014).

Another interesting point for further investigation is the use of several theoretical distributions. Then, the use of the suitability measure F and the value of the objective function J can be seen as the criteria according to which the selection of the suitable theoretical probability distribution can be done. Spiliotis and Papadopoulos, 2017 have already studied a similar methodology in case of lognormal theoretical probability distribution. In general, the proposed methodology could be expanded based on the frequency factor method which includes also the proposed analysis based on the standardized variable Z for normal probability distribution.

As aforementioned in case that a significant fuzzy band is produced, then the first think will be the change of the theoretical pdf and this can be seen as the main challenge for further investigation. However, the use of new techniques of fuzzy regression should also be investigated. For instance, it could be investigated the use of more sophisticated objective functions which will contain both the centers and the widths of the fuzzy coefficients. Hence, a new objective function could try either to increase the possibility of equality between the observations and fuzzy intervals (Shakouri et al., 2017) or could incorporate simultaneously with the fuzzy spreads the sum of possibility grade in the objective function (Yoshiyuki, 2017). Alternatively, ideas from goal programming could be applied in order to reduce the impact of outliers (Kitsikoudis et al., 2016) and hence, the objective function is changed accordingly.

5   Concluding remarks

A hybrid combination between the problem of coupling between the observed probabilities and the normal distribution with the aim of fuzzy linear regression analysis is proposed in this article. According to the proposed methodology, the exploitation of the empirical probability function can be achieved together with a fuzzy approach. The key of the proposed methodology is the application of the fuzzy linear regression model of Tanaka in order to achieve a linear relation with fuzzy numbers as coefficients. In addition, based on the proposed fuzzy regression, the mean value and the standard deviation are determined as fuzzy coefficients.

Hence, the proposed innovative methodology provides the opportunity to achieve simultaneously a fuzzy assessment of the mean value and the standard deviation based on the solution of the fuzzy linear regression. In contrast, the application of the conventional crisp regression products crisp numbers as coefficients, which differ

from the unbiased estimation of the mean value and the standard deviation. The proposed fuzzy estimation of the mean value and the standard deviation is desirable to include the unbiased estimation of the mean value and the standard deviation.

Indeed, the measure of the distance between the central values of the mean value and the standard deviation with the unbiased (usual statistical) estimations together with its fuzziness, are proposed also as measures of suitability in order to test the validation of the method and consequently the selected probability pdf.

Another interesting property of the proposed methodology is that the proposed fuzzy band includes all the data at least at some degree. In general, the proposed methodology improves the matching between the sample of the hydrological variable and the used probability distribution.

References

1. Tanaka, H.: Fuzzy data analysis by possibilistic linear models. Fuzzy Sets and Systems, **24**, 363-375 (1987).
2. Ganoulis, J.: Engineering Risk Analysis of Water Pollution: Probabilities and Fuzzy Sets. Wiley-Blackwell (2008).
3. Kitsikoudis, V., Spiliotis, M. and Hrissanthou, V.: Fuzzy regression analysis for sediment incipient motion under turbulent flow conditions. Environmental Processes, **3**(3), 663–679 (2016).
4. Tsakiris, G., Tigkas, D., Spiliotis, M.: Assessment of interconnection between two adjacent watersheds using deterministic and fuzzy approaches. European Water, **15/16**, 15-22 (2006).
5. Spiliotis, M. and Bellos, C.: Flooding risk assessment in mountain rivers. European Water **51**, 33-49 (2016).
6. Papadopoulos, B. and Sirpi, M.: Similarities and distances in fuzzy regression modeling. Soft Computing **8**(8), 556–561 (2004).
7. Chow, V., Maidment, D., and Mays, L.: Applied Hydrology. McGraw -Hill International editions (1988).
8. Spiliotis, M. and Papadopoulos, B.: A hybrid fuzzy probabilistic assessment of the extreme hydrological events. In 15th International conference of numerical analysis and applied mathematics 2017 (ICNAAM 2017) (in press), 25-30 September, Thessaloniki, Greece (2017).
9. Buckley, J. and Eslami, E.. An Introduction to Fuzzy Logic and Fuzzy Sets, Advances in Soft Computing, Volume 13, Springer-Verlag Berlin Heidelberg (2002).
10. Angelidis, P., Maris, F., Kotsovinos, N., Hrissanthou, V. Computation of drought index SPI with alternative distribution functions. Water Resources Management, **26**(9), 2453-2473 (2012).
11. Sfiris, D., Papadopoulos, B.: Non-asymptotic fuzzy estimators based on confidence intervals. *Information Sciences*, **279**, 446-459 (2014).
12. Viertl, R.: Statistical Methods for Fuzzy Data. John Wiley & Sons, Ltd, p. 256. (2011).
13. Shakouri, H., Nadimi, R., Ghaderi, S.-F.: Investigation on objective function and assessment rule in fuzzy regressions based on equality possibility, fuzzy union and intersection concepts. *Computers and Industrial Engineering*, **110**, 207-215 (2017).
14. Yabuuchi, Y.: Possibility Grades with Vagueness in Fuzzy Regression Models. *Procedia Computer Science*, **112**, 1470-1478 (2017).