Probabilistic analysis of coincident sums of precipitation at two measurement stations. Introduction to the method and an example

Gabriela Biel[1,*]

[1]Wroclaw University of Environmental and Life Sciences, Department of Mathematics, ul.Grunwaldzka 53, 50-357 Wrocław, Poland

Abstract. This article proposes the use of copula (copula function) for the purpose of two-dimensional analysis of the sums of precipitation as measured with a Hellman rain-gauge. The sums of precipitation are characterized by a two-dimensional random variable: the sum of uninterrupted sequence of rainfalls which were measured in Jelcz-Laskowice and the corresponding (coincident) sum of precipitation at the Botanical Garden in Wrocław. Several problems occur from the very start: debonding from time and lack of precipitation on one of stations. For the purpose of greater precision and correction it should be stated that in order to apply the two-dimensional copula functions we will use a random vector determining the sum of uninterrupted sequences of rainfalls at two simultaneous stations. In that way, this will not be a characteristics of the phenomenon, but rather the definition of two-dimensional random variable under analysis. Data for analysis has been derived from observational logs of the Institute of Meteorology and Water Management, branch in Wrocław. The results obtained in years 1980-2014 were subject to analysis. The aim of the work was to find the best two-dimensional probability distribution of a random variable ($Opad_{Jelcz}$, $Opad_{Ogród}$). The following were analysed from among the known copulas: the Archimedean copulas (the Gumbel copula, the Frank copula and the Clayton copula) and the Gaussian elliptical copula. The study of fitting of copulas to observed variables was carried out using the Spearmann's rank correlation coefficient and the best fitting was obtained for the Frank's copula.

1 Introduction

A multidimensional analysis of the amount of rainfall and the idea to use the copula function have defined a sequence of steps to be followed [3, 6]. The data for analysis and the research area have been identified in the first place, followed by a list of control stations along with the reasoning behind the choice. There are multiple factors affecting the amount of rainfall. These factors have a direct impact on the next step, which is the selection of stations that are to be subject to the analysis [2, 4, 8, 10, 12, 15, 17].
The main objective of the project is to develop a method for determining the amount of precipitation (and, at the same time, the probability of precipitation) in the area in which the measurement stations are located.

The use of the copula function is one of the key elements of this methodology. Correlation coefficients such as, for example, Pearson, Kendall or Spearman coefficients, are usually used for evaluating the relationship between the features under analysis [3]. It is connected with certain probabilistic assumptions of the phenomenon. The distribution of rainfall in the area determined by the measurement stations is far from normal distribution, which is why a simple assessment would be limited only to the Spearman coefficient. A number of articles have appeared in literature in recent years, discussing the use of the copula function and providing an overview of the structure of connections between variables – to a greater extent than traditional methods [3]. The coefficients that we know have numerical values, whereas the copula function represents a probability distribution.
The expected contribution of the research project to the development of science will be the comparison of the developed method with existing engineering methods such as the isohyetal method, the Thiessen polygon method, the method of inverse distances and with other methods, especially the ones taking into account the location of the station above sea level [2, 10].
The use of Spearmann's rank correlation test as a method of testing the goodness of fit in a two-dimensional case is not a standard approach. We have a two-dimensional empirical distribution on the one hand, and an estimated copula function on the other hand. The author will use a method used for analyzing high-water stage storms measured at two stations simultaneously and check whether or not it is a good way to analyse the amount of precipitation.
The expected result is to obtain the best compatibility with one of the copulas functions, to use this copula function to estimate the multivariate probability

* Corresponding author: gabriela.biel@upwr.edu.pl

distribution of random variable and use the multivariate cumulative distribution function to determine the degree of bulk tails out and the interdependence of marginal distribution's tails. These aspects are rarely given attention and found in traditional engineering methods.

2 Material and methods

Data for analysis has been obtained from the atmospheric precipitation observation logs from 1980-2014. They were collected for Hellmann Rain-Gauges at measuring stations located at a similar heights above the sea level. As a result of analyses, coincidental episodes of precipitation were obtained, which constitute input data for the model that has been developed.
The copula function has been chosen as a method of statistical estimation parameters of multivariate probability distributions because it enables:
- the combination of any boundary distributions,
- the compilation of building multivariate statistics based on estimated marginal distributions [3].
This is particularly important when random variables are considered, which are statistically dependent on one another. Moreover, the first feature is of key importance in the case of the data from winter months, when the amount rainfall is likely to be different than in other periods of the year and - consequently – it poses certain difficulties in choosing a transformation leading to normal distribution [7, 9, 14, 16].

The use of the copula function in practice, i.e. for empirical data, requires the estimation of marginal distribution parameters and a parameter or parameters of the copula function. The estimation usually takes place with the use of the maximum likelihood method [3, 13]. Archimedean copulas will be subject to the study, including the Gumbel-Hougaard copula.

I. The estimation of marginal parameters with the use of the maximum likelihood (ML) method will have the following stages [3, 13, 18, 19]:
1. The study of the homogeneity of random variables.
We assume that the distribution of random variable is unknown at this stage.
2. Determination of probability distribution measures for each marginal random variable: skewness, kurtosis, dispersion and position, the estimation of unknown parameters.
3. The use of non-parametric statistical tests and graphic methods.

II. The process of estimating the parameter or parameters of the copula function with the maximum likelihood (ML) method will have the following stages [3, 13, 18, 19]:
1. Investigating the relationship between analysed random variables (the heights of coincident precipitation episodes in the analyzed stations).
2. Estimation of the copula function parameter.
3. Use of non-parametric statistical tests.

The following has been observed and applied in the analyses carried out so far:

I.1. The most common characteristics of measurement chains subject to change are the mean value, variance (changes of these values occur either as discrete stepped changes, or as changing trends). In order to detect disturbances of hydrological phenomena, the most effective tests were used: the Kruskal-Wallis rank sum test, the Spearman rank correlation test on mean value's trend as well as on the variance's trend of the random variable. Outliers will be identified using the Grubbs-Beck test [3]. In the cited work [3] which discusses high-water stage storms, these methods are considered to be the most effective. It is also certain that other measures must be taken into account in future studies. Attention must be given to the fact that the results obtained from these tests have a low power.

I.2. Choosing the appropriate theoretical probability distributions, which are a mathematical model of random variable characteristics (separately for each marginal distribution). Perhaps the selection will be limited to distributions: the two-parameter Gamma distribution and the Weibull distribution. As a result of the estimation of probability distribution parameters (carried out by means of the maximum likelihood method), a function which depends on elements of a random sample will be identified.
II.3. Comparing the goodness of fit of adopted theoretical distribution with empirical distribution. The goodness of fit will be checked with the use of the goodness of fit tests and graphic methods used to verify the hypothesis concerning the goodness of fit (of the empirical probability distribution with the hypothetical distribution (bar diagram, probability-probability graph [P-P] of the random variables studied)).

II.1. In the first scheduled step, the testing of dependency will be performed with the use of rank's tests.

II.2 For the Gumbel-Hougaard Archimedean copula the parameter estimation will be performed using the IFM method (Inference Functions for the Margins), consisting of two stages: estimation of marginal parameters of the probability distribution and estimation of copula parameters (based on a set of estimated marginal parameters of probability distributions).

II.3 The same procedure as in the case of one-dimensional. Goodness of fit tests will be used to check the goodness of the copula's compatibility with empirical data.

The calculations and charts will be made using the R and Statistica package [13, 18, 19].

3 Example

An example of searching and choosing the copula function for precipitations measured in two stations: in

Jelcz-Laskowice and at Ogród Botaniczny in Wrocław is presented below. Firstly, we divided a year into equal periods (for example six or five weeks). Then we proceeded to calculate the amount of precipitations in rain sequences. We chose non-standard divisions in order to avoid a situation in which it is raining for a few days on the cusp of months and we divided this rainfall in an unnatural way, as a result of which the information about the total sum has been lost [11].

Table 1. Sum of rain sequences measured at two stations.

JELCZ		year	month	day	OGRÓD	
1.4		1980	1	1	0.5	
1.5		1980	1	2	2.4	
0.1		1980	1	3	0.2	
0.1		1980	1	4	0.0	
1.0		1980	1	5	3.2	6.3
0.6	4.7	1980	1	6		
		1980	1	7		
1.7		1980	1	8	1.3	
0.2	1.9	1980	1	9	0.0	1.3
0.0		1980	1	10		
0.0		1980	1	11	0.0	
		1980	1	12		
		1980	1	13		
		1980	1	14		

As in the example above, we are not interested in the duration of rain. What we are interested in is the total sum of rain sequence observed at the station and a coincidental sequence at the second station. We prepared all the data in the same way, obtaining two-dimensional variable where the sum from one station is the first coordinate, and the sum from the second station is the second coordinate [3, 11].

Table 2. Rain sequence lengths and precipitation sums in respective precipitation events (part of a data stream).

1980-01-01	**1980-01-06**	**4.70**	:	**1980-01-01**	**1980-01-05** **6.30**
1980-01-08	1980-01-11	1.90	:	1980-01-08	1980-01-11 1.30
1980-01-16	1980-01-16	0.00	:	**********	
1980-01-22	1980-01-27	6.30	:	1980-01-22	1980-01-27 7.30
1980-01-29	1980-02-08	**13.40**	:	1980-01-29	1980-02-07 **18.90**
1980-02-10	1980-02-13	8.00	:	1980-02-10	1980-02-13 9.60
1980-02-16	1980-02-18	1.30	:	1980-02-16	1980-02-18 2.90
**********			:	1980-02-21	1980-02-21 0.00
1980-02-27	**1980-03-03**	**9.10**	:	**1980-02-25**	**1980-03-04** **10.80**
1980-03-07	1980-03-13	7.90	:	1980-03-07	1980-03-13 10.20
1980-03-24	1980-03-25	0.00	:	1980-03-24	1980-03-24 0.00
1980-03-28	1980-03-30	0.00	:	**********	
1980-04-01	1980-04-06	12.50	:	1980-04-01	1980-04-06 16.90

The same procedure has been applied in the case of three-dimensional variables or more. A number of problems has been encountered that had to be addressed prior to the start of the analysis:
1. Rain occurs at one station (no rain at another station)
2. How many days without rain would interrupt the rain sequence?
3. How to proceed with the pairs where there is at least one zero?

The following assumptions have been made:
The rain sequence is interrupted when there is no rain at all any of the stations. It is shown below in the table.

Table 3. The sum of rain sequences measured at two stations including a situation in which there is no rain at one station (during one day or during rainy days at the other station).

	JELCZ	year	month	day	OGRÓD	
		1981	1	14		
	0.0	1981	1	15	3.7	
	2.3	1981	1	16	1.2	
	0.0	1981	1	17	0.7	
	1.0	1981	1	18	0.7	
	1.5	1981	1	19	2.3	
	1.7	1981	1	20	0.3	
	1.0	1981	1	21		
7.5	0.0	1981	1	22	0.2	**9.1**
		1981	1	23		
		1981	1	24		
	1.9	1981	1	25		
2.5	0.6	1981	1	26	0.0	**0.0**
		1981	1	27		

When it comes to point three, we decided to leave out the pairs with at least one zero, because the Gamma distribution does not work with such pairs of data [6, 9, 16], which meant getting back to the data that had been excluded and including it into the analysis.

Gamma distribution was selected for boundary distributions on both stations at the period of time under analysis – based on: graphical data presentation, tests of the goodness of fit of empirical distributions with theoretical distributions, previous analyses and the study of literature,.

Parameters of boundary theoretical distributions were estimated using the maximum likelihood method (ML). The obtained values of boundary distribution parameters for both variables ($Opad_{Jelcz}$, $Opad_{Ogród}$) can be found in the table below:

Table 4. The obtained values of boundary distribution parameters for both variables (Gamma distribution).

	α /shape/	**β /rate**
Jelcz	0.86171	0.09815
Ogród	0.81280	0.08837

In keeping with Sklar's statement, the estimation of the parameters of the copula function was possible after all our data has been prepared and after marginal

distributions have been described with the estimation of their parameters (and also after testing the fitting).

In this example, as in the cited research of high-water stage storms, our choice is substantiated by Spearman's rank correlation coefficient, calculated (with use of appropriate package of R) between generated realizations of the variable random ($Opad_{Jelcz}$, $Opad_{Ogród}$)..

Eight charts of simulated and observed data are presented below. Simulated data comes from the different copula functions: the normal-elliptic copula, the Gumbel-Hougaard copula, the Clayton copula, and the Frank copula for data from first six week of each year (this is the one of the analysed periods) and for all data. Of course, this data was previously prepared in the manner that has been mentioned above. Before charts, there is a table of estimated coefficients of each copula function. The estimation of these parameters was conducted with use of appropriate package of R.

Table 5. The obtained values of chosen copula functions.

	first six weeks of each year (1980-2014)
normal-elliptic	0.8
Gumbel	4.43
Clayton	3.14
Frank	18.03

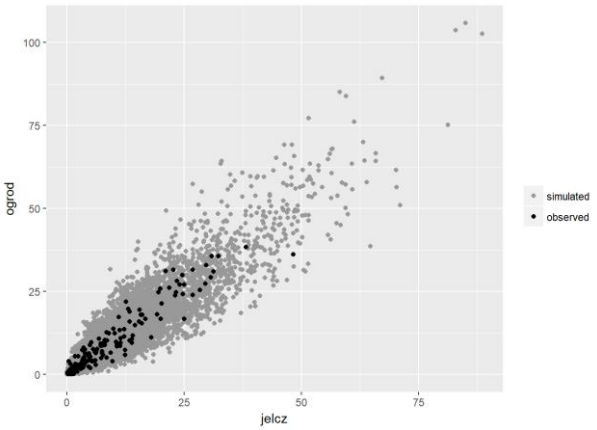

Fig. 1. The normal-elliptic copula. Observed points are from the first six weeks of each year (1980-2014). Observed points are marked in black and simulated are marked in grey.

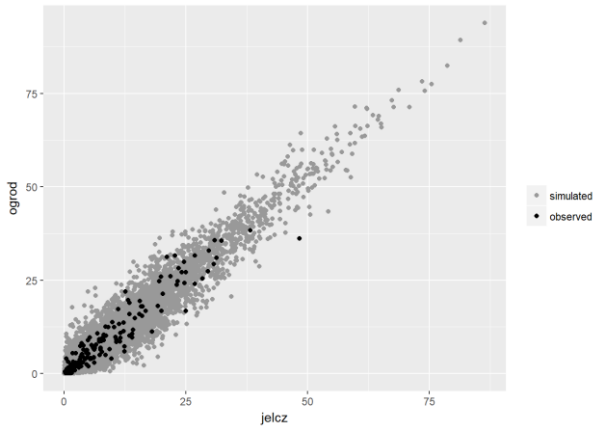

Fig. 2. The Gumbel copula. Observed points (black) are from the first six weeks of each year (1980-2014).

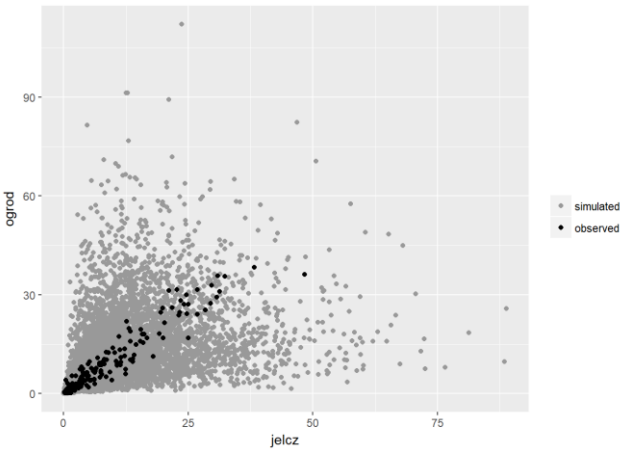

Fig. 3. The Clayton copula. Observed points are from the first six weeks of each year (1980-2014).

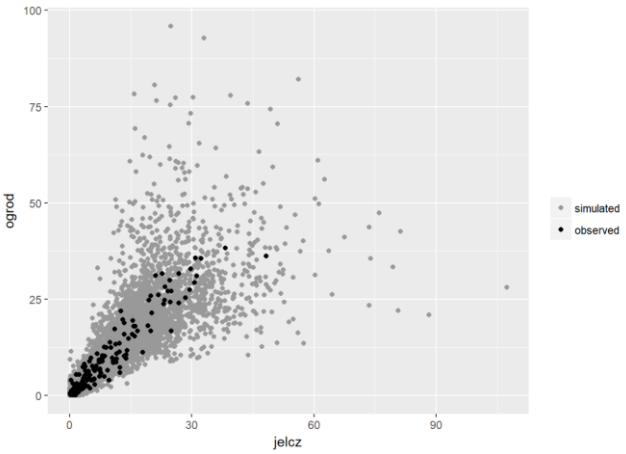

Fig. 4. The Frank copula. Observed points are from the first six weeks of each year (1980-2014).

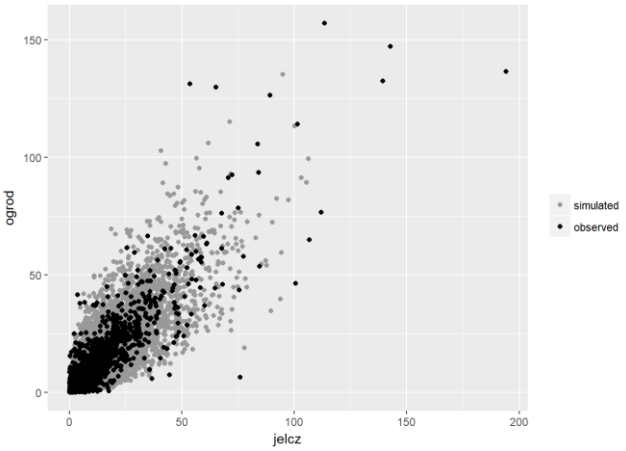

Fig. 5. The normal-elliptic copula. Observed points are all data from analysed years (1980-2014).

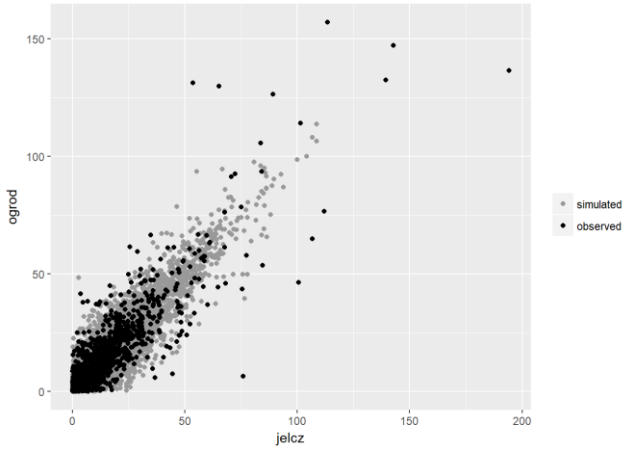

Fig. 6. The Gumbel copula. Observed points are all data from analysed years (1980-2014).

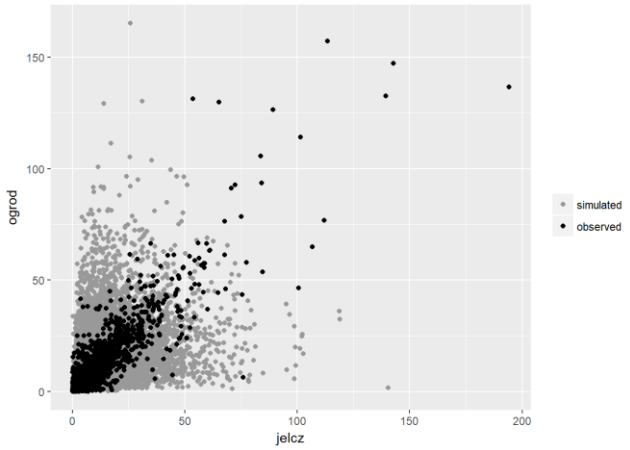

Fig. 7. The Clayton copula. Observed points are all data from analysed years (1980-2014).

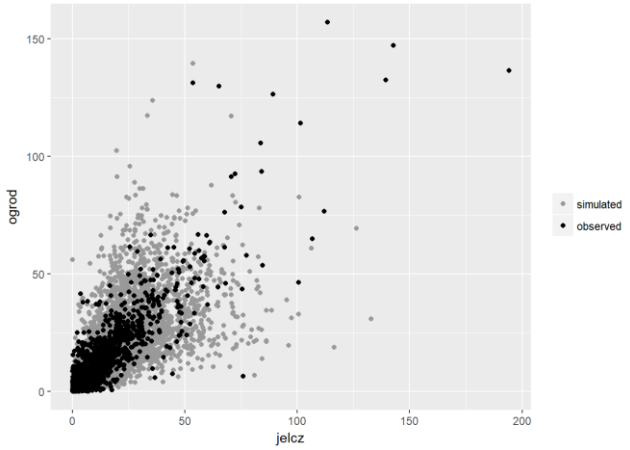

Fig. 8. The Frank copula. Observed points are all data from analysed years (1980-2014).

When we look only at charts, it is difficult to determine which copula function gives the best fitting. The second step that follows the creation of charts is choosing the right option based on Spearman's rank correlation coefficient, as in the cited research of high-water stage storms. Higher Spearman's rank correlation coefficient indicates a better fitting in that case [3].

Table 6. Spearman's rank correlation coefficient.

	first six weeks of each year (1980-2014) N=10 000	**all data from analysed years (1980-2014) N=10 000**
normal-elliptic	0.93	0.84
Gumbel	0.93	0.86
Clayton	0.79	0.70
Frank	**0.95**	**0.89**

When we make an inference from the Spearman's rank correlation coefficient it suggests that the best-fitting data is Frank's copula. This information should be treated as a mere signal, however. It is also recommended to test the goodness of fit for all of these copula functions, which seems to be the next step.

It must be realized that the R provides an opportunity to analyse more than forty copula functions and this example is only a point of departure. An excerpt from all of the analyses which will be taken into consideration, including yet other divisions of year, and, of course, other parts of the year – not only, the first six weeks [3, 6, 13, 18].

4 Conclusions

Data for analysis comes from the logs of atmospheric precipitation observations from the last 30 or 40 years. It was collected using Hellmann rain gauges from the measuring stations located at a similar height above the sea level [2, 15]. These are diel data, and the strategy of their development, in particular the method of counting rainfall sums in rainfall sequences, is an important element of the methodology as well as a new proposal – compared to the existing engineering methods [2].
The use of the copula function for the proposed analyses necessitates the estimation of marginal distributions and the comparison of empirical and theoretical marginal distributions [3, 6]. It is worth mentioning that the choice of copula as a method of estimating the parameters of multidimensional probability distributions was determined by its properties: the copula enables the combining of any marginal distributions and the development of multidimensional statistics while having marginal statistics at its disposal [3]. With marginal distributions (for example the gamma distribution) and the copula function (selected for example among the Archimedean copulas) it is possible to not only compare precipitation amounts measured at several stations, but also to determine the precipitation sum distribution for any point located in the area designated by measuring stations [3]. This point can be referred a hypothetical station.
The expected contribution of this research project to the development of science is a comparison of the discussed method with already existing engineering methods, such

as the isohyetal method, the Thiessen polygon method, the method of inverse distances and others (especially those that consider the location of the station above sea level).

One of the inspirations for this paper was the following extract [15]:
Precipitation belongs to the ungrateful estimation objects due to their temporary and spatial discontinuity, relative to seasonality and variability of precipitation conditions.

(A. Stach, J. Tamulewicz, Efficiency of chosen algorithms of monthly and yearly precipitation totals spatial estimation. Preliminary assessment, 2003)

The main goal of the project is a development of a multivariate rainfall analysis method on the area designated by measurement stations, which will enable a more straightforward estimation of the precipitation, and will become more feasible to apply also in areas where there are no measuring stations. The developed method will allow the determination of the amount of rainfall and the probability of occurrence of a certain amount of precipitation (and the occurrence of precipitation in general) anywhere in the area designated by measuring stations [2, 4, 14, 15].
It is true that the study of the adequacy of multidimensional models is not easy, both from the statistical and a practical point of view. The next step must therefore be to include more measures for comparing these models. It is worth mentioning and emphasising that the method of observing the occurrence of precipitation in two places that has been described does not address the issue of prediction, since it is detached from time. The key issue is to get insight into the relationship between the amount of precipitation at two stations.

References

1. A. Bardossy, G.G.S. Pergram, Hydrol., Earth Syst. Sci **13** (2009)
2. A. Byczkowski, *Hydrologia* (Wyd. SGGW, Warszawa, 1999)
3. M. Ciupak, K. Rokiciński, Zeszyty naukowe Akademii Marynarki Wojennej **4 (187)**, 15-34 (2011)
4. M. Czarnecka, J. Nidzgorska-Lencewicz, Woda-Środowisko-Obszary wiejskie **12**, 45-60 **(**2012)
5. P. J. Danaher, M. S. Smith, *Modeling Multivariate Distribution Using Copulas: Applications in Marketing* (Melbourne Business School, University of Melbourne, 2009)
6. Ch. Genest, A.C. Favre, J. Hydrol. Eng. **12**, 4, 347) (2007)
7. T. Izawa, Papers in Meteorology and Geophysics **15**, 167-200 (1964-1965)
8. Z. Kaczmarek, *Metody statystyczne w hydrologii i meteorologii* (Wyd. Komunikacji i Łączności, Warszawa, 1970)
9. K. Krishnamoorthy, T. Mathew, S. Mukherjee, Technometrics **50**, 69–78 (2008)
10. M. Ozga-Zielińska, J. Brzeziński, *Hydrologia stosowana* (PWN, Warszawa, 1997)
11. L. Perini, M. Beltrano, *Linking of traditional and automatic stations data: operational experience of UCEA* (3ª Conferencia International Sombre Experiencias con Estaciones Meteórologicas Automáticas, Torremolinos, Malaga 2003)
12. J. Pociask-Karteczka, *Zlewnia. Właściwości i procesy* (UJ, Kraków, 2003)
13. V. Ricci, *Fitting distributions with R*, https://cran.r-project.org/doc/contrib/Ricci-distributions-en.pdf (2005)
14. M. Smalley, T. L'Ecuyer, J. Appl. Meteorol. Clim. **54** (2015)
15. A. Stach, J. Tamulewicz, *Obieg wody – uwarunkowania i skutki w środowisku przyrodniczym*, 87-111 (Inst. Badań Czwartorzędu i Geoekologii, UAM, Poznań, 2003)
16. H.C.S Thom, Mon. Weather Rev. **86**, 117–122 (1958)
17. S. Węglarczyk, *Statystyka w inżynierii środowiska* (Wyd. Politechniki Krakowskiej, Kraków, 2010)
18. R Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistcal Computing, Vienna, Austria, 2018), URL https://www.R-project.org/
19. StatSoft, Inc. Statistica (data analysis software system), Version 12,5 (2011)