International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

An Evaluation of Probability Distributions of Synthetic Storms

Betül Saf

Hydraulic Division, Department of Civil Engineering, Pamukkale University, Turkey

Abstract: ***The study examines whether 56 data sets consisting of 100 synthetic storms with the same probability distribution (Gumbel) are different than the distribution provided for at the beginning. For this purpose, synthetic synthetic storms of Gumbel distribution with a specific time distribution and random effective durations, of which population averages and variances are known, are being derived with Monte Carlo simulation method. The parameters of the storm values derived were determined using the maximum possibility method for 7 probability distribution widely used in hydrology, and their compliance was examined using Chi-square (χ^2) and probability plot correlation coefficient tests (PPCC). It was seen that the probability distribution of the precipitation input can be different from the main distribution (Gumbel) provided for. The reason for this is that the precipitation inputs created are in the form of synthetic storms of different periods and the sample statistics of these series are different from the main distribution based on sampling.***

Keywords: Monte Carlo method, probability distribution, methods for parameter estimation, compliance tests

1. Introduction

In order to make future estimations in water resources planning, information on hydrometeorological data such as precipitation, flow, evaporation and temperature of random feature that changes by location and time, and the definition of probability distributions of these data. The main objective in peak precipitation and flow estimations with any repetition periods is to design engineering structures such as dam, crossbar, bridge and berm safely. When the probability distributions of hydraulic data are determined, the frequency and magnitude of formation of these events are also defined.

The most important hydrologic data used in the design and management of water structures are severe rain storm and flood events that are of extreme feature and significant in terms of risk. The existence of this type of data ensures that the design is correct and reliable. However, in certain cases, the basin to be analyzed and designed does not have a flow and/or precipitation observation station, thus no precipitation and flow records. When such data exist, they can be very short in terms of making extreme event estimations, or the current data length is shorter than the data length taken into consideration in the design of the water structure. For example, it is necessary to know the peak flow value with a recurrence of 100 years when designing the dam spillway. However, the current data length in many basins is 40-50 years.

In practice, the distribution that best fits the data in any station or region is not known. In order to be able to determine the distribution model, it is asked for the data length that is the basis of the analysis is at least twice the recurrence period to be estimated. In this case, one of the methods that are widely used to increase the data length is the Monte Carlo method that produces random data using the statistical features of current data.

The Monte Carlo method, also named as statistical trials, is a probability method that is widely used in the modelling of natural systems. It is used in order to simulate natural processes under the effect of random factors and solve mathematical problems with no clear solution. In this study, heavy rain series of different periods are randomly derived with Monte Carlo method depending on the physiographical properties of the basin (L; main branch length, S; harmonic slope).

2. Monte Carlo Method and its Use in Hydrology

Monte Carlo method is widely used in order to minimise the modelling and sampling errors especially in the estimations with extreme values, and the examine the relative model performances of certain statistical magnitudes with different simulations. There are many studies in hydrology literature related to Monte Carlo methods that are parametric and non-parametric aiming to examine the relative model performances of certain statistical indices. In this study, only the studies on Monte Carlo method were summarized as Monte Carlo method is used.

First the frequency distribution and the parameters of this distribution are primarily chosen in parametric Monte Carlo simulation trials. Then, synthetic data sets longer than the existing data length are produced with the help of the distribution chosen. Synthetic data are obtained with the random number generation technical with the help of the magnitudes with a certain probability (Haan, 1977). Benson (1952), Nash and Amorocho (1966) assessed the data derived in Gumbel distribution with the help of Monte Carlo technical with the aim of examining the effect of sample size on standard errors in flood estimations. With the aim of examining the estimation performance, Wallis et al. (1975) derived sample clusters with various distributions such as log-pearson 3 and extreme value 3 using the average, standard deviation and skewness coefficients, and they reached the conclusion that the coefficient of skewness has major sampling errors, and it also yields biased and limited estimations.

Adamowski (1989) investigated the parametric Monte Carlo simulation method that produces synthetic flood data with non-parametric flood frequency distributions that are more reliable than homogenous distributions for major recurrence periods, and concluded that parametric and non-parametric distributions have the same reliability for small recurrence periods. Lettenmaier and Potter (1985) created random samples with Gumbel 2-parameter lognormal and 3-parameter log pearson distributions in order to define a regional flow model connected to flood statistics drainage area. Rossi et al. (1984) and Beran et al. (1986) examined the statistical properties of the extreme value distribution with 2 components based on the observed flood data of U.K. that are derived with Monte Carlo method, and concluded that this distribution is suitable for the regional flood model. Arnell and Beran (1987) conducted simulation trials in order to compare the regional estimations of index type that also contains the generalized extreme value and the estimations obtained with the probability weighted moments method of Wakeby distributions and the difficulty of the extreme value distribution with 2 components. In their study, they found that the 2-component extreme value distribution is better in terms of bias, however Wakeby distribution is more successful in terms of variance terms.

3. Generation of Synthetic Storms

The way of deriving 56 synthetic storm samples of 100 length with a certain time distribution and random effective durations, each of which are of Gumbel distribution, and of which population arithmetic means and variances are known using the Monte Carlo technical is as follows:

First, the effective durations of the maximum precipitations were associated with basin characteristics such as the main branch length and harmonic slope using the concentration time equation of Kirpich. For this, it was considered within the change interval in the form of main branch length (L, m) $(8000 \le L \le 20000)$ and harmonic slope (S) $(0.0001 \le S \le 0.02)$, and the concentration time was calculated with the help of Kirpich equation for each (L, S) couple. Meanwhile, the critical storm duration is taken into consideration independently from the basin area.

The basin concentration time (t_c) was calculated using the below-mentioned Kirpich equation by main branch length (L) and harmonic slope (S), and the effective precipitation duration was calculated using the equation (2) by the concentration period.

$$t_c=0.00032(L)^{0.77}/S^{0.385} \quad (1)$$

$$De=2(tc)0.5 \quad (2)$$

In these equation; L means the length of the main branch (m), S means the harmonic slope, t_c means the concentration time of the basin (hour), and D_e means the effective precipitation duration (hour). After the concentration time (t_c) and effective precipitation duration (D_e) were calculated, the critical precipitation duration (D) was taken into consideration as $D_e \le D \le 2D_e$ subject to the duration of effective precipitation. In cases when the concentration time is over 4 hours ($t_c \ge 4$), the effective precipitation duration is considered as equal to the basin concentration time ($D_e=t_c$).

This way, after the storm durations (D) are randomly derived, the moment and standard deviation values of Gumbel distribution considered as the main distribution in the study are calculated with the following equations using the relationships established between standard period precipitation values and statistical parameters of Uşak precipitation station, chosen as a sample, between 1929 and 1988 (Benzeden, 2001). As the duration of precipitation is random, the moment and standard deviation values are also of random feature.

Precipitation mean-duration relationship:

$$M_D = 4.154[\ln(D/1.1763)]^{1.0263} \quad (3)$$

Precipitation standard deviation-duration relationship:

$$S_D = \exp\left[-1.46047 + 1.79839\ln(D) - 0.3155\ln^2(D) + 0.01944\ln^3(D)\right] \quad (4)$$

In these equations, D has the units of minute, M_D and S_D mm.

After equations are created for moment and standard deviation parameters, random precipitation values (Y_D) of which probability (P_T=1-1/T) ranges between 0 and 1 are derived using equation 6.

$$K_T = - \{0.45 + 0.7797 \ln[-\ln(1-1/T)]\} \quad (5)$$

$$Y_{D,T} = M_D + S_D.K_T \quad (6)$$

In these equations K_T is a coefficient depending on the recurrence time (T) and distribution type.

4. Parameter Estimation Methods and Goodness of Fit Tests

There are many probability distributions used for hydrometrological data. In this study, 7 probability distribution model (normal (NOR), lognormal with two and three parameters (LN2 and LN3), gumbel (GUM), loggumbel (LGUM), and gamma distributions with 2 and 3 parameters (G2 and G3) that are well-known and widely used in hydrology were used.

Population characteristics of the random variable are characterized by the sample statistics estimated from the current data sample. Estimations related to the sample statistics must be unbiased and effective. Many parameter estimation methods such as moments, maximum likelihood, L-moments and entropy are used in the parameter estimations of probability distributions deemed to fit any data set (Kite, 1977). In this study, the parameters of probability distributions are determined using the moments and maximum likelihood methods. Method of moments is widely used in hydrology as it is simple. However, it does not yield effective and unbiased estimations for skew distributions. In addition, the maximum probability method yields unbiased and effective estimations in samples with a high data length.

There are several tests for checking the frequency function obtained from a sample observed with a chosen theoretical probability distribution function, such as Chi-square (χ^2 test), Kolmogorov-Smirnov (K-S) and probability plot correlation coefficient test (PPCC). In this study, the Chi-square (χ^2) and probability plot correlation coefficient test (PPCC) were used in this study.

In χ^2 test, a sample with N elements of a random variable is classified into m classes, and the number of elements (N_i) in each class is calculated. When the probability of being within the same class intervals p_i is expressed in accordance with the chosen probability intensity function, below statistic is obtained. The sampling distribution of this statistic is the χ^2 distribution with a degree of freedom of (m-1). When the χ^2 value calculated in accordance with this equation (7) is lower than χ^2_α value with a probability of exceedance of (α) in the degree of freedom of (m-1), it is deemed that the observed distribution is equal to the theoretical distribution that is chosen (Kite, 1977; Bayazıt and Oğuz, 1994).

$$\chi^2 = \sum_{i=1}^{m} \frac{(N_i - Np_i)^2}{Np_i} \qquad (7)$$

In the probability plot correlation coefficient test, the probability of remaining smaller than the adjoint probability distribution in the form of F(x) is calculated for each element (x_i) in the sample; and the z_i standard normal variable value that is equal to this calculated value is then calculated. $r_{x,z}$ correlation coefficient is calculated between (x_i,z_i) couples thus determined. If the value of this coefficient is higher than the critical $r_{kr;x,z}$ value, it is considered as a theoretical distribution.

5. Assessment of Probability Distributions of Synthetic Storm Samples

For the purposes of assessing the probability distributions of synthetic precipitation samples derived in the study, the qui-square test was applied in accordance with the class interval taken into consideration as α=5% probability of exceedance, and k=8. The critical qui-square value of normal, 2-parameter lognormal, gumbel, loggumbel and 2-parameter gama distributions was determined as 11.7; while it was determined as 9.49 for 3-parameter lognormal and gama distributions.

The relative frequency values ($f_{i,k}$) of each probability distribution model were calculated with the following equation in order to assess the probability distributions:

$$f_{i,k} = 100\ (TNCH)/56 \qquad (8)$$

In this equation, the TNCH value is the total number of the series passing the χ^2 test for the distribution of which compliance is examined.

Similarly, the compliance of the probability distributions is also tested using the probability plot correlation test widely used in hydrology. The critical correlation coefficient value for this test was taken into consideration as $r_{k;x,z}$= 0.95. The following equation was created similarly to (8) equation for this test, and the partial frequency values were calculated.

$$f_{i,p} = 100\ (TNPP)/56 \qquad (9)$$

In this equation, the TNPP value is the total number of heavy rain series that are higher than r_c=0.95.

6. Result

When 56 synthetic precipitation series derived using the Monte Carlo technique were tested using the χ^2 compliance test, it was found out that the acceptable relative frequencies of LN2, G2 and GUM distributions were high. In addition, it is observed that the relative frequencies of LN3 and G3 distributions are also high (Figure 1 and Table 1). As is understood from Figure 1 and Table 1, it was determined that the synthetic storm series, each of which are derived with Gumbel distribution, may have different distribution types than the projected distribution and only one of these distributions does not exhibit significant difference than the others. In other words, when the synthetic precipitation series that are individually derived with Gumbel distribution are assessed in combination, they may have other distributions that the projected distribution (Gumbel).

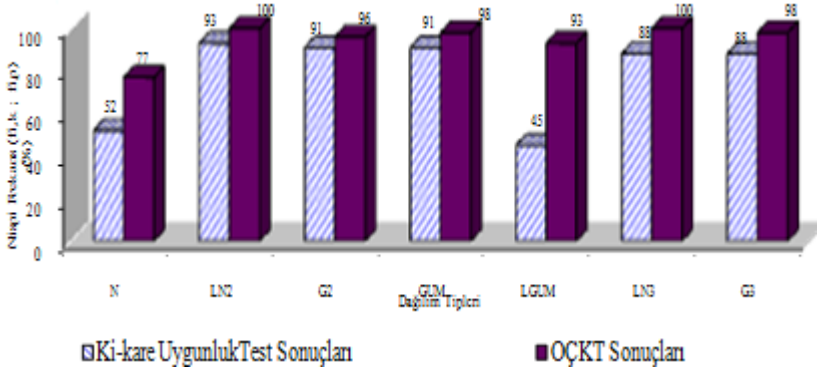

Figure 1: Relative Frequencies of the Distributions by χ^2 and Probability Plot Correlation Tests

Similar results were obtained in probability plot correlation test, and it was observed that the mixed storm series may have the LN2, LN3, GUM and G2 distributions (Figure 1 and Table 1). Furthermore, it is also observed that the type of compliance test is not effective in determining the suitable distribution in terms of the highest frequency value. Considering the distribution with the highest frequency value; LN2 has the highest relative frequency value in qui-square test, while LN2 and LN3 distributions have the highest relative frequency value in PPCC test.

The results of the study put forth that the probability distributions of the precipitation inputs may have different distributions than the projected distribution (Gumbel) as a result of the sample and the fact that the derived synthetic precipitation series are in the form of the precipitation events with different durations; however, there is no clear and significant difference between these distributions when deciding on which distribution to prefer. At this point; the length of the existing data, the physical formation type of the event, and its definition by position and time become important. In addition to these properties of the hydrological data, its ability to define the properties and data of the probability distributions statistically easily and correctly is another important issue to take into consideration.

Table 1: Results of χ^2 and PPCC Compliance Tests

Relative frequency (%)	Compliance test	Type of Probability Distribution						
		NOR	LN2	G2	GUM	LGUM	LN3	G3
$f_{i,k}$ and $f_{i,p}$	χ^2	29/56	52/56	51/56	51/56	25/56	49/56	49/56
	PPCC	43/56	56/56	54/56	55/56	52/56	56/56	55/56

References

[1] Adamowski, K. (1989), A Monte Carlo Comparison of Parametric and Nonparametric Estimation of Floods, Journal of Hydrology, 108, 295-308.
[2] Arnell, N., Beran, M. (1987), Testing the Suitability of the Two Component Extreme Value Distribution for Regional Flood Estimation, Regional Flood Frequency Analysis, Ed. By V. P. Singh, pp.159-175.
[3] Bayazıt, M., Oğuz, B. (1994), Mühendisler için İstatistik [Statistics for Engineers], Birsen Yayınevi, İstanbul.
[4] Benson, M. A. (1952), Characteristics of Frequency Curves Based on a Theoretical 1000-Year Record, USGS Open-File Report, U.S. Geological Survey, Reston, Virginia, USA.
[5] Benzeden, E (2001, Standart Süreli Maksimum Yağışların Frekans Analizinde Karşılaşılan Sorunlar [Problems Encountered in the Frequency Analysis of Maximum Precipitations with Standard Duration], DSI Teknoloji D. Bşk. Basım ve Foto-Film Şb. Md., III. Ulusal Hidroloji Kongresi Bildirileri, İzmir, p.11- 18.
[6] Beran, M. ve diğerleri (1986), Comment on "Two-Component Extreme Value Distribution for Flood Frequency Analysis, by Fabio Rossi, Mauro Fiorentino ve Pasquale Versace, Water Resources Analysis, 22, 263-6.
[7] Bras R. L., ve Rodriguez-Iturbe I. (1985), Random Functions and Hydrolog,. Addision-Wesley, Reading, Mass.
[8] Haan, C. T. (1977), Statistical Methods in Hydrology, Iowa: Iowa State University Press, Ames.
[9] Kite G. W. (1977), Frequency and Risk Analysis in Hydrology, Water Resources Publications, Fort Collins, Colorado, USA.
[10] Krajewski, W. F. et al.(1991), A Monte Carlo Study of Rainfall Sampling Effect on a Distributed Catchment Model, Water Resources Research, 27, 119-128.
[11] Lettenmaier , D. P., Potter, K. W. (1985), Testing Flood Frequency Estimation Methods Using a Regional Flood Generation Model, Water Resources Research, 21 (12), 1903-1914.
[12] Mtiraoui, A. (2004), Mixture Distributions and Spatial Scale Effects on Flood Hydrology, Ph. D. Thesis, The University of British Columbia.
[13] Rao A. R. and Hamed K. H. (2000), Flood Frequency Analysis, CRC Press, New York, Washington, D.C.
[14] Rossi ve diğerleri (1984), two-Component Extreme Value Distribution for Flood Frequency Analysis, Water Resources Research, 20 (7), 847-856.
[15] Wallis, J. R., Matalas, N. V., ve Slack, J. R. (1974). Just a Moment, Water Resources Research, 10 (2), 211-219.