Definition of the best probability distribution functions for annual minimum flows in the rivers of the Upper Euphrates River Basin, Turkey

Fatih Tosunoglu¹ and Ercan Kahya²
¹Erzurum Technical University, Engineering and Architecture Faculty, Civil Engineering Department, Erzurum, Turkey
²Istanbul Technical University, Civil Engineering Department, Istanbul, Turkey

Abstract. In this study, it is aimed to define the most suitable probability distribution functions for modelling of annual minimum flows (AM7) data for the duration 7-day that obtained from daily streamflow data of the four gauge stations located in upper part of the Euphrates River. To achieve this purpose, ten widely used distributions, namely, Gamma, Gumbel, Generalized Extreme Value, two parameter Lognormal, three parameter Lognormal, Logistic, LogLogistic, Normal, Pearson type-III and Weibull were fitted and compared. The maximum likelihood method was used to estimate parameters of the distributions. To select the most suitable distributions for AM7 series of the stations, five widely used model selection criterions (namely, the Akaike Information Criterion (AIC), Anderson Darling (AD), Bayesian Information Criterion (BIC), Cramér–von Mises (CvM) and Kolmogorov–Smirnov (KS) tests) were applied. Based on these tests/criteria, Gamma, Gumbel, Generalized Extreme Value, three parameter Lognormal, Pearson type-III were found to be the most suitable distributions for modelling data series. Finally, using the best fitted distributions that derived from these model selection methods, minimum flow estimates for various return periods (e.g. 10, 25, 50, 100, 200 and 500 years) were performed and results were discussed.

Keywords: Annual minimum flows, Euphrates River Basin, probability distribution functions, Turkey

1. Introduction

Estimation of annual minimum flow statistics is crucial for various hydrologic studies such as water quality management, planning water supplies, hydropower, cooling, irrigation system design evaluating the effect of prolonged droughts on aquatic ecosystems [1, 2]. Researchers usually need to determine the magnitude of minimum flow for a probability of exceedance or recurrence interval. To achieve this purpose, univariate frequency analysis, which is usually concerned with fitting a suitable probability distribution to observed data series, is commonly applied. Numerous studies have been carried out to estimate various types of annual minimum flows of rivers in the different regions. For instance, Zaidman et al. [3] considered four distribution types including Generalised Extreme Value, Generalised Logistic, Pearson Type-3 and Generalised Pareto distributions for modelling annual minima D-day average flows (e.g. 1, 7, 30, 60, 90, and 365 days) in British rivers. Yue and Pilon [4] used the Pearson type-III, the three parameter lognormal and log Pearson Type-III distributions for describing annual minimum streamflow observed in eleven climatic regions of Canada. Chen et al. [5] applied various probability distributions for modelling annual 7-day low flows in southern China. The three parameter lognormal distribution provided the best fit among the candidates (namely generalized extreme value, Pearson Type-3 and generalized Pareto distributions). Recently, Keshtkar [6] employed the log-normal, three-parameter lognormal, Gumbel, Pearson type III, and log Pearson type III distributions for describing the annual minimum stream flow of 20 rivers in Iran. The results showed that the Pearson type III, log Pearson type III, and Gumbel distributions were better than the other distributions. Eris et al. [7] investigated the most suitable probability distributions for modelling time...
series of different D-day low flows (e.g. 1, 7, 14, 30, 90 and 273) calculated from rivers in four hydrological basins from different regions in Turkey. They found that the 3-parameter log-normal (LN3) probability distribution function fits quite well to most of the D-day low flows in three basins, namely Meric-Ergene, Gediz and Ceyhan, in northwestern, western and southern parts of Turkey, respectively. Moreover, Weibull distribution provided the best fit to the majority of the low flow time series in Seyhan basin in the south of the country. In order to obtain the best flow estimation, a wide range of probability distributions have been used in the literature. However, only a few family of the model selection methods has been considered. It is important to remember that there are many model selection methods that can be effectively in hydrology field (e.g. the Akaike Information Criterion(AIC), the Anderson-Darling Criterion (AD), the Bayesian information criterion (BIC), the Chi-squared test ($X^2$), the Cramér–von Mises (CvM) and the Kolmogorov Smirnov (KS) tests). As each test/method has own pros and cons, the performance of different model selection methods should be considered for reaching the best solution [8].

The objective of this study is to determine the most appropriate probability distribution functions for modelling of annual minimum flows for the durations 7-day of the stations located in upper part of the Euphrates River Basin, Turkey. Gamma, Generalized Extreme Value, Gumbel, Logistic, Log-Logistic, Normal, Pearson type-III, two and three parameter Lognormal and Weibull distributions are considered and compared. Five commonly used model selection methods (namely AIC, BIC, AD, KS and CvM) are used to define the most suitable distribution type.

2. Methods

2.1. Definition of annual minimum flows

The annual minimum (AM) D-day flow is the minimum average flow of D consecutive days within one year. The mostly used averaging intervals are 1, 7, 14, 30 and 60 days. An AM (D-day) can easily be computed by using a moving average of D days on a daily flow data series and subsequently selecting the minimum of the considered series. In this study, in order to meet of requirements of various management and design problems, we consider 7-day annual minimum flows (AM7).

2.2. Probability distributions and the model selection methods

In the present study, ten widely used probability distributions, which are described in Table 1, are considered. To define the most appropriate type among candidates, we employ five powerful model selection criterions (tests), namely Akaike Information Criterion (AIC), Anderson-Darling Criterion (ADC), Bayesian information criterion (BIC), Cramér–von-Mises (CvM), and Kolmogorov–Smirnov (KS) test. Detail information on computing test statistic for each method can be easily found in many statistical textbooks or published papers [7].

2.3. Quantile estimation under different return periods

Once the most suitable probability distribution for the considered data series has been defined and validated, the return period of specified events or, conversely, the quantile magnitude corresponding to a given return period can be easily computed [2]. The return period that related to the probability of non-exceedence ($F$) can be calculated by:

$$F = 1 - \frac{E(L)}{T}$$  \hspace{1cm} (1)
where, $F = F(Q)$ indicates the probability of having a data of magnitude $Q_T$ or smaller and $E(L)$ is the expected inter-arrival time. The value of $\mu$ must be equal to 1 year if annual data is considered. Then, quantile values can be easily computed by using inverse form $Q_T = \Phi(F)$

### Table 1. Description of the probability distribution functions considered in this study

| Distribution Type | Probability Distribution Function | Parameters |
|-------------------|-----------------------------------|------------|
| Weibull           | $f(x) = \frac{k x^{k-1}}{\alpha} \exp\left[-\left(-\frac{x}{\alpha}\right)^k\right]$ | $k =$ shape parameter ($k>0$)  
 $\alpha =$ scale parameter ($\alpha>0$) |
| Two Parameter Lognormal | $f(x) = \frac{1}{\sqrt{2\pi} \sigma x} \exp\left[-\frac{1}{2} \left(\log x - \mu_y\right)^2\right]$ | $\mu_y =$ shape parameter ($\mu_y>0$)  
 $\sigma_y =$ scale parameter ($\sigma_y>0$)  
 $t =$ threshold parameter |
| Three Parameter Lognormal | $f(x) = \frac{1}{(x-t)\sqrt{2\pi}\sigma_y} \exp\left[-\frac{1}{2} \left(\log(x-t) - \mu_y\right)^2\right]$ | $k =$ shape parameter ($k>0$)  
 $\alpha =$ scale parameter ($\alpha>0$)  
 $t =$ threshold parameter  
 $\Gamma =$ gamma function |
| Gamma             | $f(x) = \frac{x^{k-1}}{\alpha^k \Gamma(k)} \exp\left[-\frac{x}{\alpha}\right]$ | $k =$ shape parameter ($k>0$)  
 $\alpha =$ scale parameter ($\alpha>0$)  
 $t =$ threshold parameter  
 $\Gamma =$ gamma function |
| Pearson Type III  | $f(x) = \frac{(x-t)^{k-1}}{(x-t)\alpha^k \Gamma(k)} \exp\left[-\frac{(x-t)}{\alpha}\right]$ | $\mu =$ mean ( location parameter )  
 $\sigma =$ standard deviation ( scale parameter ) ($\sigma>0$)  
 $t =$ threshold parameter  
 $\Gamma =$ gamma function |
| Normal            | $f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma}\right)^2\right]$ | $\mu =$ mean ( location parameter )  
 $\sigma =$ standard deviation ( scale parameter ) ($\sigma>0$) |
| Gumbel            | $f(x) = \frac{1}{\sigma} \exp\left[-\frac{(x-\mu)}{\sigma} - \exp\left(-\frac{(x-\mu)}{\sigma}\right)\right]$ | $\mu =$ mean ( location parameter )  
 $\sigma =$ standard deviation ( scale parameter ) ($\sigma>0$)  
 $\xi =$ standard deviation ( scale parameter ) ($\sigma>0$) |
| Logistic          | $f(x) = \frac{1}{\alpha} \exp\left(\frac{x-\xi}{\alpha}\right) \left[1 + \exp\left(\frac{x-\xi}{\alpha}\right)\right]^{-2}$ | $\xi =$ location parameter  
 $\alpha =$ scale parameter ($\alpha>0$)  
 $k =$ shape parameter |
| Log-logistic      | $f(x) = \frac{1}{\alpha} \exp\left[-(1 + k\xi)^{-1/k}\right] \left(1 + k\xi\right)^{-1-1/k}$ | $\xi =$ location parameter  
 $\alpha =$ scale parameter ($\alpha>0$)  
 $k =$ shape parameter  
 $z =$ location parameter |
| Generalized Extreme Value | $f(x) = \frac{1}{\alpha} \exp\left[-\left(1 + k\xi\right)^{-1/k}\right] \left(1 + k\xi\right)^{-1-1/k}$ | $\xi =$ location parameter  
 $\alpha =$ scale parameter ($\alpha>0$)  
 $k =$ shape parameter  
 $z =$ location parameter |

### 3. Study area and data used

The study area Euphrates basin, which is located in southeastern Anatolia region of Turkey, is the longest river in Southwestern Asia with 2,700 km and the biggest of 26 river basins in the country. The basin has the highest mean annual streamflow in Turkey and it is divided into three sub-basins as; namely, upper, middle, and lower. To provide irrigation, hydropower, and socio-economic development in Turkey, many important dams and hydropower projects (e.g. Atatürk Dam, Karakaya Dam, Keban Dam) were constructed on the basin. In the present study, daily streamflow records from eighteen gauge stations, which are located at different parts of lower, middle and upper Euphrates were selected for the analysis. In this study, daily streamflow time series from four gauge stations, which are located in upper Euphrates, were selected (see Figure 1). Summary information on the stations and daily streamflow data is given in Table 2.
Table 2. Summary of statistics for stations and daily streamflow series

| Station No | Altitude (m) | Observation Period | Daily mean flow (m³/s) | Standard Deviation (m³/s) |
|------------|--------------|--------------------|------------------------|--------------------------|
| 2151       | 1355         | 1964-2013          | 59.59                  | 76.67                    |
| 2119       | 1123         | 1961-2013          | 81.55                  | 90.15                    |
| 2156       | 865          | 1969-2008          | 150.35                 | 117.83                   |
| 2149       | 900          | 1963-2011          | 52.94                  | 46.60                    |

Figure 1. Map of Upper Euphrates basin and location of the selected stations

Daily streamflow records obtained from these stations were used to calculate annual minimum (AM) flows for duration 7 days (AM7). Prior to fit any probability distributions to AM data series, the data’s stationary must be proven. To this end, we used the Mann-Kendall trend test and the results are given in Table 3.

Table 3. Mann-Kendall test results of AM7 series

| Station | Test Statistic (S) | The computed Z value |
|---------|--------------------|----------------------|
| 2119    | 251                | 1.814                |
| 2149    | -102               | -0.871               |
| 2156    | 117                | 1.352                |
| 2151    | 303                | 2.383                |

According the trend test results, AM7 series of station 2151 showed an increasing trend as the computed Z value is higher than the 95% confidence limits (±1.96). Therefore, this station was excluded from the study.

3.1. Defining the most suitable probability distribution for AM7 series

Gamma, Generalized Extreme Value (gev), Gumbel (gumbel), Logistic (logis), Log-Logistic (llog), Normal (norm), Pearson type-III (P3), two and three parameter Lognormal (lnorm and lnorm3) and Weibull (weibull) distributions were fitted to AM7 series of each station. Parameters of the considered distributions were estimated via Maximum Likelihood Method (MLM). In order to define the most
suitable one, AD, AIC, BIC, CvM and KS model selection criteria were used. The model selection results are presented in Table 4.

Table 4. Results of the model selection methods for the candidate probability distributions

| Station | Probability Distribution | AD   | AIC   | BIC   | CvM   | KS    | P3    | lnorm3 |
|---------|--------------------------|------|-------|-------|-------|-------|-------|--------|
| 2119    | weibull                  | 0.4839 | 355.52 | 359.53 | 0.0753 | 0.0965 | 0.2393 | 0.1476 |
|         | lnorm                    | 0.2333 | 351.24 | 355.25 | 0.0365 | 0.0712 | 0.0709 | 0.0244 |
|         | gamma                    | 0.1557 | 350.56 | 354.57 | 0.0255 | 0.0561 | 0.0873 | 0.0268 |
|         | logis                    | 0.1960 | 352.38 | 356.40 | 0.0415 | 0.0790 | 0.0862 | 0.0257 |
|         | norm                     | 0.2393 | 353.09 | 357.11 | 0.0608 | 0.0577 | 0.0873 | 0.0250 |
|         | gumbel                   | 0.3960 | 352.68 | 358.70 | 0.0265 | 0.0581 | 0.0873 | 0.0250 |
|         | gev                      | 0.1645 | 351.60 | 351.64 | 0.0257 | 0.0566 | 0.0873 | 0.0250 |
|         | llog                     | 0.1861 | 352.55 | 352.55 | 0.0244 | 0.0563 | 0.0873 | 0.0250 |

| Station | Probability Distribution | AD   | AIC   | BIC   | CvM   | KS    | P3    | lnorm3 |
|---------|--------------------------|------|-------|-------|-------|-------|-------|--------|
| 2149    | weibull                  | 1.2833 | 275.70 | 279.49 | 0.2022 | 0.1223 | 0.0729 | 0.3415 |
|         | lnorm                    | 0.7140 | 265.87 | 269.65 | 0.1120 | 0.1188 | 0.0496 | 0.4180 |
|         | gamma                    | 0.8165 | 267.25 | 271.03 | 0.1289 | 0.1249 | 0.1091 | 0.0244 |
|         | logis                    | 0.9897 | 273.31 | 277.09 | 0.1295 | 0.1133 | 0.0922 | 0.0472 |
|         | norm                     | 1.0910 | 271.23 | 275.02 | 0.1744 | 0.1348 | 0.0780 | 0.0646 |
|         | gumbel                   | 0.5927 | 263.50 | 267.28 | 0.0922 | 0.0834 | 0.1011 | 0.0696 |
|         | gev                      | 0.7291 | 264.97 | 270.64 | 0.0780 | 0.1059 | 0.0834 | 0.0780 |
|         | llog                     | 0.7291 | 269.23 | 273.01 | 0.0873 | 0.1059 | 0.0834 | 0.0780 |
|         | P3                       | 0.3415 | 269.03 | 273.01 | 0.1011 | 0.1059 | 0.0834 | 0.0780 |

| Station | Probability Distribution | AD   | AIC   | BIC   | CvM   | KS    | P3    | lnorm3 |
|---------|--------------------------|------|-------|-------|-------|-------|-------|--------|
| 2156    | weibull                  | 0.6208 | 310.81 | 314.19 | 0.0903 | 0.1131 | 0.2033 | 0.1817 |
|         | lnorm                    | 0.1982 | 304.01 | 307.39 | 0.0297 | 0.0611 | 0.3333 | 0.2497 |
|         | gamma                    | 0.2277 | 304.51 | 307.88 | 0.0344 | 0.0675 | 0.3344 | 0.1844 |
|         | logis                    | 0.3333 | 307.67 | 311.05 | 0.0434 | 0.0691 | 0.3344 | 0.1823 |
|         | norm                     | 0.3444 | 306.19 | 309.57 | 0.0513 | 0.0861 | 0.3444 | 0.2033 |
|         | gumbel                   | 0.2033 | 303.77 | 307.14 | 0.0302 | 0.0830 | 0.2033 | 0.1844 |
|         | gev                      | 0.2497 | 305.39 | 310.46 | 0.0270 | 0.0609 | 0.0355 | 0.1844 |
|         | llog                     | 0.1844 | 306.20 | 309.58 | 0.0355 | 0.0627 | 0.0355 | 0.1823 |
|         | P3                       | 0.1823 | 304.68 | 309.75 | 0.0270 | 0.0814 | 0.0355 | 0.1823 |
|         | lnorm3                   | 0.1844 | 305.32 | 310.39 | 0.0736 | 0.0814 | 0.0355 | 0.1823 |

**The most suitable distributions are presented in bold.

From the results, it can be seen that the gamma and three parameter lognormal distributions were found to be the most suitable models for AM7 series of the station 2119. Pearson type-III distribution were the best fitted model for AM7 series of the station 2149 while GEV, Gumbel and two parameter lognormal distributions provided the best fit for AM7 series of station 2156. For visual evaluation, probability density function (PDF) and cumulative distribution function (CDF) of the selected most suitable distributions against the empirical PDF and CDFs were plotted in Figure 2. It is clearly seen from the figure that the selected distributions provided a good agreement with empirical data.

3.2. Low flow quantile estimation

Having determined the most suitable probability distributions for AM7 series, it can be easily used to calculate quantile estimates for different return periods. To achieve this aim, the selected distributions were applied to calculate AM7 series for the return periods 10, 20, 50, 100, 200 and 500 years (Table 5).
Figure 2. Empirical and theoretical PDF and CDF plots of the selected distributions

| Station No | Best fitted distribution | 10  | 20  | 50  | 100 | 200 | 500 |
|------------|--------------------------|-----|-----|-----|-----|-----|-----|
| 2119       | Gamma                    | 16.59 | 14.99 | 13.33 | 12.29 | 11.40 | 10.37 |
|            | Lnorm3                   | 16.60 | 14.96 | 13.22 | 12.12 | 11.16 | 10.04 |
| 2149       | Pearson type-III         | 14.37 | 13.96 | 13.66 | 13.53 | 13.45 | 13.39 |
| 2156       | GEV                      | 57.29 | 54.67 | 51.95 | 50.26 | 48.78 | 47.07 |
|            | Gumbel                   | 57.38 | 55.05 | 52.69 | 51.25 | 50.01 | 48.60 |
4. Conclusion

This study aimed to define the most appropriate probability distribution functions for modelling of 7-day annual minimum flows (AM7) of four stations (2119, 2149, 2156 and 2151) located in rivers of the upper Euphrates basin, Turkey. To achieve this aim, Gamma, Gumbel, Generalized Extreme Value, two parameter Lognormal, three parameter Lognormal, Logistic, Log-Logistic, Normal, Pearson type-III and Weibull distributions were considered and compared. To determine the most suitable distributions, AD, AIC, BIC, CvM, KS tests were used. According to the results of these tests, Gamma and three parameter Lognormal distributions provided the most suitable fit to AM7 series of the station 2119. Pearson type-III distribution showed better performance than other candidates for modelling AM7 series of the station 2146. Moreover, AM7 series of the station 2156 were well described by GEV and Gumbel distributions. Finally, the most suitable distributions were applied to calculate AM7 quantiles for the various return periods (10, 20, 50, 100, 200 and 500).

Acknowledgments

The authors sincerely thank the General Directorate of State Hydraulic Works, Turkey for providing the daily streamflow records used in this study.

References

[1] Kroll C N and Vogel R M 2002 Probability distribution of low streamflow series in the United States. J Hydrol Eng. 7(2), pp 137-46.
[2] Meylan P Favre A C and Musy A 2012 Predictive Hydrology: A Frequency Analysis Approach: CRC Press.
[3] Zaidman M D Keller V Young A R and Cadman D 2003 Flow-duration-frequency behaviour of British rivers based on annual minima data. J Hydrol. 277(3-4), pp 195-213.
[4] Yue S and Pilon P 2005 Probability distribution type of Canadian annual minimum streamflow. Hydrolog Sci J. 50(3) pp 427-38.
[5] Chen Y D Huang G Shao Q X and Xu C Y 2006 Regional analysis of low flow using L-moments for Dongjiang basin, South China. Hydrolog Sci J. 51(6) pp 1051-64.
[6] Keshtkar A R 2015 Low flow frequency analysis by L-moments method (Case study: Iranian Central Plateau River Basin). Desert 20(2) pp 167-75.
[7] Eris E Aksoy H Onoz B Cetin M Yuce I M Selek B Aksu H Burgan I H Esit M Ece Y I Karakus U 2018 Frequency analysis of low flows in intermittent and non-intermittent rivers from hydrological basins in Turkey. Water Supply 19(1) pp 30-39.
[8] Tosunoglu F 2018 Accurate estimation of T year extreme wind speeds by considering different model selection criterions and different parameter estimation methods. Energy 162 pp 813-24.