

TESTING THE GOODNESS OF FIT OF PROBABILITY FUNCTIONS BY INFORMATIONAL ENTROPY IN THE CASE OF GEDIZ RIVER BASIN

Türkay BARAN¹, Filiz BARBAROS^{2*}, Ali GÜL³, Özgür BOZOĞLU⁴

Dokuz Eylül University, Faculty of Engineering, Department of Civil Engineering, Tinaztepe Campus, Buca, Izmir, Turkey, Tel: +90 (232) 301 7056, ¹*turkay.baran@deu.edu.tr*; ²*filiz.barbaros@deu.edu.tr*; ³*ali.gul@deu.edu.tr*; ⁴*zgr_tmz@hotmail.com*; * *Corresponding author*

Abstract

In water systems planning, it is essential to use an objective criterion to determine the information content of hydrological data. The Entropy concept, defined by Shannon in information theory, has been applied in hydrology and water resources for measuring the information content of hydrologic processes. The presented study aims to determine the best fitting probability distribution function for the observed time series of precipitation data. There are numerous of tests to evaluate the goodness of fit of probability functions such as Chi-squared, Kolmogorov Smirnov and Anderson Darling tests. An alternative method for the mentioned tests is the Informational Entropy method. Informational Entropy method has been studied for the precipitation data of the selected Precipitation Gaging Stations of Turkey. Results showed that Informational Entropy method can be a valuable alternative method to test the goodness of fit. For a further step, in the presented study, it is studied for a selected case area and execution is done for the 22 gaging stations of State Hydraulic Works (DSI), which have long-term precipitation data located in the Gediz River Basin. Results by testing the goodness of fit of probability functions by Informational Entropy method show that Informational Entropy can be applied for fitting the probability function according to the investigated datasets.

Keywords: Informational Entropy, Gediz Basin, Probability Distribution.

1 INTRODUCTION

In the management of water resources, it has become a necessity to use water resources efficiently as a result of several researches. Getting necessary and quick information about water resources is become a fundamental issue for the researches and implementation of more useful and easily applicable methods. The entropy concept, defined by Shannon in information theory (Shannon and Weaver 1949), is a powerful tool for maximizing information gain monitored time-series processes and is applied in hydrology and water resources for measuring the information content of hydrologic processes to a similar end. Informational Entropy method has been previously studied for the precipitation data of the selected precipitation gaging stations of Turkey. In a similar manner, the approach brings in potential benefits in dealing with data at a much smaller scale but with increased level of detail.

Turkey has several river basin systems that are suitable for practicing the entropy logic. Gediz River basin, which is located in West Anatolia between the Aegean Sea and Küçük Menderes and Bakırçay river basins, is a good example of such systems for executing the approach. The basin is the largest in the Aegean region, covering 2.3% of the total surface area of Turkey. The length of the river is 276 km and the main tributaries are Deliiniş, Selendi, Demirci, Nif, Alaşehir and Kumçay stream (Baran and Barbaros 2015). The downstream reaches of the Gediz River remain within the borders of the Metropolitan Municipality of Izmir before its discharge into Izmir Bay as given in Figure 1 (Barbaros and Harmancioglu 2013). The drainage area of the river is 16775 km². The basin is engineered into extensive water resources systems, the major one being irrigation over 110,000 ha of agricultural land. The basin is in the typical Mediterranean climate. Summers are hot and dry; winters are warm and rainy. January and February are the rainiest months whereas July and August are the driest months. The annual average temperature is 15.6°C. The Gediz River Basin along the Aegean coast of Turkey is a typical case where two major problems, water scarcity and pollution, need to be addressed for sustainable management of its water resource (Baran and Barbaros 2015). Besides water scarcity and pollution of surface and ground waters, secondary problem is the periodically recurring droughts. While water scarcity can be attributed to natural hydrological conditions (e.g., droughts), increases in the number and types of water demands and significant competition for water among users better explain the current status of the basin. There are; however, serious institutional, legal, social and economic drawbacks, which enhance water allocation and water pollution problems. Water scarcity contributes more to these problems, and is basically due to competition for water among various uses and water pollution. It is not, however, a mature basin, in the

sense that the institutional set-up is not yet fully developed. Both surface and groundwater use are largely unregulated, and groundwater extraction is growing rapidly in response to urban and industrial demand (Barbaros and Harmancioglu 2013).

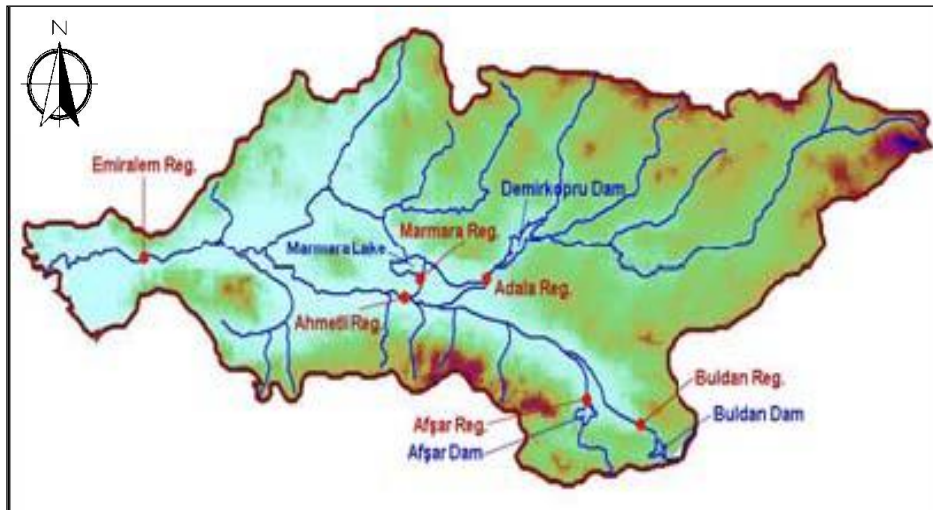


Figure 1. Gediz Basin and its boundaries.

All meteorological data are provided from State Water Works - DSI. In the presented study, total monthly precipitation data of 22 DSI Gaging Station in Gediz River Basin are investigated. Precipitation gauging stations map in Gediz Basin is given Figure 2. The observed period of Precipitation Gauging Stations - PGS is from 1961 to 2005.

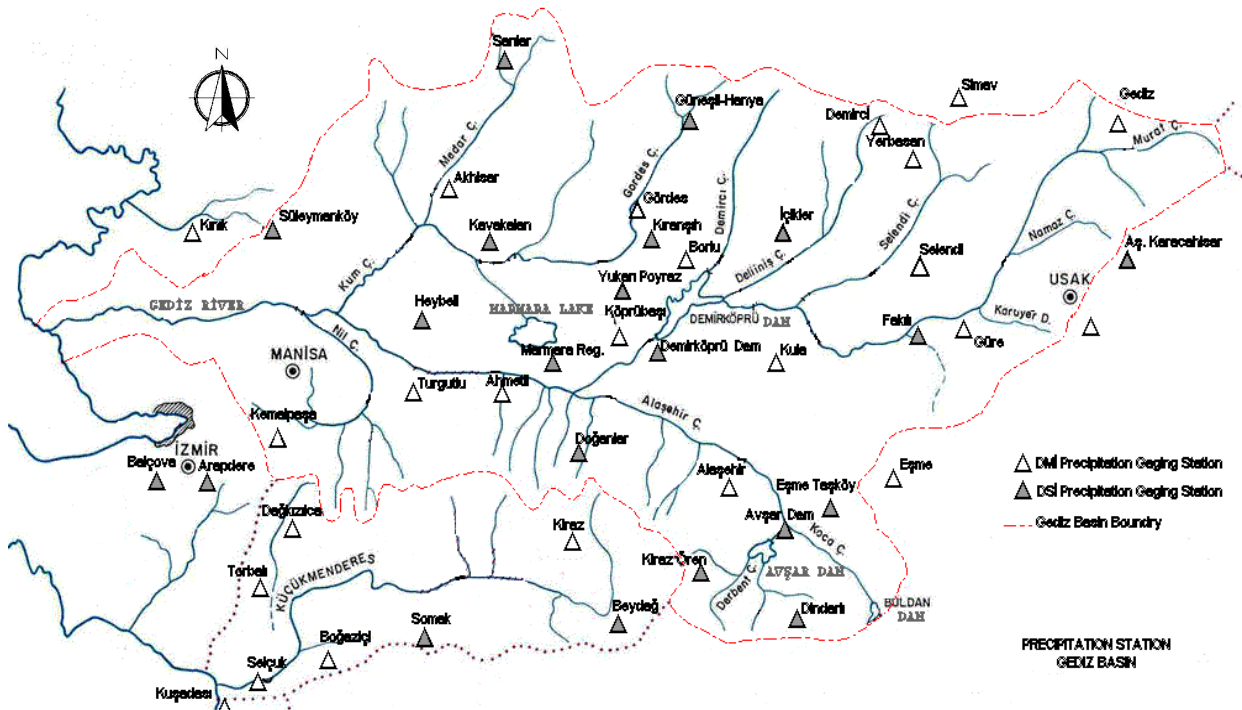


Figure 2. Precipitation gauging stations in Gediz Basin (Baran and Barbaros 2015).

The presented study tests the best fit of probability distribution function of precipitation data throughout the Gediz River Basin. For this purpose, 22 gauging stations having long-term precipitation data are investigated which located in the basin. The best fitting probability distribution is investigated by Informational Entropy method.

2 INFORMATIONAL ENTROPY METHOD

Informational entropy concept, which indirectly helps measure the information content of a given data series based on the rationale that any reduction of uncertainty through observations is basically equal to the amount of information gained, is a dedicated measure for the degree of uncertainty of a random hydrologic process in this respect. Entropy is defined as *the variation of information content instead of reduction of uncertainty*, which is in the end equivalent to *the amount of information gained*, thus it provides the opportunity for extracting results that are more reliable in a way to allow effective use of the entropy concept in solving water resources engineering problems related to information content and uncertainty (Bozoglu and Baran 2010, 2012; Baran et al., 2017a,b).

In introducing the informational entropy concept, Shannon (1949) considered in much general sense that average information content of any data source holds bigger significance than any derived description represented by each corresponding symbol. In such a case, Shannon defines the information content, $H(n)$, of a message sent by the transmitter as: was named as entropy (Shannon and Weaver, 1949) as

$$H(n) = -\sum p_n \log p_n \quad (1)$$

with units as (bit/symbol), based on the consideration that the source is described by independent and discrete signs/symbols, each associated with probabilities p_n ($n=1, \dots, N$). Later, the information content as in Shannon's entropy definition is simulated to the entropy function described in statistical mechanics (Cherry 1957; Pierce 1961; Pfeiffer 1965).

The design and operation of water resources systems require thorough understanding and thus monitoring of hydrologic processes to make optimum decisions. These data collection practices require keeping the logic of extracting the identity, location, timing and duration of the monitored process (Baran, 1993). Then, water resources planners started using the terms *expected information*, *increase of information*, or *deficiency of information* to relate design parameters to the information content conveyed by monitored data. Information has often been expressed indirectly by selected statistical parameters that included variance, standard error or correlation coefficient instead of quantitative terms (Harmancioglu and Singh 1998).

The main difficulty associated with the applicability of the entropy concept in hydrology arises from the lack of a precise definition in dealing with continuous variables. A new definition describes the concept as a measure of *variation of information* rather than an absolute measure of information. In searching for the *Variation of Information* [$H(X/X^*)$], basic statistical parameters are calculated and then the maximum entropy and the range (R) of each data set are calculated by Eqs. (2) and (3), respectively (Baran 1993):

$$H_{\max} = \log R \quad (2)$$

$$R = b - a \quad a < x < b \quad (3)$$

Marginal entropy is calculated by using the Eq. (4) and then informational entropy $H(X/X^*)$ is calculated by the Eq. (5).

$$H(X) = \log \sqrt{2\pi} + \log \sigma + 1/2 \quad (4)$$

$$H(X/X^*) = H(X) - H_{\max} \quad (5)$$

For the half range value, the acceptable entropy value for normal probability density function can be obtained as in Eq. (6) by using normal logarithms and replacing the appropriate values in Eq. (4):

$$H(X/X^*)_{cr} = 0.6605 \quad (6)$$

When the entropy $H(X/X^*)$ of the variable, which is assumed to be normal, remains below the above value, the normal probability density function can be considered acceptable to indicate that sufficient amount of information has been collected about the process (Baran 1993; Baran and Barbaros 2015).

If the a-posteriori distribution function is treated to be lognormal $LN(\mu_y, \sigma_y)$, the variation of information for the variable x can be determined as in Eq. (7). In such a case, it may be considered that the $-a$ term equals 0, since lognormal values will be all positive. Then the acceptable value of $H(X/X^*)$ for the lognormal distribution function will be as defined in Eq. (8):

$$H(X/X^*) = \log[2\text{Sinh}(a\sigma_y)] - \log \sigma_y - 1.4189 \quad (7)$$

$$H(X/X^*) = a\sigma_y - \log \sigma_y - 1.4189 \quad (8)$$

No single constant value exists to describe the confidence limit for lognormal distribution. Even when critical halfrange is determined, the confidence limits will vary according to the variance of the variable. In the least case of the known variance x , it becomes possible to compute the confidence limits (Baran 1993; Bozoglu and Baran 2010; Baran and Bacanlı 2006, 2007 a,b; Baran et al., 2017c).

3 ANALYSIS & RESULTS

In the presented study, total monthly precipitation data is analysed in terms of Informational Entropy in detail using the available data sets for 22 districts throughout the Gediz River Basin. Observation period and the cumulative executed statistical parameters for each station is given in Table 1. Analysis is done for 22 stations and the results of execution for normal and log-normal distributions are given respectively in Tables 2 & 3. In both distributions, informational entropy method results are accepted to be used to test the goodness of fit. Acceptability of the test can also be seen in Figures through 3 and 7 for the selected 5 stations of 22 total stations in the basin, where informational entropy values are below the critical values for both normal and log-normal series.

Table 1. Observation period and the cumulative executed statistical parameters.

Station Name	Observation Period	Mean	Std. Deviation	Skewness	Excess
Avşar Dam	1980-2005	35.957	35.6891	1.2871	1.4343
Beşyol	1976-2005	64.167	75.6427	1.7278	3.4639
Bozdağ	1961-2005	104.526	121.0376	1.9606	5.5397
Buldan Dam	1967-2005	39.005	37.5777	1.3424	1.9004
Çınardibi	1961-2002	78.168	90.4259	1.6368	3.2308
Demirköprü Dam	1962-1989	41.275	31.4090	1.1837	1.0299
Dindarlı	1962-2005	36.828	36.8438	1.4307	2.3572
Doğanlar	1970-2005	52.182	57.1638	1.8497	4.4735
Eşmataşköyü	1962-2005	38.404	36.1632	1.2619	1.7676
Fakılı	1962-2005	37.301	33.2821	1.1479	1.1183
Göynükören	1966-2003	38.211	36.8309	1.3356	2.0887
Hacırahmanlı	1961-1997	40.187	44.5202	1.6355	3.7035
Hanya	1961-1995	52.294	56.5573	1.5874	2.7962
İçikler	1961-2005	47.384	45.9726	1.2977	1.7176
Kavakalan	1962-1998	52.045	57.1394	1.5938	3.1945
Kıraşlı	1962-2005	49.049	48.9309	1.2576	1.3681
Marmara Lake Regulator	1961-2001	35.839	36.6187	1.2075	1.1557
Ören	1961-2005	60.131	67.5776	1.6197	2.9071
Sarılar	1962-2005	47.710	54.7744	1.6711	3.6113
Süleymanköy	1962-1997	40.449	43.9724	1.6377	3.4330
Üçpınar	1961-2005	45.576	54.2819	1.7268	3.4824
Yukarı Poyraz	1962-2003	49.105	49.2131	1.3048	1.6311

Table 2. Informational Entropy Analysis Results for Normal Distribution.

Station Name	Range	H _{max}	H _x	H(XIX')	H _{cr}	Test
Avşar Dam	176,790	5,1750	4,9938	0,1812	0,6605	Accepted
Beşyol	409,690	6,0154	5,7450	0,2704	0,6605	Accepted
Bozdağ	892,390	6,7939	6,2150	0,5789	0,6605	Accepted
Buldan Dam	195,190	5,2740	5,0454	0,2286	0,6605	Accepted
Çınardibi	579,490	6,3622	5,9235	0,4387	0,6605	Accepted
Demirköprü Dam	188,090	5,2369	5,1414	0,0955	0,6605	Accepted
Dindarlı	226,590	5,4231	5,0256	0,3975	0,6605	Accepted
Doğanlar	363,390	5,8955	5,4649	0,4306	0,6605	Accepted
Eşmataşköyü	208,190	5,3385	5,0070	0,3315	0,6605	Accepted
Fakılı	159,890	5,0750	4,9240	0,1505	0,6605	Accepted
Göynükören	214,490	5,3683	5,0253	0,3430	0,6605	Accepted
Hacırahmanlı	296,890	5,6934	5,2149	0,4785	0,6605	Accepted
Hanya	315,990	5,7557	5,4542	0,3015	0,6605	Accepted
İçikler	242,590	5,4914	5,2470	0,2444	0,6605	Accepted
Kavakalan	368,990	5,9108	5,4644	0,4463	0,6605	Accepted
Kıranşih	254,590	5,5397	5,3093	0,2303	0,6605	Accepted
Marmara Lake Regulator	166,590	5,1155	5,0195	0,0960	0,6605	Accepted
Ören	391,190	5,9692	5,6322	0,3370	0,6605	Accepted
Sarılar	362,890	5,8941	5,4222	0,4719	0,6605	Accepted
Süleymanköy	273,890	5,6127	5,2025	0,4102	0,6605	Accepted
Üçpınar	333,090	5,8084	5,4131	0,3953	0,6605	Accepted
Yukarı Poyraz	274,890	5,6164	5,3151	0,3013	0,6605	Accepted

Table 3. Informational Entropy Analysis Results for Log-Normal Distribution.

Station Name	H _{max}	H _x	H(XIX')	H _{cr}	Test
Avşar Dam	5,1750	3,5879	1,5870	2,0663	Accepted
Beşyol	2,3628	1,0912	1,2716	2,0917	Accepted
Bozdağ	2,4335	1,0818	1,3517	2,0875	Accepted
Buldan Dam	5,2740	3,6405	1,6335	2,0654	Accepted
Çınardibi	6,3621	4,4967	1,8654	2,0872	Accepted
Demirköprü Dam	5,2369	3,7350	1,5019	2,0669	Accepted
Dindarlı	5,4231	3,6194	1,8038	2,0668	Accepted
Doğanlar	5,8955	4,0486	1,8468	2,0767	Accepted
Eşmataşköyü	5,3385	3,6021	1,7363	2,0653	Accepted
Fakılı	5,0745	3,5169	1,5576	2,0675	Accepted
Göynükören	5,3683	3,6204	1,7479	2,0654	Accepted
Hacırahmanlı	5,6934	3,7968	1,8966	2,0786	Accepted
Hanya	5,7557	4,0400	1,7158	2,0747	Accepted
İçikler	5,4914	3,8420	1,6494	2,0655	Accepted
Kavakalan	5,9108	4,0479	1,8629	2,0771	Accepted
Kıranşih	5,5397	3,9032	1,6364	2,0666	Accepted
Marmara Lake Regulator	5,1155	3,6118	1,5038	2,0682	Accepted
Ören	5,9692	4,2115	1,7577	2,0812	Accepted
Sarılar	5,8941	3,9971	1,8970	2,0856	Accepted
Süleymanköy	5,6127	3,7875	1,8252	2,0755	Accepted
Üçpınar	5,8084	3,9793	1,8291	2,0943	Accepted
Yukarı Poyraz	5,6164	3,9087	1,7076	2,0669	Accepted

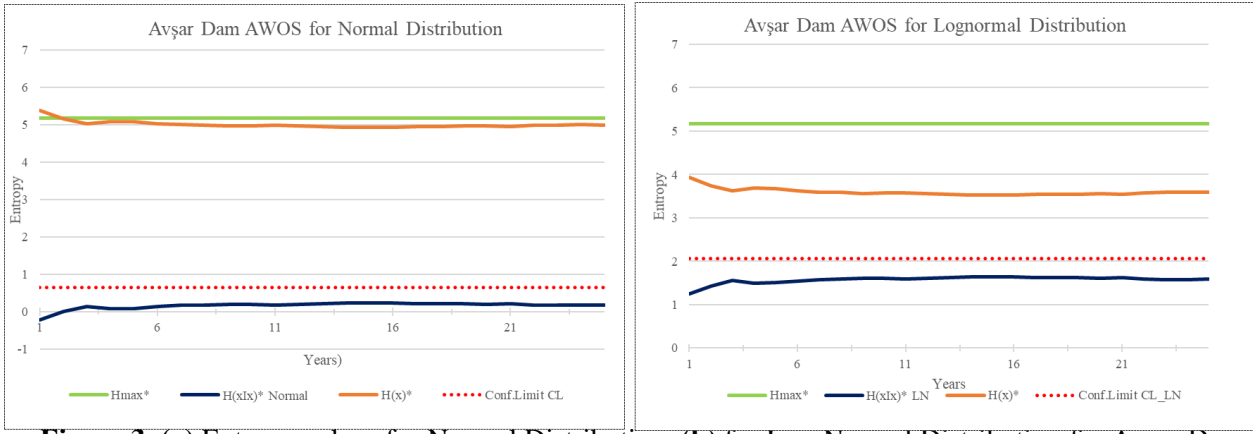


Figure 3. (a) Entropy values for Normal Distribution; (b) for Log-Normal Distribution for Avşar Dam.

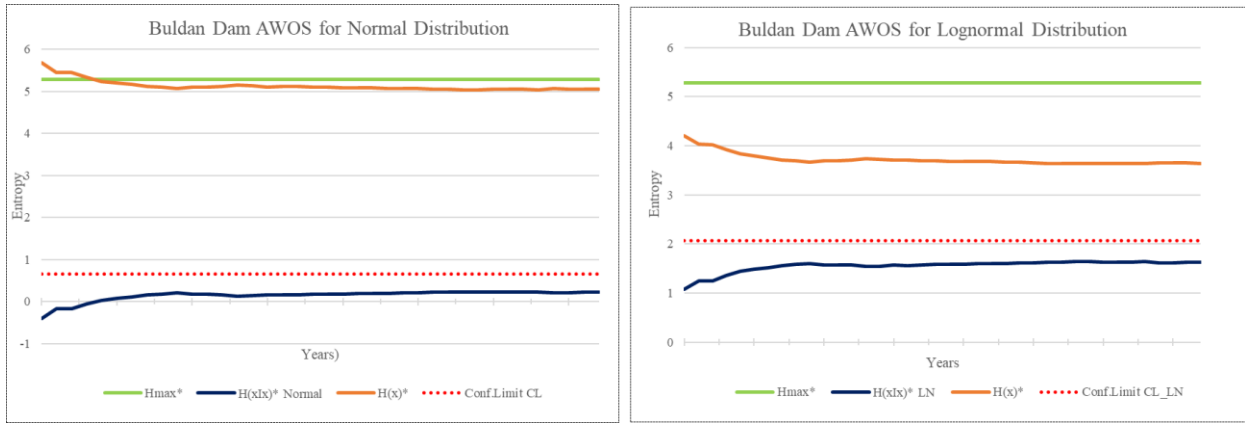


Figure 4. (a) Entropy values for Normal Distribution; (b) for Log-Normal Distribution for Buldan Dam.

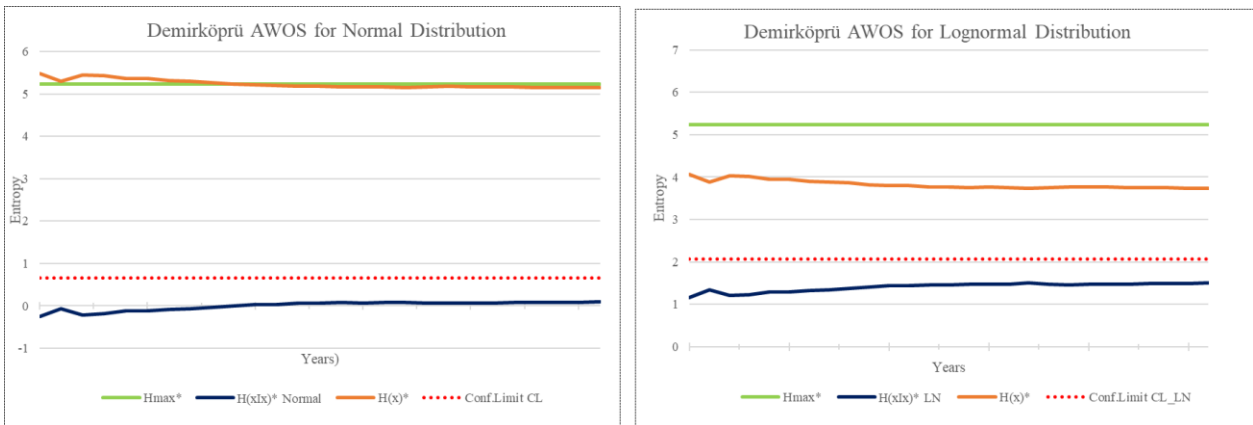


Figure 5. (a) Entropy values for Normal Distribution; (b) for Log-Normal Distribution for Demirköprü Dam.

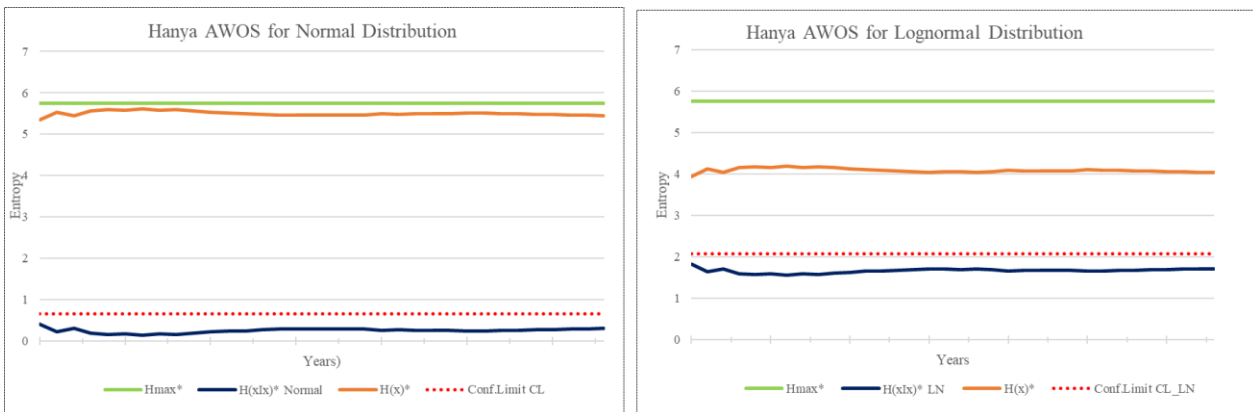


Figure 6. (a) Entropy values for Normal Distribution; (b) for Log-Normal Distribution for Hanya.

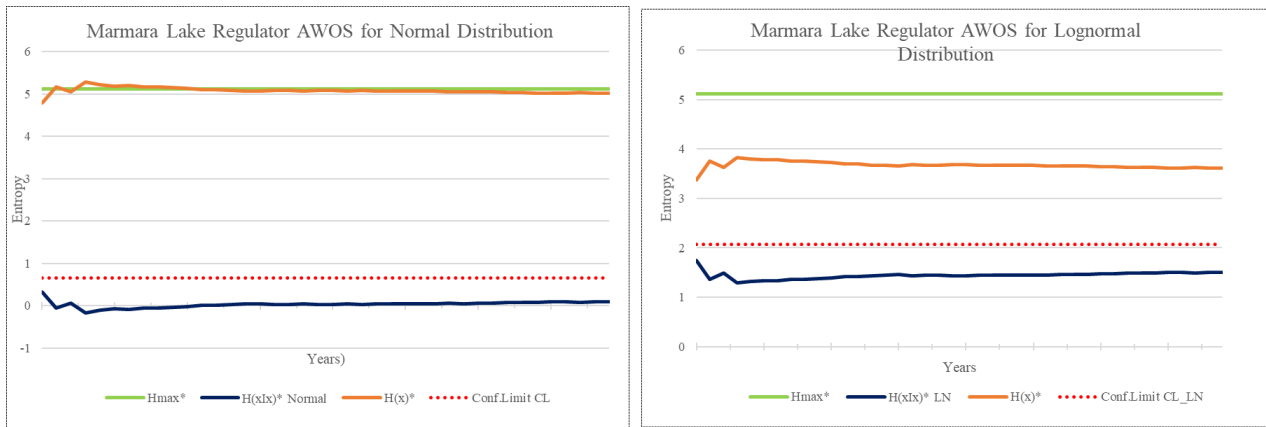


Figure 7. (a) Entropy values for Normal Distribution; **(b)** for Log-Normal Distribution for Marmara Lake Regulator.

4 CONCLUSIONS

In the presented study, the best fitting probability distribution is investigated by the Informational Entropy method and it is determined that both normal and log-normal distributions are suitable for all investigated Precipitation Gaging Stations – PGSs throughout the Gediz River Basin. Similar results are also obtained by the conventional tests such as Chi-Squared, Kolmogorov Smirnov and Anderson Darling methods. The similar testing studies previously carried out for the PGSs throughout the country and precipitation data of 60 PGSs of Turkey has been studied to prove the availability to use the Informational Entropy method to test the goodness of fit. As the execution results obtained for the selected PGSs of Turkey, the execution for PGSs of Gediz River Basin gave the same results. As the presented application of Informational Entropy method stated as one of the effective tools to evaluate hydrological data besides the other testing methods, in case of acceptance for both normal and log-normal distributions, furthermore studies should be carried out to fix the best fitting distribution.

REFERENCES

- Baran, T. (1993). Hidrolojik Süreçlerin Bilgi İçeriğindeki Değişim Miktarı olarak Entropi Tanımı [*Entropy Definition as the amount of Change in Information of Hydrological Processes*], Dokuz Eylül University, The Graduate School of Natural and Applied Sciences, Department of Civil Engineering, Department of Hydraulic – Hydrology and Water Resources, Ph. D. Thesis (Advisor: N. Harmancioglu), 153 p. [in Turkish].
- Baran, T., Bacanlı, Ü.G. (2006). Evaluation of Suitability Criteria in Stochastic Modeling, *European Water*, **13-14**, pp. 35-43.
- Baran, T., Bacanlı, Ü.G. (2007a). An Entropy Approach for Diagnostic Checking in Time Series Analysis, *WaterSA*, **33**, No. 4, pp. 487-496, ISSN 0378-4738.
- Baran, T., Bacanlı, Ü.G. (2007b). Evaluation of Goodness of Fit Criterion in Time Series Analysis, *Digest 2006*, V.17, pp. 1089-1102.
- Baran T., Barbaros F. (2015). Testing the Goodness of Fit by Informational Entropy, In Nilgun Harmancioglu (ed.) *Conference Proceedings European Water Resources Association 9th World Congress Water Resources Management in a Changing World: Challenges and Opportunities, 10-13 June 2015, Istanbul, Turkey.*
- Baran, T., Harmancioglu, N. B., Cetinkaya, C. P., Barbaros, F. (2017a). An Extension to the Revised Approach in the Assessment of Informational Entropy, *Entropy 2017*, **19 (12)**, 634, DOI: 10.3390/e19120634
- Baran, T., Barbaros, F., Gül, A., Onusluel Gül, G. (2017b). An Informational Entropy Application To Test the Goodness of Fit of Probability Functions, *European Water*, 2017, 1792-085X, 3, **59**, 39-44.
- Baran, T., Barbaros, F., Gül, A., Onusluel Gül, G. (2017c). An Informational Entropy Application to Test the Goodness of Fit of Probability Functions, pp. 403 – 408, In George Tsakiris, Vassilios A. Tsihrintzis,

- Harris Vangelis, Dimitris Tigkas (eds.) *Conference Proceedings 10th World Congress of EWRA "Panta Rhei"*, 5-9 July 2017, Athens Greece, 2111 p.
- Barbaros, F, Harmancioglu, N. (2013). Assessment of Water Quality in a Mediterranean Basin: The Case of the Gediz Basin in Turkey, pp. 507-525, In Rodrigo Maia, António Guerreiro de Brito, Abílio Seca Teixeira, José Tentúgal Valente, João Pedro Pêgo (eds.) *Conference Proceedings 8th International Conference of EWRA, 26-29 June 2013, Porto Portugal*, 1466 p.
- Bozoglu, Ö. T, Baran, T. (2010). Gediz havzası örneğinde beklenen aylık toplam yağışların entropi yöntemiyle tayini [*Determination of Expected Monthly Total Precipitation Data by Entropy Method in the case of Gediz Basin*], 6th National Hydrology Congress, Denizli [in Turkish].
- Bozoglu, Ö. T., Baran, T. (2012). Determination of Expected Value for Monthly Total Precipitation by Entropy Based Method, 10th International Congress on Advances in Civil Engineering, Ankara.
- Cherry, C. (1957). *On Human Communication: A Review, A Survey and A Criticism*. Massachusetts, the Technology Press of Massachusetts Institute of Technology, 333 p, ISBN-13: 978-0262530385.
- Harmancioglu, N, Singh, V. P. (1998). Entropy in Environmental and Water Resources. In: Herschy R. W. and Fairbridge R. W. (eds) *Encyclopedia of Hydrology and Water Resources*, Kluwer, Dordrecht, Netherlands, pp. 225-241, ISBN: 978-1-4020-4497-7.
- Pierce, J. R. (1961). *Symbols, Signals and Noise: The Nature and Process of Communication*. New York, Harper and Row Publisher INC, ISBN-13: 978-0061392320.
- Pfeiffer, P. E. (1965). *Concept of Probability Theory*. New York, McGraw-Hill Book Company, 399 p, ISBN-13: 978-0486636771.
- Shannon, C.E., Weaver, W. (1949). *The mathematical theory of communication*. Urbana, University of Illinois Press, ISBN: 9780252725487 (Release year 1998).