Effectiveness of log-logistic distribution to model water-consumption data

Seevali Surendran[1,2]* and Kiran Tota-Maharaj[1]

[1]Faculty of Environment and Technology, University of the West of England, Bristol (UWE Bristol), Bristol, BS16 1QY, UK; [2]Environment Agency, Kings Meadow House, Kings Meadow Road, Reading, RG1 8DQ, UK. *Corresponding author Email: seevali.surendran@environment-agency.gov.uk

ABSTRACT

Water consumption varies with time of use, season and socio-economic status of consumers, and is defined as a continuous random variable. Incorporating probabilistic nature in water-consumption modelling will lead to more realistic assessments of performance of water distribution systems. Furthermore, fitting water-consumption patterns into a suitable statistical distribution will assist in determining how often peaks will occur, or the probability of exceeding the peaking factor in a system, for incorporation into design calculations. There are few studies in the literature where the random variations of consumption have been considered. The purpose of this study is to evaluate real water-consumption data from the United Kingdom (UK) and North America and to investigate the possibility of establishing a standard probability distribution function to apply in simulating water consumption in developed countries. Daily water-consumption data for five years (2009–2013) were obtained from water companies in the UK and North America and analysed by fitting into normal, log-normal, log-logistic and Weibull distributions. Statistical modelling was performed using MINITAB version 18 statistical package. The Anderson-Darling goodness-of-fit test was used to show how well the selected statistical distribution fits the water-consumption data.

KEYWORDS | Anderson-Darling statistical test, log logistic distribution, MINITAB, probability distribution function, random nature, water demand

INTRODUCTION

A major unresolved problem in water-consumption modelling is the identification of an appropriate statistical distribution which best represents the water-consumption pattern. Fitting water-consumption patterns into a suitable statistical distribution will assist in finding how often the peaks will occur, or the probability of exceeding the peaking factor, in the system to incorporate into design calculations with a scientifically proven method. The aim of this research is to study real water-consumption data and to find a standard statistical distribution to use in water-consumption modelling to address the probabilistic nature of water consumption. The advantage of modelling real water-consumption data is that it will permit forecasting of the probability of occurrence in any consumption value and provide confidence in future projections.

There are relatively few studies that have considered the random variations of water consumption. It is often assumed that variation in water consumption in distribution systems follows the normal distribution, usually with insufficient justification. Furthermore, there is inadequate reliable data regarding the suitability of various statistical distributions for modelling water consumption. Goulter and Bouchart (1990), Bao and Mays (1990), Xu and Goulter (1997, 1998, 1999), Syntetos et al. (2001, 2005) and Kwietniewski (2003) made assumptions that water consumption has a normal distribution. Mays (1994) used randomly generated water-consumption data using a range of distributions to study the sensitivity of a system's performance to changes in water-consumption patterns. Khomsi et al. (1996) stated that the consumption of water has a normal distribution based on the Kolmogorov-Smirnov test (KS). However, the KS test is more sensitive near the centre of the distribution than at the tails and was not suitable to validate the water-consumption data, as the high consumption data points lie on the tail of the distribution. In the technical literature further research papers written by De Marinis et al. (2007); Tricarico et al. (2007) and Gato-Trinidad and Gan (2012) support the effectiveness of the normal distribution by means of rigorous statistical inferences on real data.

The American Water Works Association (AWWA) Research Foundation sponsored a study (Bowen et al., 1993) in residential water-use patterns in the USA, and results revealed that the demand data was not distributed normally. Several data transformations to improve the data analysis were investigated and it was found that the log transformation was only mildly effective in reducing the positive skewness of the frequency distributions of the data.

Surendran and Tanyimboh (2004), and Tanyimboh et al. (2004) addressed the issue of the modelling of short-term consumption variations in a comprehensive way, using UK water-consumption data and concluded that data fitted better with a long tail distribution rather than a normal distribution. However, the findings were limited to UK water-consumption data.

The log-logistic distribution resembles the log-normal in shape, it has a more tractable form and is one of the few distributions for which the probability distribution, cumulative density and quartile functions exist in simple closed form (Kleiber, 2004). Furthermore, it can cope well with outliers in the upper tail (Dey & Kundu 2004). The log logistic distribution has been used by Swamee (2002), El-Saidi et al. (1990) and Rowinski et al. (2001), in hydrological studies (frequency analysis of multi-year drought durations, precipitation data and flood frequency analysis) and survival (reliability) analysis, which have the outliers in upper tail. Gargano et al. (2016, 2017) stated that log-logistic distribution is the best fit for real water-consumption data.

Ashkar and Mahdi (2006), Cordeiro et al. (2012) and Ramos et al. (2013) described the log-logistic distribution in detail and concluded that the log-logistic model is suitable for positive skewed data and positive random variables. As water consumption is a random variable and is positively skewed, it defines the suitability of log-logistic distribution in modelling water consumption.

Identifying a suitable statistical distribution

It is a general assumption that water consumption follows a normal distribution and literature review shows that the studies undertaken in the past supported this assumption. Water consumption will vary as a result of weather patterns, fire incidents and leakages, and these scenarios will lead to extreme usage conditions. Consequently, due to high consumption from time to time, the data would fit better in a positively skewed distribution than in a normal distribution

As a preliminary check, to identify the distribution patterns of the real water-consumption data obtained from the two water companies, normal graphs were drawn to check the normality of the data. The normal graphs were drawn for the 20 data sets (yearly data) and they show that out of 20 data sets, 2 sets follow a normal distribution and the other 18 sets follow a positively skewed distribution.

It can be concluded that water-consumption data will fit well into positively skewed distributions such as log-normal, log-logistic and Weibull distributions.

This study used these distributions to select the best fit for water-consumption modelling. The normal distribution was used for comparison. These four distributions are described in Table 1, providing their suitability on application to modelling water consumption.

Table 1. Suitability for using normal, log-normal, log-logistic and Weilbull distributions in modelling

Distribution	Description	Suitability for water-consumption data modelling	Probability distribution function (PDF)
Normal	Has the familiar symmetrical bell shape and its sample space extends from minus to plus infinity.	The water-consumption data contains only positive values and since the data typically show skewed frequency patterns, the	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad -\infty<x<\infty$ where, μ is mean of the distribution and σ is

		normal distribution must be approached with caution, particularly if inferences will focus on the tails of the distribution.	the standard deviation.
Log-normal	Logarithmically related to normal distribution and shows considerable flexibility of shape, which is always skewed to the right.	In log-normal distribution its sample space admits only positive values and suitable to use in analysing water-consumption data.	$f(x)=\frac{1}{\sigma x\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right] \quad x,\mu,\sigma>0$ where μ is the mean and σ is the standard deviation.
Log-logistic	The log-logistic distribution resembles the log-normal in shape, has a more tractable form. It can cope well with outliers in the upper tail.	It is a uni-model, defined only for positive random variables and positively skewed which is best representing the water consumption pattern.	$f(t)=\frac{\lambda\kappa(\lambda t)^{\kappa-1}}{[1+(\lambda t)^{\kappa}]^2} \quad t\geq 0$ where κ is called a shape parameter, as κ increases the density become more peaked. The parameter λ is a scale parameter.
Weibull	Depending on the values of the parameters, the Weibull distribution can be used to model a variety of life behaviors and provides better distribution for life length data.	To use Weibull distribution in analysis, it is essential to have a very good justifiable estimate for the shape parameter to replicate the accurate distribution pattern.	$f(t)=\frac{\beta}{\eta}\left(\frac{t}{\eta}\right)^{\beta-1}\exp\left[-\left(\frac{t}{\eta}\right)^{\beta}\right]$ where β is the shape parameter and η is the scale parameter.

Description of data and the relative water distribution systems

The daily water-consumption data for the five years from 1st April 2009 to 1st April 2013 were obtained from a water utility company in North West Region of England to use in this research. The water works system delivers water to approximately 6.7 million households and businesses in the UK. The data were collected at the supply end of the network system using flow meters which are either connected to telemetry (which are live) or data loggers. The data loggers record the number of pulses within a 15-minute interval. This was then used to calculate and average the flow rate during the 15-minute period, depending on the pulse setting (i.e. how many litres per pulse). This raw data is imported daily into the data management system and converged to hourly and daily volumes.

To analyse the North American daily water consumption, data for three demand zones from a Canadian city in Manitoba Province were obtained. The data were collected using flow meters connected to data loggers at the water treatment plant by the water services division. The data were received between 1st January 2009 and 31st December 2014. The water supply system delivers an average of 225 million liters per day of water to approximately 270,000 households and businesses across approximately 297 square kilometers (114 square miles) of developed area in Canada.

METHODOLOGY

In this research, a suitable statistical distribution was selected using a descriptive analysis. Data were screened and sorted by plotting raw demand data against time. This provided a quick reference to check the abnormality of data. If the points were homogeneously distributed and there were no negative points, this meant that the data were all most acceptable to use in the analysis. Similarly, if there were any inconsistencies in the distribution, these time series graphs would show the abnormal data points to be removed prior to analysis.

The data were then analysed using MINITAB version 18 statistical package, and was fitted into a suitable probability distribution. As previously described, the descriptive analysis show that data fit well into a positively skewed distribution and log-normal, Weibull and log-logistic distributions were applied to find an appropriate distribution. The normal distribution was used for comparison purposes.

Analysing data

Once data has been fitted in to any distribution, the 'goodness-of-fit test' should be used to see how well the data fit into the distribution. The parameters of distribution such as location, shape and scale are also essential to describe the distribution.

The Anderson-Darling test

The appropriateness of the distribution for water-consumption data was assessed by comparison to the normal, log-normal, log-logistic and Weibull distributions using the Anderson-Darling goodness of fit test. The Anderson-Darling Test (Stephens, 1974) is used to test if a sample of data came from a population with a specific distribution. It is a modification of the K-S test and gives more weight to the tails than the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling Test makes use of the specific distribution in calculating critical values. This has an advantage of allowing a more sensitive test and the disadvantage is that critical values must be calculated for each distribution. The critical values were calculated, tabulated and published by Stephens (1974), for a few specific distributions, including log-logistic distribution.

The equation for the Anderson-Darling parameter, A^2, is

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln\omega_i + \ln(1-\omega_{n+1-i})\right]$$

where, n is the number of observations and ω_i is the value of the distribution in question at the ith largest observation. A smaller AD value indicates that the distribution fits the data better. The critical value of the Anderson-Darling parameter at the 95% confidence interval is 2.492 the 1% point is 3.857 for $n \geq 5$ (Johnson, 2000). The Anderson-Darling test was preferred to the Kolmogorov-Smirnov test because of the latter's lack of sensitivity in tails (Ahmad et al., 1988; Johnson, 2000).

Parameter estimates

The location and scale parameters are associated with central tendency and dispersion, respectively, and are essential to describe the distribution. The parameters for normal distribution are the mean and standard deviation and they are directly related to the location and scale parameters (Rigby, 2004). The log-normal, log-logistic and Weibull distributions use location, shape or scale as their parameters and unlike normal distribution they need to transform the location and scale parameters to represent mean and standard deviation using complex equations. These parameters have allowed the distribution to have flexibility and effectiveness in modelling applications. In simple terms the shape parameter allows a distribution to take on a variety of shapes depending on the value of the shape parameter. The effect of the location parameter is to shift the graph to the left or right of the horizontal axis. The scale parameter describes the stretching capacity of the probability distribution function.

The graphical method

There are various numerical and graphical methods used in the literature for estimating the parameters of a probability distribution. In this study, graphical methods were selected for the analysis along with the maximum likelihood method to draw the probability plots (see Figures 1 and 2 in the following section). The data were analysed using 95% confidence intervals (5% significance level) and were fitted to normal, log-normal, Weibull and log-logistic distributions to establish the parameters for the distribution. The middle line in the probability plot shows the normal line and the other two lines show the 95% confidence intervals. Montgomery and Runger (2002) stated that normal probability plots are useful in identifying distributions that fit into the normal distribution and those which have skewed distributions with long tails. If the data falls below the normal line then data has a positively skewed distribution (Montgomery & Runger, 2002).

RESULTS AND DISCUSSION

The normal probability plots for the UK and North American water-consumption data for the year 2009 is shown in Figure 1and Figure 2, respectively. The data shown in Figure 1 and Figure 2 shows that values on both ends tend to fall below the normal line. This demonstrates that the data has a positively skewed distribution. Further to this, the graphs of four distributions for the UK and North American data show that more data points are within the 95% confidence level for log-logistic distribution than log-normal, Weibull and normal distributions.

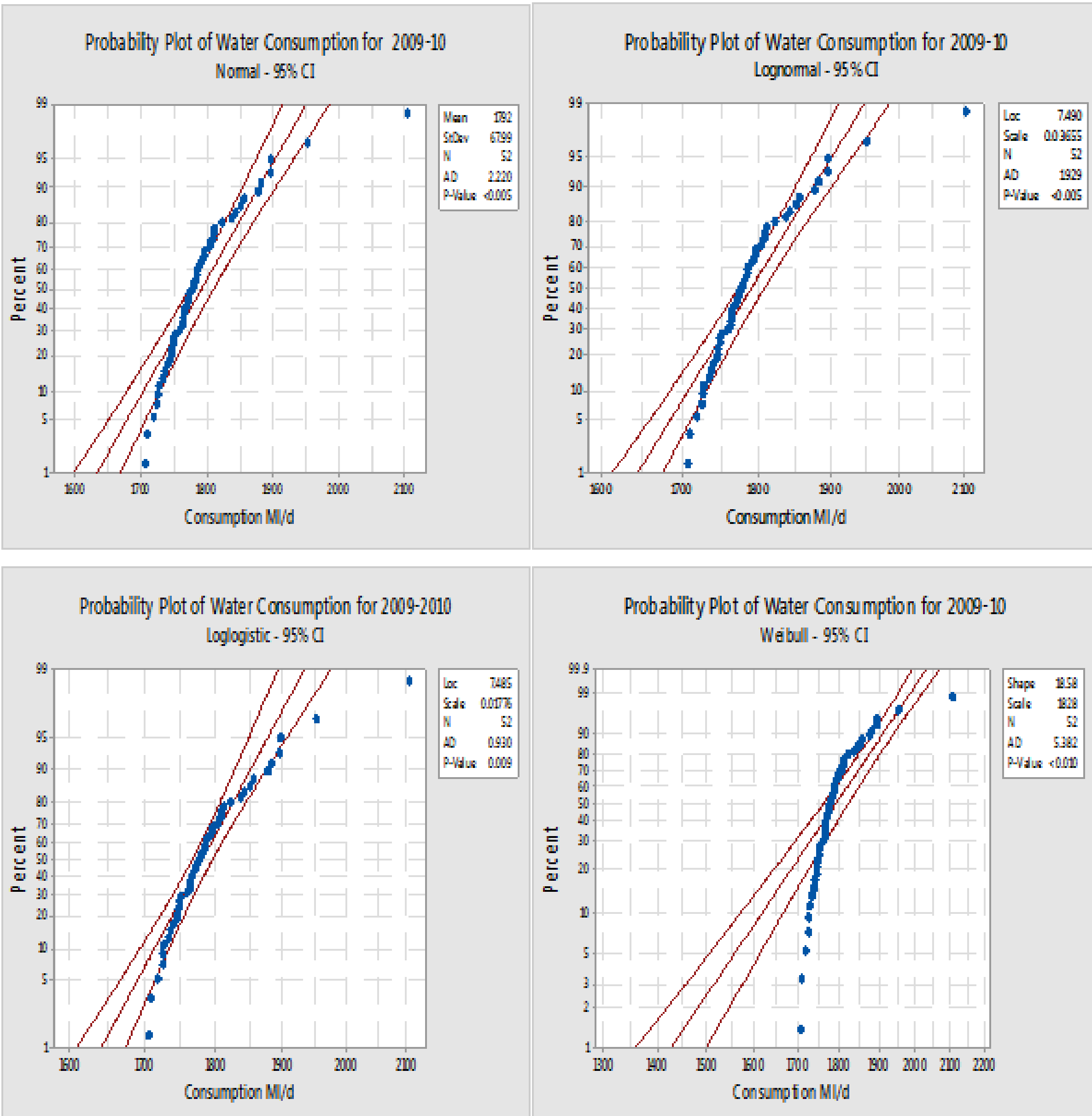

Figure 1. Probability plots for normal, log-normal, log-logistic and Weibull – UK

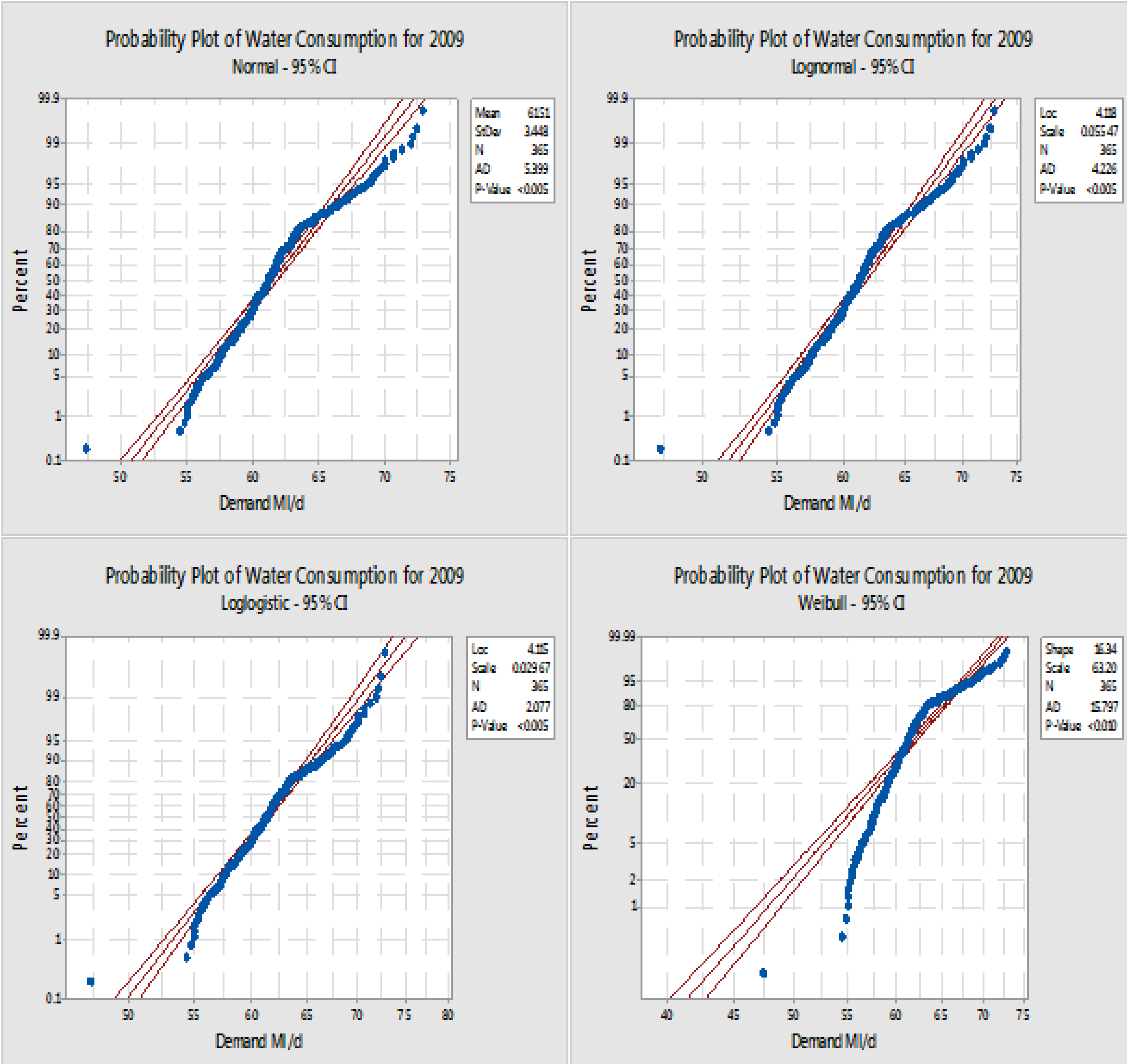

Figure 2. Probability plots for normal, log-normal, log-logistic and Weibull – North American data.

The goodness-of-fit test

The Anderson-Darling (AD) goodness-of-fit test was used to confirm the best fit of data for normal, log-normal, log-logistic and Weibull distribution. The AD values for normal, log-normal, log-logistic and Weibull distributions for the UK and North American data are shown in Figures 3 to 6. The data used in this study has shown that the log-logistic distribution has the lowest AD values when compared with the normal, Weibull and log-normal distributions.

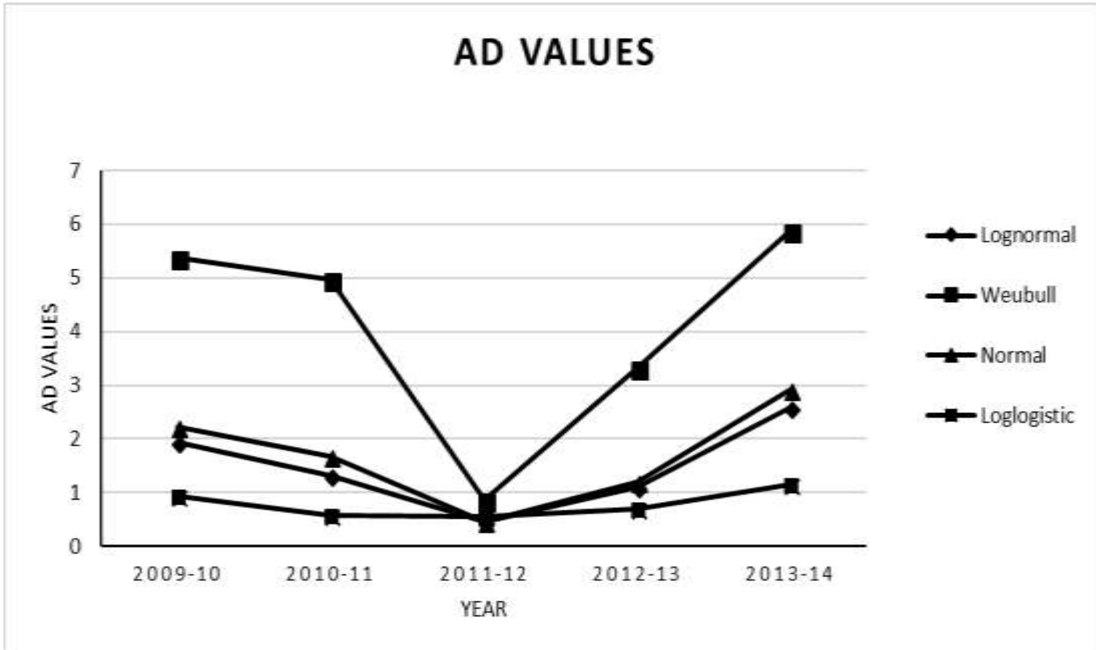

Figure 3. Anderson-Darling (AD) test values – UK.

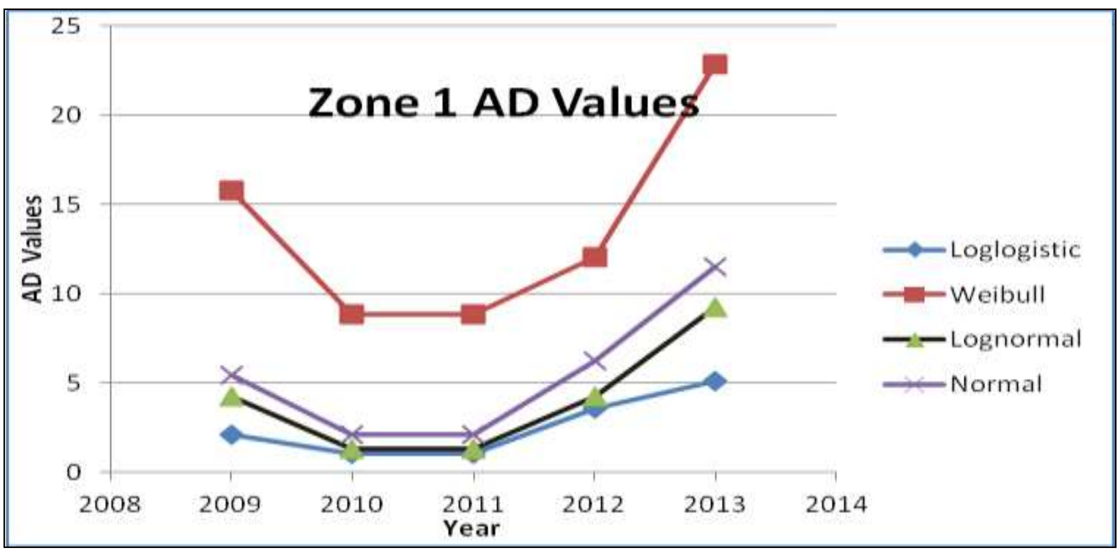

Figure 4. Anderson-Darling (AD) test values for Zone 1 – North America.

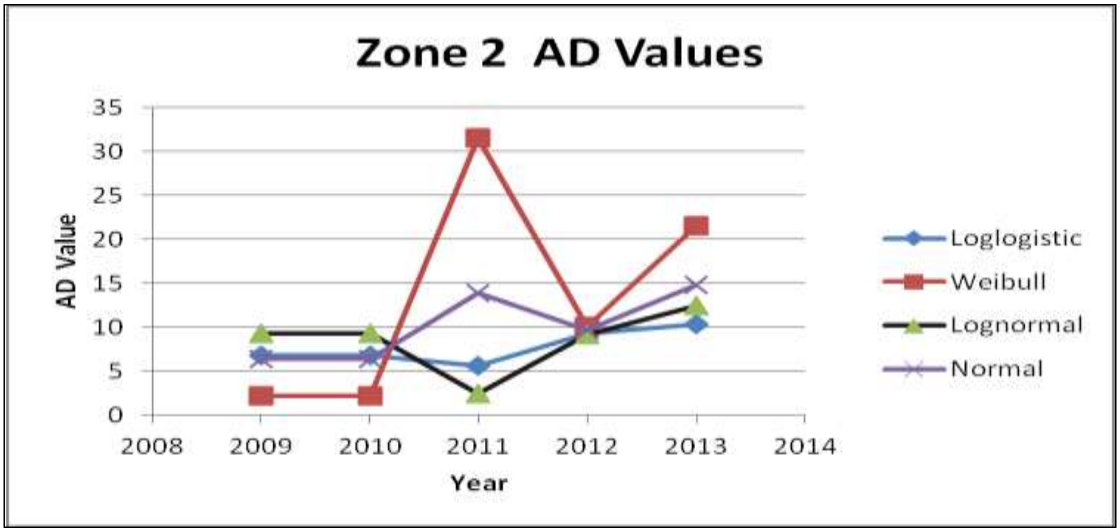

Figure 5. Anderson-Darling (AD) test values for Zone 2 – North America.

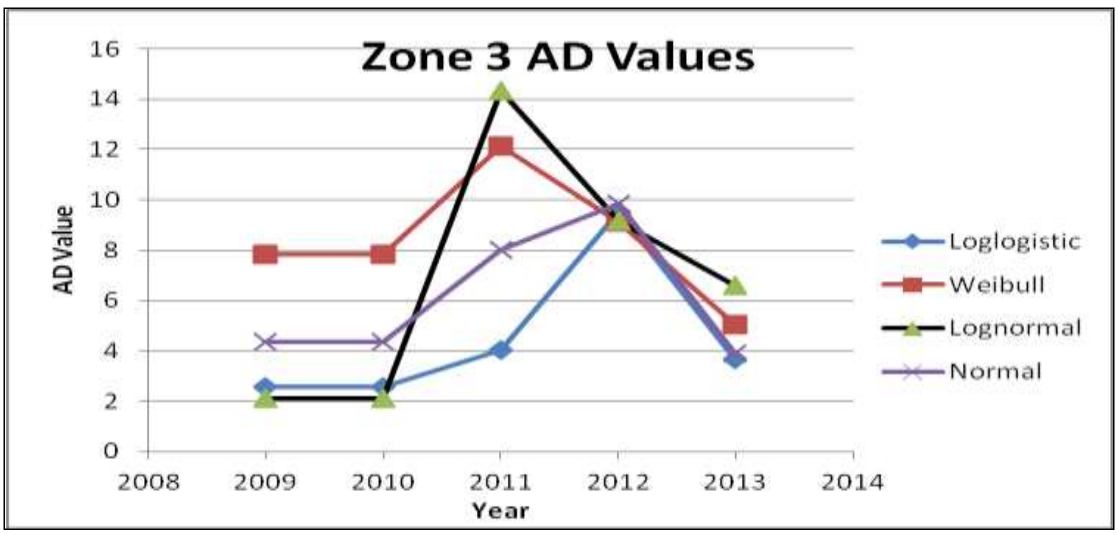

Figure 6. Anderson-Darling (AD) test values for Zone 3 – North America.

Parameter estimates

The location and scale parameters are associated with central tendency and dispersion, respectively, and are essential to describe the distribution. The parameters for normal distribution are the mean and standard deviation and they are directly related to the location and scale parameters (Rigby, 2004). The log-normal, log-logistic and Weibull distributions use location, shape or scale as their parameters and unlike normal distribution they need to transform the location and scale parameters to represent mean and standard deviation using complex equations.

These parameters have allowed the distribution to have flexibility and effectiveness in modelling applications. In simple terms the shape parameter allows a distribution to take on a variety of shapes depending on the value of the shape parameter. The effect of the location parameter is to shift the graph to the left or right on the horizontal axis. The scale parameter describes the stretching capacity of the probability distribution function.

The location parameter obtained in this study for the log-logistic distribution is approximately 7.4 for the UK's water-consumption data. The scale parameter is in between 0.0107 and 0.026 (Table 2). With regard to the North American water-consumption data, the location parameter obtained for log-logistic distribution is between 3.92 and 4.562. Similarly, the scale parameter is between 0.0296 and 0.389 (Table 3). The standard deviation is in a range of 1720 to 1792 and mean value is 35 to 96.51 for the UK's consumption data (Table 4).

Table 2. Location and scale parameters for log-logistic distribution for UK water demand data

Date	**Location**	**Scale**
2009 data	7.485	0.01776
2010 data	7.482	0.0256
2011 data	7.464	0.01517
2012 data	7.448	0.01068

Table 3. Location and scale parameters for log-logistic distribution for Canadian water demand data

Date	Zone 1		Zone 2		Zone 3	
	Location	Scale	Location	Scale	Location	Scale
2009 data	4.089	0.0415	4.545	0.0457	3.920	0.105
2010 data	4.089	0.0415	4.545	0.0457	3.920	0.105
2011 data	4.187	0.055	4.446	0.063	4.308	0.142
2012 data	4.098	0.389	4.361	0.041	4.152	0.092
2013 data	4.115	0.0296	4.562	0.054	3.992	0.088

Table 4. Standard deviation and mean values for water-consumption data for UK water demand

Data	Standard Deviation	Mean
2009 data	67.9	1792
2010 data	96.51	1787
2011 data	45.31	1746
2012 data	35	1720

CONCLUSIONS

It was observed that by analysing water-consumption data, 88% of the water-consumption data has a positively skewed distribution. This means that data would fit better for positively skewed distributions such as log-normal, log-logistic and Weibull. Following detailed analysis of data, the study shows that from the four selected distribution patterns studied, the log-logistic distribution provided the lowest AD values and was the most suitable water-distribution pattern to standardise when modelling the water demand.

The findings in this study are in accordance with the literature which stated that log-logistic distribution is the best fit for real water-consumption data. Although log-normal and log-logistic distributions may be similar for moderate sample sizes, it is still desirable to choose a more suitable model to obtain an accurate probability values at tails.

Moreover, the normal and log-normal distributions produced marginally acceptable AD values. The AD values obtained for the Weibull distribution have higher values when compared with the other three distributions (log-logistic, log-normal and normal) and were found not to be suitable in simulating the water demand data.

To the best of the authors' knowledge, there are no prior studies which have incorporated the probability of occurrence using real water-consumption data built upon a statistically analysed method focused on the upper tails. Using AD test to validate the data, this study focused on the data on upper tails which best represents the water-consumption data.

The log-logistic distribution could be used as a standard statistical distribution in quantifying the probability of exceedence of the water consumption. Additionally, this work also has the potential to provide, significant information to help policy makers forecast future demands using a fully probabilistic method.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to the Environment Agency for supporting this research initiative, and water utility companies from the UK and North America for providing the data sets for this study.

REFERENCES

[1] Ahmad, M. I., Sinclair, C. D. & Spurr, B. D. (1988) Assessment of flood frequency models using empirical distribution function statistics. Wat. Resour. Res. 24 (8), 1323-1328

[2] Ashkar F and Mahdi S, (2006). Fitting the log-logistic distribution by generalized moments. Journal of Hydrology, 328, 694-703

[3] Bao, Y. and Mays, L.W., (1990). Model for water distribution system reliability. J. of Hydraul. Eng., Vol.116(9), 1119-1137

[4] Bowen P T, Harp J F, Baxter W J and Shull R D, (1993). Residential water use patterns. American Water Works Association - Research Foundation, USA.

[5] Cordeiro G M, Santana T V F, Ortega E M M and Silva G O. (2012). The Kumaraswamy-LogLogistic distribution. Journal of Statistical Theory and Applications, 11, 265-291.

[6] De Marinis, G., Gargano, R. and Tricarico, C. (2007). Water demand models for a small number of users. ASCE Proceedings of the 8th Annual International Symposium on Water Distribution Systems Analysis, Cincinnati, OH, doi: 10.1061/40941(247)41.

[7] Dey, A. K. & Kundu, D. (2004), Discriminating between Log normal and Log logistic distributions, Journal of Statistical Computation and Simulation vol.74, no.2, 107–121.

[8] El-Saidi M A, Singh K P and Bartolucci A A. (1990), A note on a characterisation of the generalised log-logistic distribution. Environmetrics; 1 (4), 337-342.

[9] Gargano, R.; Tricarico, C.; Del Giudice, G.; Granata, F. (2016). A Stochastic Model for Daily Residential Water Demand. Water Sci. Technol. Water Supply 2016, 16, 1753-1767.

[10] Gargano, R., Tricarico, C., Granata, F., Santopietro, S., de Marinis, G. (2017). Probabilistic Models for the Peak Residential Water Demand. Water (Switzerland), 9, 417.

[11] Gato-Trinidad, S. and Gan, K., (2012). Characterizing maximum residential water demand. Urban Water - WIT Transactions on The Built Environment, Vol.122, 15-24.

[12] Goulter I C and Boulchart F, (1990). Reliability constrained pipe network model, Journal of hydraulic engineering, ASCE, 116, (2), 211- 227.

[13] Johnson R A, (2000). Probability and statistics for Engineers, Prentice Hall, London.

[14] Khomsi D, Walters G A, Thorly A R D and Ouazar D, (1996). Reliability tester for water distribution networks, Journal of Computing in Civil Engineering, ASCE, 10, (1), 11-19.

[15] Kleiber C (2004). 'Lorenz ordering of order statistics from log-logistic and related distributions', Journal of Statistical Planning and Inference', 120 (1-2),13-19

[16] Kwietniewski M, (2003). ‘Reliability Modelling of Water Distribution System (WDS) for Operation and Maintenance Needs’, Journal of Hydro-Engineering and Environmental Mechanics Vol. 51 (2004), No. 1, pp. 85–92.
[17] Mays L W, (1994). Computer Modelling of Free Surface and Pressurised flows, Chaudray M H and Mays L W (editors), Kluwer Academic Publishers, Netherland, 485-517.
[18] Montogomary D C and Runger G C, (2002). ‘Applied statistics and probability for Engineers’, 3rd edition, Printed in USA.

[19] Ramos M W A, Cordeiro G, Marinho P, Dias C (2013). The Zografos-Balakrishnan Log-Logistic Distribution: Properties and Applications, Journal of Statistical Theory and Applications, Vol. 12, No. 3 (September 2013), 225-244

[20] Rowinski P M, Strupczewski W G and Singh V P (2001), ‘A note on the applicability of log-Gumbel and log-logistic probability distributions in hydrological analyses’: I. Hydrological Sciences Journal, 47 (1), 107-122.

[21] Stephens M A (1974), ‘EDF statistics for goodness of fit and some comparisons’, J. American Statistical Association, Vol.69, pp. 730-737.
[22] Surendran, S., Tanyimboh, T. and Tabesh, M. (2005). Peaking demand factor-based reliability analysis of water distribution systems. Advances in Engineering Software, 36(11-12), pp.789-796.96.
[23] Swamee, P.K. (2002). Near lognormal distribution. J. Hydrol. Eng., 7(6), 441-444.
[24] Syntetos A. A. & Boylan, J E, (2001). On the bias of intermittent demand estimates. International Journal of Production Economics, 71, 457– 466.
[25] Syntetos A.A, and Boylan, J E, (2005). The accuracy of intermittent demand estimates, International Journal of Forecasting 21 (2005) 303– 314
[26] Tanyimboh T T and Surendran S, (2002). Log-logistic Distribution Models for Water Demands, 4th International Conference on Engineering and Technology Civil-Comp press, Sterling.
[27] Tricarico, C., de Marinis, G., Gargano, R. and Leopardi, A., (2007). Peak residential water demand. Water Management Journal, Vol.160(WM2), pp.115-121.

[28] Xu C and Goulter I C, (1997). A New model for reliability based optimal design of water distribution networks, The 27th Congress of the International Association for Hydraulic Research, ASCE, 423-428.
[29] Xu C and Goulter I C, (1998). Probabilistic model for distribution reliability, Journal of Water Resources Planning and Management, ASCE, 124, (4), 218-228.
[30] Xu C and Goulter I C, (1999). Reliability based optimal design of water distribution networks, Journal of Water Resources Planning and Management, ASCE, 125, (6), 352-362.