PAPER 116 – PROBABILISTIC MODELLING OF FLOOD FLOW ALONG MGENI RIVER AT TABLE MOUNTAIN, SOUTH AFRICA

O. O. Olofintoye[1*], B. F. Bakare[2], A. W. Salami[1,3], A. M. Ayanshola[1], S. O. Bilewu[1], T. S. Abdulkadir[1,4], M. O. Jimoh[5], O. Oyebode[6]

[1]Department of Water Resources and Environmental Engineering, University of Ilorin, Ilorin, Nigeria.
[2]Department of Chemical Engineering, Mangosuthu University of Technology, Durban, South Africa.
[3]National Centre for Hydropower Research and Development, University of Ilorin, Ilorin, Nigeria.
[4]Department of Civil Engineering, Kabale University, Republic of Uganda.
[5]School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom.
[6]Department of Civil Engineering, University of South Africa, Johannesburg, South Africa.
[*]Email: olofintoye.oo@unilorin.edu.ng

ABSTRACT
Accurate flood prediction is indispensable for effective environmental protection. This study establishes best-fit probability distribution models for estimating annual maximum flows at Table Mountain along the Mgeni River. Six probability distribution functions (Normal, Log-Normal, Pearson III, Log-Pearson III, Gumbel and Log-Gumbel) were fitted to the annual maximum flow series of the river. Reliability of raw data was verified using the Spearman-brown reliability check. The non-parametric Chi-square goodness-of-fit statistical test was employed to determine the suitability of the functions in representing the observed data at a specified level of significance. The Kruskal-Wallis non-parametric H-Test was employed to determine if there were significant differences in the performances of the models. Finally, a test score statistics was used to establish the most appropriate model for application in the study area. It was found that the Log-Pearson III probability distribution function, which is a 3-parameter model taking cognizance of a skew coefficient in estimation, is the most suitable model. The Log-Normal and Pearson III models were also found suitable while the Normal, Gumbel and Log-Gumbel models were not suitable at the tests' level of significance. The models found suitable are recommended for predicting flood flows in the study area.

KEYWORDS: Annual maximum flow, Probability distribution function, Flood prediction, Goodness of fit test, H-test, Test score

8. INTRODUCTION

Disasters result in loss of lives, destruction of infrastructure, damaged ecosystems and undermined societal development (Sudmeier-Rieux et al., 2019). Floods are natural disasters resulting from unusual surplus or excess of water in which the resulting higher than usual water levels extend out onto the floodplains. They occur in river basins, usually as the result of a lot of rain (Loucks & Bee, 2005). These natural events have always been a part of the geologic history of the earth, and because human settlements and activities have always tended to use floodplains, their uses has frequently interfered with the natural processes in the floodplain, causing inconvenience and catastrophe to humans (Mays, 2011).

According to Witwatersrand (2022), flooding events in areas around the Mgeni River in the KwaZulu-Natal province in South Africa have doubled over the last century. Noteworthy is the disastrous flood that hit Durban city in the province in April 2022. This is considered the most catastrophic natural disaster yet recorded in the region in collective terms of lives lost, homes and infrastructure destroyed and economic losses. The calamitous flood in its wake resulted in loss of 459 human lives, 88 people still missing by the end of May 2022, and over 4000 destroyed homes rendering 40,000 people homeless and 45,000 people temporarily unemployed. Also, substantial damages were caused to some of Mgeni raw water abstraction and conveyance infrastructure resulting in shortfall in potable water supply in some districts (Umgeni, 2022). The cost of infrastructure and business losses amounted to an estimated 2 billion USD. Several historical flooding events have been recorded in the region notable among which are the Durban flood in September 1987 and the great Deluge of 1856.

Studies suggest that an increase in the severity and frequency of extreme hydrological events is expected worldwide. Recent anthropogenic global climate warming may have contributed to the trend in increased flooding and this trend may continue into the future (NZCCO, 2008; Witwatersrand, 2022). Hence, it is pertinent to establish or re-establish accurate flood prediction models for effective environmental protection in flood prone areas through frequency analysis of flows, to establish best-fit probability distribution functions that can facilitate the prediction of future flood events.

Several researchers (Topaloglu, 2002; Olukanni, 2003; Khudri & Sadia, 2013; Paul et al., 2014) have analysed historical hydrologic events using methods of statistical frequency analysis, and developed appropriate best-fit models to forecast extreme events. For instance, Salami (2007) developed probability distribution models

for predicting annual peak and low flows at selected gauging stations along some South-West Rivers in Nigeria. Probability distribution functions such as Gumbel, Normal, Log-Pearson type III and Log-normal distributions were fitted to the flow series. The Gumbel (Extreme value type I) and Log-normal models were found suitable for estimating flows in the study area. Adeyemo and Olofintoye (2014) demonstrated the application of the parametric Normal, Lognormal, Pearson III, Log-Pearson III, Gumbel and Log-Gumbel probability distribution models in estimating monthly river flows into the Vanderkloof dam in South Africa. The models were found appropriate for flow computations in the reservoir.

This study analyzes the frequency distribution of flood flows along the Mgeni river in South Africa using data obtained from the Table Mountain gauging station. It employs statistical methods to determine the fitness of the developed models. Establishing relevant functions for predicting flood flow in the river is germane in articulating proactive measures towards addressing the challenges associated with flooding in the watershed.

9. THEORETICAL ANALYSIS

Statistical methods were adopted for the analysis in this work. The mathematical principles applicable are discussed in the following subsections.

2.1 Data Reliability Check:

The Spearman-Brown coefficient was used to check the reliability of data. Formula for computing the coefficient is given in Equation 1 as:

$$r_{SB} = \frac{2 \times r}{1 + r} \qquad (1)$$

Where r_{SB} is the Spearman-Brown correlation coefficient and r is the Pearson Moment correlation coefficient between the data set split in half. A value of the coefficient greater than 0.8, implies good reliability of data.

2.2 Probability Curve Fitting:

Six probability curves were fitted to flow data in the study area. The probability density functions and the generalized equations of the models are presented in Table 1.

Table 1: Probability density functions and generalized models of fitted curves.

SN.	Distribution	Probability density function	Generalized model equation
1	Normal	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{(x-\mu)^2/2\sigma^2}$	$Q_T = Q_{av} + k\sigma$
2	Log-Normal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{(\ln x - \mu_{\ln x})^2/2\sigma^2_{\ln x}}$	$\log Q_T = \overline{\log Q} + k\sigma_{\log Q}$
3	Pearson III	$f(x) = \frac{1}{\beta\Gamma(p)}\left(\frac{x-\alpha}{\beta}\right)^{p-1} e^{-(x-\alpha)/\beta}$	$Q_T = Q_{av} + k'\sigma$
4	Log-Pearson III	$f(x) = \frac{1}{x\beta\Gamma(p)}\left(\frac{\ln x - \alpha}{\beta}\right)^{p-1} e^{-(\ln x - \alpha)/\beta}$	$\log Q_T = \overline{\log Q} + k'\sigma_{\log Q}$
5	Gumbel	$f(x) = \beta \exp\{-\beta(x-u) - e^{-\beta(x-u)}\}$	$Q_T = Q_{av} + \sigma(0.78Y_T - 0.45)$
6	Log-Gumbel	$f(x) = (\beta/x)\exp\{-\beta(\ln x - u) - e^{-\beta(\ln x - u)}\}$	$\log Q_T = \overline{\log Q} + \sigma_{\log Q}(0.78Y_T - 0.45)$

Source: Olofintoye (2007).

Where x is the standardized variate, μ, u and α are location (mean) parameters while σ and β are scale parameters (variance). $\Gamma()$ is the gamma function and p is a shape parameter. Q_{av} is the mean of the observed data while Q_T is the estimated flow for a specified return period T. Y_T is the reduced variate computed for a specified return period. The variable k is a frequency factor obtained from Normal distribution tables while k' is obtained from tables of frequency factors for the Pearson distribution.

2.3 Goodness-of-fit Test:

The Chi-square (χ^2) goodness-of-fit test was used to check if an experimental distribution follows a hypothesized one. This test is based on the sum of the squares of differences between the observed and estimated frequencies. Equation 2 may be used to compute a χ^2 value.

$$\chi^2 = \sum_{j=1}^{N} \frac{(o_j - e_j)^2}{e_j} \qquad (2)$$

Where o_j is the observed frequency, e_j is the estimated frequency, and N is the total frequency. If the calculated value is greater than the critical value at the specified level of significance, then the null hypothesis of a good fit is rejected.

10. METHODOLOGY

Details on the methodologies applied herein are presented in the following sub-sections.

3.1 Data and Validation:

Annual maximum flow data at Table Mountain gauging station, along the Mgeni River was collected from the Department of Water and Sanitation (DWS), Pretoria, South Africa. DWS is the custodian of hydrological data in the study area. The unit of the data is cubic metres per second. Sixty (60) years of data (1951 – 2022) was used for the analyses in this study. Statistics were computed to provide an insight into the nature of the data and the Spearman-Brown split-half technique was used to check the validity of the data.

3.2 Fitting Probability Distribution Functions:

The flow series was ranked according to Weilbull's plotting position and the corresponding return periods were estimated. Weibull's plotting position is an efficient formula for computing plotting position for unspecified distribution (Viessman et al., 1989). The Weibull plotting position may be computed using Equation 3 as:

$$P = \frac{m}{n+1} \quad (3)$$

Where m is the series of event ranking, 1 for highest value and so on in descending order, n is the number of events and P is an estimate of the probability of values being equal or greater than the ranked value. The return period T is estimated as the reciprocal of the probability using Equation 4 as follows:

$$T = \frac{1}{P} \quad (4)$$

The flow series was evaluated using six methods of probability distribution functions, Normal, Log-Normal, Pearson III, Log-Pearson III, Gumbel and Log-Gumbel. Curve fitting was carried out following standard procedure (Topaloglu, 2002; Salami, 2007; Olofintoye & Adeyemo, 2012; Khudri & Sadia, 2013; Adeyemo & Olofintoye, 2014; Paul et al., 2014).

3.3 Evaluating the Suitability of the Developed Models:

The suitability of the established probability models was examined using the statistical non-parametric chi-square (χ^2) goodness-of-fit test. If the computed value of the statistics using Equation (2) is greater than the corresponding critical value (or if the ratio $\chi^2_{Computed}/\chi^2_{Critical}$ is greater than one), the null hypothesis is rejected at the specified level of significance and it is concluded that there is significant discrepancy between experiment and theory (Olofintoye, 2007; Gupta, 2013). A 10% level of significance was adopted for the test in this study following the suggestion of (Haktanir et al., 2012).

3.4 Comparison of Flood Flow Models:

Since flood flow data are known to be skewed and generally not to follow a Normal distribution, comparisons between the performances of the models were made using the Kruskal-Wallis H-Test. This is the non-parametric equivalence of the Analysis of Variance (ANOVA) technique. The H-Test was performed at 5% level of significance as suggested by (Gupta, 2013). Only models that were found suitable by the Chi-square test were included in this test. Probability plots were also made to aid in visual inspection of model fit.

3.5 Test Scoring and Selection of the Most Suitable Flood Flow Model:

A test scoring scheme that has been successfully applied in the literature (Olofintoye, 2007; Olofintoye et al., 2009; Khudri & Sadia, 2013; Paul et al., 2014) was adopted to evaluate the fitness of a model relative to the others. Scoring of models was based on the Mean Relative Deviation (MRD), Mean Square Relative Deviation (MSRD) (Jou et al., 2009) and $\chi^2_{Computed}/\chi^2_{Critical}$ fit statistics. Lower values of these statistics indicate better fit. In the scheme, the assessment of a model is based on total test score obtained on all the fit statistics. A maximum score of six (6) is awarded a function best supported by a fitness test while a minimum of one (1) is awarded to the least performing model. A zero (0) score is awarded in all cases when the chi-square test indicates that a model is inappropriate. The model with the maximum overall score is selected as the most suitable model. Elaborate information on statistical test scoring may be found in the mentioned literature and other statistical texts.

11. RESULTS AND DISCUSSION

The flood data collected was analyzed in order to determine its reliability. The correlation coefficient (r) was computed and adjusted for the split half method using the Spearman's brown correlation coefficient (r_{SB}). The value of r was found to be 0.86 while r_{SB}= 0.92. These values lie in the excellent range and suggest that the data

is reliable within the acceptable limits. A summary of statistics of the annual maximum flow series in the study area is present in Table 2.

Table 2: Summary of statistics of annual maximum flows in Mgeni River at Table Mountain (1951 – 2022)

Mean value $\bar{x}$, (m^3/s)	*Median* (m^3/s)	*Standard deviation* σ, (m^3/s)	*Skewness coefficient, G*	*Minimum* (m^3/s)	*Maximum* (m^3/s)
135.29	94.19	134.18	2.42	8.82	760.82

It is observed from examination of statistics in Table 2, that the range and standard deviation of the series are high. Also, the mean lies to the right of the median with a positive coefficient of skewness. This indicates that the frequency distribution of maximum annual flow in the river is positively skewed. This corroborates the observations of Viessman et al. (1989) and other pundits who have noted that the distribution of flood flows are generally markedly right skewed partially due to the influence of natural phenomena.

The developed model equations, results of chi-square goodness-of fit test and its significance status and the computed values of the fit statistics of each function are presented in Table 3. The test scores obtained on each fit statistic are emphasized in italics and enclosed in parentheses. The probability fitness plots for the models are presented in Figure 1.

Table 3: Summary of developed probability distribution models and statistical tests

		Test Statistics				
Model	*Model Equation*	*MRD*	*MSRD*	$\chi^2_{Computed}$ / $\chi^2_{Critical}$	*Total Test Score*	*Significance Status*
Normal	Qp = 135.29 + 134.18K	5.8050 *(0)*	105,273.76*(0)*	1.0120 *(0)*	0	Significant
Log-Normal	Qp = Antilog (1.93 + 0.45K)	3.6659 *(5)*	303.30*(5)*	0.7836 *(6)*	16	Not Significant
Pearson III	Qp = 135.29 + 134.18K'	2.8938 *(4)*	1,631.63*(3)*	0.9107 *(4)*	11	Not Significant
Log-Pearson III	Qp = Antilog (1.93 + 0.45K')	2.8613 *(6)*	202.82*(6)*	0.7772 *(5)*	17	Not Significant
Gumbel	Qp = 74.91 + 104.66YT	3.0927 *(0)*	2,854.19*(0)*	0.7507 *(0)*	0	Significant
Log-Gumbel	Qp = Antilog (1.73+ 0.35YT)	3.4269 *(0)*	1,504.36*(0)*	0.4575 *(0)*	0	Significant

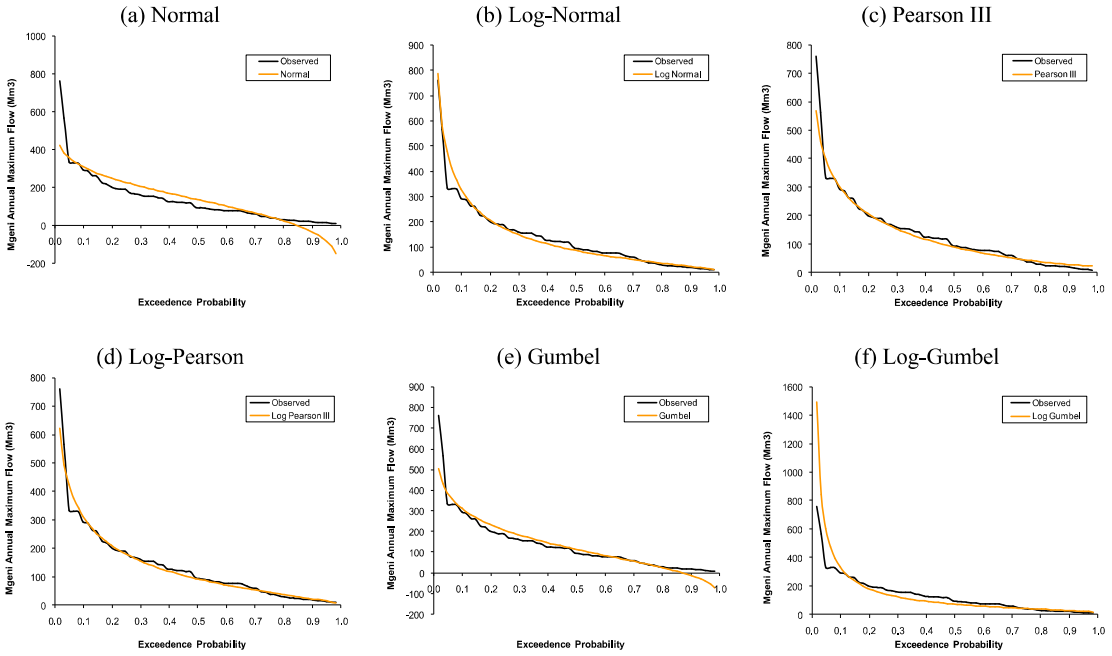

Figure 1: Fitness plots for Models

It is evident from Table 3 and Figure 1 that only 3 models viz. Log-Normal, Pearson III and Log-Pearson III distributions exhibited good fit to the data at the level of significance, and are therefore suitable for predicting flood flows in the study area. The Normal and Gumbel curves showed large deviations at the tails of the distribution. Also, some part of the curves lies in the negative range below the horizontal axis (Figures 1a and 1e).

While the Log-Gumbel was entirely restricted to the positive range of estimation, it showed very large deviation at the left tail of the distribution which rendered it unsuitable. Consequently, the Normal, Gumbel and Log-Gumbel distribution models were not compared further in the Kruskal-Wallis Test.

The Normal distribution is generally known to provide reasonable approximation in the centre of a distribution but is often inadequate at one or both tails of skewed distributions (Olofintoye, 2007; Salami, 2007). This is evident from the results in this study as the Normal distribution model produced a poor fit and exhibited significant deviations at the tails of the model (Figure 1a). This phenomenon has also been observed by Olofintoye et al. (2009) who reported that hydrologic series are generally skewed and found the Normal distribution model inadequate at most of the hydrologic stations in the study. Applications of normalization procedures before fitting Normal models to flow series may however be beneficial in resolving this problem in future studies.

The significance statuses of the Log-Normal, Pearson III and Log-Pearson III models indicate that they are suitable for estimating future flood flow in the study area. Based on total test scores obtained (Table3), Log-Pearson III is the best performing model, Log-Normal the second and Pearson III the third best. However, the value of the computed H-statistic in the Kruskal-Wallis test (H = 0.064) was less than the critical value of the statistic ($H_{critital}$ = 5.99) at the 5% level of significance. This suggests that there are no significant differences in the performance of the models and hence the 3 models are applicable in estimating flood flow in the area.

Since flood flow data are known to be generally skewed, 3-parameters distributions like the Pearsonian models, which employ the computation of a skew coefficient in estimation of hydrologic random variables, are often found to provide good fits in flood modelling. The selection of Log-Pearson III function as the best-fit model in this study further corroborates the findings of (Olofintoye, 2007; Olofintoye & Sule, 2009; Olofintoye et al., 2009) that the distributions hydrologic data are positively skewed and Log-Pearson III probability distribution model may conveniently be used to estimate future events.

In particular, the United States (U.S.) Interagency Advisory Committee on Water Data has recommended the use of Log-Pearson III distribution in calculating flood recurrences. It is also the default distribution adopted by the U.S. Geological Survey for flood studies. Likewise, the U.S. Water Resources Council (WRC) has also recommended that Log-Pearson III distribution be used as a base method for flood flow frequency studies in an attempt to promote a consistent, uniform approach to flood flow frequency determination for use in all federal planning involving water related land resources (Mays, 2011). Therefore, the findings in the study are in accord with international suggestions.

12. CONCLUSION

The need for reliable hydrologic studies cannot be overemphasized particularly in this era of climatic vicissitude. Increases in severities and frequencies of extreme daily rainfall events with attendant flooding are speculated worldwide due to recent anthropogenic global warming (IPCC, 2011). Therefore, establishing methods germane in effective management of hydrologic events is of great practical value in flood control.

Frequency analysis of annual maximum flow of Mgeni River measured at Table Mountain Gauging Station was performed to establish probability distribution models apposite in estimating flood. Six probability curves were fitted. It was found that the Log-Normal, Pearson III and Log-Pearson III distribution models were suitable for modelling flood flow in the area. These models may be useful in developing strategies for mitigating the adverse effects of floods experienced in the Mgeni watershed in South Africa.

A 3-parameter Log-Pearson III model which takes cognizance of a skew coefficient in estimations was found to be most appropriate in modelling flood flow in the study area. This is consistent with results from earlier studies and in accord with suggested international standards. Also, the length of data used in the analyses is statistically large and sufficient to provide good approximations to the population parameters and support the various statistical tests (Gupta, 2013). This suggests that the established models are appropriate and are therefore recommended for estimating flood flows in the Mgeni watershed, South Africa.

ACKNOWLEDGEMENT

The authors acknowledge the Department of Water and Sanitation, South Africa for making river flow data available for this study.

REFERENCES

Adeyemo, J., & Olofintoye, O. (2014). Optimized Fourier Approximation Models for Estimating Monthly Streamflow in the Vanderkloof Dam, South Africa. In A. Tantar, E. Tantar, J. Sun, W. Zhang, Q. Ding, O. Schütze, M. Emmerich, P. Legrand, P. D. Moral, & Coello Coello C.A. (Eds.), *EVOLVE - A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation V* (pp. 293-306). Springer International Publishing.

Gupta, S. C. (2013). *Fundamentals of Statistics* (I. Gupta, Ed. Seventh ed.). Himalaya Publishing House.

Haktanir, T., Bajabaa, S., & Masoud, M. (2012). Stochastic analyses of maximum daily rainfall series recorded at two stations across the Mediterranean Sea. *Arabian Journal of Geosciences*, *2012*, 1 - 16. https://doi.org/10.1007/s12517-012-0652-0

IPCC. (2011). *Managing the risks of extreme events and disasters to advance climate change adaptation*.

Jou, P. H., Akhoond-Ali, A. M., Behnia, A., & Chinipardaz, R. (2009). A Comparison of Parametric and Nonparametric Density Functions for Estimating Annual Precipitation in Iran. *Research Journal of Environmental Sciences*, *3*(1), 62-70.

Khudri, M. M., & Sadia, F. (2013). Determination of the Best Fit Probability Distribution for Annual Extreme Precipitation in Bangladesh. *European Journal of Scientific Research*, *103*(3), 391-404. http://www.europeanjournalofscientificresearch.com

Loucks, D. P., & Bee, E. v. (2005). *Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications*. UNESCO and Delft Hydraulics.

Mays, L. W. (2011). *Water Resources Engineering* (Second ed.). John Wiley & Sons, Inc.

NZCCO. (2008). *Climate change effects and impacts assessment: a guidance manual for local governments in New Zealan–2nd edition, .*

Olofintoye, O., & Adeyemo, J. (2012, 05 – 09 May, 2012). Development and Assessment of a Fourier Approximation Model for the Prediction of Annual Rainfall in Ilorin, Nigeria. Water Footprint: Water Institute of Southern Africa Biennial Conference and Exhibition, International Conference Center, Cape Town.

Olofintoye, O. O. (2007). *Frequency Analysis of Maximum Daily Rainfall for Selected Urban Cities in Nigeria* [M.Eng Thesis, University of Ilorin]. Ilorin, Nigeria.

Olofintoye, O. O., & Sule, B. F. (2009). Best–Fit Probability Distribution Model for Peak Daily Rainfall in Southern Nigeria. 1st Annual Civil Engineering Conference University of Ilorin, Department of Civil Engineering, University of Ilorin, Nigeria.

Olofintoye, O. O., Sule, B. F., & Salami, A. W. (2009). Best–fit Probability distribution model for peak daily rainfall of selected Cities in Nigeria. *, New York Science Journal*, *2*(3), 1-12.

Olukanni, D. O. (2003). *Development of Statistical Model for Reservoir Inflow at Kainji, Jebba, and Shiroro Hydropower Stations* University of Ilorin]. Ilorin, Nigeria.

Paul, S., Hasan, M. A., & Shopan, A. A. (2014). *AN ASSESSMENT OF BEST FITTED RAINFALL DISTRIBUTION FOR PROJECTED CLIMATE CHANGE OVER BANGLADESH USING STATISTICAL DOWNSCALING METHOD* 2nd International Conference on Civil Engineering for Sustainable Development, KUET, Khulna, Bangladesh.

Salami, A. W. (2007). Fitting probability distribution models to peak and low flows in selected rivers in South-West Nigeria. *Nigerian Journal of Technological Development*, *5*, 50-63.

Sudmeier-Rieux, K., Nehren, U., Sandholz, S., & Doswald, N. (2019). *Disasters and Ecosystems, Resilience in a Changing Climate - Source Book* (U. N. E. Programme, Ed.). UNEP and Cologne: TH Köln - University of Applied Sciences.

Topaloglu, F. (2002). Determining Suitable Probability Distribution Models for Flow and Precipitation Series of the Seyhan River Basin. *Turk Journal of Agric*, *26*, 189 - 194.

Umgeni. (2022). *STATEMENT: Heavy downpours cause damage to Umgeni Water infrastructure*. Retrieved 24th April, 2023 from https://www.umgeni.co.za/statement-heavy-downpours-cause-damage-to-umgeni-water-infrastructure/

Viessman, W., Krapp, J. W., & Harbough, T. E. (1989). *Introduction to Hydrology* (Third ed.). Harper and Row Publishers Inc.

Witwatersrand. (2022). *The 2022 Durban floods were the most catastrophic yet recorded in KwaZulu-Natal*. Retrieved 24th April, 2023 from https://www.wits.ac.za/news/latest-news/general-news/2023/2023-04/the-2022-durban-floods-were-the-most-catastrophic-yet-recorded-in-kwazulu-natal.html