Selection of Best Fit Probability Distribution for Flood Frequency Analysis in South West Western Australia

Benjamin P Brash[1] and Ataur Rahman[2]

[1]Student, Western Sydney University, NSW, Australia
[2]Associate Professor, Western Sydney University, NSW, Australia
Corresponding author's E-mail: 17882910@student.westernsydney.edu.au

Abstract

Design floods are needed to design safe and adequate infrastructure such as bridges, levees and dams. The estimation of design floods is generally performed by at-site flood frequency analysis (FFA) when adequate historical streamflow data is available at the site of interest. The 3rd edition of Australian Rainfall & Runoff (ARR) 1987 recommended the use of the log-Pearson type 3 (LP3) distribution to estimate design floods for all Australian catchments including South West Western Australia. The 4th edition of ARR no longer prescribes a particular probability distribution for FFA in Australia. In this study six different probability distributions are used to assess which probability distribution could be recommended when performing FFA for catchments located in South West Western Australia. The distributions that are assessed in this study are: normal distribution, log-normal (LN) distribution, LP3 distribution, generalised extreme value (GEV) distribution, extreme value type 1 (EV1) distribution and generalized Pareto (GP) distribution. Use of the software package TUFLOW Flike, recommended in the 4th edition of ARR, is made to fit the LN, LP3, GEV, EV1 & GP distributions, while the R software package is used to fit the normal distribution. Goodness of fit testing is performed to assist in the ranking of the various distributions for the selected catchments. From this study of the six distributions on 103 catchments located in South West Western Australia we find that the GEV distribution provides the best estimation of design floods from FFA. However, some catchments in South West Western Australia do not follow the GEV distribution where other distribution should be applied. This study highlights the difficulties in recommending a single probability distribution for FFA in a given region.

***Keywords*:** Floods, LP3 distribution, GEV distribution, flood frequency analysis, ARR

1. INTRODUCTION

Floods are one of the most devastating natural disasters which may cause significant damage to property and crop, and loss of life. For example, 2010-11 Brisbane flood caused economic damage over $10 billion and loss of over 30 human lives. Estimation of design flood is one of the most researched topics in hydrology. One of the most effective methods of design flood estimation is through the application of flood frequency analysis (FFA). Haddad & Rahman (2010) noted that FFA is one of the most challenging topics in hydrology as there are many probability distributions and parameter estimation methods to choose for a given application. FFA assumes that the observed flood discharges can be described by a selected probability distribution (Merz & Thieken, 2009). The aim of FFA is to estimate the flood discharge for a given annual exceedance probability (AEP) through the application of a selected probability distribution to a given flood data set. A design flood is defined as the peak flood discharge for an associated AEP (Caballero & Rahman, 2014). Wang (2015) noted that FFA is one of the most efficient methods of design flood estimation when there is adequate flood data available at the location of interest. Schendel & Thongwichian (2015) noted that estimated design floods may not perfectly match the recorded flood data due to uncertainty introduced by short data length and hence understanding of inherent uncertainty in FFA is important (Mirzaei et al., 2015; Yan & Moradkhani, 2015).

TUFLOW Flike is a software that has been specifically developed for FFA in Australia as a part of ARR (Kuczera and Franks, 2016) and has been adopted in this study. Determination of the probability distribution that best suits a given dataset is important so that informed decisions can be made as to the most appropriate probability distribution to apply. This is commonly performed through the application of goodness of fit (GoF) tests. A GoF test describes how well a probability distribution fits a given flood dataset through analytical and graphical methods (Heo et al., 2013).

2. STUDY AREA

This study focuses on South West (SW) of Western Australia (WA) located within Drainage Division 6 (BOM, 2012). The study area is presented in Figure 1. Drainage Division 6 has an area of approximately 326,000 km^2, which is bordered by the Indian Ocean to the West, the Great Australian Bight to the South, the Pilbara-Gascoyne region to the North and the South-Western Plateau region to the East. Drainage Division 6 is generally considered to have a flat topography, with a maximum elevation of 1000 m above the mean sea level. Drainage Division 6 is generally found to be made up of sandy soil and dune systems along the coastal regions. The climate of this region is considered to be temperate with warm summers and cool winters. A total of 103 gauged catchments from SW WA are used in this study. These data were collated as a part of ARR Revision 'Project 5 Regional flood methods' (Rahman et al., 2016).

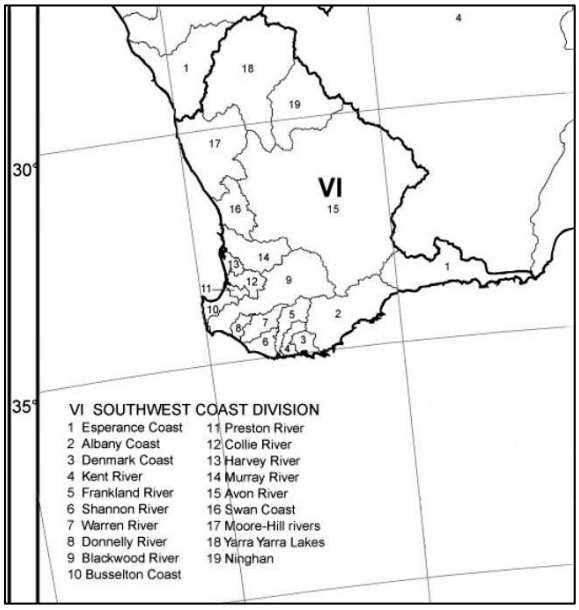

Figure 1 Drainage Division 6 in South West of Western Australia

3. SELECTION OF CANDIDATE PROBABILITY DISTRIBUTIONS

The principal aim of FFA is determination of the relationship between flood magnitude and the return period (Cunnane, 1978; Cunnane, 1985). The probability distributions that are applied in this study are as follows; normal (Norm), log-normal (LN), log-Pearson type 3 (LP3), generalised extreme value (GEV), extreme value type 1 (EV1) and generalised Pareto (GP) distributions.

The non-parametric method (Bardsley, 1989) is also used in this study to visualise the goodness-of-fit for a given distribution. Cunnane's plotting position formula is used here, which can be described by:

$$T = (n+0.2)/(m-0.4) \quad (1)$$

where T is the return period in years, n is the sample size and m is the rank of the data.

4. GOODNESS OF FIT TESTING

In this study, the Easyfit software is adopted to assess the goodness of fit (GoF). Three GoF tests are built into this software; the Kolmogorov-Smirnov (KS) test, Anderson-Darling (AD) test and the Chi-squared (Chi) test. All the three tests are applied to the AM flood data to select the best fit probability distribution for each of the selected catchments.

The Kolmogorov-Smirnov (KS) test is based on the largest difference between the sample cumulative distribution function and the hypothesized cumulative distribution function. The sample cumulative distribution function is described as:

$$F_n(x) = \frac{1}{n}.(Number\ of\ observation\ \leq x) \tag{2}$$

where n is the sample size and x is a random sample of the dataset being assessed.

The AD test compares the fit of an observed cumulative distribution function to that of an expected cumulative distribution function. This test gives more weight to the extreme data values. To perform an AD test the Anderson-Darling statistic (A^2) is determined as below:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\cdot\left[lnF(X_n) + \left(ln\left(1 - F(X_{n-1+i})\right)\right)\right] \tag{3}$$

where n is the sample size and X_n is a sample from the dataset to be assessed.

The Chi-squared test is a simple method to determine if a sample comes from a population that belongs to a specific distribution. This GoF test is performed on binned data, where binned data divides the entire range of the dataset into specific intervals for analysis. Each of these bins must contain a minimum of 5 data points for the Chi-squared test to be valid. After the data has been binned the Chi-squared statistic (χ^2) is determined by the following equation:

$$\chi^2 = \sum_{i=1}^{k}\frac{(O_i - E_i)^2}{E_i} \tag{4}$$

where O_i is the observed frequency for a bin i and E_i is the expected frequency for a bin i which is calculated by:

$$E_i = F(x_2) - F(x_1) \tag{5}$$

where F is the cumulative distribution function of the probability distribution being assessed and x_1, x_2 are the limits for bin i.

5. RESULTS AND DISCUSSION

Estimation of the AEP flood quantiles Q_{50}, Q_{20}, Q_{10}, Q_5, Q_2 & Q_1 for the 103 catchments located in SW WA was performed using the R software package for normal distribution, and TUFLOW Flike, for the log normal (LN), log-Pearson type 3 (LP3), generalised extreme value (GEV), extreme value type 1 (EV1) & generalised pareto (GP) distributions. The Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and Chi-squared goodness of fit (GoF) tests were applied to each catchment to determine the best fit distributions for each catchment. The probability distributions were then ranked according to the GoF test results. On review of the GoF test results, it was decided to exclude the results obtained from the Chi-squared test as it was inconsistent.

Ranking of the distributions based on the GoF test results was performed through the use of a consistent ranking system. Liu (2011) describes that there are many ranking systems available but to ensure consistency it is best to maintain one ranking system for a project. For this study, the

competitive ranking system has been used for the ranking of the GoF test results. The competitive ranking system ranks results numerically in descending order with ties in results being given the same ranking but any results that follow these tied results receiving a skipped ranking value. For example, in a ranking system of 5, 6, 6 and 7, the first result would receive a rank of 1, the next two results would receive a rank of 2 and the final result would receive a rank of 4, skipping rank 3. This system has been chosen due to its simplicity and that there are only 6 distributions for ranking. If more distribution were to be used then a more in-depth ranking system may be required. For the ranking scores found within this system a rank of 1 refers to the best fitting distribution result and a rank of 6 refers to the worst fitting distribution result.

The overall ranking for the six probability distributions for the 103 catchments is summarised in Table 1. From these rankings, it can be seen that the GEV provides the overall best choice of probability distribution for FFA in the SW of WA. This is followed by the LP3, the LN, the GP, the EV1 and normal distribution.

Table 1. Overall ranking of probability distributions

Distribution	Average Ranking	Number of Catchments to Achieve Rank 1 with Associated Distribution
GEV	1.7	63
LP3	2.4	33
LN	3.1	7
GP	3.71	16
EV1	3.73	9
Norm	5.3	1

From the results presented in Table 1 it can be seen that GEV is the best fit distribution for 63 out of the 103 sites, i.e. for 61% of the sites. This finding is consistent with Vogel, McMahon & Chiew (1993) who found (from their study of 7 catchments located in Drainage Division 6) that the GEV was the best-fit distribution. In sub-region 2 (Albany coast) we find that the LP3 distribution performs the best with a rank of 1 for the three sites, and GEV is ranked 2nd in this sub-region. Sub-regions 6 (Shannon River), 11 (Preston River) and 13 (Harvey River) have GEV as the rank 1 distribution. Sub-regions 6 and 11 have EV1 as the 2nd ranking distribution, and sub-region 13 has LP3 as the 2nd ranking distribution. We can see from Table 1 that the GP distribution is the best fit distribution for 16 sites, and all of these are located along the South West coastline of Drainage Division 6. Only one site shows normal distribution as the best-fit distribution, which is as expected as normal distribution is generally a poor descriptor of annual maximum flood data.

6. CONCLUSION

A study of the selection of the best fit probability distributions for flood frequency analysis is performed for South West of Western Australia (Drainage Division 6). For this study, the observed annual maximum (AM) streamflow data are analysed through the use of the R software and Tuflow FLIKE to estimate flood quantiles. A total of six probability distributions are examined: normal (Norm), log-normal (LN), log-pearson type 3 (LP3), generalised extreme value (GEV), extreme value type 1 (EV1) and generalised Pareto (GP) distributions. The Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) goodness of fit (GoF) tests are applied at each of the study catchments to assess the goodness of fit of the selected distributions. It is found that the GEV distribution provides the best outcome where 63 out of the 103 sites, i.e. for 61% of the sites have GEV as the best-fit distribution. The second best performer is the LP3 distribution, which is the best-fit distribution for 33 sites (i.e. for 32% of the sites). The results of this study should be used with caution as there are 39% of the catchments where GEV is not the best choice in South West of Western Australia and hence an alternative distribution like LP3 may provide more accurate design flood estimates.

REFERENCES

Bardsley WE (1989). Using historical data in nonparametric flood estimation, Journal of Hydrology, 108, 249-255.

Bureau of Meteorology (BOM) (2012). Australian Water Resources Assessment 2012, <http://www.bom.gov.au/water/awra/2012/swcoast.shtml>.

Cunnane C (1978). Unbiased plotting position - a review, Journal of Hydrology, 37, 205-222.

Cunnane C (1985). Factors affecting choice of distribution for flood series, Hydrological Sciences Journal, 30, 25-36.

Caballero WL, Rahman A (2014). Application of Monte Carlo simulation technique for flood estimation for two catchments in New South Wales, Australia, Natural Hazards, 74, 1475-1488.

Haddad K, Rahman A (2010). Selection of the best fit flood frequency distribution and parameter estimation procedure: a case study for Tasmania in Australia, Stochastic Environmental Research and Risk Assessment, 25, 415-428.

Heo J-H, Shin H, Nama W, Oma J, Jeong C (2013). Approximation of modified Anderson–Darling test statistics for extreme value distributions with unknown shape parameter, Journal of Hydrology, 499, 41-49.

Institution of Engineers Australia (I. E. Aust.) (1987). Australian Rainfall and Runoff: A Guide to Flood Estimation, Editor: Pilgrim, D.H., Engineers Australia, Canberra.

Kuczera G, Franks S (2016). At-site flood frequency analysis. In: Australian Rainfall & Runoff, Chapter 2, Book 2, edited by Ball et al., Commonwealth of Australia.

Liu TY (2011). Statistical Learning Theory for Ranking, Learning to Rank for Information Retrieval, Springer-Verlag, Berlin, pp. 195-200.

Merz B, Thieken AH (2009). Flood risk curves and uncertainty bounds, Natural Hazards, 51, 437-458.

Mirzaei M, Huang YF, El-Shafie A, Chimeh T, Lee J, Vaizadeh N, Adamowski J (2015). Uncertainty analysis for extreme flood events in a semi-arid region, Natural Hazards, 78, 1947-1960.

Rahman AS, Rahman A, Zaman MA, Haddad K, Ahsan A, Imteaz M (2013). A study on selection of probability distributions for at-site flood frequency analysis in Australia, Natural Hazards, 69, 1803-1813.

Rahman A, Haddad K, Kuczera G, Weinmann PE (2016). Regional flood methods. In: Australian Rainfall & Runoff, Chapter 3, Book 3, edited by Ball et al., Commonwealth of Australia.

Schendel S, Thongwichian R (2015). Flood frequency analysis: confidence interval estimation by test inversion bootstrapping, Advances in Water Resources, 83, pp. 1-9, DOI 10.1016/j.advwatres.2015.05.004.

Vogel RM, McMahon TA, Chiew FHS (1993). Floodflow frequency model selection in Australia, Journal of Hydrology, 146, 421-49.

Wang C (2015). A joint probability approach for coincidental flood frequency analysis at ungauged basins confluences, Natural Hazards, 82, 1727-1741.

Yan H, Moradkhani H (2015). Toward more robust extreme flood prediction by Bayesian hierarchical and multimodeling, Natural Hazards, 81, 203-225.