Flood Frequency Analysis

Marie Biolková
(s1653813, B097372)
and
Theodora Karaoli
(s1624799, B104830)

Year 4 Project
School of Mathematics
University of Edinburgh
September 2019

Abstract

The design of structures for flood control and water management depends heavily on estimation of peak streamflow, typically using observed extreme values from the past. The probability that a flood of a given magnitude will occur is inversely related to its return period, so predicting high return periods requires an extensive length of records which is often unavailable. Statistical procedures, such as regional flood frequency analysis, have been developed to deal with this issue and hence provide reliable risk assessment for regions affected by floods. The aim of this study was to perform at-site flood frequency analysis to determine the most appropriate fit for estimating the annual peak discharge at Pearl river in the United States. Eight candidate probability distributions and three methods of estimation were considered and compared. The goodness-of-fit tests adopted were the Kolmogorov-Smirnov test, the Anderson-Darling test, Root-mean-square error, the Akaike information criterion, and the Bayesian information criterion, in addition to visual assessment made by the L-moments ratio diagram. It was found that Log-Pearson III and Gumbel distribution provided the most accurate fit for three out of eight stations each. However, the sample L-moment ratios suggest that Weibull distribution, which was selected as most appropriate for two stations, is also a reasonable choice and therefore should also be considered for predicting design floods in the region.

Contents

Chapter 1

Introduction

Flood is a natural disaster that usually occurs after a momentary overflow of water from rivers, lakes, oceans onto typically dry land. It may happen after rain, snow, coastal storms, storm surges, overflows of dams and other water systems. A flood may result in damage of buildings, outage of electricity, disruption of transportation, creation of landslides, injuries or even death. For instance, a flood of Mississippi river in 1927 resulted in the death of 246 people and in 1993, the cost after the Midwest flooding arose to $30.2 billion and 48 deaths [6]. Floods may also have impact on the environment through the disruption of natural ecosystems and pollution. Particularly, in their publication on floods, the Queensland government in Australia reports that crops can be damaged after flooding in agricultural production areas, habitat may be lost and pollutants may be released. Moreover, they warn that indirect costs may also arise after a flood, such as stress, anxiety and disruption of living of the residents of the flooded area. However, floods can also be beneficial: they recharge dry areas by providing water for irrigation and resources of drinking water, and improve the quality of soil for agriculture. For instance, the fish production can be enhanced, provided that the land is not polluted after human's interruption, as the nutrients supplied from the land during a flood are beneficial for the fish [49]. Kipkemboi et al. (2010) have specifically investigated how the fish production can be increased, taking advantage of nature and floods [33]. Since the agricultural sector is the main source of economic growth of numerous and particularly developing countries, many populations rely on regular flooding because it makes their land more fertile and consequently increases their production. Flood frequency analysis (FFA) is widely used to help sidestep all the devastating consequences as well as to maximise the benefits that arise from floods in arid regions.

Flood frequency analysis is a method for predicting design floods along a river and the corresponding water discharge through the modelling of observed peak streamflow from the past. This involves finding the distribution that best fits the historical data which is then used to determine the likelihood of a future flood of a given magnitude. The information on the magnitude and recurrence of floods provided by frequency analysis is crucial for the design of bridges, dams, culverts, flood control structures, sewage disposal plants and for many other engineering and economic purposes. Inaccurate estimates are bound to cause safety issues or

additional costs may arise due to necessary repairs. The design of most projects is based on past extreme events measured at multiple sites within a region of interest. This ensures that the conclusions are more robust and allows us to create a regional curve that can be applied to any site in the basin.

In this report, we will lay theoretical grounds and subsequently discuss flood frequency analysis at-site which was performed using data provided by the U.S. Geological Survey's (USGS) National Water Information System. Flood frequency analysis comprises of the two main components: parameter estimation and goodness-of-fit assessment. These steps are repeated for all candidate distributions in order to find the one that describes the data most accurately, i.e. achieves the best results in goodness-of-fit tests. The most common distributions used in flood frequency analysis were considered here – these are described in detail in Chapter 4. Chapter 3 is dedicated to three different estimation methods that were explored to identify the parameter estimates, namely maximum likelihood estimation, method of moments and probability weighted moments method. For evaluation we used five goodness-of-fit tests which we introduce in Chapter 5. The purpose of this report was to conduct a regional analysis for the basin of Pearl River, Mississippi in the United States by repeating the above process for multiple stations and hence finding a frequency curve that generally provides the most accurate estimates of recurrence periods for the region of interest.

1.1 Historical background

In this section, we aim to explore the historical background of FFA, outline major milestones and figures in the research as well as highlight key literature. Some parts of this section were based on Rumsey (2015), a paper about the evolution of FFA in the United States.

Several engineers in the early 1900s were trying to predict flood magnitudes and decide what safety precautions could be taken. For instance, five flood control dams were contracted in Dayton, after the great flood of 1913 in the area. Many civil engineers of the time were trying to guess when and how large the next floods would be and were suggesting further constructions to be made. However, no mathematical justification was used to examine the floods, nor statistical predictions were made to help motivate the necessary precautions.

The first paper on the estimation of flood frequencies and magnitudes of rivers, based on statistical approaches, was published in 1914 by the civil engineer W. Fuller. In contrast to previous publications, Fuller's work was revolutionary because of the way it dealt with the challenge of very low number of data. Since, at that time, historical data of floods over a large enough time period were not available, Fuller derived formulae for predicting flood magnitudes which could be used with a limited amount of data from many rivers. The estimates were based on the yearly average flood of a river, so that predictions could be made even with data of only 15 years time. Fuller's work was later refined by one

of his close colleagues, A. Hazen (1930) who was himself an important contributor to the FFA research.

The government, as well as many universities, have then expressed their interest in flood probability. The Flood Control Act of 1936 stated that the Federal Government should get involved in flood control through the construction of the required infrastructure if the costs are exceeded by the benefits or if the security of the people is put at risk. National Flood Insurance Program was created following changes in legislation in 1968, making protection against losses from flooding damage available for the first time. Later in 1973, the Flood Disaster Protection Act stated that flood insurance was required for all projects located in areas with high risk of flooding or mudflow, as identified by the Federal Government [47]. These changes in legislation made FFA attractive for actuarial purposes in the United States.

Flood frequency analysis has been a widely researched topic. An important figure was White (1945) who emphasised how humans should adjust to the flooding danger. Moreover, Hosking (1985 & 1986) discussed the theory of parameter estimation via probability weighted moments and its application to the generalized extreme value distribution, before introducing L-moments as an alternative for parameter estimation in frequency analysis [27]. Hosking also co-authored a popular guide on regional frequency analysis based on L-moments [28]. Kite (1977) first introduced and popularized the use of frequency factors and derived the experssions for some of them. V. T. Chow has also played a major role in the field of hydrology, as he wrote several, widely known books on the topic and he was the founder of the International Water Resources Association [1]. We will often refer to the publications of the these scientists in this project.

Chapter 2

Flood Frequency Analysis

Frequency analysis of hydrologic data aims to relate the magnitude of extreme events to their frequency of occurrence through the use of probability distributions [12]. To apply this analysis, we use the annual peak discharge series, and we assume that the collected data are independent, identically distributed and stationary. This chapter will cover the terminology and methodology of flood frequency analysis.

2.1 Terminology

Before we begin with any theoretical and technical aspects we introduce the terminology relevant to flood frequency analysis and establish the corresponding mathematical notation that we will use in this project. This section is based on Chow (1988).

First, define the recurrence interval τ as the time it takes for an event with magnitude x_T to be repeated or exceeded. In other words it represents the time between occurrences of $X \geq x_T$. The corresponding return period T is the average length of the recurrence interval, i.e. $T = \mathbb{E}(\tau)$.

The probability of recurrence interval τ is similar to a geometric distribution with probability function $p(\tau) = (1-p)^{\tau-1}p$ for $0 \leq p \leq 1$, with mean $\mathbb{E}(\tau) = \frac{1}{p}$ (although τ does not need to be discrete). Therefore, the exceedance probability $p = \mathrm{P}(X \geq x_T)$ is equal to the inverse of the return period, $\frac{1}{T}$.

The event magnitude x_T can be determined using quantile functions or frequency factors. If we express x_T as the mean μ of the historical data with some difference Δx_T from the mean, we can calculate Δx_T as the product of the frequency factor K_T and the standard deviation σ. Hence the magnitude can be represented as

$$x_T = \mu + K_T\sigma \quad \text{or estimated by} \quad \widehat{x}_T = \bar{x} + K_T s. \tag{2.1}$$

where $\bar{x} = \sum_{i=1}^{n} x_T^{(i)}$, $s = \sqrt{\sum_{i=1}^{n} \left(x_T^{(i)} - \bar{x}\right)^2 / (n-1)}$ is the sample standard deviation and n is the sample size.

The value of K_T depends on the return period T and the distribution choice. If the cumulative distribution function (cdf) of a distribution has a closed form inverse function, K_T can be found analytically, otherwise numerical approximations are used to calculate K_T. An example of deriving K_T analytically is given for the exponential distribution in Section 4.2.1.

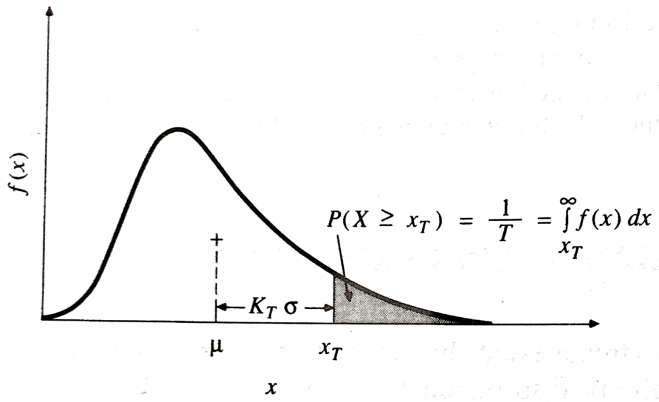


Figure 2.1: Diagram showing the event magnitude x_T, the frequency factor K_T and the exceedance probability p [12].

2.2 Methodology

Any task involving data, including FFA, starts by exploratory analysis. It is important to familiarise ourselves with the structure of the data through plots and numerical summaries, and detect possible anomalies that could have a detrimental effect on the analysis. There are multiple possible approaches that have been used in FFA case studies for detecting outliers. We will describe two commonly used methods.

The U.S. Water Resources Council (1981) recommends the following approach. Since the distribution of annual maximum discharge is significantly positively skewed, we first consider high outliers and then low ones. A high outlier is determined as an observation exceeding the threshold

$$x_H = \bar{x} + K_N s, \tag{2.2}$$

where $\bar{x}$ is the mean of the observations transformed to log-scale (not considering zero magnitudes or previously detected outliers), s is their standard deviation and K_N is a constant that depends on the sample size proposed by the U.S. Water Resources Council (Appendix 4, 1981) – it is essentially a one-sided value for the normal distribution with significance level $\alpha = 0.1$. Threshold for low outliers is

$$x_L = \bar{x} - K_N s. \tag{2.3}$$

This approach gives very high thresholds. Alternatively, the interquartile range (IQR), which was chosen to be used in this case, may be used for outlier detection.

If we denote Q_1 and Q_3 the first and third quartiles of the data, respectively, then IQR equals $Q_3 - Q_1$ and an outlier would satisfy

$$x < Q_1 - C \times \text{IQR} \quad \text{or} \quad x > Q_3 + C \times \text{IQR}. \tag{2.4}$$

The constant C is typically set to 1.5, however in the case of annual peak values, only very extreme outliers should be removed. Therefore, we use $C = 3$.

The exceedance probabilities are estimated based on the observed data ranked in order of magnitude to obtain the plotting position. These are subsequently plotted and used to estimate the return periods which is a standard procedure in analysis of extreme meteorological events. Many different plotting position formulas have been proposed in the past, for instance by Hazen (1914). The standard method adopted by the U.S. Water Resources Council (1981), which is the Weibull plotting position, was applied here. As proposed by Weibull (1939), the return period is estimated by:

$$P(X \geq x_i) = \frac{i}{n+1} \quad \Longrightarrow \quad T = \frac{n+1}{i}, \tag{2.5}$$

where x_i is the i^{th} largest value and n is the number of years for which data is available after removing outliers. An example of application of the formula can be found in Table 2.1. It has been proven that other estimation formulas are biased and consequently underestimate the exceedance probabilities, thus the Weibull formula is a natural choice in extreme value analysis [41].

x_i	$p_i = \frac{i}{n+1}$	$T = \frac{1}{p_i}$	$\widehat{x}_T = \bar{x} + K_T s$
22600	0.02	50	22338
⋮	⋮	⋮	⋮
2700	0.98	1.02	2953

Table 2.1: Example of the use of the Weibull formula with the normal distribution.

The next step is parameter estimation for which three different methods were considered: maximum likelihood, method of moments and probability weighted moments. Chapter 3 contains more details about the theory behind each of these methods. The estimated parameters determine the fitted distribution and we can predict the event magnitude by finding the quantile corresponding to the return period T. In other words, we find x_T such that

$$P(X \geq x_T) = p \iff P(X \leq x_T) = 1 - p = 1 - \frac{1}{T}. \tag{2.6}$$

An alternative approach is to use the frequency factor of the distribution as discussed in Section 2.1.

In order to visually assess the fit, quantile plots may be considered. These compare the empirical quantiles to the theoretical quantiles of the fitted distribution. Alternatively, the predicted discharge can be plotted against the return period

together with the collected data to which we have assigned the plotting position. However, we also need a formal and systematic way to compare the quality of fitted models and therefore we conducted various goodness-of-fit tests, to which Chapter 5 is dedicated. These tests are used to determine the best combination of distribution and estimation method for a specific dataset.

Finally, the aim is to determine whether there exists a combination of distribution and estimation method that can be reliably applied to all (or most) stations in the region. A common way to do is by plotting the L-moments ratios, as described in Section 5.6. As a result, we obtain a regional curve that can be used to estimate the magnitude of floods in the catchment area.

In practice, the sample size of flood data is typically not sufficient to guarantee reliable estimates. It is therefore of great importance to evaluate the accuracy of the predictions for design purposes. The uncertainty of estimated flood magnitude increases with the return period – this is obvious since a 100-year flood is rare and thus more difficult to estimate than a 5-year flood. Rao and Hamed (2000) described the process of quantifying the standard error of estimates for an event with a return period T, for each of the parameter estimation methods used here. They also noted that while the standard error accounts for the lack of data, it does not take the choice of an unsuitable distribution into consideration. Hu et al. (2013) applied bootstrap to the original extreme values in order to determine the uncertainty of the point estimates. Although the importance of confidence intervals for hydro-engineering is undeniable, it is outside the scope of this project due to time constraints and could potentially extend the research on this topic.

In summary, we gave an overview of the process of carrying out flood frequency analysis. The next few chapters provide a background on the theory of the key two components, parameter estimation and goodness-of-fit evaluation.

Chapter 3

Estimation Methods

Maximum likelihood, ordinary moments and probability weighted moments (or L-moments) are methods widely used in frequency analysis to estimate parameters of the distributions that are selected for the given extreme data. Some work better with specific distributions in terms of accuracy and efficiency. The best combination of distribution and estimation method is likely to vary between different datasets. Consequently, multiple gauging stations should be considered in order to find the distribution and estimation method that best models the regional floods. For example, an analysis of hundreds of Norwegian stations showed that the estimation through L-moments is recommended, in particular with the Gumbel distribution. It was also concluded that the method of moments was the most stable one, unlike the maximum likelihood method which should be avoided due to its instability with three-parameter distributions [35]. In this chapter we introduce the aforementioned methods for obtaining parameter estimates and their standard errors, as well as highlight their key properties, advantages and disadvantages.

3.1 Maximum Likelihood Estimation (MLE)

Maximum likelihood is probably the most popular approach for estimating parameters of a distribution. Although the concept of maximum likelihood had been known and used by mathematicians such as Gauss or Laplace, it was first introduced as a method of estimation in 1922 by Ronald A. Fisher after spending almost a decade finding justifications for its validity [5]. Formally, he defined the MLE of a parameter θ as the value $\widehat{\theta}$ that maximises the likelihood $L(\theta|\mathbf{x})$. In other words, the method finds the parameter that makes the observed values $x_1, ..., x_n$ as likely as possible. If the data are independent, it is easier to work with the joint log-likelihood rather than the joint likelihood – thanks to the logarithm, we work with a sum instead of a product. In order to find $\widehat{\theta}$ we set the first derivative of the (log)likelihood to zero and solve for θ [69]. For some distributions there is no analytical solution and one has to resort to numerical optimisation.

The resulting estimators are consistent and therefore provide reliable estimates if

the sample size n is sufficiently large. In addition, there is no consistent estimator that is more efficient – MLE has the smallest standard error for large enough n. In particular, the variance is bounded below by the Cramér–Rao bound

$$\text{Var}(\widehat{\theta}) \geq I(\theta)^{-1} \tag{3.1}$$

where $I(\theta) = \mathbb{E}\left(\left[\frac{d \log L(\theta|\mathbf{x})}{d\theta}\right]^2\right)$ is the observed Fisher information. Conveniently, under very general conditions, it can be evaluated as $I(\theta) = -\mathbb{E}\left(\frac{d^2 \log L(\theta|\mathbf{x})}{d\theta^2}\right)$. Note that in the multiparameter case, $I(\theta)$ becomes a matrix with $(i,j)^{\text{th}}$ entry equal to $-\mathbb{E}\left(\frac{\partial^2 \log L(\theta|\mathbf{x})}{\partial\theta_i \partial\theta_j}\right)$.

As $n \to \infty$, the variance of the MLE approaches the bound in (3.1) and therefore the Cramér–Rao bound can be used as an approximation for $\text{Var}(\widehat{\theta})$ for large samples:

$$\widehat{\text{Var}}(\widehat{\theta}) = I(\widehat{\theta})^{-1}. \tag{3.2}$$

It can also be shown that $\widehat{\theta}$ is asymptotically normally distributed around the true value θ_0 such that

$$\widehat{\theta} \sim N\left(\theta_0, I(\theta_0)^{-1}\right) \quad \text{as } n \to \infty.$$

Another important property of the maximum likelihood is the invariance property, which states that if $\widehat{\theta}$ is the MLE for θ then $g(\widehat{\theta})$ is the MLE for $g(\theta)$ (if g is bijective) [69].

3.2 Method of Moments (MOM)

The method of moments involves equating the theoretical moments of a distribution with the sample moments and subsequently solving for the parameters. It was first introduced by Karl Pearson in 1902 who defined the the k^{th} theoretical moment about the origin (raw moment) as

$$\mu_k = \mathbb{E}\left(X^k\right) = g_k(\theta_1, ..., \theta_k).$$

Note that the first raw moment is the mean $\mu = \mathbb{E}(X)$. The method of moments estimators (or the sample moments) $\widehat{\theta}_1, ..., \widehat{\theta}_k$ can be found by solving the equations

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

This method yields consistent estimators, however they may not belong to the parameter space and may not provide a sufficient statistics of the data. It is also required that the distribution has finite moments. Although straightforward to apply, this method can produce severely biased estimates if the data is limited and the number of parameters to estimate is large (more than two) [51]. If the

likelihood function is complicated, MOM may be computationally cheaper than MLE [68].

The variance of the MOM estimate for the parameter τ can be approximated by the Delta method. Using a (first-order) Taylor expansion, it estimates the standard errors of transformations of random variables. Suppose we wish to find the error of $\tau = \varphi(\theta)$ and assume we know the standard error of the estimate of θ, $\widehat{\theta}$. This is typically obtained by maximum likelihood estimation, since the inverse of the Fisher information provides an estimated variance of the parameter estimate (Equation 3.2). Then for an estimate of τ the Delta method approximates its variance as [62]

$$\mathrm{Var}(\widehat{\tau}) = \left|\varphi'(\widehat{\theta})\right|^2 \widehat{\mathrm{Var}}(\widehat{\theta}).$$

For the multiparameter case where $\tau = \varphi(\theta_1, ..., \theta_k)$ this becomes

$$\mathrm{Var}(\widehat{\tau}) = (\widehat{\nabla}\varphi)^T I(\widehat{\theta})^{-1}(\widehat{\nabla}\varphi),$$

where $\nabla\varphi = \begin{pmatrix} \frac{\partial\varphi}{\partial\theta_1} & \dots & \frac{\partial\varphi}{\partial\theta_k} \end{pmatrix}^T$ is the gradient of φ, $I(\widehat{\theta})$ is the expected Fisher information matrix at the MLE and $\widehat{\nabla}\varphi = \nabla\varphi|_{\theta=\widehat{\theta}}$.

Although straightforward, evaluating the moments involves calculating integrals which can be problematic if they are not tractable. An alternative in such situations is to find the moment-generating function

$$M_X(t) = \mathbb{E}\left(\exp(tX)\right) = \int_{-\infty}^{\infty} \exp(tx) f_X(x)\mathrm{d}x.$$

It is defined for all t for which the expected value exists. Furthermore, provided that the k^{th} moment exists, it is equal to the k^{th} derivative of the moment-generating function evaluated at $t = 0$, i.e.

$$M_X^{(k)}(0) = \mathbb{E}\left(X^k\right).$$

The moment-generating function uniquely defines a distribution. It also gives us all moments of the distribution.

For higher order-moments, a central moment (centered around the mean) is preferred. The k^{th} central moment about the origin is given by

$$\mu_k^{(c)} = \mathbb{E}\left((X - \mathbb{E}(X))^k\right).$$

Consequently, $\mu_0^{(c)} = 1, \mu_1^{(c)} = 0$ and $\mu_2^{(c)} = \mathrm{Var}(X) = \sigma^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. The skewness and kurtosis are defined using the third and fourth central moments, respectively. In particular, the skewness is the third standardized moment defined

as

$$\tilde{\mu}_3 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3^{(c)}}{\sigma^3} = \frac{\mathbb{E}\left[(X-\mu)^3\right]}{\left(\mathbb{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\mathbb{E}\left(X^3\right) - 3\mu\sigma^2 - \mu^3}{\sigma^3}.$$

The sample skewness C_s can be obtained by replacing the moments in the above formula by sample moments [51].

3.3 Probability Weighted Moments (PWM)

The probability weighted moments were first introduced by Greenwood et al. (1979) as an alternative to the traditional moments, particularly for distributions that can be expressed in an inverse form. The PWMs of a distribution $F(x) = P(X \leq x)$ with inverse $x(F)$[1] are formally defined as

$$M_{i,j,k} = \mathbb{E}\left[X^i F^j (1-F)^k\right] = \int_0^1 [x(F)]^i F^j (1-F)^k dF, \quad i,j,k \in \mathbb{R}.$$

Notice that $M_{i,0,0}$ gives the i^{th} conventional (raw) moment for $i > 0$. Usually, i is chosen to be 1 for this ensures x only enters through the first power, resulting in a simpler form than the standard moments. Furthermore, if $x(F)$ is continuous and $M_{i,0,0}$ exists, then for non-negative integers j and k, $M_{i,j,k}$ exists. In addition, if the inverse function has a closed form, it can be solved analytically.

Naturally, Hosking (1986) suggested the use of either $M_{1,j,0}$ or $M_{1,0,k}$ [26]. This is because for $i, j \in \mathbb{Z}^+$, the term $F^j(1-F)^k$ reduces to powers of F of $1-F$ and therefore $M_{1,0,k}$ can be expressed as a function of $M_{1,j,0}$ and vice versa:

$$\alpha_k = M_{1,0,k} = \sum_{j=0}^{k} \binom{k}{j} (-1)^j M_{1,j,0} \tag{3.3}$$

$$\beta_j = M_{1,j,0} = \sum_{k=0}^{j} \binom{j}{k} (-1)^k M_{1,0,k} \tag{3.4}$$

The name of the method refers to the use of F^j or $(1-F)^k$ for weighting the ordered observations. If the powers of the cumulative distribution are used, more weight is assigned to large observations. Using the powers of the complement of the cumulative function $1-F$ to apply weighting results in the opposite, i.e. large weight is put on small observations. Both approaches yield the same parameter estimates [52]. The choice of either (3.3) or (3.4) depends on the specific distribution, as one might be easier to work with than the other.

Although similar to MOM, PWMs are less sensitive to outliers, as the observed values are only taken to the first power. While PWM estimators are not severely

[1]The inverse is probably more commonly denoted $F^{-1}(x)$ but the notation $x(F)$ chosen by Greenwood is convenient for PWMs. We will use these notations interchangeably.

biased, have low variance and tend to outperform maximum likelihood estimates [25], they do not have any interpretable meaning. For this reason L-moments, which are essentially linear combinations of PWMs, were introduced. Unlike PWMs, L-moments are useful quantities for summarizing the location, scale and shape of a distribution.

3.3.1 Linear moments (L-moments)

Linear moments were first introduced by Hosking (1990) who defines them as the expectation of linear combinations of order statistics, $X_{1:n} \leqslant X_{2:n} \leqslant \ldots \leqslant X_{n:n}$. Provided that the mean of the random variable X is finite, the L-moments exist and are given by

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} \mathbb{E}(X_{r-k:r}), \quad r = 1, 2, \ldots \tag{3.5}$$

These quantities fully characterize a distribution. The first four L-Moments (λ_i) can be written as a linear combination of the PWMs α_j, β_k, for $i \in \{1, 2, 3, 4\}$, $j, k \in \{0, 1, 2, 3\}$ and are given by

$$\lambda_1 = \alpha_0 = \beta_0 \tag{3.6}$$
$$\lambda_2 = \alpha_0 - 2\alpha_1 = 2\beta_1 - \beta_0 \tag{3.7}$$
$$\lambda_3 = \alpha_0 - 6\alpha_1 + 6\alpha_2 = 6\beta_2 - 6\beta_1 + \beta_0 \tag{3.8}$$
$$\lambda_4 = \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0. \tag{3.9}$$

We shall denote l_i the sample estimate of the i^{th} L-moment. Analogously to standardised moments, we may define standardised L-moments $\tau_k = \lambda_k/\lambda_2$ for $k = 2, 3, \ldots$. These are referred to as L-moment ratios and we call the first three (τ_2, τ_3 and τ_4) coefficient of L-variation, L-skewness and L-kurtosis, respectively. Selection of an appropriate distribution in regional frequency analysis often involves plotting L-moment ratio diagrams of the candidate distributions and comparing them to the sample L-moment ratios [48]. This visual goodness-of-fit method will be discussed in detail in Chapter 5.

Similarly to PWMs, the benefits of L-moments lie in their robustness with respect to outliers (compared to traditional moments) and their reliability for inference-making when the sample size is small, due to their low bias. In some cases the resulting estimates may be more efficient than the ones obtained by MLE [27].

Distributions without an explicit definition for the quantile function $x(F)$, such as normal or Pearson, can not be easily defined over PWMs or L-moments. Hosking (1990, Table 1 and Table 2) derived expressions for L-moments of some common distributions and provided methods for estimating their parameters.

Chapter 4

Distributions

The choice of distribution for flood frequency analysis poses a challenge for hydrologists, and it has been a popular topic in research. Annual peak streamflow data tend to be skewed. A variety of distributions can be fitted to skewed data, however each of these also provide different tail shapes. Because the tail of the distribution corresponds to large return periods, failing to model it accurately could result in poor choices in practice, such as building insufficient infrastructure and risking costly damages, or overspending on unnecessary measures in the case of a false positive.

In 1989, World Meteorological Organisation published a report which summarises the relevant research concerning the problematics of distribution choice in flood frequency analysis, and provides a guidebook for hydrologists who study annual maxima or partial duration series. Some of the candidate distributions include log-normal, (log)-Pearson type III, gamma, generalized extreme value (GEV), generalized logistic and Wakeby [14].

In recent at-site analyses, the GEV distribution was found to provide the most reliable results in an on-site case study in Norway [35], log-Pearson III and GEV were in the top three best fitting distributions following a case study in Australia [50], EV1 distribution (Gumbel) was deemed most suitable for Malakkara and Neeleswaram in India [64], log-Pearson III performed best on data from Minab river in Iran [32], and the normal distribution had the highest coefficient of determination and minimum root mean square error (RMSE) when the Weibull plotting position was used in the flood analysis of Osun river in Nigeria [2].

Here we considered three families of distributions: Normal, Gamma and GEV. The intention was to perform the analysis with some well-known distributions first (normal or exponential) to become confident with the methods, before exploring the potential of distributions that are specific to hydrology. For each estimation method, we will derive the parameter estimates for one distribution from each family, focusing on the less common ones. Particularly, the maximum likelihood and moments estimators have been derived for log-normal distribution from the Normal family, for Pearson distribution from the Gamma family and for Gumbel distribution from the GEV family. The probability weighted moments estimators

have been derived for Weibull distribution from GEV family. The results for the remaining distributions have only been stated since the derivations are often similar.

4.1 Normal family

4.1.1 Normal distribution

The normal distribution is the most important and common statistical distribution. It can represent for instance the distribution of height or IQ of people. Many natural phenomena approximately follow or can be transformed to follow a normal distribution. It is very well tabulated and easily accessible through statistical software, therefore convenient to use. Furthermore, the validity of many statistical methods, such as regression models and hypothesis testing, depends on the assumption that the data are normal, or alternative approaches must be sought. Often it is more convenient to perform transformations in order to make the data approximately normal.

The probability density of the normal distribution is

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}$$

where μ is the mean or the expected value of the distribution and σ^2 is the variance. The standard deviation, $\sigma > 0$, is the distance of the mean from the inflection point of the curve. When the mean, μ, and the variance, σ^2, are 0 and 1 respectively then the distribution is called standard normal. Standardization of distributions is commonly used as it is easier to calculate probabilities. The aim is to transform the distribution to a standard normal with mean $\mu = 0$, and variance $\sigma^2 = 1$, which can be done using $Z = \frac{X-\mu}{\sigma}$.

The normal distribution has some very useful properties. It is symmetric about the mean, so the skewness is zero and consequently the mean, median and mode coincide. Furthermore, the Central Limit Theorem tells us that if we add together many independent random outcomes with finite mean and variance, the resulting distribution is approximately normal. The individual random variables do not need to be normally distributed. Consequently, some variables in hydrology are assumed to follow the normal distribution because they come from the addition of independent variables, e.g. annual precipitation, annual average streamflow [12].

Despite the quantile function being intractable, there are many resources for obtaining the desired values, including tables or computer software. The frequency factor is the standard normal variable, $K_T = z$. It can be found in `R` using the command `qnorm(p, 0, 1)`. Then the estimated discharge is $\widehat{x}_T = \widehat{\mu} + \widehat{\sigma}z$, obtained from Equation 2.1 in Chapter 2.

Maximum Likelihood Estimation

The ML estimates of the normal distribution $N(\mu, \sigma^2)$ parameters are [69]:

$$\widehat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Method of Moments

The parameter estimates obtained via the moments method yield identical results to the MLEs above. A detailed derivation of this result can be found in Rao and Hamed (2000).

Probability Weighted Moments

The inability to explicitly define the quantile function $x(F)$ makes PWM estimation difficult. Nevertheless, Hosking (Table 1, 1990) derived the following properties of the normal distribution:

$$\mu = \lambda_1 \quad \sigma = \sqrt{\pi}\lambda_2.$$

This allows us to estimate the parameters through sample L-moments, i.e. $\widehat{\mu} = l_1$, $\widehat{\sigma} = \sqrt{\pi} l_2$.

4.1.2 Log-normal distribution

The log-normal distribution is a continuous statistical distribution of variables whose natural logarithms are normally distributed. Mathematically, if we let X follow a log-normal distribution then $Y = \log X \sim N(\mu, \sigma^2)$. This also means that a log-normally distributed variable can be converted to a normal distribution using $X = \exp(\mu + \sigma z)$ where z is the standard normal variable and μ, σ are the mean and standard deviation of $\log X$, respectively.

The ability to represent skewed data, which are common in real-life applications, is what makes this distribution so important. Moreover, its variance is proportional to its mean which makes it a good candidate for modelling economic phenomena, such as distribution of wealth in the society. The Central Limit Theorem (CLT) also implies that the product of random variables is asymptotically log-normal, if the conditions of CLT are satisfied. Because the population growth of animals and plants is multiplicative, the log-normal distribution plays an important role in ecology. The log-normal distribution also appears frequently in various biological mechanisms, it can model for example lengths of latent periods of infectious diseases or even the distribution of mineral resources in the Earth's crust [38]. Mingyang et al. (2017) used the log-normal distribution to automatically detect epilepsy [37]. Interestingly, the length of internet discussion

fora posts were also shown to closely follow the log-normal distribution [58].

We consider the two-parameter log-normal distribution, LN(2). The probability density function is given by

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{\frac{-\left(\log x-\mu\right)^2}{2\sigma^2}\right\} \quad x>0,$$

where $-\infty < \mu < \infty$ is the location, $\sigma > 0$ is the scale of the natural logarithm of the variable. The estimated flood magnitude is given by $\widehat{x}_T = \exp\left(\widehat{\mu} + z\widehat{\sigma}\right)$, using the standard normal frequency factor z.

Maximum Likelihood Estimation

To derive the likelihood estimates of the parameters, $\widehat{\mu}$ and $\widehat{\sigma}^2$ we first compute the joint likelihood function:

$$\begin{aligned} L(\mu,\sigma^2|X) &= \prod_{i=1}^{n}[f(X_i|\mu,\sigma^2)] \\ &= \prod_{i=1}^{n}\left(\frac{1}{X_i\sigma\sqrt{2\pi}}\exp\left\{\frac{-\left(\log X_i-\mu\right)^2}{2\sigma^2}\right\}\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\prod_{i=1}^{n}\left(\frac{1}{X_i}\exp\left\{\frac{-\left(\log X_i-\mu\right)^2}{2\sigma^2}\right\}\right). \end{aligned}$$

The log-likelihood is given by

$$l(\mu,\sigma^2|X) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\log X_i - \frac{\sum_{i=1}^{n}\left(\log X_i-\mu\right)^2}{2\sigma^2}.$$

Taking partial derivatives with respect to μ and σ^2,

$$\begin{aligned} \frac{\partial l}{\partial \mu} &= -\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}-2\left(\log X_i-\mu\right)\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\log X_i - \frac{n\mu}{\sigma^2} \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} - \frac{1}{2}\sum_{i=1}^{n}\left(\log X_i-\mu\right)^2\left(\frac{-1}{(\sigma^2)^2}\right) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(\log X_i-\mu\right)^2 \end{aligned}$$

Setting $\frac{\partial l}{\partial \mu} = 0$ and $\frac{\partial l}{\partial \sigma^2} = 0$ and solving for $\widehat{\mu}$ and $\widehat{\sigma}^2$,

$$\begin{aligned} \frac{1}{\sigma^2}\sum_{i=1}^{n}\log X_i - \frac{n\mu}{\sigma^2} = 0 \Rightarrow \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\log X_i, \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(\log X_i-\mu\right)^2 = 0 \Rightarrow \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\log X_i-\widehat{\mu}\right)^2. \end{aligned}$$

Method of Moments

If $X \sim LN(2)$ then $\log X = Y \sim N(\mu, \sigma^2)$, so the m^{th} moment $\mathrm{E}[X^m]$ satisfies

$$\begin{aligned}
\mathrm{E}\left[\exp(mY)\right] &= \int_{-\infty}^{\infty} \exp(my) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \mathrm{d}y \\
&= \exp\frac{m\left(2\mu + m\sigma^2\right)}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-(\mu+m\sigma^2))^2}{2\sigma^2}\right) \mathrm{d}y \\
&= \exp\frac{m\left(2\mu + m\sigma^2\right)}{2}.
\end{aligned}$$

Above we used the moment generating function of the normal distribution to find the moments of the log-normal distribution, which in fact does not have a moment generating function itself – it diverges for any positive number [23]. However, all moments of the log-normal distribution are finite.

Equating this with the first two moments

$$\begin{aligned}
\mu_1 = \mathbb{E}[X] &\Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i = \exp\left(\mu + \frac{\sigma^2}{2}\right) \\
\mu_2 = \mathbb{E}\left[X^2\right] &\Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \exp\left(2\mu + 2\sigma^2\right)
\end{aligned}$$

Solving for μ we obtain,

$$\widehat{\mu} = \ln\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \frac{\widehat{\sigma}^2}{2} = \frac{1}{2}\ln\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \widehat{\sigma}^2.$$

Solving for $\widehat{\sigma}^2$,

$$\begin{aligned}
\ln\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \frac{\widehat{\sigma}^2}{2} &= \frac{1}{2}\ln\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \widehat{\sigma}^2 \\
\widehat{\sigma}^2 &= \ln\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - 2\ln\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
\widehat{\sigma}^2 &= \ln\left(\sum_{i=1}^{n} X_i^2\right) - 2\ln\left(\sum_{i=1}^{n} X_i\right) + \ln(n)
\end{aligned}$$

Substituting back,

$$\begin{aligned}\widehat{\mu} &= \ln\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \frac{1}{2}\left(\ln\left(\sum_{i=1}^{n} X_i^2\right) - 2\ln\left(\sum_{i=1}^{n} X_i\right) + \ln(n)\right) \\ &= 2\ln\left(\sum_{i=1}^{n} X_i\right) - \frac{1}{2}\ln\left(\sum_{i=1}^{n} X_i^2\right) - \frac{3}{2}\ln(n)\,.\end{aligned}$$

Probability Weighted Moments

Although Hosking (1990) found expressions for the parameters using L-moments, it is more convenient to transform the data and apply the estimation for normal distribution and then follow the formulas given in Section 4.1.1.

4.2 Gamma family

4.2.1 Exponential distribution

The exponential distribution is the continuous counterpart of the geometric distribution and it is used to model inter-event times of stochastic processes (in particular homogeneous Poisson processes). Typical examples of its applications would be modelling the time until a radioactive particle decays or the arrival time of customers to a store. The probability density function is

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad \text{for } x > 0,$$

where $\lambda > 0$ determines the rate. The exponential distribution has the memoryless property, i.e. for a random variable T, it satisfies

$$\Pr(T > s + t | T > s) = \Pr(T > t), \quad \forall s, t \geq 0.$$

In other words, the distribution of the remaining waiting time stays the same no matter how much time has already elapsed. The sum of independent and identically distributed exponential random variables results in the gamma distribution.

The cumulative function of the exponential distribution is $F(x) = 1 - \exp(-\lambda x)$, so the inverse can be found analytically. Therefore following Equation 2.6 we have that $x_T = F^{-1}(1 - \frac{1}{T}) = \frac{1}{\lambda}\log T$. The frequency factor can be computed from Equation 2.1 by equating

$$\frac{1}{\lambda}\log(T) = \mu + K_T\sigma$$

Substituting the mean and standard of deviation of the exponential distribution, $\mu = \sigma = \frac{1}{\lambda}$, it can be deduced that $K_T = \log(T) - 1$.

Maximum Likelihood Estimation

The ML estimate of the parameter is given by [69]

$$\widehat{\lambda} = \frac{1}{\bar{x}} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

Method of Moments

Using the moments for estimation gives the same result as the MLE. The proof is omitted but the derivation can be found in Rao and Hamed (2000).

Probability Weighted Moments

The distribution has only one parameter, it is enough to obtain α_0 (Equation 3.6) in order to derive the parameter estimate. Since the probability weighted moment α_0 is the same as the conventional moment, the PWM estimate coincides with the MOM estimate.

4.2.2 Gamma distribution

Gamma distribution is commonly used in science and engineering to model positive continuous variables which have skewed distributions. We will use the parametrization with the probability density function

$$f(x;\alpha,\beta) = \frac{x^{\beta-1}\exp\left\{-\dfrac{x}{\alpha}\right\}}{\Gamma(\beta)\alpha^{\beta}}, \quad x > 0$$

where $\alpha > 0, \beta > 0$ are the scale and shape parameters, respectively. However, the probability density function of Gamma distribution is often defined with the rate parameter $1/\beta$. When $\alpha < 1$ it has en exponential shape, and when $\alpha = 1$ it coincides with exponential distribution with mean β. If $\alpha > 1$, the shape of the Gamma distribution is skewed and unimodal, and the skewness decreases as α increases. The chi-squared distribution with ν degrees of freedom arises from the gamma distribution with $\alpha = \nu/2$ and $\beta = 2$.

Besides applications in engineering, it is frequently used in survival analysis. The Erlang distribution, which is a special case of gamma distribution, describes the distribution of cancer incidence by age [9]. Aksoy (2000) fitted the gamma distribution to daily rainfall data. Furthermore, the gamma distribution is popular in Bayesian statistics because it acts as a conjugate prior for the rate of the exponential distribution or the precision parameter of the normal distribution.

The frequency factor was derived by Kite (1977) and it is given by

$$K_T = z + (z^2 - 1)k + \frac{1}{3}(z^3 - 6z)k^2 - (z^2 - 1)k^3 + zk^4 + \frac{1}{3}k^5, \qquad (4.1)$$

where z is the standard normal variable and $k = \dfrac{C_s}{6}$, using the coefficient of skewness as defined in Section 3.2.

Maximum Likelihood Estimation

There is no closed-form solution for the parameter estimates and the use of iterative methods is required for solving the expression below involving the digamma function: [69]:

$$-n\ln(\beta) + \sum_{i=1}^{n} \ln(x_i) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0.$$

Method of Moments

The parameter estimates of Gamma distribution using the method of moments have been derived by Rao and Hamed (2000) and are given by

$$\widehat{\alpha} = \frac{1}{n\bar{x}} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$
$$\widehat{\beta} = \frac{\bar{x}}{\widehat{\alpha}} = \frac{n\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Probability Weighted Moments

It is generally difficult to derive the parameter estimates for distributions in the Gamma family, since a closed form of the the inverse cumulative function $x(F)$ does not exist. We should therefore resort to the application of numerical methods. The L-Moments of the two parameter Gamma distribution have been derived by Hosking (1990) and are given by

$$\lambda_1 = \alpha\beta,$$
$$\lambda_2 = \frac{\alpha\Gamma(\beta + \frac{1}{2})}{\sqrt{\pi}\Gamma(\beta)}.$$

Using the sample estimates, λ_1 and λ_2 can be replaced by l_1 and l_2 and for $0 < \dfrac{l_2}{l_1} < \dfrac{1}{2}$, the parameter estimate for β can be found by calculating

$$\widehat{\beta} = \frac{1 - 0.3080z}{z - 0.05812z^2 + 0.01765z^3}, \quad \text{where} \quad z = \pi\left(\frac{l_2}{l_1}\right)^2.$$

For $\frac{1}{2} \leq \frac{l_2}{l_1} < 1$, β should be estimated using

$$\widehat{\beta} = \frac{0.7213z - 0.5947z^2}{1 - 2.1817z + 1.2113z^2}, \quad \text{where} \quad z = 1 - \frac{l_2}{l_1}.$$

Finally, we set $\widehat{\alpha} = \frac{l_1}{\widehat{\beta}}$.

4.2.3 Pearson III

The Pearson family of distributions was named after Karl Pearson who introduced it as a tool for modelling asymmetric data that are so common in many areas of science [56]. The family consists of several distributions that provide coverage of curves with a variety of different locations, variances, skewness and kurtosis (the distributions have up to 4 parameters). Note that Beta and Gamma distributions are members of the family, in particular Pearson Type III is a generalized (shifted) gamma distribution [36]:

$$f(x; \alpha, \beta, \varepsilon) = \frac{1}{\alpha\Gamma(\beta)} \left(\frac{x - \varepsilon}{\alpha}\right)^{\beta-1} \exp\left\{-\frac{x - \varepsilon}{\alpha}\right\}, \quad \varepsilon \leq x < \infty.$$

It is easy to see that the two-parameter Gamma distribution is a special case of Pearson III, where $\varepsilon = 0$.

Pearson III is commonly used in hydrology, for instance to model total rainfall depths of storm events with their return periods – in a study involving five Brazilian weather stations it outperformed the Gamma distribution in long-term predictions and was therefore recommended for calculating the Standardized Precipitation Index (SPI) in the region [10]. Furthermore, Haoran et al. (2010) used Pearson III distribution to predict the volume of freight at a port in China for logistics purposes.

The frequency factor K_T is the same as for Gamma distribution (Equation 4.1), but the estimate of the quantile x_T is given by

$$\widehat{x}_T = \widehat{\alpha}\widehat{\beta} + \widehat{\varepsilon} + K_T\sqrt{\widehat{\alpha}^2\widehat{\beta}}.$$

Maximum Likelihood Estimation

The joint likelihood function is given by

$$\begin{aligned} L(\alpha, \beta, \varepsilon | X) &= \prod_{i=1}^{n} \left(\frac{1}{\alpha\Gamma(\beta)} \left(\frac{X_i - \varepsilon}{\alpha}\right)^{\beta-1} \exp\left\{-\frac{X_i - \varepsilon}{\alpha}\right\} \right) \\ &= \frac{1}{\alpha^n\Gamma(\beta)^n} \prod_{i=1}^{n} \left(\frac{X_i - \varepsilon}{\alpha}\right)^{\beta-1} \exp\left\{ -\sum_{i=1}^{n} \left(\frac{X_i - \varepsilon}{\alpha}\right) \right\}. \end{aligned}$$

Taking logarithms we obtain the log-likelihood

$$\log L(\alpha,\beta,\varepsilon|X) = -n\log\alpha - n\log\Gamma(\beta) + (\beta-1)\sum_{i=1}^{n}\log(X_i-\varepsilon) - n(\beta-1)\log\alpha - \sum_{i=1}^{n}\frac{X_i-\varepsilon}{\alpha}.$$

Taking partial derivatives with respect to α, β and ε,

$$\begin{aligned}
\frac{\partial l}{\partial\alpha} &= \frac{-n}{a} - \frac{n(\beta-1)}{\alpha} + \frac{1}{\alpha^2}\sum_{i=1}^{n}(X_i-\varepsilon) = -\frac{n\beta}{\alpha} + \frac{1}{\alpha^2}\sum_{i=1}^{n}(X_i-\varepsilon),\\
\frac{\partial l}{\partial\beta} &= -n\frac{\Gamma'(\beta)}{\Gamma(\beta)} + \sum_{i=1}^{n}\log(X_i-\varepsilon) - n\log\alpha,\\
\frac{\partial l}{\partial\varepsilon} &= -(\beta-1)\sum_{i=1}^{n}\frac{1}{X_i-\varepsilon} + \frac{n}{\alpha}.
\end{aligned}$$

Setting the partial derivatives equal to zero and solving for $\widehat{\alpha}$, $\widehat{\beta}$ and $\widehat{\varepsilon}$,

$$\begin{aligned}
\frac{\partial l}{\partial\alpha} = 0 \Rightarrow n\alpha\beta &= \sum_{i=1}^{n}(X_i-\varepsilon) \Rightarrow \widehat{\alpha} = \frac{1}{n\widehat{\beta}}\sum_{i=1}^{n}(X_i-\widehat{\varepsilon})\\
\frac{\partial l}{\partial\beta} = 0 \Rightarrow \frac{\Gamma'(\widehat{\beta})}{\Gamma(\widehat{\beta})} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\log(X_i-\widehat{\varepsilon})}{\widehat{\alpha}},\\
\frac{\partial l}{\partial\varepsilon} = 0 \Rightarrow \sum_{i=1}^{n}\frac{1}{X_i-\widehat{\varepsilon}} &= \frac{n}{\widehat{\alpha}(\widehat{\beta}-1)},
\end{aligned}$$

which should be solved numerically.

Method of Moments

For the first moment,

$$\mu_1 = \mathbb{E}(X) = \frac{1}{\alpha\Gamma(\beta)}\int_{\alpha}^{\infty} x\left(\frac{x-\varepsilon}{\alpha}\right)^{\beta-1} e^{-\frac{(x-\varepsilon)}{\alpha}}dx$$

Using the substitution $y = \frac{x-\varepsilon}{\alpha} \implies x = \alpha y + \varepsilon \implies dx = \alpha dy$, we get

$$\begin{aligned}
\mu_1 &= \frac{1}{\Gamma(\beta)}\int_0^{\infty}(\alpha y+\varepsilon)y^{\beta-1}e^{-y}dy\\
&= \frac{1}{\Gamma(\beta)}\left[\int_0^{\infty}\alpha y^{\beta}e^{-y}dy + \varepsilon\int_0^{\infty}y^{\beta-1}e^{-y}dy\right]\\
&= \frac{1}{\Gamma(\beta)}[\alpha\Gamma(\beta+1) + \varepsilon\Gamma(\beta)]
\end{aligned}$$

Using the property of the Gamma function $\Gamma(\beta+1)=\beta\Gamma(\beta)$, we get

$$\mu_1=\alpha\beta+\varepsilon.$$

For the second moment,

$$\begin{aligned}
\mu_2=\mathbb{E}(X^2)&=\frac{1}{\alpha\Gamma(\beta)}\int_\alpha^\infty x^2\left(\frac{x-\varepsilon}{\alpha}\right)^{\beta-1}e^{-\frac{(x-\varepsilon)}{\alpha}}dx\\
&=\frac{1}{\Gamma(\beta)}\int_0^\infty(\alpha y+\varepsilon)^2y^{\beta-1}e^{-y}dy\\
&=\frac{1}{\Gamma(\beta)}\left[\alpha^2\int_0^\infty y^{\beta+1}e^{-y}dy+2\alpha\varepsilon\int_0^\infty y^{\beta}e^{-y}dy+\varepsilon^2\int_0^\infty y^{\beta-1}e^{-y}dy\right]\\
&=\frac{1}{\Gamma(\beta)}[\alpha^2\Gamma(\beta+2)+2\alpha\varepsilon\Gamma(\beta+1)+\varepsilon^2\Gamma(\beta)]\\
&=\frac{1}{\Gamma(\beta)}[\alpha^2(\beta+1)\beta\Gamma(\beta)+2\alpha\varepsilon\beta\Gamma(\beta)+\varepsilon^2\Gamma(\beta)]\\
&=\alpha^2\beta(\beta+1)+2\alpha\varepsilon\beta+\varepsilon^2\\
&=\alpha^2\beta+(\alpha\beta+\varepsilon)^2.
\end{aligned}$$

Similarly for the third moment,

$$\begin{aligned}
\mu_3=\mathbb{E}(X^3)&=\frac{1}{\alpha\Gamma(\beta)}\int_\alpha^\infty x^3\left(\frac{x-\varepsilon}{\alpha}\right)^{\beta-1}e^{-\frac{(x-\varepsilon)}{\alpha}}dx\\
&=\frac{1}{\Gamma(\beta)}\int_0^\infty(\alpha y+\varepsilon)^3y^{\beta-1}e^{-y}dy\\
&=\frac{1}{\Gamma(\beta)}\left[\alpha^3\int_0^\infty y^{\beta+2}e^{-y}dy+3\alpha^2\varepsilon\int_0^\infty y^{\beta+1}e^{-y}dy+3\alpha\varepsilon^2\int_0^\infty y^{\beta}e^{-y}dy\right.\\
&\quad\left.+\varepsilon^3\int_0^\infty y^{\beta-1}e^{-y}dy\right]\\
&=\alpha^3\beta(\beta+1)(\beta+2)+3\alpha^2\varepsilon\beta(\beta+1)+3\alpha\varepsilon^2\beta+\varepsilon^3
\end{aligned}$$

The skewness γ is then given by

$$\begin{aligned}
\gamma&=\frac{\alpha^3\beta(\beta+1)(\beta+2)+3\alpha^2\varepsilon\beta(\beta+1)+3\alpha\varepsilon^2\beta+\varepsilon^3-3(\alpha\beta+\varepsilon)\alpha^2\beta-(\alpha\beta+\varepsilon)^3}{(\alpha^2\beta)^{3/2}}\\
&=\frac{2\alpha^3\beta}{(\alpha^2\beta)^{3/2}}=\frac{2}{\sqrt{\beta}}.
\end{aligned}$$

Thus, the method of moments estimates are

$$\widehat{\beta}=\left(\frac{2}{C_s}\right)^2,\quad \widehat{\alpha}=\sqrt{\frac{\widehat{\mu}_2}{\widehat{\beta}}},\quad \widehat{\varepsilon}=\widehat{\mu}_1-\sqrt{\widehat{\mu}_2\widehat{\beta}}.$$

where C_s is the sample coeffiecient of skewness, as explained in Section 3.2.

Probability Weighted Moments

Similarly to Gamma distribution, Pearson III does not have an inverse cumulative function that could be written in closed form, therefore it is impossible to obtain the parameter estimates explicitly. Song and Ding (1988) have computed the probability weighted moments for Pearson III distribution, obtaining the following equations to be solved simultaneously:

$$\begin{aligned} M_0 &= \alpha\beta + \varepsilon \implies \varepsilon = M_0 - \alpha\beta \\ M_1 &= \alpha S_1^1(\beta) + \varepsilon S_1^0(\beta) \implies \alpha = \frac{M_1 - M_0 S_1^0(\beta)}{S_1^1(\beta) - \beta S_1^0(\beta)} \\ M_2 &= \alpha S_2^1(\beta) + \varepsilon S_2^0(\beta), \end{aligned}$$

where $S_l^k(\beta) = \int_0^\infty \left[\int_0^x \frac{1}{\Gamma(\beta)} t^{\beta-1}\mathrm{e}^{-t}\mathrm{d}t\right]^l \frac{1}{\Gamma(\beta)} x^{\beta-1+k}\mathrm{e}^{-x}\mathrm{d}x$.

The equation for the second moment M_2 can therefore be written as a function of β and solved by numerical methods to obtain the estimated parameters. Song and Ding (1988) prepared tables estimating the values for $S_l^k(\beta)$. Following up on their work, Ding, Song and Yang (1989), further expanded on the parameter estimations and they proved that $S_1^0(\beta) = \frac{1}{2}$ and $S_2^0(\beta) = \frac{1}{3}$. Updated tables can be found in their paper.

4.2.4 Log-Pearson III

Analogously to the normal/log-normal relationship, if the natural logarithm of a variable X, $\log X$, follows the Pearson III distribution, then X is from the Log-Pearson III (LP3) distribution. If $\log X$ has zero skew, the LP3 distribution reduces to the log-normal distribution. LP3 is widely used in hydrology and it is very specific to this particular field. As much as its applications outside this sector are rare, the importance of this distribution for hydrology is undeniable. For example, it has been adopted by U.S. government and also recommended for flood frequency analysis in Australia [66].

The density is defined as [12]

$$f(x;\alpha,\beta,\varepsilon) = \frac{1}{\alpha x \Gamma(\beta)} \left(\frac{\log x - \varepsilon}{\alpha}\right)^{\beta-1} \exp\left\{-\frac{\log x - \varepsilon}{\alpha}\right\}, \quad \varepsilon \le \log x$$

where α and β are both shape parameters and ε is the threshold value. The parameter and quantile estimation is done by applying the Pearson III estimation methods to the transformed variable $Y = \log X$.

4.3 Generalized Extreme Value (GEV) family

Extreme value theory provides the building blocks for inference about the probability of very rare events. Sometimes known as the Fisher–Tippett distribution,

GEV is often used to model the smallest or largest value within a set of independent and identically distributed random variables.

The probability density function is

$$f(x;\mu,\sigma,\xi) = \frac{1}{\sigma}t(x)^{\xi+1}\exp\left(-t(x)\right),$$
$$\text{where } t(x) = \begin{cases} \left(1+\xi\dfrac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ \exp\left\{-\dfrac{(x-\mu)}{\sigma}\right\} & \text{if } \xi = 0. \end{cases}$$

The cumulative distribution function can therefore be written as

$$F(x;\mu,\sigma,\xi) = \exp\left(-t(x)\right).$$

It is required that $\sigma > 0$ (scale), as well as $\xi, \mu \in \mathbb{R}$ (shape and location, respectively). A negative ξ indicates a heavy tail while a positive value of ξ describes a light tail. The GEV with $\xi = 0$ forms the Gumbel family, Weibull family if $\xi < 0$ and Fréchet family if $\xi > 0$. Note that reverse Weibull is the distribution obtained by setting $\xi < 0$ [17].

GEV has many applications in various sectors, including the environmental research and meteorology for rainfall and wind data, non-life insurance and actuarial for large claims in automobile insurance, risk management, metallurgy, geology and seismic analysis for predicting the magnitudes of earthquakes [30, 8]. For example, according to Marimoutou et al. (2009), models using extreme value theory applied to the increased price of oil gave the best results in their case study. Furthermore, GEV was used in a case study to examine the price of stock data, resulting in a very good fit comparing with the empirical density of near-extreme events [39].

4.3.1 Gumbel distribution (EV Type I)

The most common type of GEV distribution is Gumbel, also known as Extreme Value Type I (EV1) distribution. The versatility of the GEV family was illustrated above, however Gumbel in particular, was used to analyse the largest earthquakes in Turkey and specifically to calculate the occurrence and the expectation of earthquakes of extreme magnitudes by Tsapanos et al. (2014). Moreover, EV1 was chosen by Vidal (2014) to predict the return periods of extreme precipitation intensities. It was also shown to fit well the annual maximum wind speeds [24]. These examples indicate that Gumbel is a popular choice in the environmental sciences.

The distribution is characterized by the probability density

$$f(x;\mu,\beta) = \frac{1}{\beta}\exp\left(-\left(\frac{x-\mu}{\beta} + \exp\left(-\frac{x-\mu}{\beta}\right)\right)\right),$$

where μ is the location parameter and $\beta > 0$ is the scale parameter. The cumulative density has an explicit inverse form:

$$F(x;\mu,\beta) = \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right) \quad \Rightarrow \quad x(F) = \mu - \beta[\log(-\log(F)],$$

which means that the quantile corresponding to the return period T can be found by calculating

$$\widehat{x}_T = \widehat{\mu} - \widehat{\beta}\log\left[-\log\left(1-\frac{1}{T}\right)\right].$$

This can also be done using the frequency factor as suggested by Chow (1953)

$$K_T = -\frac{\sqrt{6}}{\pi}\left[\gamma + \log\left\{\log\left(\frac{T}{T-1}\right)\right\}\right],$$

where $\gamma = 0.5772$ is Euler's constant.

Maximum Likelihood Estimation

The joint likelihood is given by

$$L(\mu,\beta;x) = \prod_{i=1}^{n}\left(\frac{1}{\beta}\exp\left(-\frac{(x_i-\mu)}{\beta}\right)\exp\left[-\exp\left(-\frac{(x_i-\mu)}{\beta}\right)\right]\right).$$

The log-likelihood is thus

$$\log L(\mu,\beta;x) = -n\log\beta - \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\beta}\right) - \sum_{i=1}^{n}\left[\exp\left(-\frac{(x_i-\mu)}{\beta}\right)\right].$$

Taking partial derivatives,

$$\frac{\partial \log L(\mu,\beta)}{\partial\mu} = \frac{1}{\beta}\left[n - \sum_{i=1}^{n}\exp\left(-\frac{x_i-\mu}{\beta}\right)\right],$$

$$\frac{\partial \log L(\mu,\beta)}{\partial\beta} = -\frac{n}{\beta} + \sum_{i=1}^{n}\left[\frac{x_i-\mu}{\beta^2}\right] - \sum_{i=1}^{n}\left[\frac{x_i-\mu}{\beta^2}\right]\exp\left(-\frac{x_i-\mu}{\beta}\right),$$

for $\beta \neq 0$. Solving $\frac{\partial \log L(\mu,\beta)}{\partial \beta} = 0$ and $\frac{\partial \log L(\mu,\beta)}{\partial \mu} = 0$, we can get the ML estimates of μ and β using numerical methods [40]

$$\widehat{\mu} = \widehat{\beta}\left[\log n - \log \sum_{i=1}^{n} \exp\left(-\frac{x_i}{\widehat{\beta}}\right)\right],$$

$$\widehat{\beta} = \bar{x} - \frac{\sum_{i=1}^{n} x_i \exp\left(-\frac{x_i}{\widehat{\beta}}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{x_i}{\widehat{\beta}}\right)}.$$

Method of Moments

The standardized distribution (letting $y = (x - \mu)/\beta$) is

$$f_Y(y) = \exp\big(-y - \exp(-y)\big).$$

Let $Z = \exp(-Y)$. Using the change of variables theorem we show that Z follows the exponential distribution. Rewrite $y = -\log z$, then

$$f_Z(z) = f_Y(y)\left|\frac{dy}{dz}\right| = \left|-\frac{1}{z}\right| \exp[\log z - \exp(\log z)] = \exp(-z), \quad z \geq 0$$

which is the pdf of the exponential distribution. It follows that

$$\mathbb{E}(Z^{-t}) = \int_0^{\infty} z^{-t} \exp(-z) dz = \Gamma(1-t), \quad t < 1.$$

Notice that

$$\mathbb{E}(\exp(t\beta Z)) = \mathbb{E}(\exp[t(X-\mu)]) = \mathbb{E}(\exp(tX))\exp(-t\mu)$$

and the moment generating function for the Gumbel distribution is therefore $M_X(t) = \mathbb{E}(\exp(tX)) = \Gamma(1-t\beta)\exp(t\mu)$ [31]. Thus, the first moment is

$$\mu_1 = \frac{dM_X(t)}{dt}\bigg|_{t=0} = \big[\mu \exp(t\mu)\Gamma(1-t\beta) - \Gamma'(1-t\beta)\exp(t\mu)\beta\big]\bigg|_{t=0}$$
$$= \mu + \gamma\beta.^{1}$$

Hence μ can be approximated by $\widehat{\mu} = \bar{x} - \gamma\widehat{\beta}$. The second moment is

$$\mu_2 = \frac{d^2 M_X(t)}{dt^2}\bigg|_{t=0} = \big[-2\beta\mu\exp(t\mu)\Gamma'(1-t\beta) + \mu^2\exp(t\mu) + \Gamma''(1-t\beta)\beta^2\big]\bigg|_{t=0}$$
$$= \mu^2 + 2\beta\mu\gamma + \beta^2\left(\gamma^2 + \frac{\pi^2}{6}\right).^{2}$$

[1] $\Gamma'(1) = \int_0^{\infty} e^{-x}\ln x dx = -\gamma \approx -0.5772$ is the negative of Euler's constant.

Substituting for μ and solving for β yields

$$\begin{aligned}\mu_2 &= (\mu_1 - \beta\gamma)^2 + 2\beta\gamma(\mu_1 - \beta\gamma) + \beta^2\left(\gamma^2 + \frac{\pi^2}{6}\right)\\ &= \beta^2\left(\gamma^2 + \frac{\pi^2}{6}\right) - \beta^2\gamma^2 + \mu_1^2.\\ &\Rightarrow \mu_2 - \mu_1^2 = \beta^2\frac{\pi^2}{6}\\ &\Rightarrow \sigma^2 = \beta^2\frac{\pi^2}{6}.\end{aligned}$$

Thus $\widehat{\beta} = \sqrt{6}s/\pi$.

Probability Weighted Moments

One can easily find that the inverse cumulative function is

$$x(F) = \mu - \beta[\ln(-\ln(F)].$$

According to Greenwood et al. (1979) and Hosking (1986), the PWMs are derived similarly to the ones for Weibull distribution and are given by

$$\beta_r = \frac{\mu}{1+r} + \frac{\beta(\log(1+r) + \gamma}{1+r},$$

where γ is Euler's constant. Using the sample estimates we can obtain b_0 and b_1 which are then used to find the parameter estimates $\widehat{\beta}$ and $\widehat{\mu}$. The estimates are also given in terms of the first and second sample linear moments, l_1 and l_2

$$\begin{aligned}\widehat{\beta} &= \frac{2b_1 - b_0}{\log(2)} = \frac{l_2}{\log(2)},\\ \widehat{\mu} &= b_0 - \gamma\widehat{\beta} = l_1 - \gamma\widehat{\beta}.\end{aligned}$$

4.3.2 Weibull distribution (EV Type III)

A versatile distribution widely used for reliability engineering and life data analysis and failure times modelling, Weibull is commonly applied in many fields including economics, biology, engineering and hydrology. For instance, it appears in various papers with topics including flight load variation in helicopters, strength of steel and glass, wind power, raindrop size, cancer clinical trials, cell survival, business failures and stock returns [67, 53]. More examples of topics on which Weibull distribution was fitted can be found in Rinne (2009). More specifically, according to Mudholkar et al. (1996), it can be effectively used to analyse survival data [45]. Furthermore, Baqerin et al. (2016), fitted Weibull distribution to estimate scheduling performance in repetitive construction projects, such as

[2] $\Gamma''(1) = \int_0^\infty (\ln x)^2 e^{-x} dx = \gamma^2 + \frac{\pi^2}{6}$.

project durations of bridge, highway and building construction.

The probability density function is

$$f(x;\mu,\beta,\alpha) = \frac{\alpha}{\beta}\left(\frac{x-\mu}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x-\mu}{\beta}\right)^{\alpha}\right], \qquad \mu \leq x$$

where μ is the location parameter, β is the scale parameter and α is the shape parameter.

The cumulative function is given by

$$F(x) = 1 - \exp\left[-\left(\frac{x-\mu}{\beta}\right)^{\alpha}\right],$$

so the inverse cumulative function can be written as

$$x(F) = \mu + \beta[-\log(1-F)]^{1/\alpha}.$$

Substituting $F = \frac{1}{T}$ we find the estimated event magnitude associated with return period T by calculating

$$\widehat{x}_T = \widehat{\mu} + \widehat{\beta}[\log T]^{1/\alpha}.$$

Alternatively, one can resort to the use of the frequency factor as suggested in Equation 2.1 [12]:

$$K_T = \frac{[\log T]^{1/\alpha} - \Gamma(1/\alpha+1)}{\left[\Gamma(2/\alpha+1) - \Gamma^2(1/\alpha+1)\right]^{1/2}}.$$

Maximum Likelihood Estimation

Equating $\frac{\partial \log L(\mu,\beta,\alpha)}{\partial \mu} = 0$, $\frac{\partial \log L(\mu,\beta,\alpha)}{\partial \alpha} = 0$ and $\frac{\partial \log L(\mu,\beta,\alpha)}{\partial \beta} = 0$ yields

$$\sum_{i=1}^{n}\left[\frac{1}{\alpha} + \log(x_i-\mu) - \log\beta - \left(\frac{(x_i-\mu)}{\beta}\right)^{\alpha}\log\left(\frac{(x_i-\mu)}{\beta}\right)\right] = 0$$

$$\sum_{i=1}^{n}\left[-\frac{\alpha}{\beta} + \left(\frac{\alpha}{\beta}\right)\left(\frac{(x_i-\mu)}{\beta}\right)^{\alpha}\right] = 0$$

$$\sum_{i=1}^{n}\left[-\frac{(\alpha-1)}{(x_i-\mu)} + \left(\frac{\alpha}{\beta}\right)\left(\frac{(x_i-\mu)}{\beta}\right)^{\alpha-1}\right] = 0$$

respectively. This should be solved iteratively [70].

Method of Moments

As suggested by Cran (1998), the k^{th} moment is given by

$$\mu_k = \mu + \frac{\beta\Gamma\left(1+\frac{1}{\alpha}\right)}{k^{1/\alpha}}.$$

The expressions for the parameters can be written as a function of lower order moments as follows:

$$\mu = \frac{\mu_1\mu_4 - \mu_2^2}{\mu_1 + \mu_4 - 2\mu_2}, \quad \beta = \frac{\mu_1 - \mu}{\Gamma\left(1+1/\alpha\right)}, \quad \alpha = \frac{\log(2)}{\log\left(\mu_1 - \mu_2\right) - \log\left(\mu_2 - \mu_4\right)}.$$

From here, replacing the moments by their sample estimates $\widehat{\mu}_m$ yields the parameter estimates.

Probability Weighted Moments

The inverse function is given by $x(F) = \mu + \beta[-\log(1-F)]^{1/\alpha}$. The PWMs are therefore

$$\begin{aligned}
\alpha_k &= \int_0^1 x(F)(1-F)^k dF = \int_0^1 \left[\mu + \beta[-\log(1-F)]^{1/\alpha}\right](1-F)^k dF \\
&= \mu\int_0^1 (1-F)^k dF + \beta\int_0^1 -\log(1-F)]^{1/\alpha}(1-F)^k dF, \quad \text{let } u = -\log\left(1-F\right) \\
&= \frac{\mu}{k+1} + \beta\int_0^\infty u^{1/\alpha}\exp\left[-(k+1)u\right]\mathrm{d}u, \quad \text{let } v = (k+1)u \\
&= \frac{\mu}{k+1} + \frac{\beta}{(k+1)^{1/\alpha}}\int_0^\infty v^{1/\alpha}\exp\left(-v\right)dv \\
&= \frac{\mu}{k+1} + \frac{\beta}{(k+1)^{1/\alpha}}\Gamma(1+1/\alpha)).
\end{aligned}$$

It follows that

$$\alpha_0 = \mu + \beta\Gamma\left(1+\frac{1}{\alpha}\right), \quad \alpha_1 = \frac{\mu}{2} + \frac{\beta\Gamma\left(1+\frac{1}{\alpha}\right)}{2^{1+1/\alpha}}, \quad \alpha_2 = \frac{\mu}{3} + \frac{\beta\Gamma\left(1+\frac{1}{\alpha}\right)}{3^{1+1/\alpha}}.$$

We can simplify these expressions so that the gamma function cancels out, and estimate the parameters by numerically solving for $\widehat{\alpha}$

$$\frac{3\widehat{\alpha}_2 - \widehat{\alpha}_0}{2\widehat{\alpha}_1 - \widehat{\alpha_0}} = \frac{3^{-\frac{1}{\alpha}} - 1}{2^{-\frac{1}{\alpha}} - 1}$$

and then substituting the result into the following expressions

$$\widehat{\mu} = \widehat{\alpha}_0 - \widehat{\beta}\Gamma\left(1+\frac{1}{\widehat{\alpha}}\right), \quad \widehat{\beta} = \frac{2\widehat{\alpha}_1 - \widehat{\alpha}_0}{\left(2^{-\frac{1}{\widehat{\alpha}}} - 1\right)\Gamma\left(1+\frac{1}{\widehat{\alpha}}\right)}.$$

Chapter 5

Goodness-of-fit Tests

Goodness-of-fit tests aim to summarise the discrepancy between a statistical model and the observed data. They compare the observed values with either the values fitted by a model of interest or theoretical quantiles of a known sampling distribution. When considering a variety of models (e.g. nested models) one might be interested in a relative model fit. A careful assessment of fit is crucial for making reliable inferences. The use of a model which is confident overall but performs poorly for some parts of the data might have terrible consequences in practice [43].

Many different metrics have been suggested for determining whether a fit is satisfactory or not. Here we consider five different techniques, namely the Root-Mean-Square Error (RMSE), Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests, as well as the Akaike and Bayesian Information Criteria (AIC and BIC, respectively). The reason for this specific choice of goodness-of-fits tests was to provide a variety of methods where each measures the discrepancy between the data and the model in a slightly different way. For instance, as we will see the RMSE measures the typical distance of the fitted value from the observed value, A-D and K-S compare the empirical distribution with the cumulative distribution function while AIC and BIC are based on the likelihood and penalise the need for estimation. We will also introduce L-moments ratio diagrams which serve as a visual aid in the selection of regional curves in flood frequency analysis.

5.1 Root-Mean-Square Error (RMSE)

The Root-Mean-Square Error calculates the dispersion of the residuals of a fitted model. If the fit is good, the typical distance from the fitted curve should be low. It is calculated according to the formula below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{x_i} - x_i)^2}{n}},$$

where $\widehat{x_i}$ are the predicted values and x_i are the observed values. Calculating the sum of the square of the residuals can sometimes be difficult, therefore the

RMSE can alternatively be computed by

$$RMSE = \sqrt{1 - r^2} s_x,$$

where r is the correlation coefficient and s_x is the standard deviation of X. Intuitively, RMSE is zero if the data are perfectly correlated with the fitted values, i.e. when $r = 1$.

5.2 Kolmogorov-Smirnov test (K-S)

The K-S test is a non-parametric test, used to examine whether a random sample of observations comes from a specific distribution. Under the null hypothesis, the K-S test assumes that the observations come from a distribution with continuous distribution function $F_0(x)$. The test statistic is a quantification of the maximum distance between the empirical distribution function (edf) of the sample and the cumulative distribution function (cdf). If the sample is in fact from a given distribution, this distance should be low. Let $x_{(1)}, x_{(2)}, ..., x_{(n)}$ be the ordered observations of an n-sized random sample. The edf $S_n(x)$ gives the ratio of the observations from the sample that are less than or equal to x. It is thus defined as

$$S_n(x) = \begin{cases} 0, & x < x_{(1)} \\ \frac{k}{n}, & x_{(k)} \leq x < x_{(k+1)} \\ 1, & x \geq x_{(n)} \end{cases},$$

for $k = 1, ..., n - 1$.

The test statistic is thus given by [57]

$$D = \max_x |S_n(x) - F_0(x)|,$$

$$D = \max_i |S_n(x_{(i)}) - F_0(x_{(i)})|, |S_n(x_{(i-1)}) - F_0(x_{(i)})|.$$

The distribution of the test statistic is known exactly and can be easily looked up in tables or in statistical software. Moreover, this distribution is the same for all (continuous) $F_0(x)$, since the test is non-parametric. As a rule of thumb, one may find the critical values by evaluating $1.358/\sqrt{n}$ for a test with significance level $\alpha = 0.05$, if the sample size n is large. The critical region consists of values that are greater than this threshold [57].

5.3 Anderson-Darling test (A-D)

The Anderson-Darling test is an alternative test to assess how well the data fits a specified distribution by comparing the observed edf to the expected cdf. The null hypothesis, which states that the observations $x_1, \ldots, x_n$ were drawn from continuous distribution with cdf $F_0(x)$, is rejected if the test statistic is greater than a critical value at significance level α. The test statistic has been given in

many different forms, for instance [57],

$$\begin{aligned} A &= \int_{-\infty}^{\infty} \frac{\left(S_n(x) - F_0(x)\right)^2}{F_0(x)\left[1 - F_0(x)\right]} dx \\ &= -n - \sum_{i=1}^{n} \frac{(2i-1)}{n} \left[\ln F_0\left(x_i\right) + \ln\left(1 - F_0\left(x_{n+1-i}\right)\right)\right], \end{aligned}$$

where F is the cdf of the distribution, S_n is the edf defined in Section 5.2, x_i are the ordered data and n is the number of observations [44]. The critical value for the $\alpha = 0.05$ significance level is 2.5. Compared to the K-S test, more weight is given to the tail of the distribution with the A-D test which makes this test more powerful. Unlike K-S, A-D is not distribution-free.

5.4 Akaike Information Criterion (AIC)

Introduced by Akaike (1973), AIC is a measure of relative goodness of fit based on information theory that is used to compare different models which use maximum likelihood to estimate parameters. It is defined as [3]

$$\text{AIC} = -2\log\left(\widehat{L(\theta_q)}\right) + 2q.$$

Above q is the number of estimated parameters in the model and $\widehat{L(\theta_q)}$ is the maximised likelihood obtained by fitting q parameters. By allowing the model to become more complex (by adding more parameters), one can typically increase the likelihood, however this might cause overfitting. AIC deals with this problem by penalising the number of parameters. The lower the AIC, the better the model compared to all other models, including non-nested models. However, it does not give any information whether a single model fits the data well.

5.5 Bayesian Information Criterion (BIC)

If maximum likelihood estimation is used, BIC is a method of model selection given by

$$\text{BIC} = -2\log\left(\widehat{L(\theta_q)}\right) + q\log n.$$

Although similar to AIC, unless the sample size n is very small, the penalty for model complexity is larger than in the case of AIC. On one hand, this means that BIC will rarely select a complicated model and hence it is less prone to overfitting than AIC. On the other hand, it is possible that, unlike AIC, BIC will select a model that is too simple due to the large penalty. They might provide conflicting results (when AIC selects a more complicated model than BIC). It is often recommended to use these two measures (AIC and BIC) together. BIC is sometimes known as SIC (Schwarz information criterion) after G. E. Schwarz who first introduced it in 1978. Although the BIC formula does not have any

Bayesian component to it, Schwarz's argument for this procedure is based on the asymptotic behaviour of Bayes estimators [55].

5.6 L-moment ratio diagrams

The relationship between L-moment ratios is unique for each distinct distribution. That is why the plots of coefficient of L-variation (τ_2) against L-skewness (τ_3) and L-kurtosis (τ_4) against L-skewness (τ_3), or the L-moment ratio diagrams, are useful tools for finding the distribution that best fits a sample of data. While the former is used with two-parameter distributions, the $\tau_4 - \tau_3$ relationship is considered when some of the candidate distributions have a third parameter.

The plots of L-kurtosis against L-skewness have been widely used for supporting the choice of probability distributions for fitting hydrological data in regional frequency analysis [27, 28, 48]. Typically, the theoretical L-moment ratios for all candidate distributions are plotted together with the sample L-moment ratios of the gauging stations of in the region. In a $\tau_4 - \tau_3$ diagram, two-parameter distributions[1] are represented as points while three-parameter distributions are shown as curves. There are therefore two options for discriminating between the suitability of candidate distributions [48]:

- The most appropriate distribution is closest to the sample average of the moment ratios. If the lengths of records vary between stations, a weighted average may be considered.
- The best-fitting line through the sample L-moment ratios determines the most appropriate distribution for the given data.

Hosking and Wallis (1997, Appendix) derived L-moments and L-moment ratios for many distributions and these were used to produce the L-moments ratio diagram in the next chapter (Figure 6.5). Note that the L-moment ratios of the Log-Pearson III distribution cover a two-dimensional region because it has two shape parameters [19].

It is important to emphasize that this method is purely graphical, therefore subjective, and should not be used as a replacement for the aforementioned goodness-of-fit tests but rather as their complement.

[1] This also concerns the exponential distribution – although we only considered the version with a single parameter (the rate), adding the location parameter does not affect the L-moment ratios.

Chapter 6

U.S. Data

In this chapter we demonstrate the application of flood frequency analysis to measurements from a specific gauging station, carefully detailing the process. Subsequently, we carried out the analysis for multiple stations in the basin which allowed us to infer which model is most suitable to this particular region. Due to the accessibility of the data and length of records, the stations of interest were selected from the U.S. Geological Survey's (USGS) National Water Information System [60]. In particular, an analysis of eight gauging stations on Pearl river, Mississippi was conducted. A map of the studied locations can be found in Figure 6.1, and the their summary statistics are shown in Table 6.1.

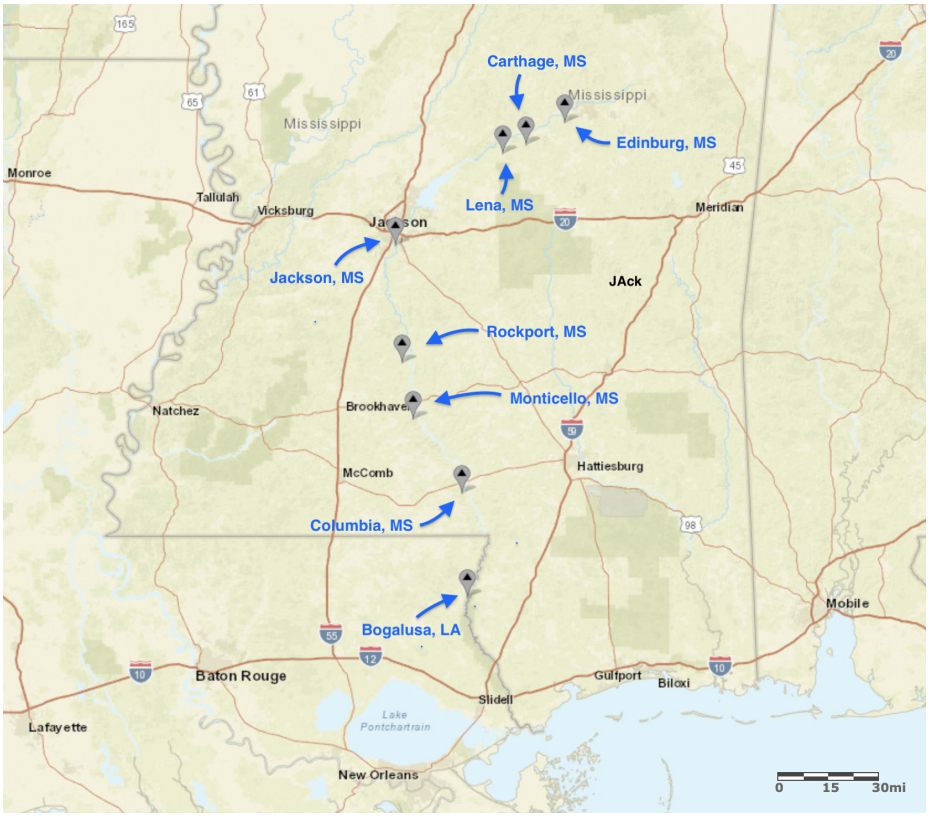


Figure 6.1: Map of the gauging stations on Pearl river [61].

Station	Sample size	Mean aps (cfs)	Std. Deviation (cfs)	Skewness
Jackson, MS	120	31,712	17,782	2.04
Edinburg, MS	109	13,596	10,198	2.90
Carthage, MS	58	18,982	15,347	3.27
Lena, MS	45	29,791	21,128	2.26
Rockport, MS	48	41,929	18,139	1.95
Monticello, MS	94	40,763	18,609	1.52
Columbia, MS	113	42,400	21,051	2.52
Bogalusa, LA	79	52,139	23,905	1.36

Table 6.1: Summary statistics for the annual peak streamflow (aps) recorded at the analysed gauging stations, given in cubic feet per second (cfs).

Recall from Chapter 2 that FFA at-site involves the following steps:

- For each gauging station:
    1. Process data and check for outliers using Formulas 2.2 and 2.3.
    2. Assign ranks to observations (in decreasing order) and obtain the estimated return period (plotting position) as given by Equation 2.5.
    3. Estimate parameters of each of the candidate distributions using MLE, MOM and PWM methods (as described in Chapter 3).
    4. Using the parameter estimates from the previous step, determine the estimated peak discharge for all return periods of interest. This is done either by calculating the frequency factor and then substituting it into Equation 2.1, or by finding the quantile of the distribution corresponding to a given return period T directly by applying the inverse, e.g. by executing the command `qnorm` in `R` for normal distribution.
    5. Assess the goodness-of-fit using the tests introduced in Chapter 5.
    6. Find the combination of distribution and estimation method that yields the best results in a majority of the goodness-of-fit tests.
- Repeat the process for multiple sites, plot the sample L-moment ratios and decide on the best-fitting distribution in the region.

We will now discuss the application of this procedure to stations on Pearl river with main focus on the one in Jackson, MS.

6.1 Data preparation

The dataset of each station includes the annual peak streamflow (aps) measured in cubic feet per second (cfs) and the date of the flood. The length of the records varies from station to station, with earliest measurements being from 1874 and the most recent ones from 2017. For consistency, only observations after the year 1900 were considered, since only a few events were recorded in 1874-1899.

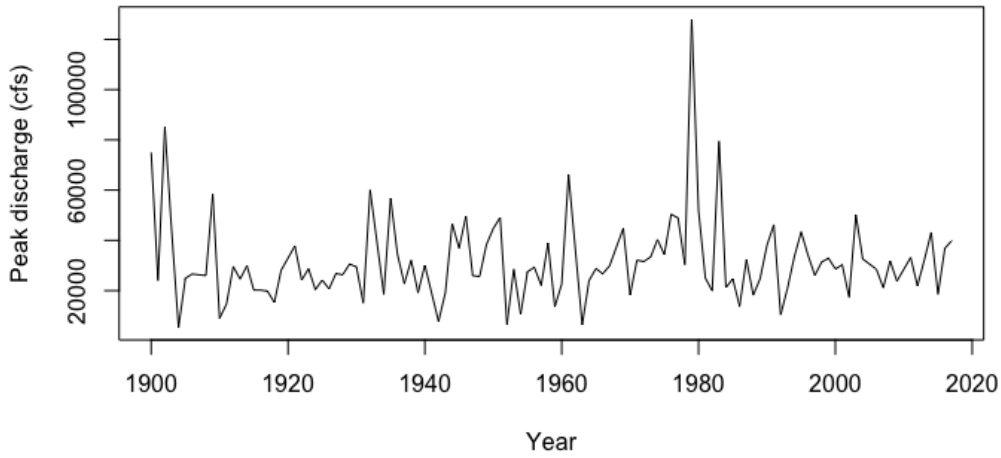


Figure 6.2: The annual peak discharge data for the station in Jackson as time series.

Note that the typical annual mean streamflow for Jackson was 4275.2 cfs between 1962 and 2017. The data from before 1962 was not available.

The data pre-processing involved ensuring that there was only one maximum value for each year, followed by visualisation of the dataset and checking for outliers. The flood peaks by year and the density of the data are shown in Figures 6.2 and 6.3 respectively. Any observations above or below the thresholds described in Section 2.2 were omitted from the analysis. For the data from Jackson, the IQR method detected two high outliers, in 1902 and 1979.

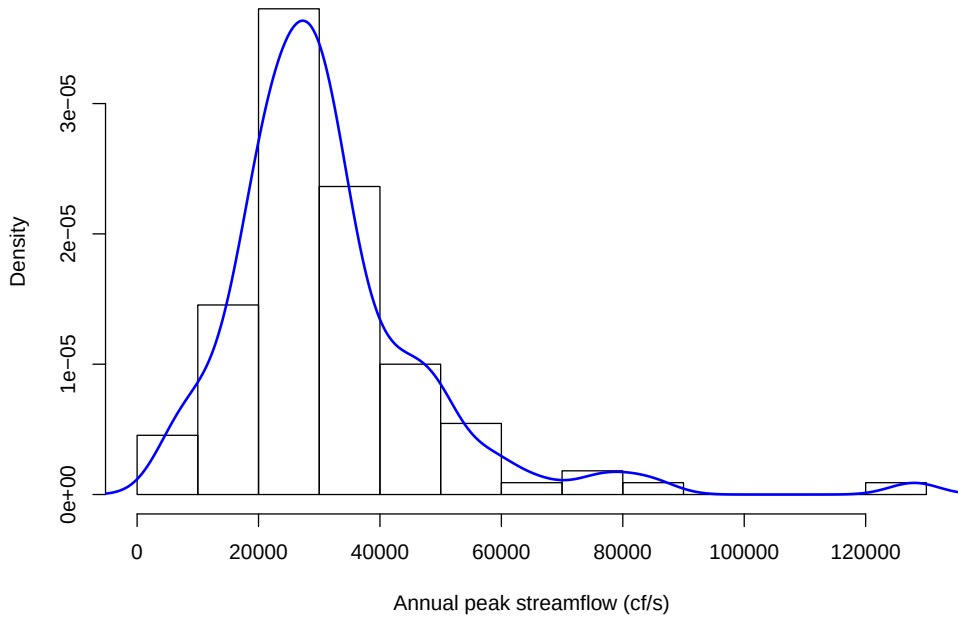


Figure 6.3: The empirical density plot of the raw data collected at Jackson.

6.2 Results and discussion

Parameter estimation was done through the use of self-defined functions and optimizer, as well as the following `R` libraries: `fitdistrib, fitdistrplus, FAdist, EnvStats, nsRFA, evd, Lmoments, lmomco`. The package `goftest` was also used for evaluating the A-D test statistic. The parameter estimates and their standard errors for the Jackson station are summarised in Table 6.2, and the goodness-of-fit tests are shown in Table 6.3.

Distribution	Estimation method	Parameter estimate (Standard error) Location	Scale/Rate	Shape
Normal	MLE	30197 (1305)	13557 (922.4)	–
	MOM	30197 (1305)	13557 (922.4)	–
	PWM	30197 (1251)	12999 (639.5)	–
Log-normal	MLE	10.209 (0.047)	0.488 (0.033)	–
	MOM	10.209 (0.047)	0.488 (0.033)	–
	PWM	10.209 (0.045)	0.468 (0.023)	–
Exponential	MLE	–	3.312×10^{-5} (5.303×10^{-6})	–
	MOM	–	3.312×10^{-5} (5.303×10^{-6})	–
	PWM	–	3.312×10^{-5} (5.303×10^{-6})	–
Gamma	MLE	–	6206.6 (872.7)	4.859 (0.648)
	MOM	–	6086.5	4.961
	PWM	–	6086.5	4.961
Pearson III	MLE	599.0 (2596)	6394.7 (1178)	4.628 (1.155)
	MOM	3181.9	6866.9	3.934
	PWM	3163.5	6645.3	4.068
Log-Pearson III	MLE	5.686 (0.845)	0.060 (0.012)	75.716 (27.83)
	MOM	8.974	0.194	6.349
	PWM	8.737	0.153	9.622
Gumbel	MLE	24034 (1117)	11022 (826.9)	–
	MOM	24096 (1099)	10570 (1067)	–
	PWM	24090 (1678)	10580 (916.5)	–
Weibull	MLE	3633.1 (1439)	29937 (2343)	2.034 (0.202)
	MOM (2-par)	–	34072	2.370
	PWM	8686.8	24069	1.663

Table 6.2: Parameter estimates, as defined in Chapter 3, and their standard errors where it was available, for the Jackson station.

The following was observed:

- Fitting Pearson III, and especially log-Pearson III by MLE gave unstable results. Particularly for Pearson III even though convergence was reached, it was observed that slightly different initial conditions produced very different estimates, especially the location. On the other hand, when Log-Pearson III was used with the MLE method, using various different starting points and optimisation methods, the optimisation often failed to converge. The unstability of MLE with three-parameter distributions was also reported by Kobierska et al. (2018).

- MOM and PWM sometimes produce estimates that are not in the parameter space (e.g. Weibull PWM gives a threshold/location higher than the minimal value).

- Recalling the formulas given in Section 4.3.2 for estimating the parameters of Weibull distribution with the method of moments, the second moment was found to be larger than the first moment, resulting in a negative number inside the logarithm. This has caused the method to fail for this specific dataset, so we have therefore opted for the two-parameter Weibull distribution, when MOM was considered[1].

- The standard errors of the estimated parameters for each distribution were recorded in Table 6.2, where available. For MLE, the standard errors were found using the Hessian matrix. The availability of standard errors for the other two methods depended on whether they were provided by the libraries used in R. The standard errors were high in general. Notice that PWM estimation yielded the most efficient results for the normal and log-normal distribution.

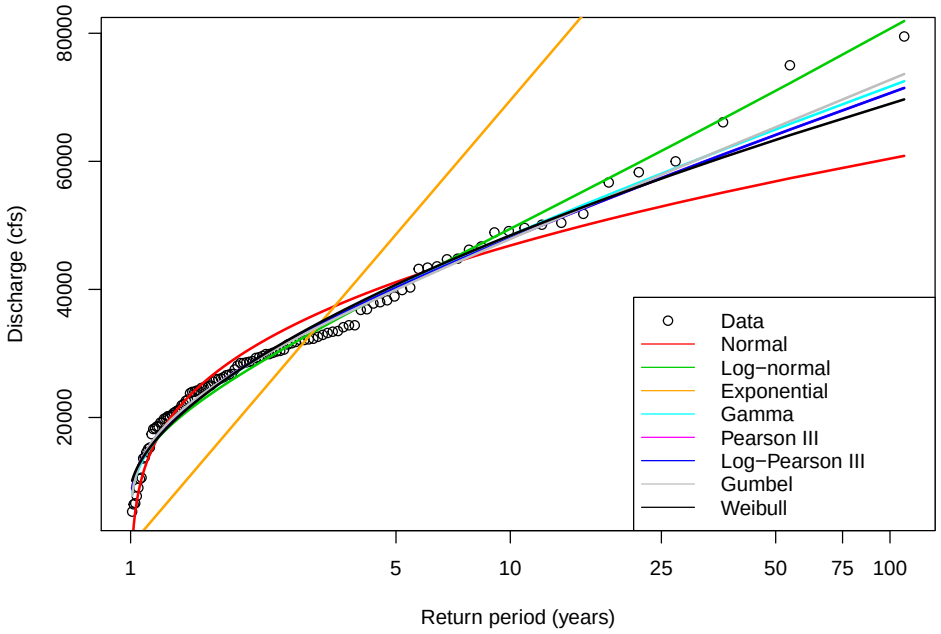


Figure 6.4: Observed and estimated peak streamflows for Jackson, MS.

Figure 6.4 compares the fits of candidate distributions, where the curve for each distribution was plotted using the estimation method that gave the best results. As expected, the one-parameter exponential cannot capture the flood peaks with enough accuracy. It can be deduced that the log-normal distribution explains

[1]The two-parameter Weibull distribution is given by $f(x;\alpha,\beta) = \frac{\alpha}{\beta^\alpha} x^{(\alpha-1)} \exp\left\{-\left(\frac{x}{\beta}\right)^\alpha\right\}$. The MOM estimate for β is given by $\widehat{\beta} = \frac{\bar{x}}{\Gamma\left(1+\frac{1}{\widehat{\alpha}}\right)}$ and $\widehat{\alpha}$ is given by solving $\frac{n\sum_{i=1}^n x_i^2}{\left[\sum_{i=1}^n x_i\right]^2} = \frac{\Gamma\left(1+\frac{2}{\widehat{\alpha}}\right)}{\Gamma^2\left(1+\frac{1}{\widehat{\alpha}}\right)}$, iteratively.

higher return periods best, followed by Gumbel. However Gumbel is preferable for low return periods. Observed and estimated peak streamflows for the remaining stations can be found in Appendix A.

Distribution	Estimation method	RMSE	K-S	A-D	AIC	BIC
Normal	MLE	3480.5	0.130	1.772	2365.7	2371
	MOM	3480.5	0.130	1.772	N/A	N/A
	PWM	3529.2	0.120	1.688	N/A	N/A
Log-normal	MLE	2060.8	0.093	1.400	2360.5	2365.9
	MOM	2060.8	0.093	1.400	N/A	N/A
	PWM	1689.0	0.093	1.264	N/A	N/A
Exponential	MLE	15014.2	0.333	16.599	2446.1	2448.8
	MOM	15014.2	0.333	16.599	N/A	N/A
	PWM	15014.2	0.333	16.599	N/A	N/A
Gamma	MLE	1849.1	0.074	0.764	2352.5	2357.9
	MOM	1875.0	0.074	0.727	N/A	N/A
	PWM	1875.0	0.074	0.727	N/A	N/A
Pearson III	MLE	1830.3	0.074	0.790	2354.8	2362.9
	MOM	1858.8	0.074	0.880	N/A	N/A
	PWM	1921.7	0.065	0.844	N/A	N/A
Log-Pearson III	MLE	2614.9	0.111	2.214	2370.8	2378.9
	MOM	2773.5	0.102	∞	N/A	N/A
	PWM	2970.4	0.093	∞	N/A	N/A
Gumbel	MLE	1533.7	0.074	0.610	2351.2	2356.6
	MOM	1679.6	0.065	0.550	N/A	N/A
	PWM	1674.9	0.065	0.551	N/A	N/A
Weibull	MLE	2491.2	1.085	0.093	2356.3	2364.3
	MOM	2805.9	0.111	1.215	N/A	N/A
	PWM	2246.3	0.074	∞	N/A	N/A

Table 6.3: Goodness-of-fit tests for Jackson station.

Comparing the results of different goodness-of-fit metrics from Table 6.3, it can be concluded that the data for the station in Jackson are best fitted by Gumbel distribution. The lowest RMSE and comparatively low K-S and A-D statistics were observed, and Gumbel distribution also gave the lowest AIC and BIC when likelihood-based estimation was applied. The best performance was achieved using the probability weighted moments method which yielded RMSE of 1674.9, K-S of 0.065 and A-D of 0.551. Although the MLE outperformed PWMs in terms of RMSE, the remaining tests were in favour of PWMs. This led to the conclusion that this distribution is suitable for modelling flood peaks at Jackson, with the preferred parameter estimation being the probability weighted moments method.

Notice that in three cases the A-D statistic diverges. This is likely to be due to the fact that the fitted distribution puts little or no mass on some observations, resulting in an attempt to divide by (or take logarithms of) a very small value or zero, as can be seen from formulas in Section 5.3. This indeed indicates a lack of

fit but other (e.g. distribution-free) goodness-of-fit tests may be more informative in such case.

Different conclusions were reached when the remaining stations were considered. The best combination for the station at Carthage was found to be the log-Pearson III distribution with the PWMs method, but for Bogalusa the log-normal distribution with the MLE method was identified as most appropriate. Table 6.4 displays the distribution that was found to fit the data best and the preferred parameter estimation method for each station.

Station	Distribution	Estimation Method
Jackson, MS	Gumbel	PWMs
Edinburg, MS	Weibull	PWMs
Carthage, MS	Log-Pearson III	PWMs
Lena, MS	Gumbel/Log-Normal	MLE/PWMs
Rockport, MS	Weibull	PWMs
Monticello, MS	Log-Pearson III	MLE
Columbia, MS	Log-Pearson III	PWMs
Bogalusa, LA	Gumbel/Log-Normal	PWMs/MLE

Table 6.4: Best combination of distribution and estimation method for analysed gauging stations.

For a visual assessment of the overall suitability for the basin of Pearl river, L-moments ratios diagram was plotted and can be found in Figure 6.5.

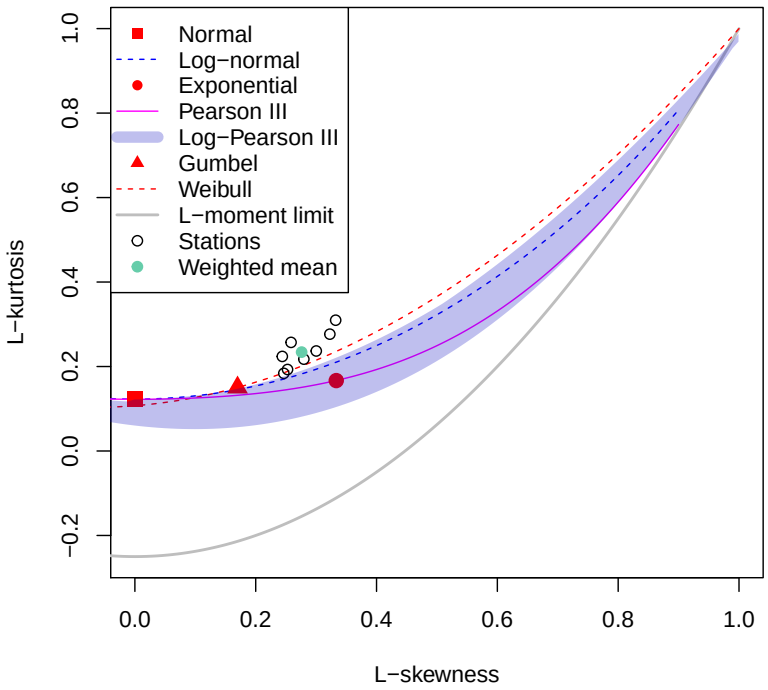


Figure 6.5: L-moments ratios diagram for the stations in the catchment area of Pearl River and the candidate distributions.

Let us examine the proximity of sample L-moment ratios from the theoretical ratios of each distribution. Considering one and two-parameter distributions first, the weighted mean of L-moment ratios seems to be closest to the exponential distribution. However, we have seen that the exponential distribution was evaluated as the poorest fit – perhaps estimating the location parameter in addition to the rate would make this distribution more adequate. Gumbel is not too far from the weighted mean either but both the L-skewness and L-kurtosis of the data was higher. As for the three-parameter distributions, the best-fitting line through the sample L-moment ratios is given by Weibull, followed by the upper limit of the log-Pearson III region.

Overall, the L-moment ratios diagram suggests that Weibull is the distribution that can be most accurately applied to the data from the gauging stations in the region. Notice that Pearson III and the Normal distribution would not be favourites for modelling the peak discharge at Pearl river, which is consistent with Table 6.4, since these distributions were not preferred for any of the stations studied.

In summary, the estimates produced by log-Pearson III were most accurate for three out of eight stations of interest and this distribution was also deemed appropriate by the L-moment ratios diagram for fitting the data collected at Pearl river. Other strong candidates were Gumbel and Weibull, in particular when estimated by the probability weighted moments.

The code used to produce figures and results in this chapter can be found on GitHub, accessed at `https://github.com/mariegold/ffa`.

6.2.1 Visualisation

An interactive web application was built using the `shiny` package in `R` to support the presentation of the results. It can be accessed at `https://mariegold.shinyapps.io/code/`.

Chapter 7

Conclusions

Flood frequency analysis is a method applied to data observed in a river system collected over a long period of time. It aims to relate the magnitude of a flood with the frequency of its occurrence and is therefore essential for water management and control.

Upon building the necessary theoretical background, focusing on the use of the annual peak flow data in flood frequency analysis, an at-site flood frequency analysis was performed. Data from eight gauging stations at Pearl river were extracted from the USGS National Water Information System and subsequently fitted by eight probability distributions – Normal, Log-Normal, Exponential, Gamma, Pearson III, Log-Pearson III, Gumbel and Weibull – using each of the three estimation methods – maximum likelihood, method of moments and probability weighted moments. The performance of each combination was assessed conducting five different statistical tests: RMSE, Kolomogorov-Smirnov, Anderson-Darling, AIC and BIC. The combination that gave the best results in a majority of goodness-of-fit tests was chosen for each station, and also the best overall frequency curve, which would enable accurate estimation across the entire region, was determined through the visualisation of L-moment ratios.

For the station in Jackson, which was our main location of interest, it was found that Gumbel distribution can predict the peaks of floods most reliably, in particular when estimated by the probability weighted moments. Out of eight stations analysed, Gumbel was determined as most suitable in three cases, similarly to Log-Pearson III (LP3).

The diagram of L-moment ratios indicates that Weibull distribution might be the best candidate distribution for the regional curve, however Log-Pearson III was also able to capture the range L-skewness and L-kurtosis of the observed data well. Since LP3 was selected as most appropriate for three of the stations, as opposed to Weibull which was only chosen in two cases, there is more evidence supporting the choice of Log-Pearson III as the regional distribution. This conclusion is consistent with the recommendation of the U.S. Water Resources Council, suggesting LP3 is the most applicable distribution to the rivers in the US. Nonetheless, we suggest that all three distributions (LP3, Gumbel, Weibull)

are considered for modelling design floods in the Pearl river basin until a more extensive and indicative analysis confirms that LP3 is indeed the best choice.

Future work may involve including more gauge stations in the analysis, or exploring additional distributions, such as the five-parameter Wakeby. Similarly, Bayesian methods have gained popularity in frequency analysis for parameter estimation and it would be interesting to see how they compare to the methods applied here. Finally, the confidence intervals for the estimated peaks were not covered in much detail and should be subject to further investigation.

Bibliography

[1] Ackermann, W.C. (1984). *Memorial Tributes.* National Academy of Engineering, Volume 2. p. 47.

[2] Adeboye, O., Alatise, M. (2007). *Performance of Probability Distributions and Plotting Positions in Estimating the Flood of River Osun at Apoje Sub-basin, Nigeria.* Agric Eng Int CIGR J. 9.

[3] Akaike, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle.* Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY

[4] Aksoy, H. (2000). *Use of gamma distribution in hydrological analysis.* Turkish Journal of Engineering and Environmental Sciences, 24(6), 419-428.

[5] Aldrich, J. (1997). *R.A. Fisher and the making of maximum likelihood 1912-1922.* Statist. Sci., 12(3), pp.162–176.

[6] American Rivers. (2020, January 15). `https://www.americanrivers.org/rivers/discover-your-river/whats-the-biggest-flood-in-history/`

[7] Baqerin, M.H., Shafahi, Y., Kashani, H. (2016). *Application of Weibull Analysis to Evaluate and Forecast Schedule Performance in Repetitive Projects.* Journal of Construction Engineering and Management, 142(2).

[8] Beirlant, J. (2004). *Statistics of extremes: theory and applications.* Chichester, England: J. Wiley.

[9] Belikov A.V. (2017). *The number of key carcinogenic events can be predicted from cancer incidence.* Scientific reports, 7(1), 12170.

[10] Blain, G. (2011). *Standardized precipitation index based on pearson type III distribution.* Revista Brasileira de Meteorologia. 26. 167-180.

[11] Chow, V.T. (1953). *Frequency analysis of hydrologic data with special application to rainfall intensities.* University of Illinois at Urbana Champaign, College of Engineering. Engineering Experiment Station..

[12] Chow, V.T., Maidment, D.R., Mays, L.W. (1988). *Applied hydrology.* New York: McGraw-Hill.

[13] Cran, G.W. (1988). *Moment estimators for the 3-parameter Weibull distribution.* IEEE Transactions on Reliability, 37(4), 360-363.

[14] Cunnane, C. (1989). *Statistical distributions for flood frequency analysis.* Geneva, Switzerland: Secretariat of the World Meteorological Organization.

[15] Ding, J., Song, D. (1988). *The application of probability weighted moments in estimating the parameters of the Pearson Type Three Distribution.* Journal of Hydrology, 101(1), pp.47–61.

[16] Ding, J., Song, D., Yang, R. (1989). *Further research on application of probability weighted moments in estimating parameters of the Pearson type three distribution.* Journal of Hydrology, 110(3), pp.239-257.

[17] Gorgoso-Varela, J.J., Rojo-Alboreca, A. (2014). *Use of Gumbel and Weibull functions to model extreme values of diameter distributions in forest stands.* Annals of Forest Science, Springer Verlag/EDP Sciences.

[18] Greenwood, J.A., Landwehr, J.M., Matalas, N.C. (1979). *Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form.* Water Resources Research, 15(5), pp.1049-1054.

[19] Griffis, V., Stedinger, J.R. (2007). *Log-Pearson type 3 distribution and its application in flood frequency analysis. I: Distribution characteristics.* Journal of Hydrologic Engineering 12(5): 482–491.

[20] Haoran, S., Wen, D. (2010). *Application of Pearson type III distribution to prediction of Port's volume of freight.* 2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM), 2, pp.732-736.

[21] Hazen, A. (1914). *Storage to be provided in impounding reservoirs for municipal water supply.* Proceedings of the American Society of Civil Engineers, 1913, Vol. 39, Issue 9, Pg. 1943-2044.

[22] Hazen, A. (1930). *Flood Flows: A Study of Frequencies and Magnitudes.* New York: J. Wiley & Sons, inc.

[23] Heyde, C.C. (1963). *On a property of the lognormal distribution.* Journal of the Royal Statistical Society, Series B, 25 (2): 392–393.

[24] Hong, H.P., Li, S.H., Mara, T.G. (2013). *Performance of the generalized least-squares method for the Gumbel distribution and its application to annual maximum wind speeds.* Journal of Wind Engineering & Industrial Aerodynamics, 119, pp.121-132.

[25] Hosking, J.R.M., Wallis, J.R., Wood, E.F. (1985). *Estimation of the generalized extreme-value distribution by the method of probability-weighted moments.* Technometrics, 27(3), 251-261.

[26] Hosking, J.R.M., (1986). *The theory of probability weighted moments.* IBM Research Report 12210.

[27] Hosking, J.R.M. (1990). *L-moments: Analysis and estimation of distributions using linear combinations of order statistics.* Journal of the Royal Statistical Society: Series B (Methodological), 52(1), 105-124.

[28] Hosking, J.R.M, Wallis, J.R. (1997). *Regional Frequency Analysis: An Approach Based on L-moments.* Cambridge University Press, UK.

[29] Hu, Y.M., Liang, Z.M., Li, B.Q., Yu, Z.B. (2013). *Uncertainty assessment of hydrological frequency analysis using bootstrap method.* Mathematical Problems in Engineering.

[30] Huang, C., Lin, J.G., Ren, Y.Y. (2013). *Testing for the shape parameter of generalized extreme value distribution based on the L_q-likelihood ratio statistic.* Metrika, 76(5), pp.641-671.

[31] Johnson, N.L., Kotz, S., Balakrishnan, N. (1994). *Extreme Value Distributions*, Continuous univariate distributions, vol 2, pp. 11-12.

[32] Khosravi, Gh., Majidi, A., Nohegar, A. (2013). *Determination of Suitable Probability Distribution for Annual Mean and Peak Discharges Estimation (Case Study: Minab River- Barantin Gage, Iran).* International Journal of Probability and Statistics. 1. 160-163.

[33] Kipkemboi, J., Kilonzi, C., Dam, A., Kitaka, N., Mathooko, J., Denny, P. (2010). *Enhancing the fish production potential of Lake Victoria papyrus wetlands, Kenya, using seasonal flood-dependent ponds.* Wetlands Ecology and Management, 18(4), pp.471–483.

[34] Kite, G.W. (1977). *Frequency and Risk Analysis in Hydrology.* Water Resources Publications. Fort Collins, Colorado.

[35] Kobierska-Baffie, F.A., Engeland, K., Thorarinsdottir, T.L. (2018). *Evaluation of design flood estimates – a case study for Norway.* Hydrology Research 49(2): 450-465.

[36] Koutrouvelis, I., Canavos, G. (1999). *Estimation in the Pearson type 3 distribution.* Water Resources Research, 35(9), pp.2693–2704.

[37] Li, M., Chen, W., Zhang, T., (2017). *Application of MODWT and log-normal distribution model for automatic epilepsy identification.* Biocybernetics and Biomedical Engineering, 37(4), pp.679-689.

[38] Limpert, E., Stahel, W. A., Abbt, M. (2001). *Log-normal Distributions across the Sciences: Keys and Clues: On the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: That is the question.* BioScience, Volume 51, Issue 5, pp.341-352.

[39] Lin J., Huang C., Zhuang Q., Zhu L. (2010). *Estimating generalized state density of near-extreme events and its applications in analyzing stock data.* Insur Math Econ 47(1):13-20 .

[40] Mahdi, S., Cenac, M. (2012). *Estimating Parameters of Gumbel Distribution using the Methods of Moments, probability weighted Moments and maximum likelihood.* Revista de Matemática: Teoría y Aplicaciones. 12. 151.

[41] Makkonen, L. (2005). *Plotting position in extreme value analysis.* Journal of Applied Meteorology and Climatology, 45, 334-345.

[42] Marimoutou V., Raggad B., Trabelsi A. (2009). *Extreme value theory and value at risk: application to oil market.* Energy Econ 31(4):519-530.

[43] Maydeu-Olivares, A., Forero, C. (2010). *Goodness-of-Fit Testing*, International Encyclopedia of Education, vol. 7, pp. 190-196.

[44] Millington, N. Das, S., Simonovic, S.P. (2011). *The Comparison of GEV, Log-Pearson Type 3 and Gumbel Distributions in the Upper Thames River Watershed under Global Climate Models.* Water Resources Research Report. London, Canada: Department of Civil and Environmental Engineering, The University of Western Ontario.

[45] Mudholkar, G.S., Srivastava, D.K.,Kollia, G.D. (1996). *A Generalization of the Weibull Distribution with Application to the Analysis of Survival Data.* Journal of the American Statistical Association, 91(436), pp.1575–1583.

[46] Nielsen, M.A. (2011). *Parameter Estimation for the Two-Parameter Weibull Distribution.* All Theses and Dissertations. 2509.

[47] Office of the General Counsel. (1997). *The National Flood Insurance Act of 1968, as amended, and the Flood Disaster Protection Act of 1973, as Amended.* 42 U.S.C. 4001 et. seq.

[48] Peel, M.C., Wang, Q.J., Vogel, R.M., McMahon, T.A. (2001). *The utility of L-moment ratio diagrams for selecting a regional probability distribution.* Hydrological sciences journal, 46(1), 147-155.

[49] Queensland Government. (2020, March 1st) *Understanding Floods: Questions & Answers.* `https://www.chiefscientist.qld.gov.au/__data/assets/pdf_file/0022/49801/understanding-floods_full_colour.pdf`

[50] Rahman, A.S., Rahman, A., Zaman, M.A., 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(3), pp.1803–1813.

[51] Rao, A.R., Hamed, K.H. (2000). *Flood Frequency Analysis.*Boca Raton, Florida, CRC Press LLC.

[52] Rasmussen, P. F. (2001). *Generalized probability weighted moments: application to the generalized Pareto distribution.* Water Resources Research, 37(6), 1745-1751.

[53] Rinne, H. (2009). *The Weibull distribution a handbook*, Boca Raton: CRC Press.

[54] Rumsey, B. (2015). *From Flood Flows to Flood Maps: The Understanding of Flood Probabilities in the United States.* Historical Social Research, 40(2), pp.134–150.

[55] Schwarz, G.E. (1978). *Estimating the dimension of a model.* Annals of Statistics 6 (2): 461–464.

[56] Shakil, M., Kibria, B.M.G, Singh, J. (2010). *A New Family of Distributions Based on the Generalized Pearson Differential Equation with Some Applications.* Austrian Journal of Statistics Volume. 39. 259-278.

[57] Sharifi Far, S. (2019). *Applied Statistics – Lecture Notes.* The University of Edinburgh.

[58] Sobkowicz, P., Thelwall, M., Buckley, K., Paltoglou, G., Sobkowitz, A. (2013). *Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?.* EPJ Data Sci. 2, 2.

[59] Tsapanos, T.M., Bayrak, Y., Cinar, H., Koravos, G.C., Bayrak, E., Kalogirou, E.E., Tsapanou, A.V., Vougiouka, G.E. (2014). *Analysis of largest earthquakes in Turkey and its vicinity by application of the Gumbel III distribution.* Acta Geophysica, 62(1), pp.59–82.

[60] U.S. Geological Survey. (2019, October 18). *Surface Water for USA: Peak Streamflow.* `https://nwis.waterdata.usgs.gov/usa/nwis/peak`

[61] U.S. Geological Survey. (2020, February 17). *National Water Information System: Mapper.* `https://maps.waterdata.usgs.gov/mapper/`

[62] Vallentin, M. (2012). *Probability and Statistics Cookbook.* `https://www.isixsigma.com/wp-content/uploads/2012/09/cookbook-en.pdf`

[63] Vidal, I. (2014). *A Bayesian analysis of the Gumbel distribution: an application to extreme rainfall data.* Stochastic Environmental Research and Risk Assessment, 28(3), pp.571–582.

[64] Vivekanandan, N. (2014). *Comparison of Probability Distributions for Estimation of Peak Flood Discharge.* Open Access Library Journal, 1(4), 1-7.

[65] Water Resources Council (US) Hydrology Committee. (1981). *Guidelines for determining flood flow frequency* (Vol. 17). US Water Resources Council.

[66] Water Resources Council (US) Hydrology Committee. (1967). *A uniform technique for determining flood flow frequencies.* Washington: US Water Resources Council.

[67] Weibull, W. (1939). *A statistical theory of the strength of materials.* Stockholm: Generalstabens litografiska anstalts förlag.

[68] Wickham, C. (2020, January 15) *Statistics - Lecture One.* https://www.stat.berkeley.edu/~vigre/activities/bootstrap/2006/wickham_stati.pdf

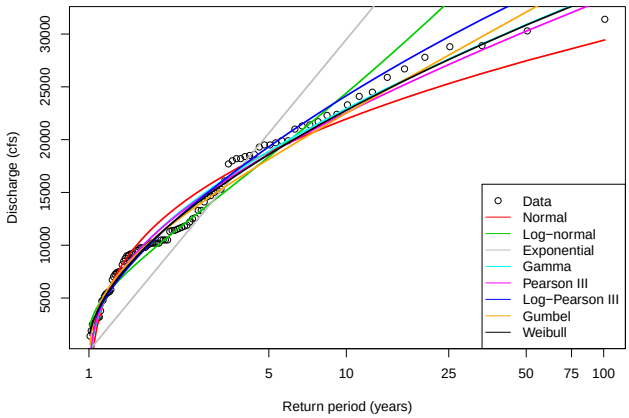
[69] Worton, B., Carvalho, M. (2018). *Statistical Methodology - Lecture Notes.* The University of Edinburgh.

[70] Yang, F., Ren, H., Hu, Z. (2019). *Maximum likelihood estimation for three-parameter Weibull distribution using evolutionary strategy.* Mathematical Problems in Engineering.

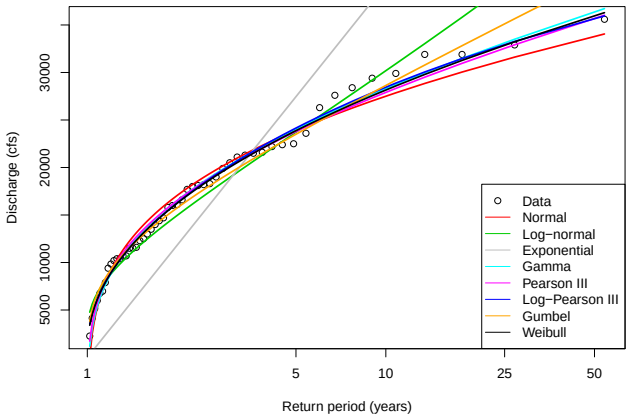
Appendix A

Additional graphical results

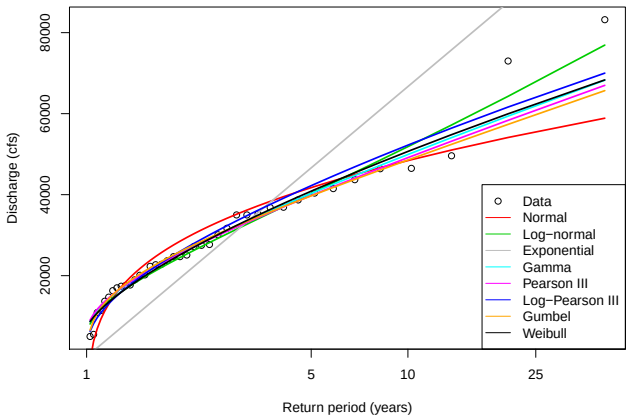
Figure A.1: Observed and estimated peak streamflows for different stations at Pearl river.



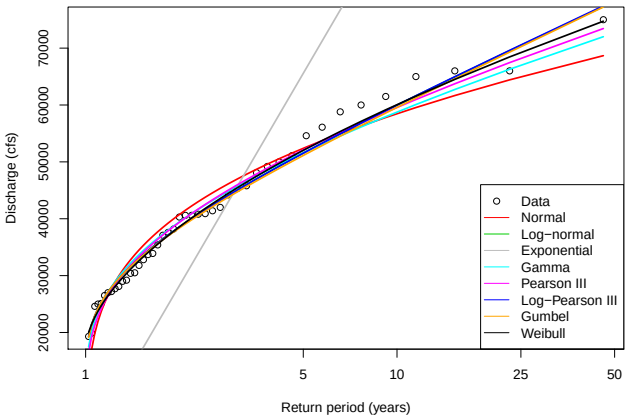
(a) Edinburg



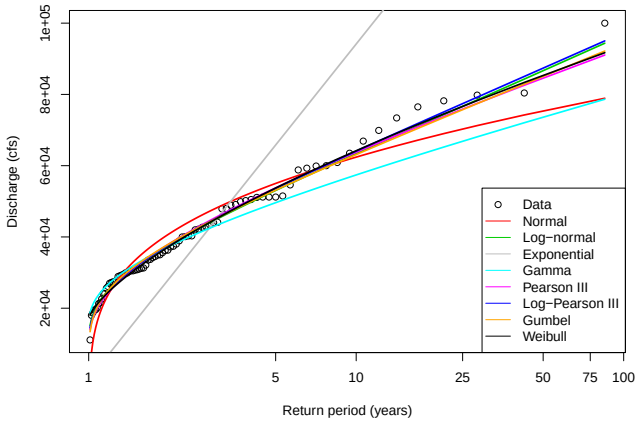
(b) Carthage



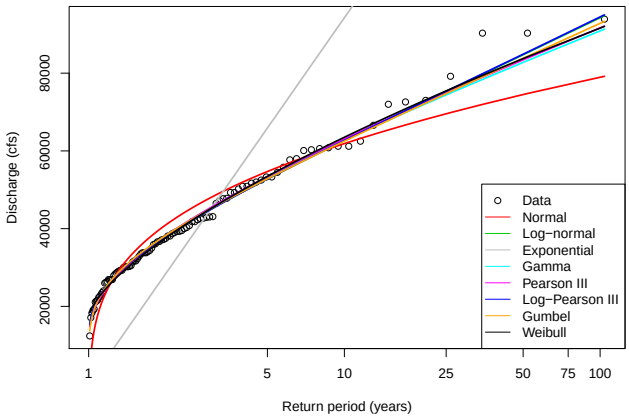
(c) Lena



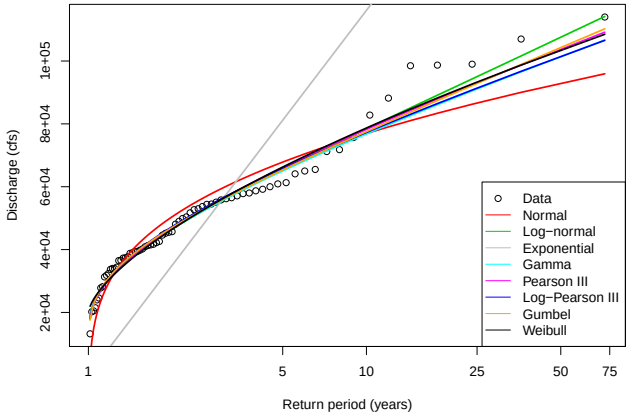
(d) Rockport



(e) Monticello



(f) Columbia



(g) Bogalusa