Final Report

On

Design flood estimation for small structures in the South Bihar area

NATIONAL INSTITUTE OF HYDROLOGY

ROORKEE – 247 667 UTTARAKHAND

September, 2023

Design Flood Estimation for Small Structures in the South Bihar Area

Study Group

Pankaj Mani, Scientist F

Sh Jagdish Prasad Patra, Scientist D

Sh Biswajit Chakravorty, Scientist G

I C Thakur, Director WALMI

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

This study presents the estimation of design basis flood and safe grade elevation for river basins in South Bihar using L-moment based Regional Flood Frequency Analysis (RFFA). Annual peak flow data from multiple gauging stations were analyzed, and inconsistent and insufficient datasets were screened to ensure reliability. Homogeneous regions were identified through hierarchical clustering based on hydrological characteristics such as coefficient of variation and mean flow. The clustering results indicated distinct groups, with the majority of stations exhibiting similar hydrological behavior.

Regional homogeneity was assessed using L-moment based heterogeneity measures, confirming an acceptably homogeneous region after exclusion of discordant sites. Various probability distributions were evaluated using L-moment ratio diagrams and goodness-of-fit statistics, with the Generalized Extreme Value (GEV) distribution identified as the best fit for the region. Regional flood frequency relationships were subsequently developed for estimating flood magnitudes corresponding to different return periods.

Further, relationships between peak flood and catchment characteristics such as area, stream length, slope, rainfall, and forest cover were established using multivariate regression techniques. Among linear, logarithmic, and exponential models, the exponential regression model incorporating catchment area and rainfall provided the best fit. These relationships were extended to ungauged catchments, enabling estimation of flood magnitudes using physiographic parameters and regional growth factors.

The study provides a comprehensive framework for flood estimation in both gauged and ungauged basins, supporting improved flood risk assessment, hydraulic design, and water resources planning in the region.

1 Introduction

Design flood estimation is a basic pre-requisite for design of any water resources projects/ hydraulic structures such as dams, spillways, road and railway bridges, culverts, urban drainage systems. Even for planning of non structural measures of flood management such as flood plain zoning and regulation, estimation of design flood is a basic and essential activity. For design flood estimation long term observed hydrological/ meteorological data are required. In India, several state and central agencies are maintaining Hydro-meteorological sites (HMS) and collecting data, still the sampling density is very scanty. Within country geographical area, the Central agency Central Water Commission altogether maintains 901 HMS sites out of which 589 sites are for discharged measurement and that too are on major rivers and their tributaries only (India-WRIS, 2022). In Bihar, CWC maintains 57 discharge sites, mostly on the main Ganga river or its tributaries in north Bihar and very few (only 11) in south Bihar (LGBO, 2022). Water Resources Department also maintains some of the discharge sites in Bihar (22 stations in south Bihar as per water year book 2018-19), although the long term data is not available for many of the sites. Thus many of the streams are still ungauged. One of the major challenges faced by field engineers/ hydrologist is to predict the design flow in ungauged basin. In absence of river flow records at or near the site of interest, it is difficult to derive reliable design flood estimates directly. In such a situation, regional flood frequency relationships developed for the region are one of the alternative methods for prediction of design floods, especially for small-to medium-size catchments.

The approaches for design flood estimation may be broadly categorized in two categories; (i) deterministic approach using design storm and unit hydrograph (UH); and (ii) probabilistic approach involving flood frequency analysis. The deterministic and probabilistic methods, which have been used for design flood estimation, are empirical methods, rational method, flood frequency analysis methods, unit hydrograph techniques, and watershed models. Pilgrim and Cordery (1993) mention that estimation of peak flows on small- to medium-sized rural drainage basins is

probably the most common application of flood estimation as well as being of greatest overall economic importance. In almost all cases, no observed data are available at the design site, and little time can be spent on the estimate, precluding use of other data in the region. The author further state that hundreds of different methods have been used for estimating floods on small drainage basins, most involving arbitrary formulas. The three most widely used types of methods are the rational method, the US Soil Conservation Service (USCS) method, and regional flood frequency methods. Regional flood frequency analysis resolves the problem of short data records or unavailability of data by "trading space for time"; as the data from several sites are used in estimating flood frequencies at any site. The choice of method primarily depends on design criteria applicable to the structure and availability of data.

As per Bureau of Indian standards (BIS) hydrologic design criteria, frequency-based floods find their applications in the estimation of design floods for almost all the types of hydraulic structures, viz., small-size dams, barrages, weirs, road and railway bridges, cross-drainage structures, flood control structures, excluding large- and intermediate-size dams. To overcome the problems of prediction of floods of various return periods for ungauged and sparsely gauged catchments, a robust procedure of regional flood frequency estimation is required to be developed. (IS-7784-1, 1993, IS 12094 (2000), IS-14815:2000)

Design flood estimation is the pre requisite for the design of various hydraulic structures. Often, this information is required at locations where stream flow series are too short to allow a robust estimation of flood quantiles corresponding to long return periods or where no data are available at all. Regional flood frequency analysis such as the index flood method offers a solution to this problem and has widely been used to estimate flood quantiles in such situations. The idea is to compensate for the lack of temporal data by spatial data, taken within a region with similar flood behavior and transfer information from gauged to ungauged sites. The underlying assumption is that flood data within a homogeneous region is drawn from the same frequency distribution, apart from a scaling factor. In this study, regional flood frequency based design flood estimation for small structures in South Bihar area is proposed with the following objectives:

(i) To test the regional homogeneity of the study area

(ii) To identify the robust distribution for the study area based on L-moment ratio diagram and Zidist statistics.

(iii) To develop RFF relationship for estimation of floods of various return period for gauged catchments of study area uning L-moment approach.

(iv) To develop regional relationship between mean annual peak floods and physiographic characteristics for estimating the mean annual peak flood for ungauged catchments of the study area .

(v) To develop regional flood formula for estimation of floods of various return periods for ungauged catchments by coupling the relationship between mean annual flood and physiographic characteristics.

2 Literature Review

A regional flood frequency relationship is developed for gauged catchments based on the robust identified frequency distribution. This relationship is coupled with the regional relationship between mean annual peak flood and catchment and a regional flood frequency relationship is also developed for ungauged catchments of the study area. Flood frequency estimates of the gauged and ungauged catchments based on data of the gauging sites constituting the homogeneous region and the available data of all the gauging sites.

2.1 Flood Frequency Analysis

In flood frequency analysis, a unique relationship between a flood magnitude (Q) and the corresponding average recurrence interval (T) is sought. The task is to extract information from a set of streamflow records to estimate the relationship between Q and T. Statistical methods are generally used for flood frequency analysis as quantifying the physical processes that determine a flood magnitude is often associated with a high degree of uncertainty. Three different models may be considered for the purpose of flood frequency analysis. These models are (1) the annual maximum flood series (AM) model, (2) the partial duration series (PD) or the peaks over threshold (POT) model, and (3) the time series (TS) model. In the (AM) series, only the peak flow in each year of record is considered. Most flood frequency analysis techniques are based on AM series. Flood peaks do not occur with any fixed pattern in time or magnitude. Time intervals between floods vary. The definition of return period is the average of these inter-event times between flood events (Cunnane, 1989). This is also called average recurrence interval (ARI).

Large floods naturally have large return periods and vice versa. The definition of the return period may not involve any reference to probability. However, a relationship between the probability of occurrence of a flood and its return period can be justified. A given flood q with a return period T may be exceeded once in T years. Hence the probability of exceedance is P(QT>q)=1/T. The cumulative probability of non–exceedance, $F(Q_T)$ is given by Equation (1)

$$\boldsymbol{F(Q_T) = P(Q_T \leq q) = 1 - P(Q_T > q) = 1 - \frac{1}{T}} \quad \textbf{(2.1)}$$

Eq 1 is the basis for estimating the magnitude of a flood, Q_T given the return period T.

Often, the observed flood series data are plotted on probability paper to check whether they follow a particular distribution, to detect data errors and to check for outliers. Probability plots require an initial estimate of the probability of non-exceedance which is called a "plotting position". A plotting position formula which is used often is that given by Cunnane (1989):

$$\boldsymbol{F = \frac{i-0.4}{N+0.2}} \qquad \textbf{(2.2)}$$

where N is the sample size and i is the rank of the observations in ascending order. Some other commonly used plotting position formulas can be found in Cunnane (1989).

The data used in flood frequency analysis is assumed to be independent and identically distributed. The flood data are considered to be stochastic. Further it is assumed that the flood data have not been affected by natural or man made changes in the hydrological regime.

In practice there are many pitfalls to these assumptions. The assumption that the data in a given system arise from a single parent distribution is subject to question especially when large catchments are being analysed. In circumstances such as this more than one type of rainfall or flow may contribute to the extreme events in a region of interest. These assumptions have been questioned and discussed extensively in Klemes, (1987a, 1987b) and Yevjevich (1986).

2.1.1 Regional Flood Frequency Analysis

The availability of streamflow data is an important aspect in any flood frequency analysis. The estimation of probability of occurrence of extreme floods is an extrapolation based on limited data. Thus the larger the data set, the more accurate the estimates will be. From a statistical view point, estimation from a small sample may give unreasonable or physically unrealistic parameter estimates, especially for distributions with a large number of parameters (three or more). Large variations associated with small sample sizes cause the estimates to be biased. In practice, however, data may be limited or in some cases may not be available for a site. In such situations, regional flood frequency analysis (RFFA) is most useful.

RFFA is a technique of transferring information from gauged suites to ungauged sites. RFFA serves two purposes. For sites where data are not available, the analysis is based on regional data (Cunnane, 1989). For sites with limited data, the joint use of data measured at a site, called at - site data, and regional data from a number of stations in a region provides sufficient information to enable a probability distribution to be used with greater reliability. This type of analysis represents a substitution of space for time where data from different locations in a region are used to compensate for short records at a single site (National Research Council, 1988; Stedinger et al., 1993).

2.1.2 Regional Homogeneity

RFFA is based on the concept of regional homogeneity which assumes that annual maximum flood populations at several sites in a region are similar in statistical characteristics and are not dependant on catchment size (Cunnane, 1989). Although this assumption may not be strictly valid, it is convenient and effective in most applications.

One of the simplest RFFA procedures that has been used for a long time is the index flood method. The key assumption in the index flood method is that the distribution of floods at different sites within a region is the same except for a site - specific scale or index flood factor. Homogeneity in regards to the index flood relies on the concept that the standardized regional flood peaks have a common probability distribution with identical parameter values.

The identification of homogenous regions is an elementary step in RFFA (Bates et al., 1998). The application typically involves the allocation of an ungauged catchment to an appropriate homogenous group and the prediction of flood quantiles using developed models based on catchment characteristics (Bates et al., 1998). That is, the RFFA based on homogenous regions can transfer the information from similar gauged catchments to ungauged catchments to allow for flood prediction.

There have been many techniques developed which attempt to establish homogenous regions. For example the PRM uses geographical contiguity as an indication of homogeneity that is the catchments which are nearby to each other should have similar runoff coefficients.

Looking at homogeneity from a theoretical point of view, two catchments may be treated as homogenous with respect to flood behaviour if they both satisfy two criteria: the inputs (such as rainfall) to the hydrological systems are identical, and the climatic and physical characteristics changing the input to flood peak are the same. No two catchments can satisfy these criteria perfectly based on the fact that each catchment has a unique physical characteristic and that each catchment has different climatic inputs. The question remains, in the search for practical "homogeneity", one has to make decisions on the degree of similarity or dissimilarity that is acceptable and deciding a cut-off point where a region is acceptably homogenous or heterogeneous, in consideration of the practical applications of the techniques.

In defining homogenous regions for use in RFFA, a balance has to be made between including more sites for increased information and maintaining an acceptable level of homogeneity. In most situations when more sites are added to a region, certainly more information is gained about the flood regime; however, sites that are hydrologically dissimilar can increase the heterogeneity in the region.

2.1.3 Inter – Site Dependence

Some RFFA methods make use of inter – site dependence while others do not. Inter – site dependence as reported by Cunnane (1988) states that streamflow data points across a region will show similar behaviour within any given timeframe. This means that;

1) In some years the annual maximum flows at all sites are due to a single widespread meteorological event.

2) In relatively dry years, peak flows are generally low over the entire region, in which case all annual maxima will be low.

To be able to counteract these trends in RFFA, previous studies have indicated that a concurrent record of sufficient length should be adopted (Stedinger, 1983a).

2.1.4 Distributional Choices

The choice of an appropriate probability distribution to be used in flood frequency analysis has been a topic of interest for a long time and is of prime importance in at-site and RFFA. It has received widespread attention by researchers. Benson (1968) and NERC (1975) devote

considerable attention to this problem. Cunnane (1989) summarised the distributions commonly used in hydrology, mentioning 14 different distributions.

In some countries, a common distribution has been selected to achieve uniformity between different design agencies. The U.S.A Interagency Advisory Committee on Water Data (IACWD, 1982) and the Institution of Engineers Australia (I. E. Aust., 1987) recommend the Log Pearson Type 3 (LP3) distribution for use in United States and Australia respectively. Other distributions that have received considerable attention include Extreme Value Types 1, 2, 3, Generalised Extreme Value (GEV) (NERC, 1975), Wakeby (Houghton, 1978), Generalised Pareto (GPA) (Smith, 1987), Two-component Extreme Value (Rossi et al., 1984) and the Log-Logistic distribution (Ahmad et al., 1988).

The use of a standard distribution has been criticised by Wallis & Wood (1985) and Lettenmaier & Potter (1990). They argue that a reassessment of the use of the LP3 distribution for practical flood design is overdue. Vogel et al. (1993) studied the suitability of a number of distributions (including the LP3) for Australia. They found that the Generalised Extreme Value (GEV) and Wakeby distributions provide the best approximation to flood flow data in the regions of Australia that are dominated by rainfall during the winter months; for the remainder of the continent, the Generalised Pareto (GPA) and Wakeby distributions provide better approximations. For the same data set, the LP3 performed satisfactorily, but not as well as either the GEV or GPA distribution. The distributions that have attracted the most interest as possible alternatives to the LP3 are the GEV and Wakeby. Studies by Rahman et al. (1999b) and Haddad and Rahman (2007) showed that GEV-LH moments method provide better results than the LP3 distribution in South–East Australia.

2.2 Methods for Identification of Homogeneous Regions

The methods for obtaining homogenous regions are based on either geographical contiguity or flood characteristics alone or catchment characteristics alone. The theoretical aspects, limitations and associated problems with identification of homogenous regions based on flood data (annual maximum series) are discussed below.

In this approach, the degree of homogeneity of a proposed group is judged on the basis of a dimensionless coefficient of the annual maximum flood series, such as the coefficient of variation (Cv), coefficient of skewness (Cs) or similar measures. Examples are given by Pilon and Adamowski (1992), Lu and Stedinger (1992), Hosking and Wallis (1993) and Fill and Stedinger (1995a, b).

Hosking and Wallis (1991, 1993) proposed a heterogeneity measure based on the L moment ratios L CV, L CS and L kurtosis. The advantages of this test are that it is based on L moments and not distribution-specific like those mentioned above. This test has received considerable attention in recent years (Rahman et al., 1999b).

Cunnane (1988) mentioned that identification of a homogeneous region is necessarily based on statistical tests of hypothesis, the associated power of which, with currently available amounts of hydrological data, is low. Thus it is not possible to divide, with great assurance, a large number of catchments into homogeneous subgroups using flow records with limited lengths.

2.3 Regional Flood Frequency Analysis Methods

There are a number of RFFA methods based on streamflow data that have been reported. Some of the most commonly used methods are discussed below.

2.3.1 Index Flood Method

Most RFFA methods involve some sort of standardisation of the data. If the standardisation is of the form X = Q/b where b is an index of the overall flood magnitudes on a catchment, then it is referred to as an index flood method. The underlying concept of this method is that the distributions of floods at sites of a homogeneous region are identical, apart from a site-specific scaling factor b (the "index flood"), which reflects the size, rainfall, and runoff characteristics of each catchment. It is assumed that all the moments of order higher than one are identical after correction for scale. Usually b is taken as the mean flood, Q , or the median flood ~Q of the site. When b = Q , then the variate X has the properties (Cunnane, 1988): E(X) = 1; s X = Cv (Q) ; gX = gQ ; where E(X) = expected value of X; sX = standard deviation of X values; Cv = coefficient of variation; and g = skew coefficient. For small samples, X is the ratio of two random variables, rather than a single scaled random variable. Stedinger (1983) showed that its effects can be quite marked in small samples, in that the distribution of X is

quite different in form to that of Q . Further, if samples are small, the variance of Q contributes appreciably to sampling variance of estimated X quantiles.

The dimensionless rescaled data Xij = Qij/bj are the basis for estimating XT, the regional growth factor of ARI of T years. The index flood method that used records of equal length for each sites which have been tested for homogeneity at the 10 year ARI level. This method, though widely used, has some limitations as mentioned in Section 2.3. A slightly different type of index flood method has been proposed by Hosking and Wallis (1990, 1993), in which it is assumed that the distribution of XT is known apart from the distribution parameters qk (k = 1, ..., p). These parameters are estimated separately at each site and combined to give regional estimates:

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{k}}^{\boldsymbol{R}} = \frac{\sum_{j=1}^{M} w_j \hat{\theta}_k^{(j)}}{\sum_{j=1}^{M} w_j} \qquad \textbf{(2.3)}$$

Where $\hat{\theta}_k^{(j)}$ is the site j estimate of θ_k and $w_j = N_j$ are the weights. Equation 2 gives a weighted average, the site estimate being given weight proportional to its record length because, for regular statistical models, the variance of $\hat{\theta}_k^{(j)}$ is inversely proportional to Nj (Hosking and Wallis, 1993). Some research (Lettenmaier and Potter, 1985; Wallis and Wood, 1985; Hosking and Wallis, 1986; Potter and Lettenmaier, 1990; Rahman, 1997 and Rahman et al., 1999b) has found that index flood procedures, coupled with probability weighted moments or L moments can yield reasonably accurate quantile estimation.

In the Index Flood Method, the dimensionless regional growth curve is used to estimate X_T.
The flood quantile having an ARI of T year is then obtained from:

$$Q_T = X_T \bar{Q} \qquad (2.4)$$

In the case of a gauged site, the at-site mean flood is used in Equation 2.4; for an ungauged site, Q is estimated using regional information. Equation in 2.4 is based on the following variables:

Q_T , which is the flood quantile at a site, with an ARI of T years; and

X_T , which is the regional growth factor, this defines the frequency distribution common to all the sites in a homogenous region.

$\bar{Q}$, which is the known as the flood index, is typically represented (in gauged catchments) by the mean of the at – site annual maximum flood series. Being used as a scale parameter, it is recognised as the term which dictates the difference in quantiles between individual sites than homogenous groups.

Equation 4 is the essence of the index flood method. In the case where a site has gauging information, the at – site mean flood is used to represent $\bar{Q}$. However, when the method is to be applied to the ungauged catchment case where there is little or no data available the difficulty in estimating $\bar{Q}$ becomes evident. Estimation such as this is typically performed via multiple regression between flood series (annual maximum series from gauged catchments in the region) and catchment/climatic characteristic within the region. The general form of the regression equation can be expressed as:

$$\bar{Q} = aB^b C^c D^d \quad \textbf{(2.5)}$$

where B, C, D, … are catchment characteristics and a, b, c, d, … are parameters of the regression equation.

A significant amount of research has been conducted in regards to the index flood method both in the past and more recently. Dalrymple (1960) was one of the first to develop an index flood technique which was used by the USGS prior to 1965. The method developed by Dalrymple was to relate annual flood series to catchments area for a particular region of interest. Relationships were then sought on geographical representation; the particular area was then divided into divisions based on similarity (Riggs, 1973).

The second part of Dalrymple's approach involved averaging the shapes of similar curves for the region to create one similar common curve; this method was relatively easy to implement as only one variable was required which was catchment area. As this approach is an empirical one a number of limitations have been identified:

1). Arbitrary decisions are required at boundaries of regions with respect to mean annual flood and the shape of the frequency curve.

2). There was no consideration of other important factors which have shown to be plausible/influential in the flood generation process (Riggs, 1973).

Australian Rainfall & Runoff (ARR) (I.E Aust, 1987) did not recognise the superiority of the index flood method (IFM) as a design flood estimation technique. ARR (I.E Aust, 1987) was also very critical of the way the IFM had been developed in the past and commented that the regression techniques used are limited in their functional form and lack of a sound physical and statistical basis. More research had been conducted by Lettenmaier and Potter (1985), Wallis and Wood (1985), Hosking and Wallis (1986) and Potter and Lettenmaier (1990) who showed that the index flood procedures, coupled with probability weighted moments or L- moments yielded more accurate quantile estimates.

The index flood method had been criticised on the grounds that the coefficient of variation of the flood series v C may vary approximately inversely with catchment area, thus resulting in flatter flood frequency curves for larger catchments. This had particularly been noticed in the case of humid catchments that differed greatly in size (Dawdy, 1961; Benson, 1962; Riggs, 1973; Smith, 1992).

There have been recent studies carried out by Bates et al. (1998) and Rahman et al. (1999a) where the development of an application for design flood estimation in ungauged catchments in South - east Australia was tested using IFM. The method involved the assignment of ungauged catchments to a particular homogenous group identified (through the use of L-moments) on the basis of catchment and climatic characteristics as opposed to geographical proximity. The relationships sought were carried out by statistical procedures such as canonical correlation analysis; tree based modelling and other multivariate statistical techniques. This allowed for the development of a RFFA method using up to 12 independent climatic and catchment characteristics variables.

Although the results of this method showed promise when compared to the Probabilistic Rational Method (PRM) its limitations were already evident in that it needed a large number of independent variables which are very time consuming to obtain. The results of this method also depend upon the correct catchment assignment to a homogenous group; thus, any wrong assignment would greatly increase error in quantile estimation.

2.3.2 Station Year Method

The standardised X values of all the sites in the region are treated as if they form a single random sample of size L from a common X parent population. The pooled standardised data are then fitted to a suitable distribution, and XT values are calculated. Since this method ignores inter-site dependence, it may lead to bias, especially at large return periods (Cunnane, 1988).

2.3.3 Bayesian Method

This method allows an unknown parameter to be estimated as a random variable rather than as a fixed unknown constant. It does not require standardisation of flood data or the assumption of regional homogeneity, like the index flood method. Cunnane and Nash (1971) and Cunnane (1988) discussed the Bayesian approach to RFFA. Kuczera (1982b) proposed a "Linear Empirical Bayes" approach for RFFA, but Lettenmaier and Potter (1985) found in a simulation study that this approach does not provide quantile estimates as precise as estimates obtained from index flood methods. Cunnane (1988) mentioned that the Bayes approach, though theoretically attractive, is not as 'precise' as the across region averaging method based on probability weighted moments (PWMs).

2.3.4 Probabilistic Rational Method

The rational method has often been regarded as a deterministic representation of the flood generated from an individual storm. However, the rational method recommended in ARR, (I. E. Aust., 1987) (see also Pilgrim and Cordery, 1993), is based on a probabilistic approach for use in estimating design floods. This "Probabilistic Rational Method" (PRM) is given by:

$$Q_T = 0.278\, C_T\, I_{tcT}\, A \qquad (2.6)$$

where Q_T is the peak flow rate (m3/s) for an ARI of T years; C_T is runoff coefficient (dimensionless) for ARI of T years; $I_{tc\ T}$ is average rainfall intensity (mm/h) for a design duration of time of concentration tc hours and ARI of T years; and A is the catchment area (km2).

The method may be regarded as a regional model, with design rainfall intensity I tc ,T and catchment area A as independent variables. The runoff coefficient C_T is a factor which lumps the effects of climatic and physical characteristics, other than catchment area and rainfall intensity. It is noteworthy that in ARR 1987 the values of C_T were estimated using conventional moment estimates from flow records of limited lengths e.g. some sites had only 10 years of records. Since conventional moment estimates are largely affected by sampling variability and extremes in the data, a higher degree of uncertainty in quantile estimation is likely to arise due to C_T. The mapping and use of runoff coefficients are based on the assumption of geographical contiguity, an assumption that is unlikely to be satisfied, as mentioned in Section 2.3.

Lay (1989) was able to report on the PRM and its application for ungauged rural catchments in Victoria, which confirmed that the PRM is in accordance with criteria set out by the National Committee on Water Resources. However, the method is only as good as the data it is derived from, hence to keep the PRM up to date and workable, it was recommended the database within ARR should be updated at regular intervals (e.g. every 5 years). It is clearly known that this recommendation has not been fulfilled as there has been no update of the PRM since 1987.

Rahman & Hollerbach, (2003) investigated the physical significance of runoff coefficients and assessed the extent of uncertainty of design flood estimates obtained by the PRM. By following the method of derivation in ARR, runoff coefficients were estimated for 104 gauged catchments in South east Australia. The mapping of these C_{10} coefficients onto a suitable map of the area indicated that C_{10} coefficients show little spatial coherence. The C coefficients are mapped according to the position of the gauging station and some interpolation is then required for areas where there is little or no data so that the contours can be developed. The error introduced into the contours is through the interpolation technique; this is due to the fact that some regions will be exposed to greater spatial changes in physical topography and other factors which will directly affect the C_{10} coefficients.

Rahman (1997) stated the underlying concept of contiguous regions, ie nearby catchments are hydrologically similar is true to the extent that contiguous areas are likely to have similar meteorological characteristics and therefore similar hydrological inputs. But geographical proximity cannot be a guarantee for hydrological similarity, as two nearby catchments may

possess quite dissimilar physical characteristics. Geographical regions may cut across geologic, climatic and topographic boundaries, causing abrupt changes in hydrological parameters at their boundaries (Wiltshire, 1986). A geographical region that is called homogeneous may include catchments exhibiting a wide variety of catchment characteristics, also with very different flood characteristics (Wiltshire, 1986c; Acreman and Wiltshire, 1989). In a very similar fashion Rahman and Holerbach (2003) also stated that while nearby catchments shows similar meteorological characteristics, they may possess quite dissimilar physical characteristics, which clearly indicates that the method of simple linear interpolation over a geographical space on the map of C_{10} in ARR (I.E Aust, 1987) has little validity.

Rahman and Hollerbach (2003) also examined the uncertainty associated with design flood estimation with use of the PRM adopting the developed C10 coefficients. This study showed that for about 40% of the catchments, the PRM underestimated the observed flood quantiles. The developed C10 coefficients however, did show reasonable correlation with pan evaporation, quaternary sediment area, stream density and mainstream slope. An attempt was made to develop prediction equations for C10 coefficients and catchment characteristics however this proved to be unsuccessful.

2.4 Quantile Regression Technique

United States Geological Survey (USGS) proposed a QRT where a large number of gauged catchments are selected from a region and flood quantiles are estimated from recorded streamflow data, which are then regressed against catchment variables that are most likely to govern the flood generation process. Studies by Benson (1962) suggested that T-year flood peak discharges could be estimated directly using catchment characteristics data by multiple regression analysis.

The quantile regression technique can be expressed as follows:

$$\boldsymbol{Q = aB^b C^c D^d} \qquad \textbf{(2.7)}$$

Where B, C, D, … are catchment characteristics variables and T Q is the flood magnitude with T year ARI (flood quantile), and a, b, c, … are regression coefficients.

This method is not based on a constant coefficient of variation (Cv) of annual maximum flood series in the region like the index flood method. It has been noted the method can give design flood estimates that do not vary smoothly with ARI; however, hydrological judgment can be exercised in situations such as these when flood frequency curves need to be adjusted to increase smoothly with T.

There have been various techniques and many applications of regression models that have been adopted for hydrological regression. Most of these methods are derived from the methodology set out by the USGS as described above.

The USGS for a long time have been applying the QRT. A well-known study using the QRT with an Ordinary Least Squares (OLS) procedure had been carried out by Thomas and Benson (1970). The study tested four regions in the United States for design flood estimation using multiple regression techniques that related streamflow characteristics to drainage-basin characteristics. This study found that the QRT was predicting quantiles estimates quite accurately as compared to previous methods adopted by the USGS. However, there was still the point made that the equations were lacking statistically sound methodology.

The OLS estimator has traditionally been used by hydrologists to estimate the regression parameters b in regional hydrological models. But in order for the OLS model to be statistically efficient and robust, the annual maximum flood series in the region must be uncorrelated, all the sites in the region should have equal record length and all estimates of T year events have equal variance.

Since the annual maximum flow data in a region do not generally satisfy these criteria, the assumption that the model residual errors in OLS are homoskedastic is violated and the OLS approach can provide very distorted estimates of the model's predictive precision (model error) and the precision with which the regression model parameters are being estimated (Stedinger and Tasker, 1985).

To overcome the above problems in OLS, Stedinger and Tasker (1985) proposed the Generalised Least Squares (GLS) procedure which can result in remarkable improvements in

the precision with which the parameters of regional hydrologic regression models can be estimated, in particular when the record length varies widely from site to site. In the GLS model, the assumptions of equal variance of the T year events and zero cross-correlation for concurrent flows are relaxed.

The GLS procedure as described by Stedinger and Tasker (1985) and Tasker and Stedinger (1989) require an estimate of the covariance matrix of residual errors $\sum(Y)$.

2.4.1 Generalised Least Squares (GLS) And Weighted Least Squares (WLS)

As discussed above, the parameters of regional hydrological models have been estimated using the OLS procedure. However, regionalisation using hydrological data violates the assumption that the residual errors associated with the individual observations are homoskedastic and independently distributed (Stedinger and Tasker, 1985). In the case of hydrological data, variations in streamflow record length and cross – correlation among concurrent flows result in estimates of the T year events which vary in precision.

What has received great attention in the US is how to best estimate the parameters of a regional hydrological model given the limitations of OLS which will not identify the efficient estimates of a regression model parameters when the residual errors are not homoscedastic and independently distributed (Stedinger and Tasker, 1985).

Moreover, as shown in the studies cited above, OLS estimates of the standard error of prediction and the estimated parameters are highly biased. Weighted and Generalised Least Squares techniques were developed to deal with situations like those encountered in hydrology where a regression models residuals are heteroscedastic and perhaps cross –correlated (Draper and Smith, 1981; Johnston, 1972). Tasker (1980), has in fact, used a Weighted Least Squares (WLS) procedure to account for unequal record lengths. Marin (1983) and Kuczera (1982a, b, 1983) developed a Bayesian and Empirical Bayesian methodology which deals with these issues.

An obstacle to the use of WLS and GLS procedures with hydrological data is the need to provide an estimate of the covariance matrix of residual errors; that covariance matrix is a

function of the precision with which the true model can predict values of the streamflow statistics of concern as well as the sampling error in the available estimates of that statistic. The discussions and examples in the works by Tasker (1980) and Kuczera (1983b) illustrate difficulties associated with estimation of this matrix.

Stedinger and Tasker (1985) showed in a Monte Carlo simulation with synthetically generated flow sequences, a comparison of the performance of the OLS procedure with that of a GLS procedure. In situations where the available streamflow records at gauged sites are of different and widely varying length and concurrent flows at different sites are cross correlated, the GLS procedure provided more accurate parameter estimates, better estimates of the accuracy with the regression models parameters were being estimated, and almost unbiased estimates of the variance of the underlying regression model residuals.

A simpler WLS procedure neglects the cross correlations among concurrent flows. The WLS algorithm is shown to do as well as the GLS procedure when the cross correlation among concurrent flows are relatively modest.

2.4.2 Application of Generalised Least Squares Regression

Tasker et al. (1986) compared the GLS estimation technique of Stedinger and Tasker (1985) with OLS estimation and polynomial estimation in a split – sample experiment, in which real data from Pima County, Arizona were used. Two conclusions were drawn from their study. First, of the data sets considered, the differences between the model parameter estimates obtained with OLS and GLS procedures were quite modest. This can be reflected in the fact that most sites had less than 20 years of streamflow data. This, coupled with large model – error variance, meant that the sampling-error term had little effect on the analysis. In addition to this, most of the cross correlations seemed very small. The second conclusion is that the GLS method provides a nearly unbiased estimate of the true variance of prediction, while the OLS approach substantially over estimates the true prediction variance. None the less Tasker et al. (1986) found from a statistical stand point, the method is satisfying because it deals with the problem of cross – correlated data and unequal variance between sites in a sound and logical manner.

Tasker and Stedinger (1987) propose an adjustment to the GLS model of Stedinger and Tasker (1985) to account for possible information about historical floods available at some stations in a region. The historical information is assumed to be in the form of observations of all peaks above a threshold during a long period outside the systematic record period. A Monte Carlo simulation experiment was performed to compare the GLS estimator adjusted for historical floods with the unadjusted GLS estimator and the OLS estimator. The results indicated that (1) using the GLS estimator adjusted for historical information significantly improves the regression model; (2) The modified GLS method described in Tasker and Stedinger (1987) outperforms the widely used OLS method in estimating regression model parameters.

2.4.3 Operational GLS Model for Hydrologic Regression

Monte Carlo simulation studies were undertaken by Stedinger and Tasker (1985) which documented the value of GLS procedures to estimate empirical relationships between streamflow statistics and physiographic basin characteristics. Tasker and Stedinger (1989) presented an extension of the GLS method that deals with the realities and complexities of regional hydrological data sets that were not addressed in the Monte Carlo simulation studies. These extensions include (1) a more realistic model of the underlying model error; (2) smoothed estimates of cross correlation of flows; (3) procedures for including historical flow data; (4) diagnostic statistics describing leverage and influence for GLS regression. Thus implementation of the GLS regression method employed by Stedinger and Tasker (1985) requires these new extensions to be incorporated into the model especially in regards to identifying the realistic model error associated with the GLS analysis.

2.4.4 Operational Bayesian GLS Regression for Regional Hydrologic Analysis

Reis et al. (2003, 2005) introduced a Bayesian approach to parameter estimation for the GLS regional regression model developed by Stedinger and Tasker (1985) for hydrological analysis. The results presented in Reis et al. (2005) show that for cases in which the model error variance is small compared to sampling error of the at – site estimates, which is often the case for regionalisation of a shape parameter, the Bayesian estimator provides a more reasonable estimate of the model error variance than the Method of Moments (MOM) and Maximum Likelihood (ML) estimators. This paper by Reis et al. (2005) also presents regression statistics for WLS and GLS models including pseudo analysis of variance, a pseudo R^2, error variance

ratio (EVR) and variance inflation ratio (VIR), and leverage and influence. The regression procedure was illustrated with two examples of regionalisation. Results obtained from OLS, WLS and GLS procedures were compared. The OLS procedure provided very misleading results because it did not make any distinction between the variance due to the model error and the variance due to the sampling error. For the examples presented, the GLS method was found to provide the best framework because the cross correlation between concurrent flows proved to be important. Both leverage and influence statistics were very useful in identifying stations that did have a significant impact on the analysis.

2.4.5 Use of GLS Regression In Regional Hydrologic Analysis

Griffis and Stedinger (2007) looked at the GLS regression method in more detail. Previous studies by the US Geological Survey using the LP3 distribution have neglected the impact of uncertainty on the weighted skew on quantile estimation. The needed relationship has been developed in this paper and its use is also illustrated in a regional flood study with 162 sites from South Carolina. The performance of this model is compared to separate models for each hydrological region tested. The results were both surprising and hydrologically reasonable. This paper also looks at new statistical diagnostic metrices such as a condition number to check for multicollinearity, a new pseudo R2 appropriate for use with GLS regression, and two error variance ratios.

2.4.6 Application of Generalised Least Squares to Low-Flow Frequency Analysis

Vogel and Kroll (1990) undertook a study to compare the GLS and OLS regression procedures in developing generalised low-flow frequency relationships for ungauged sites in Massachusetts. The GLS regression procedures let to almost identical regional regression model parameter estimates when compared to the OLS procedures. Although the GLS procedures led to only marginal gains in the prediction errors associated with low flow regional regression equations, that result only reflects the fact that all sites had at least eleven years of data, and most had more than twenty years of data. In addition, the large model error component of the total prediction errors implies that the sampling error had only a small impact on the analysis. Vogel and Kroll (1990) made note of the fact, that the GLS procedure will have significant advantages over OLS procedures in studies which seek to include very short

records such as at partilal record sites. In such instances, GLS procedures can lead to significant improvements because the number of sites included in the analyses can be increased considerably.

Kroll and Stedinger (1999) examined the development of regional regression relationships with censored data for low-streamflow statistics. The basic problem was when no discharge record is available for a site; a regional regression relationship can be developed to estimate the low flow quantiles. The problems that arise in the derivation of such models are when some at-site estimates are reported as zero. One concern is that quantile estimates reported as zero may be in the range from zero to the measurement threshold. A second concern is that logarithmic transformation cannot be used with zero quantile estimates, so traditional log linear least squares estimators cannot be computed. The study by Kroll and Stedinger (1999) uses visual examples and Monte Carlo simulation to compare the performance of techniques for estimating the parameters of a regional regression model when some at-site quantile estimates are zero.

The OLS techniques employed in practice include adding a small constant to all at-site quantile estimates, or neglecting all observations reported as zero. Both these approaches performed poorly when compared to the use of a Tobit model, which is a maximum likelihood estimator (MLE) procedure that represents the below threshold estimates as a range from zero to the threshold level. A weighted Tobit model that accounts for the heteroscedasticity of the residuals in the regional regression model was also examined, but provided relatively little gain over the ordinary Tobit model.

Hewa et al. (2003) stated that the model inferences using the OLS method would be misleading for the highly correlated dependent variables. Hewa et al. (2003) points out the error structure of a regional model (error covariance matrix) is a powerful tool in deciding the most appropriate regression procedure. The methodology of the GLS procedure presented in this analysis is capable of estimating a more realistic error covariance matrix for regional hydrological models. Hewa et al. (2003) found that the estimated sampling variance of the 7D10YR (7-day 10 year ARI) extreme low flows over the study area varied by four orders of magnitude and 92.3% of the inter correlation values are significantly different from zero at the 5% level of significance, indicating that 7D10YR values over the study area are not equally reliable and are inter correlated. Hence the GLS procedure selected as the most appropriate

methodology for this regionalisation study. Regional prediction equations based on the OLS analysis were objectively evaluated. It was found that the GLS approach outperforms the OLS method.

Hailegeorgis and Alfredsen (2016) performed regional flood frequency analysis (RFFA) using the L-moments method and annual maximum series (AMS) of mean daily streamflow observations for reliable prediction of flood quantiles. The similarity in at-site and regional parameters of distributions, high flow regime and seasonality, and runoff response from precipitation-runoff models were used to identify homogeneous catchments, bootstrap resampling for estimation of uncertainty and regression methods for prediction in ungauged basins. The study reveals that a linear regression between index-flood and catchment area (R2 = 0.95) performed superior to a power-law (R2 = 0.80) and a linear regression between at-site quantiles and catchment area (e.g. R2 = 0.88 for a 200 year flood). The study also found that there is considerable uncertainty in regional growth curves (e.g. −6.7% to −13.5% and +5.7% to +24.7% respectively for 95% lower and upper confidence limits.

Quardra et al. 2007 discussed the adaptation of some regional estimation approaches to tropical climates and a comparison of their performance on the basis of their application to data selected rivers in Mexico. The author presented four approaches for the delineation of homogeneous regions; (i) hierarchical cluster analysis approach which leads to fixed hydrologic regions, (ii) canonical correlation analysis (CCA) which allows the determination of hydrologic neighborhoods that are specific to the site of interest, (iii) revised canonical correlation analysis approach that is free of parameter optimization and (iv) technique of canonical kriging which consists in interpolating hydrological variables over the canonical physiographical space. Regional estimation is carried out based on a multiple regression approach. A data set of 29 stations in the region is used to show the advantages and weaknesses of each method and to demonstrate their usefulness in the context of regional flood quantile estimation. This study allows also to test the robustness of these methods through their application to a real world case study with a relatively limited number of stations. While all methods performed quite adequately, results indicate clearly the advantages of the neighbourhood type of approach and the superiority of the canonical correlation analysis based methods. The hierarchical clustering seems generally to lead to less biased quintile estimates,

however, the lowest root mean square error values are almost consistently obtained for the CCA-based methods.

Pan et al., 2023 presented the peaks-over-threshold (POT) based RFFA techniques for south-eastern Australia using data from 151 catchments. A comparison is made between ordinary least squares (OLS) and weighted least squares (WLS) methods in developing POT-based RFFA techniques. The OLS based method is found to perform better than the WLS. The median relative error values of the developed prediction equations range 31–38%. The new POT-based RFFA technique overcomes the limitations of the current Australian Rainfall and Runoff, which does not have any RFFA technique for very frequent floods.

Dubey 2019 carried out Regional Flood Frequency Analysis in Narmada Basin located in central India. Index Flood method utilizing Gumbel's EV-1 distribution is used in the study to develop the RRF relationship. The Annual Peak Flood data of 16 gauging sites of Narmada Basin, having record length of 12 to 17 years, is utilized for flood estimation of different return period flood.

Kumar 2019 developed RFF relationships using L-moments approach for seventeen hydro-meteorological Subzones of India covering about 78.52% of geographical area of India. The annual maximum peak floods data of each of the Subzones were screened using the Discordancy measure (Di) and homogeneity of the region is tested employing the L-moments based heterogeneity measure (H). After screening of the data and testing the regional homogeneity, the data of 184 streamflow gauging sites, out of the available data of 261 streamflow gauging sites are considered suitable for conducting regional flood frequency analysis. Based on the L-moments ratio diagram and $|Z_i^{dist}|$ -statistic criteria, robust frequency distributions are identified for each of the Subzones. Out of the 17 Subzones, PE3 distribution is identified as suitable for seven Subzones, GNO for four, GPA for three, GEV for two and GLO for one of the Subzones. Regional flood frequency relationships are developed for gauged catchments, using the robust identified distribution for each of the Subzones. Also, for estimation of floods of various return periods for ungauged catchments, the regional flood frequency relationships developed for gauged catchments are coupled with the regional relationship between mean annual maximum peak flood and catchment area of the respective Subzones.

Rakesh et al, 2003 developed RFF relationships for Middle Ganga Plains Subzone 1(f) of India using L moment approach. The screening of the data has been carried out based on the discordancy measure (Di) while homogeneity of the region has been tested using heterogeneity measure, H. For computing the heterogeneity measure, 500 simulations were carried out using the four parameter Kappa distribution. Based on this test, it has been observed that the data of 8 out of 11 bridge sites constitute a homogeneous region. Hence, the data of these 8 sites have been used in this study. Catchment areas of these 8 sites vary from 32.89 to 447.76 km2 and their mean annual peak floods vary from 24.29 to 555.21 m3 s−1. Comparative regional flood frequency analysis studies have been carried out using the various frequency distributions. Based on the L-moment ratio diagram and and $|Z_i^{dist}|$-statistic criteria, GEV distribution has been identified as the robust distribution for the study area. For estimation of floods of various return periods for gauged catchments of the study area, regional flood frequency relationship has been developed using the L-moments based GEV distribution. Also, for estimation of floods of desired return periods for ungauged catchments, regional flood frequency relationship has been developed by coupling the regional flood frequency relationship with the regional relationship between mean annual maximum peak flood and catchment area.

Kumar and Chatterjee (2011) carried out the RFF analysis for Mahanadi subzone 3d based on the L-moment approach. Based on the L-moment ratio diagram and $|Z_i^{dist}|$-statistic criteria, GNO is identified as the robust frequency distribution for the study area. For estimation of floods of various return periods for gauged catchments of the study area, the regional flood frequency relationship is developed using the L-moment based GNO distribution. Also, for estimation of floods of various return periods for ungauged catchments, the regional flood frequency relationships developed for gauged catchments is coupled with the regional relationship between mean annual maximum peak flood and catchment area.

Rijal and Rahman (2005) presented the comparison of the performances of the Probabilistic Rational Method and Quantile Regression Technique using streamflow and catchment characteristics data from 98 catchments in southeast Australia. A total of 20 catchments were selected randomly from the 98 catchments and put aside for independent testing of the Quantile

Regression Technique and the Probabilistic Rational Method. The 20 test catchments and the 78 catchments used for the model development were found to have very similar catchment characteristics. It has been found that the Quantile Regression Technique in general provides more accurate design flood estimates than the Probabilistic Rational Method. The 75th percentile values of the relative errors in design flood estimates for the average recurrence intervals of 2, 5, 10, 20, 50 and 100 years were in the range of 45 to 62% for the Quantile Regression Technique as compared to 61% to 80% for the Probabilistic Rational Method. It has also been found that there is a chance of about 10% that the error in design flood estimates will exceed 100% with both the Quantile Regression Technique and the Probabilistic Rational Method. Hence, the users of these techniques should be aware of this large error and provision should be made accordingly.

Requena et al, 2017 carried out RFF analysis in which the daily stream flow series at the ungauged site is regionally estimated from daily information at the gauged sites through a regional flow duration curve approach. Then, a local flood frequency analysis is performed on the extracted maximum peak flow series. The approach, referred to as regional streamflow-based frequency analysis (RSBFA), is applied to a case study in the province of Quebec, Canada. Results indicate that the performance of the RSBFA approach is comparable to traditional methods. However, the proposed method has the advantages of being simple, flexible, and of providing the whole daily streamflow series at the ungauged site, which allows the direct estimation of a large number of other flow characteristics, such as low-flow features. The RSBFA approach also avoids performing a complete at-site flood frequency analysis at each gauged site. The fact that all the regional information is included in the regionally estimated daily stream flow series implies a number of benefits: annual or seasonal, absolute or specific, stationary or nonstationary, and univariate or multivariate flood quantiles corresponding to any return period may then be obtained through the estimated series without reconducting a regional analysis.

3 STUDY AREA

3.1 Hydrological Data Requirement for the study

The Ganga river is the main drainage flowing west to east in Bihar. It divides the state in two parts; (i) north Bihar and South Bihar. The river basins of Bihar are shown in Figure 3.1.

Figure 3.1: River basins of Bihar

The extent of the study area for this study includes the watersheds of all the rivers in the south Bihar except Sone river. Thus, the major rivers systems flowing in south Bihar east of Son river are:

1. Punpun
2. Falgu
3. Kiul Harohar
4. Badua and
5. Chandan

Further within the above major river system, several small stream flows and finally meets the Ganga river. Central Water Commission (CWC) maintains the Gauge and Discharge (GD) sites on major rivers while Water Resources Department (WRD) Bihar maintains the gauging on small rivers. The extent of study area is shown in Figure 3.2. The geographical extent of the study area is 83.8˚-88˚ E and 23.7˚-26˚ N. This figure also shows the GD sites maintained by Central Water Commission.

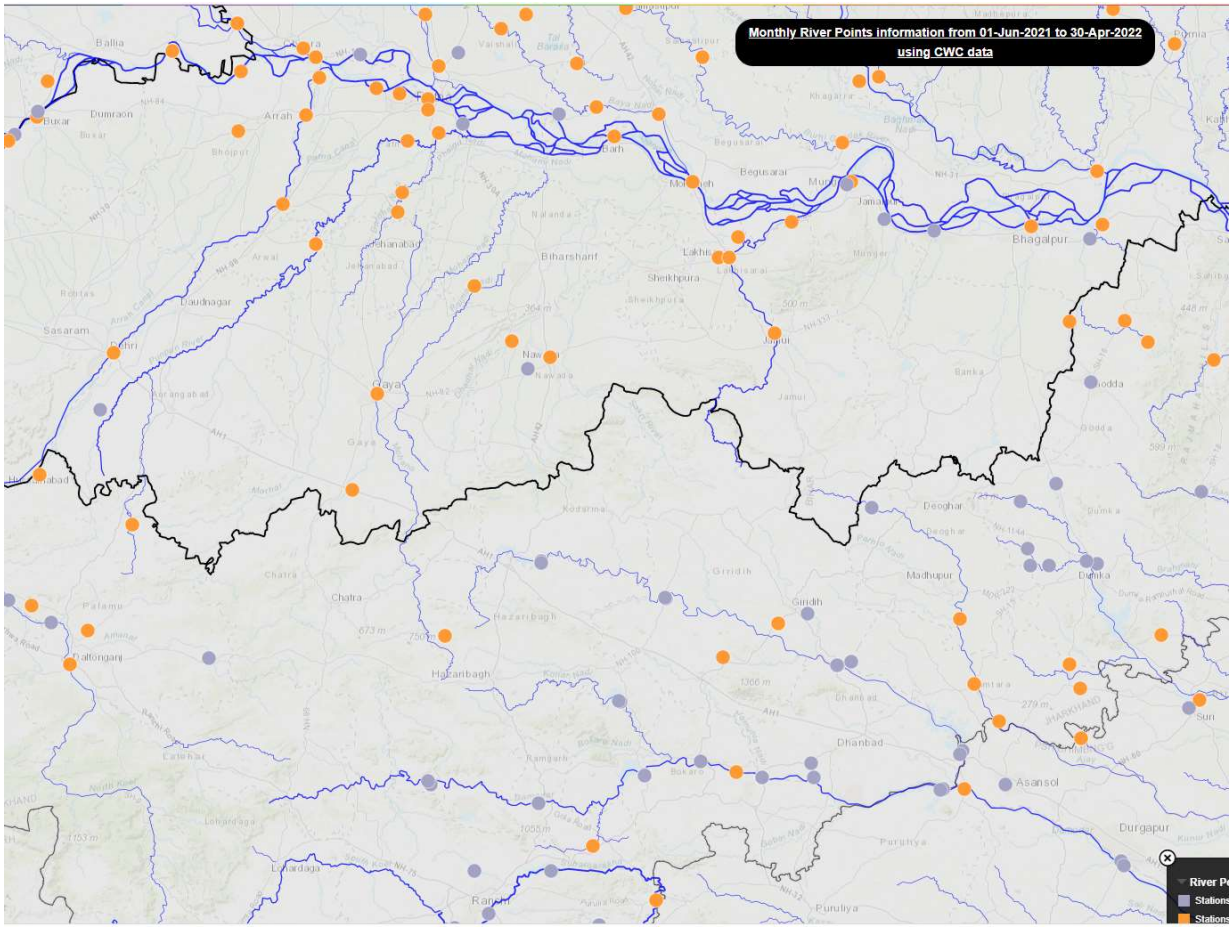

Figure 3.2: Extent of study area and CWC GD sites

(Source: India WRIS)

4 DATA AVAILABILITY

4.1 Digital Elevation Model (DEM)

Three online DEM data have been used in the study. The are namely; (i) SRTM data, (ii) Carto DEM and (iii) ALOS PALSAR DEM. The DEM is used to extract the drainage line and the watershed. The drainage line extracted from SRTM, Carto DEM and ALOS DEM are shown in Figure 4.1, Figure 4.2 and Figure 4.3, respectively. The extracted drainage lines are matched with the drainage line extracted from high resolution satellite images. The performance of online DEM is very poor for low relief terrain. Gross mismatch in drainage line is observed from all three online DEM derivatives, so drainage lines in low relief area near Ganga river are extracted from satellite images.

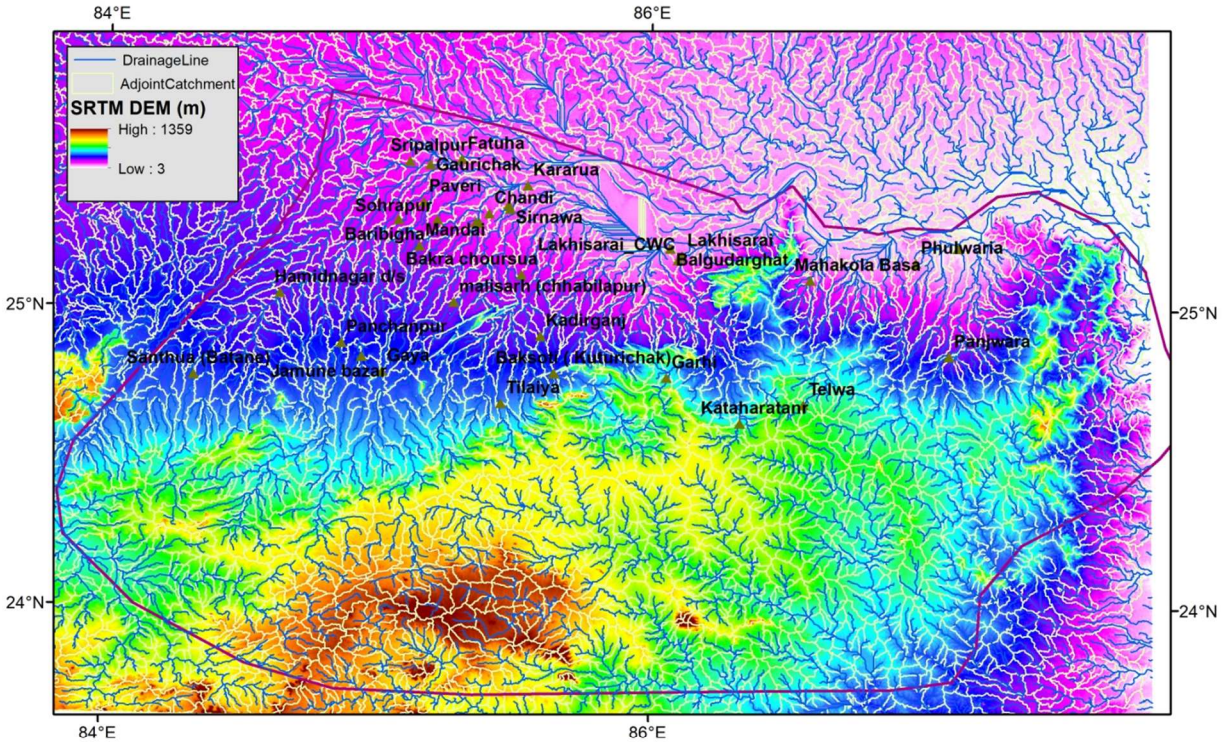

Figure 4.1: Drainage line extracted from SRTM data

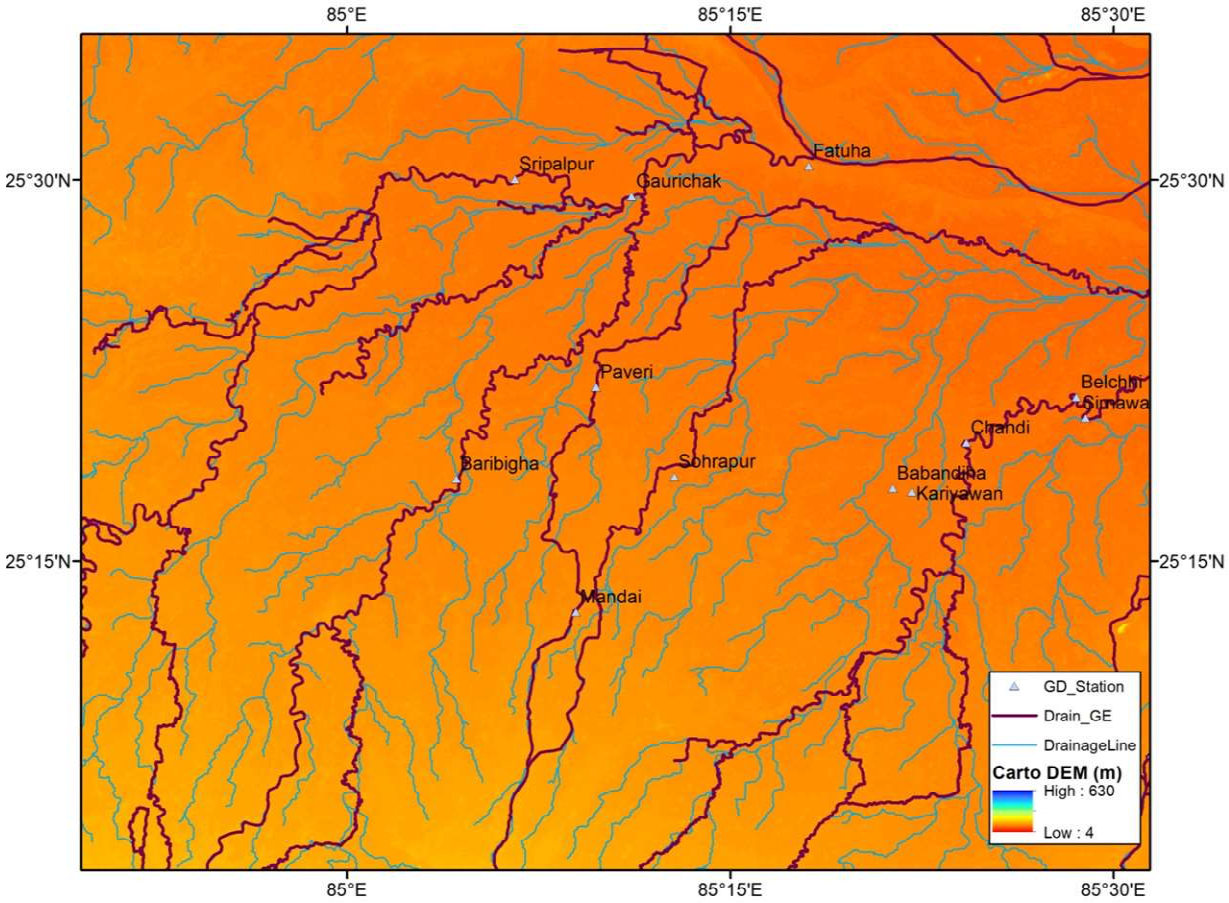

Figure 4.2: Drainage line extracted from Carto DEM data

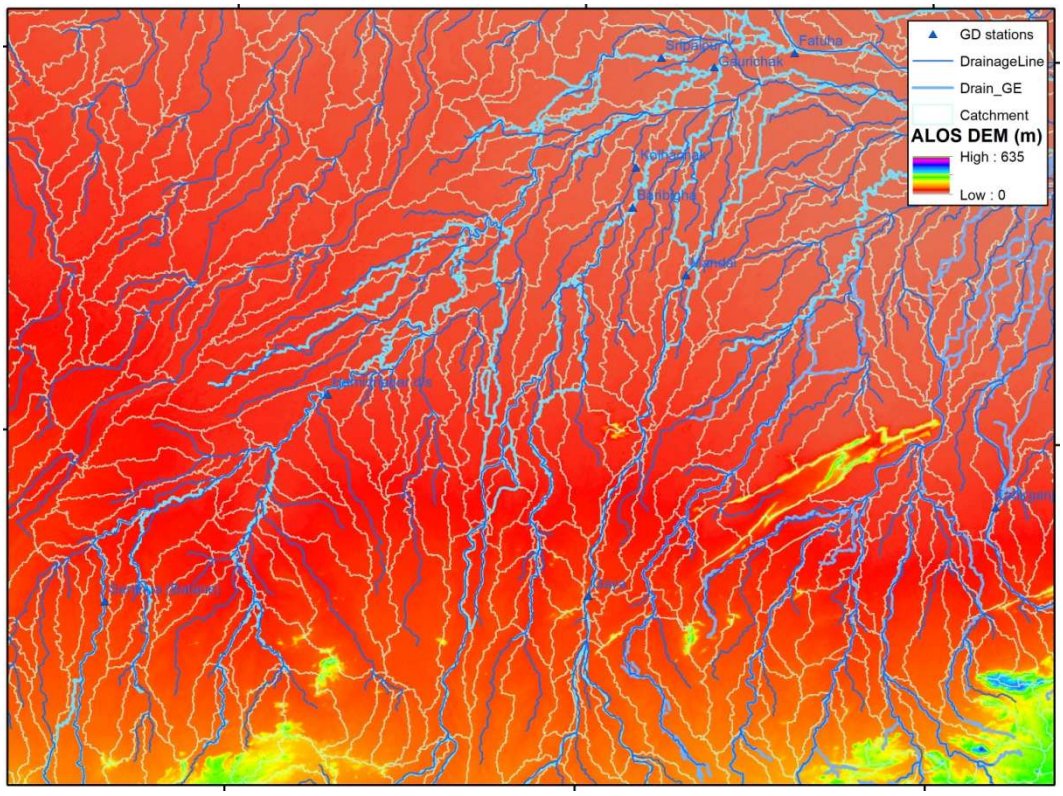

Figure 4.3: Drainage line extracted from ALOS DEM

4.2 Annual Maximum flood Data

The annual maximum flood data for various gauging sites are collected from CWC and WRD Bihar. The availability of peak discharge data for stations of south Bihar from WRD Bihar and CWC are shown in Figure 4.4 and Figure 4.5. The data at availability at WRD sites are very scant and short-term series are available. For CWC sites, long term data are available (1960 2020), although very limited sites are maintained by CWC in this region.

Drainage network is extracted from ALOS DEM for part of South Bihar rivers for which annual maximum peak flow of more than 10 years are available. Table 4.1 shows the length of years for which annual maximum peak discharge is available. The extracted watersheds are shown in Figure 4.6. The catchment characterises for these watersheds are computed as shown in Table 4.2.

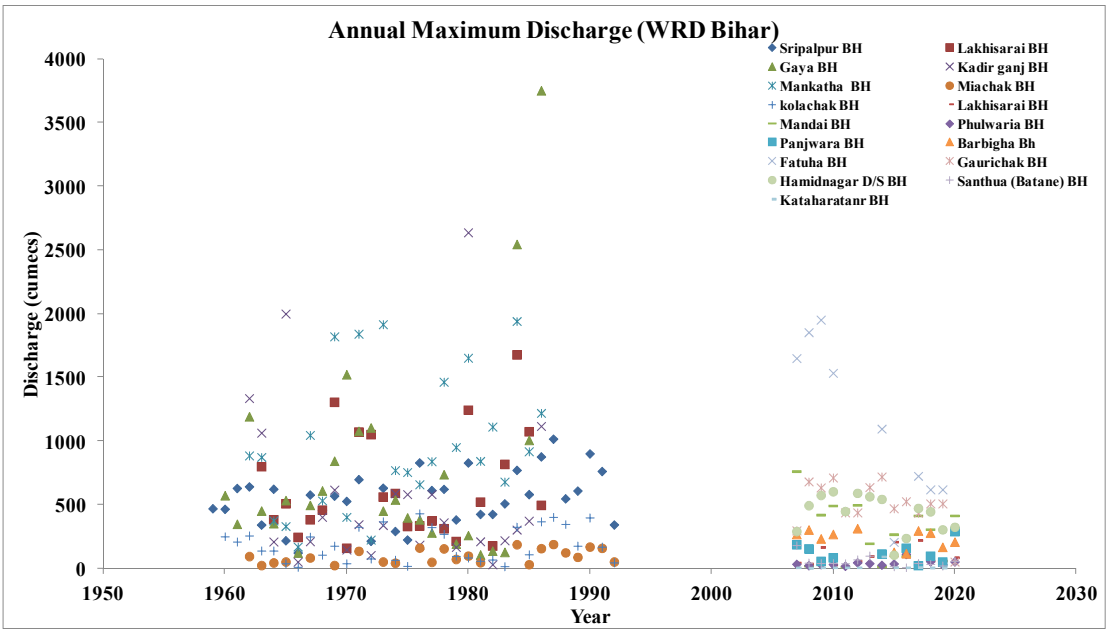

Figure 4.4: Annual maximum flood series obtained from WRD Bihar

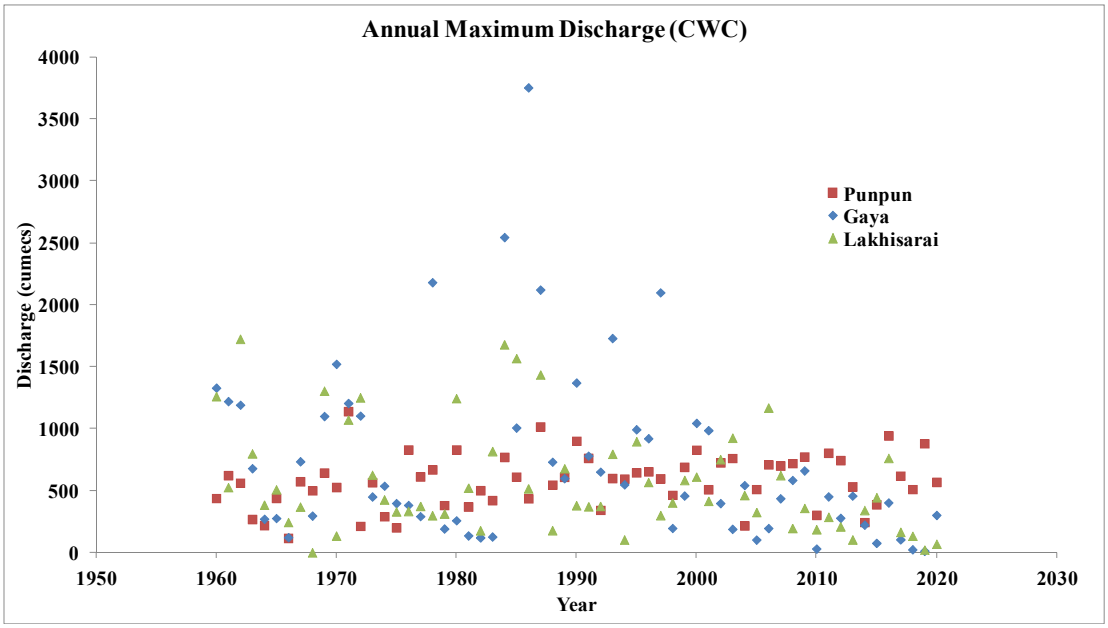

Figure 4.5: Annual maximum flood series obtained from WRD Bihar

Table 4.1: Availability of annual maximum flood and its characteristics

SN	Basin	River	Site	No. of year	Discharge data (m3/s)				CV
					Mean	Max	Min	SD	
1	punpun	Dardha	Barbigha Bh	12	232.92	313.02	115.78	67.10	3.471426
2		Punpun	Sripalpur	61	576.17	1138.45	115.90	215.08	2.678878
3		Dardha	Kolha Chak	33	185.18	429.22	10.14	132.83	1.394108
4		Punpun	Fatuha BH	12	940.59	1949.41	62.61	659.15	1.426974
5		Punpun	Gaurichak BH	15	540.95	764.83	302.93	140.77	3.842858
6		Punpun	Hamidnagar D/S BH	15	429.92	602.15	104.28	148.96	2.886149
7		Punpun	Panchanpur	9	283.20	493.36	151.27	119.87	2.362597
8		Punpun	Santhua (Batane) BH	15	51.08	179.16	6.02	42.16	1.211708
9		Jamune	Jamune Bazar	6	87.45	103.55	55.99	16.25	5.382706
10	Kiul harohar	Barnar	Kathara tand	12	12.65	91.67	4.07	24.90	0.508034
11		Kiul	Lakhisarai BH (CWC)	61	576.64	1722.93	21.52	420.84	1.370219
12		Falgu	Gaya (CWC)	61	723.11	3751.46	12.17	707.15	1.022572
13		Falgu	Mandai BH	15	384.07	1059.36	11.54	265.61	1.446002
14		Paimar	Malisarh (Chhabilapur)	6	135.89	312.54	1.90	141.21	0.96236
15		Tilaiya	Tilaiya	3	59.82	75.92	31.61	24.51	2.440448
16		Dhowa	Kharuara(Harnaut)	8	300.92	544.36	84.77	181.73	1.655825
17		Sakri	Kadirganj	31	520.02	2635.48	28.87	572.48	0.908364
18		Harohar	Baigudar	6	204.18	773.36	72.51	279.10	0.73156
19		Upper Kiul	Garhi	5	213.65	486.81	95.95	156.65	1.363936
20		Mohane	Chandi	7	44.11	75.74	21.23	24.30	1.815546
21		Sakri	Baksoti (Kuturichak)	7	428.66	1654.73	94.99	553.69	0.774193
22		Goithwa	Bakra chorsua	6	83.94	161.09	47.70	44.90	1.869591
23		Chiraiya	Kariyawan	7	46.83	69.06	17.85	20.09	2.33048
24		Nunain	Babhandeeha	6	18.60	47.24	2.84	16.94	1.09824
25		Lokaian	Sohrapur	7	303.99	505.23	126.28	136.40	2.228661
26	Badua Belharna	Upper Badua	Telwa	7	63.65	104.25	9.87	37.02	1.719444
27		Mohane	Mahakola Basa	2	47.23	55.27	39.18	11.38	4.150792
28	Chandan	Khalkhaliya	Phulwaria BH	15	41.01	110.86	18.07	21.76	1.884882
29		Ghogha	Kahalgaon road crossing	8	303.08	365.31	193.42	55.45	5.465794
30	Chir Gerua	Chiraiya	Panjwara BH	8	119.27	189.19	21.99	62.52	1.907891

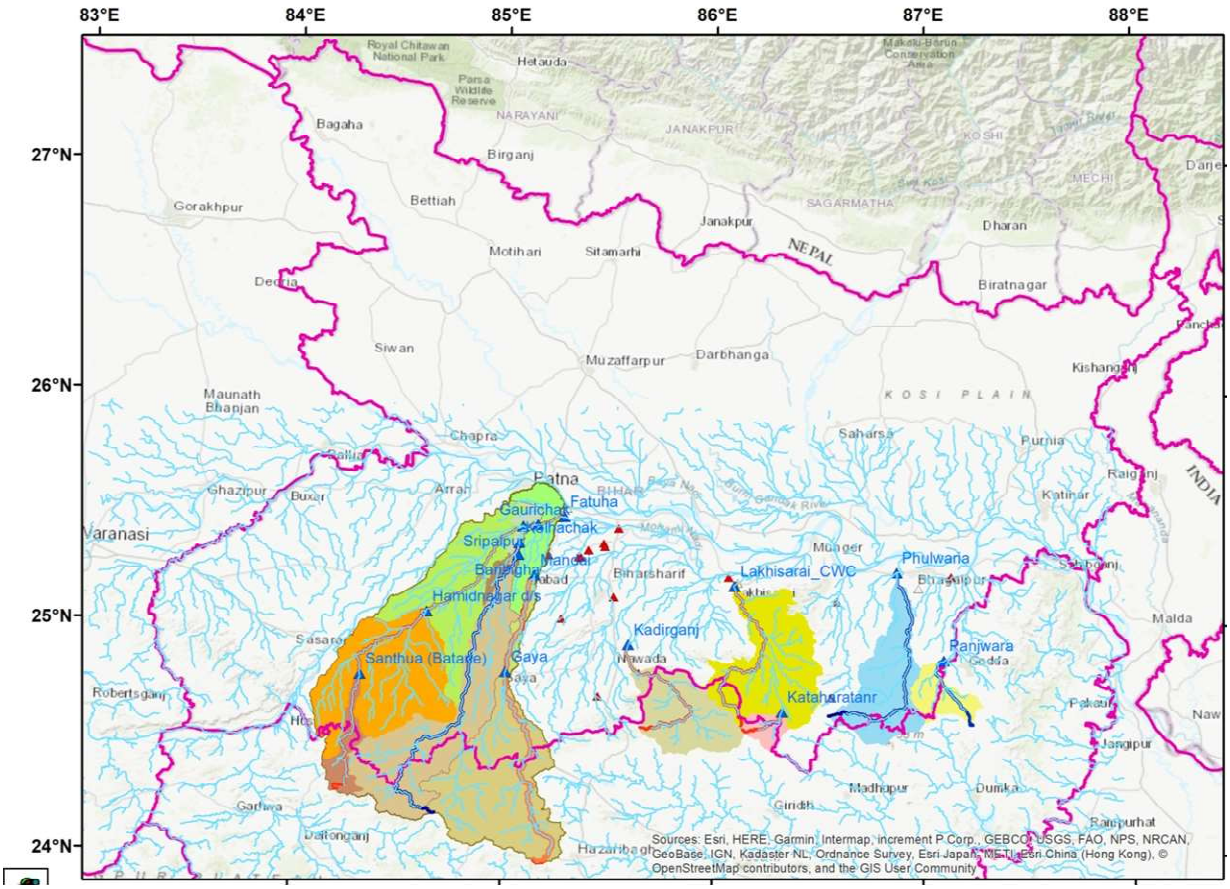

Figure 4.6: Availability of annual maximum flood series

Table 4.2: Catchment characteristics of streams extracted from ALOS DEM

SN	Basin	River	Site	C Area (sq. Km)	LFP_L (km)	LFP_S (m/m)
1	punpun	Dardha	Barbigha Bh	2984.79	203.41	0.00119
2		Punpun	Sripalpur	5380.09	233.28	0.00069
3		Dardha	Kolha Chak	3124.38	212.68	0.00112
4		Punpun	Fatuha BH	8854.94	245.81	0.00065
5		Punpun	Gaurichak BH	8565.19	236.53	0.00421
6		Punpun	Hamidnagar D/S BH	3393.64	149.85	0.00145
7		Punpun	Panchanpur	2011.43	139.76	0.00158
8		Punpun	Santhua (Batane) BH	455.51	84.88	0.00251
9		Jamune	Jamune Bazar	283.74	49.24	0.00125
10	Kiul harohar	Barnar	Kathara tand	267.46	35.43	0.00598
11		Kiul	Lakhisarai BH (CWC)	2607.62	123.41	0.00224

12		Falgu	Gaya (CWC)	3131.13	141.91	0.00414
13		Falgu	Mandai BH	3363.45	199.10	0.00259
14		Paimar	Malisarh (Chhabilapur)	650.94	95.85	0.00118
15		Tilaiya	Tilaiya	465.2	47.15	0.00580
16		Dhowa	Kharuara(Harnaut)	310.57	69.77	0.00038
17		Sakri	Kadirganj	1500.03	99.32	0.00252
18		Harohar	Baigudar	1560.45	82.19	0.00127
19		Upper Kiul	Garhi	217.49	40.76	0.00448
20		Mohane	Chandi	1656.34	141.75	0.00085
21		Sakri	Baksoti (Kuturichak)	1427.43	81.28	0.00038
22		Goithwa	Bakra chorsua	2497.58	126.98	0.00162
23		Chiraiya	Kariyawan	348.73	78.86	0.00073
24		Nunain	Babhandeeha	122.15	29.34	0.00082
25		Lokaian	Sohrapur	3390.36	212.71	0.00246
26	Badua Belharn a	Upper Badua	Telwa	95.05	18.87	0.00389
27		Mohane	Mahakola Basa	490.24	52.85	0.00242
28		Khalkhali ya	Phulwaria BH	1433.35	119.72	0.00030
29	Chanda n	Ghogha	Kahalgaon road crossing	614.39	57.01	0.00096
30	Chir Gerua	Chiraiya	Panjwara BH	552.31	47.38	0.00279

5 Methodology

In the study, L-moment based regional flood frequency analysis has been used to estimate the design flood for gauged and ungauged catchments. The important work elements in the study are (i) identification of hydro-meteorological homogenous region, (ii) Estimation of best fit distribution for annual flood peak series, (iii) Computation of physiographic characteristics using online DEM in GIS environment, (iv) Regional flood frequency analysis for estimation of floods of various return periods for the gauged catchments and (v) Developing relationship for peak flood with physiographic characteristics of the catchment to develop flood estimation for ungauged catchment.

5.1 Hierarchical Clustering

Hierarchical clustering is an unsupervised machine learning algorithm that groups data into a tree of nested clusters. The main types include agglomerative and divisive. Hierarchical cluster analysis helps find patterns and connections in datasets. Results are presented in a dendrogram diagram showing the distance relationships between clusters.

Clustering is an unsupervised machine learning technique used in data analysis to detect and group similar objects. Hierarchical cluster analysis (HCA), or hierarchical clustering, groups objects into a cluster hierarchy without enforcing a linear order within them. Many disciplines, such as biology, image analysis and the social sciences, use hierarchical clustering methods to explore and recognize patterns in datasets.

5.1.1 Main Types of Hierarchical Clustering

Agglomerative (AGNES): A bottom-up approach where each data point starts as its own cluster and merges with the closest neighbour until only one cluster remains.

Divisive (DIANA): A top-down approach where all data points start in one cluster, which is recursively split into smaller clusters.

5.1.2 Hierarchical Clustering Algorithms

The algorithm used by all eight of the clustering methods is outlined as follows. Let the distance between clusters *i* and *j* be represented as d_{ij} and let cluster *i* contain n_i objects. Let ***D*** represent the set of all remaining d_{ij}. Suppose there are *N* objects to cluster.

1. Find the smallest element d_{ij} remaining in ***D***.
2. Merge clusters *i* and *j* into a single new cluster, *k*.
3. Calculate a new set of distances d_{km} using the following distance formula.

   $$d_{km} = \alpha_i d_{im} + \alpha_j d_{jim} + \beta d_{ij} + \gamma\gamma\blacklozenge d_{im} - d_{jm}\blacklozenge$$

   Here *m* represents any cluster other than *k*. These new distances replace d_{im} and d_{jm} in ***D***. Also let $n_k = n_i + n_j . j$

   Note that the eight algorithms available represent eight choices for α_i, α_j, β, and $\gamma\gamma$.

4. Repeat steps 1 - 3 until ***D*** contains a single group made up of all objects. This will require *N-1* iterations.

The brief comments about each of the eight techniques are as follows:

Single Linkage: Also known as *nearest neighbor* clustering, this is one of the oldest and most famous of the hierarchical techniques. The distance between two groups is defined as the distance between their two closest members. It often yields clusters in which individuals are added sequentially to a single group.

The coefficients of the distance equation are

$$\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma\gamma = -0.5.$$

Complete Linkage: Also known as furthest neighbor or maximum method, this method defines the distance between two groups as the distance between their two farthest-apart members. This method usually yields clusters that are well separated and compact.

The coefficients of the distance equation are

$\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma\gamma = 0.5.$

Simple Average: Also called the weighted pair-group method, this algorithm defines the distance between groups as the average distance between each of the members, weighted so that the two groups have an equal influence on the final result.

The coefficients of the distance equation are

$$\alpha_i = \alpha_j = 0.5, \beta = 0, \gamma\gamma = 0.$$

Centroid: Also referred to as the unweighted pair-group centroid method, this method defines the distance between two groups as the distance between their centroids (center of gravity or vector average). The method should only be used with Euclidean distances.

The coefficients of the distance equation are

$$\alpha_i = \frac{n_i}{n_k}, \qquad \alpha_j = \frac{n_j}{n_k}, \qquad \beta = -\alpha_i\alpha_j, \qquad \gamma = 0$$

Backward links may occur with this method. These are recognizable when the dendrogram no longer exhibits its simple tree-like structure in which each fusion results in a new cluster that is at a higher distance level (moves from right to left). With backward links, fusions can take place that result in clusters at a lower distance level (move from left to right). The dendrogram is difficult to interpret in this case.

Median: Also called the weighted pair-group centroid method, this defines the distance between two groups as the weighted distance between their centroids, the weight being proportional to the number of individuals in each group. Backward links (see discussion under Centroid) may occur with this method. The method should only be used with Euclidean distances.

The coefficients of the distance equation are

$$\alpha_i = \alpha_j = 0.5, \ = -0.25, \ = 0.$$

Group Average: Also called the unweighted pair-group method, this is perhaps the most widely used of all the hierarchical cluster techniques. The distance between two groups is defined as the average distance between each of their members.

The coefficients of the distance equation are

$$\alpha_i = \frac{n_i}{n_k},\ \alpha_j = \frac{n_j}{n_k}, \beta = 0, \gamma = 0$$

Ward's Minimum Variance: With this method, groups are formed so that the pooled within-group sum of squares is minimized. That is, at each step, the two clusters are fused which result in the least increase in the pooled within-group sum of squares.

The coefficients of the distance equation are

$$\alpha_i = \frac{n_i + n_m}{n_k + n_m}, \qquad \alpha_j = \frac{n_j + n_m}{n_k + n_m}, \qquad \beta = \frac{-n_m}{n_k + n_m}, \qquad \gamma = 0$$

Flexible Strategy: The coefficients of the distance equation should conform to the following constraints

$$\alpha_i = 1 - \beta - \alpha_j, \alpha_j = 1 - \beta - \alpha_i, -1 \leq \beta \leq 1, \gamma\gamma = 0.$$

5.1.3 Goodness-of-Fit

Given the large number of techniques, it is often difficult to decide which is best. One criterion that has become popular is to use the result that has largest *cophenetic correlation coefficient*. This is the correlation between the original distances and those that result from the cluster configuration. Values above 0.75 are felt to be good. The Group Average method appears to produce high values of this statistic. This may be one reason that it is so popular.

A second measure of goodness of fit called *delta*. This statistics measure degree of distortion rather than degree of resemblance (as with the cophenetic correlation). The two delta coefficients are given by

5.1.4 Linkage Methods (Measuring Distance)

Single Linkage: Shortest distance between two points in different clusters.

Complete Linkage: Longest distance between two points in different clusters.

Average Linkage: Average distance between all pairs of points in different clusters.

Centroid Linkage: Distance between the centroids of two clusters.

Hierarchical Clustering Algorithms

The algorithm used by all eight of the clustering methods is outlined as follows. Let the distance between

clusters i and j be represented as $ddiiii$ and let cluster i contain $nnii$ objects. Let D represent the set of all

remaining $ddiiii$. Suppose there are N objects to cluster.

1. Find the smallest element $ddiiii$ remaining in D.

2. Merge clusters i and j into a single new cluster, k.

3. Calculate a new set of distances $ddkkkk$ using the following distance formula.

$ddkkkk = \alpha\alpha iiddiiii + \alpha\alpha jjddjjjj + \beta\beta ddiiii + \gamma\gamma\text{\textopenbox}ddiiii - ddjjjj\text{\textopenbox}$

Here m represents any cluster other than k. These new distances replace $ddiiii$ and $ddjjjj$ in D. Also let

$nnkk = nnii + nnjj$.

Note that the eight algorithms available represent eight choices for $\alpha\alpha ii$, $\alpha\alpha jj$, $\beta\beta$, and $\gamma\gamma$.

4. Repeat steps 1 - 3 until D contains a single group made up of all objects. This will require N-1

iterations.

We will now give brief comments about each of the eight techniques.

There are two types of hierarchical clustering:

- Agglomerative or bottom-up approach1 that repeatedly merges clusters into larger ones until a single cluster emerges.

- Divisive or top-down approach that2 starts with all data in a single cluster and continues to split out successive clusters until all clusters are singletons.

Hierarchical clustering analysis has high computational costs. While using a heap can reduce computation time, memory requirements are increased. Both the divisive and agglomerative types of clustering are “greedy,” meaning that the algorithm decides which clusters to merge or split by making the locally optimal choice at each stage of the process. It is also possible to apply a stop criterion, where the algorithm stops agglomeration or splitting clusters when it reaches a predetermined number of clusters.

A tree-like diagram called a dendrogram3 is often used to visualize the hierarchy of clusters. It displays the order in which clusters have been merged or divided and shows the similarity or distance between data points. Dendrograms can also be understood as a nested list of lists4 with attributes.

5.2 Regional Flood frequency Analysis

Some of the commonly used parameter estimation methods for most of the frequency distributions include: (i) method of least squares, (ii) method of moments, (iii) method of maximum likelihood, (iv) method of probability weighted moments, (v) method based on principle of maximum entropy, and (vi) method based on L-moments. L-moments are a recent development within statistics (Hosking, 1990). In a wide range of hydrologic applications, L-moments provide simple and reasonably efficient estimators of characteristics of hydrologic data and of a distribution's parameters (Stedinger et al., 1992). Like the ordinary product moments, L-moments summarize the characteristics or shapes of theoretical probability distributions and observed samples. Both moment types offer measures of distributional location (mean), scale (variance), skewness (shape), and kurtosis (peakedness).

Recently a number of regional flood frequency analysis studies have been carried out based on the L-moments approach. The L-moment methods are demonstrably

superior to those that have been used previously, and are now being adopted by many organizations worldwide (Hosking and Wallis, 1997). The L-moments offer significant advantages over ordinary product moments, especially for environmental data sets, because of the following (Zafirakou-Koulouris et al., 1998).

i. L-moment ratio estimators of location, scale and shape are nearly unbiased, regardless of the probability distribution from which the observations arise (Hosking, 1990).
ii. L-moment ratio estimators such as L-coefficient of variation, L-skewness, and L-kurtosis can exhibit lower bias than conventional product moment ratios, especially for highly skewed samples.
iii. The L-moment ratio estimators of L- coefficient of variation and L-skewness do not have bounds which depend on sample size as do the ordinary product moment ratio estimators of coefficient of variation and skewness.
iv. L-moment estimators are linear combinations of the observations and thus are less sensitive to the largest observations in a sample than product moment estimators, which square or cube the observations.
v. L-moment ratio diagrams are particularly good at identifying the distributional properties of highly skewed data, whereas ordinary product moment diagrams are almost useless for this task (Vogel and Fennessey, 1993).

5.2.1 Probability Weighted Moments and L-Moments

The L-moments are an alternative system of describing the shapes of probability distributions (Hosking and Wallis, 1997). They arose as modifications of probability weighted moments (PWMs) of Greenwood et al. (1979). Probability weighted moments is defined as:

$$M_{p,r,s} = E\left(x^p \{F\}^r \{1-F\}^s\right) = \int_0^1 \{x(F)\}^p F^r \{1-F\}^s \, dF \qquad \textbf{(5.1)}$$

where, F = F(x) is the cumulative distribution function (CDF) for x, x(F) is the inverse CDF of x evaluated at the probability F, and p, r and s are real numbers. If p is a

nonnegative integer, $M_{p,0,0}$ represents the conventional moment of order p about the origin. If p = 1 and s = 0,

$$M_{1,r,0} = \beta_r = \int_0^1 x(F) F^r dF \tag{5.2}$$

For an ordered sample $x_1 \leq x_2 \ldots \leq x_N$, N > r, the unbiased sample PWM's are given by

$$\hat{\beta}_r = \frac{1}{N} \frac{\sum_{i=1}^{N} \binom{i-1}{r} x_i}{\binom{N-1}{r}} \tag{5.3}$$

For any distribution the r^{th} L-moment λ_r is related to the r^{th} PWM (Hosking, 1990), through:

$$\lambda_{r+1} = \sum_{k=0}^{r} \beta_k (-1)^{r-k} \binom{r}{k} \binom{r+k}{k} \tag{5.4}$$

These L-moments are linear functions of PWMs. For example, the first four L-moments are related to the PWMs using:

$$\begin{aligned} \lambda_1 &= \beta_0 \\ \lambda_2 &= 2\beta_1 - \beta_0 \\ \lambda_3 &= 6\beta_2 - 6\beta_1 + \beta_0 \\ \lambda_4 &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \end{aligned} \tag{5.5}$$

The L-moments are analogous to their conventional counterparts as they can be directly interpreted as measures of scale and shape of probability distributions and hence, are more convenient than the PWMs. Hosking (1990) defined L-moment ratios which are analogous to conventional moment ratios as:

$$\begin{aligned} &\text{L-coefficient of variation, L-CV: } \tau_2 = \lambda_2 / \lambda_1 \\ &\text{L-coefficient of skewness, L-skew: } \tau_3 = \lambda_3 / \lambda_2 \\ &\text{L-coefficient of kurtosis, L-kurtosis: } \tau_4 = \lambda_4 / \lambda_2 \end{aligned} \tag{5.6}$$

Analogous to the conventional moment ratios, λ_1 is a measure of location, τ_2 is a measure of scale and dispersion, τ_3 is a measure of skewness and τ_4 is a measure of kurtosis. Hosking (1990) showed that for $x \geq 0$, the value of τ_2 lies between 0 and 1, while the absolute values of τ_3 and τ_4 lie between 0 and 1. This restriction in the values of the L-coefficients works out to be an advantage in their interpretation as opposed to the conventional moments which do not have any bounds (Rao and Hamed, 2000).

5.2.2 Screening of Data Using Discordancy Measure Test

The objective of screening of data is to check that the data are appropriate for performing the regional flood frequency analysis. In this study, screening of the data was performed using the L-moments based Discordancy measure (D_i). Discordancy is measured in terms of the L-moments of the sites' data and the aim is to identify those sites that are grossly discordant with the group as a whole. The sample L-moment ratios (t_2, t_3 and t_4) of a site are considered as a point in a three-dimensional space. A

$$\bar{u} = N^{-1} \sum_{i=1}^{N} u_i \qquad (15.7)$$

(5.7)

group of sites form a cluster of such points in the three-dimensional space. A site is considered discordant if it is far from the centre of the cluster.

Hosking and Wallis (1997) defined the Discordancy measure D_i for a site i in a group of N sites. Let $u_i = [t_2^{(i)} \; t_3^{(i)} \; t_4^{(i)}]^T$ be a vector containing the sample L-moment ratios

$$A_m = \sum_{i=1}^{N} (u_i - \bar{u})(u_i - \bar{u})^T \qquad (15.8)$$

(5.8)

t_2, t_3 and t_4 values for site i, analogous to their regional values termed as τ_2, τ_3, and τ_4, expressed in Eq. (5.6). T denotes transposition of a vector or matrix. Let

be the (unweighted) group average. The sample covariance matrix is defined as:

The Discordancy measure for site i is defined as:

$$D_i = \frac{1}{3} N(u_i - \bar{u})^T A_m^{-1}(u_i - \bar{u}) \qquad (15.9)$$

(5.9)

The site i is declared to be discordant, if D_i is greater than the critical value of the Discordancy statistic D_i, given in a tabular form by Hosking and Wallis (1997).

5.2.3 Test of Regional Homogeneity

For testing regional homogeneity, a test statistic H, termed as heterogeneity measure was proposed by Hosking and Wallis (1993). It compares the "inter-site variations in sample L-moments for the group of sites" with "what would be expected of a homogeneous region". The inter-site variations in sample L-moments are evaluated based on any of the three measures of variability V_1 (based on L-CV), V_2 (based on L-CV and L-Skew) and V_3 (based on L- Skew and L-Kurtosis). These measures of variability are computed as follows:

(i) V_1 is the weighted standard deviation of at site L-CV's ($t_2^{(i)}$)

$$V_1 = \left[\sum_{i=1}^{N} n_i \ (t_2^{(i)} - t_2^R)^2 \Big/ \sum_{i=1}^{N} n_i \right]^{1/2} \qquad (5.10)$$

where, n_i is the record length at each site and t_2^R is the regional average L-CV weighted proportionally to the sites' record length as given below.

$$t_2^R = \sum_{i=1}^{N} n_i \ t_2^{(i)} \Big/ \sum_{i=1}^{N} n_i \qquad (5.11)$$

(ii) V_2 is the weighted average distance from the site to the group weighted mean on a graph of t_2 versus t_3

$$V_2 = \sum_{i=1}^{N} n_i \ \left\{ (t_2^{(i)} - t_2^R)^2 + (t_3^{(i)} - t_3^R)^2 \right\}^{1/2} \Big/ \sum_{i=1}^{N} n_i$$

(5.12)

where, t_3^R is the regional average L-Skew weighted proportionally to the sites' record length.

(iii) V_3 is the weighted average distance from the site to the group weighted mean on a graph of t_3 versus t_4

$$V_3 = \sum_{i=1}^{N} n_i \left\{(t_3^{(i)} - t_3^R)^2 + (t_4^{(i)} - t_4^R)^2\right\}^{½} \Big/ \sum_{i=1}^{N} n_i \qquad (5.13)$$

where, t_4^R is the regional average L-Kurtosis weighted proportionally to the sites' record length.

To establish "what would be expected of a homogeneous region", firstly simulations are used to generate homogeneous regions with sites having same record lengths as those of observed data. In order to generate the simulated data, a four parameter Kappa distribution is used. The four parameter Kappa distribution is chosen so as not to commit to a particular two or three parameter distribution. Further, the four parameter Kappa distribution includes as special cases the generalised logistic, generalised extreme value and generalised pareto distributions and hence, acts as a good representation of many of the probability distributions occurring in environmental sciences.

The parameters of the Kappa distribution are obtained using the regional average L-moment ratios t_2^R, t_3^R, t_4^R and mean = 1. A large number of data regions are generated (say N_{sim} = 500) based on this Kappa distribution. The simulated regions are homogeneous and have no cross-correlation or serial correlation. Further, the sites have the same record lengths as the observed data. For each generated region, V_j (i.e. any of V_1, V_2 or V_3) is computed using Eqns. 15.10 to 15.13. Subsequently, their mean (μ_v) and standard deviation (σ_v) are computed.

The heterogeneity measure H(j) (i.e. H(1), H(2) or H(3)) is computed as:

$$H(j) = \frac{V_j - \mu_v}{\sigma_v} \qquad (5.14)$$

If the heterogeneity measure is sufficiently large, the region is declared to be heterogeneous. Hosking and Wallis (1997) suggested the following criteria for assessing heterogeneity of a region: if $H(j) < 1$, the region is acceptably homogeneous; if $1 \leq H(j) < 2$, the region is possibly heterogeneous; and if $H(j) \geq 2$,

the region is definitely heterogeneous. These boundary values of H(j) being 1 and 2 were determined by Hosking and Wallis (1997) by performing a series of Monte Carlo experiments in which the accuracy of quantile estimates corresponding to different values of H(j) were computed. The authors further observed that for both real world data and artificially simulated regions, H(1) has much better power to discriminate between homogeneous and heterogeneous regions as compared to H(2) and H(3).

15.3.4 Identification of Robust Regional Frequency Distribution

The best fit frequency distribution for a homogeneous region is determined by how well the L-skewness and L-kurtosis of the fitted distribution match the regional average L-skewness and L-kurtosis of the observed data (Hosking and Wallis, 1997). The procedure adopted (Hosking and Wallis, 1997) is briefly stated below.

Initially, several three parameter distributions are fitted to the regional average L-moments t_2^R, t_3^R and mean =1. Let τ_4^{Dist} be the L-kurtosis of the fitted distribution which may be GEV, GLO, GNO, PE3 etc. Using the N_{sim} number of simulated regions of the Kappa distribution (as obtained for the heterogeneity measure described in section 15.3.3 above), the regional average L-kurtosis, t_4^m is computed for the m^{th} simulated region. The bias of t_4^R is computed as:

$$B_4 = N_{sim}^{-1} \sum_{m=1}^{N_{sim}} \left(t_4^m - t_4^R \right) \qquad (5.15)$$

The standard deviation of t_4^R is computed as:

$$\sigma_4 = \left[\left(N_{sim} - 1 \right)^{-1} \left\{ \sum_{m=1}^{N_{sim}} \left(t_4^m - t_4^R \right)^2 - N_{sim} B_4^2 \right\} \right]^{1/2} \qquad (5.16)$$

The goodness-of-fit measure for each distribution is computed as (Hosking and Wallis, 1997):

$$Z^{dist} = \left(\tau_4^{dist} - t_4^R + B_4 \right) / \sigma_4 \qquad (5.17)$$

The fit is considered to be adequate if $|Z^{dist}|$–statistic is sufficiently close to zero, a reasonable criterion being $|Z^{dist}|$ –statistic less than 1.64. Hosking and Wallis (1997) states that the $|Z^{dist}|$ –statistic has the form of a normal distribution under suitable assumptions. Thus the criterion $|Z^{dist}|$ –statistic less than 1.64 corresponds to acceptance of the hypothesized distribution at a confidence level of 90%.

The regional flood frequency analysis is based on the most restrictive fundamental hypothesis of existence of homogeneous regions within which the statistical properties of dimensionless flood flows do not vary with location (e.g., dimensionless statistical moments such as the coefficients of variation and skewness are constant).

The region boundary starts with from administrative boundaries to physiographic and meteo-climatic boundaries.

.

6 Analysis and results

In this chapter, the approach for identification of homogeneous region, L-moment based Regional Flood Frequency Analysis (RFFA) for gauged and ungauged watershed has been described.

6.1 Hierarchical clustering technique

Altogether, the annual peak flow data for 30 rivers stations were collected. However the stations with minimum 6 annual values were considered in the analysis. Further, the annual peak flow data of Fatuha and Gaurichak also shows inconsistency. The Histogram of CV plots of GD stations combined with a cumulative percentage (Pareto-style) curve is shown in Figure 6.1. The X-axis (CV) represents different coefficient of variation (CV) classes (e.g., 1.5, 2, 1, 2.5, 3, etc.) while the left Y-axis (No. of station) shows the frequency (number of stations) in each CV class. The right Y-axis (%) shows the cumulative percentage. The figure shows that the highest frequencies are concentrated in the lower CV range (roughly 1 to 2.5), highlighted by the red box. As CV increases, the number of stations decreases significantly. The cumulative curve shows the running total percentage of stations. the figure shows that about 80% of stations (23 stations) fall within the lower CV range (around up to ~2.5–3). Thus, the distribution is right-skewed (many low CV values, few high ones) and most stations exhibit low variability (low CV).

The highlighted region emphasizes that a large majority (~80%) of stations are concentrated in the lower CV classes, which is useful for understanding variability patterns or selecting representative stations.

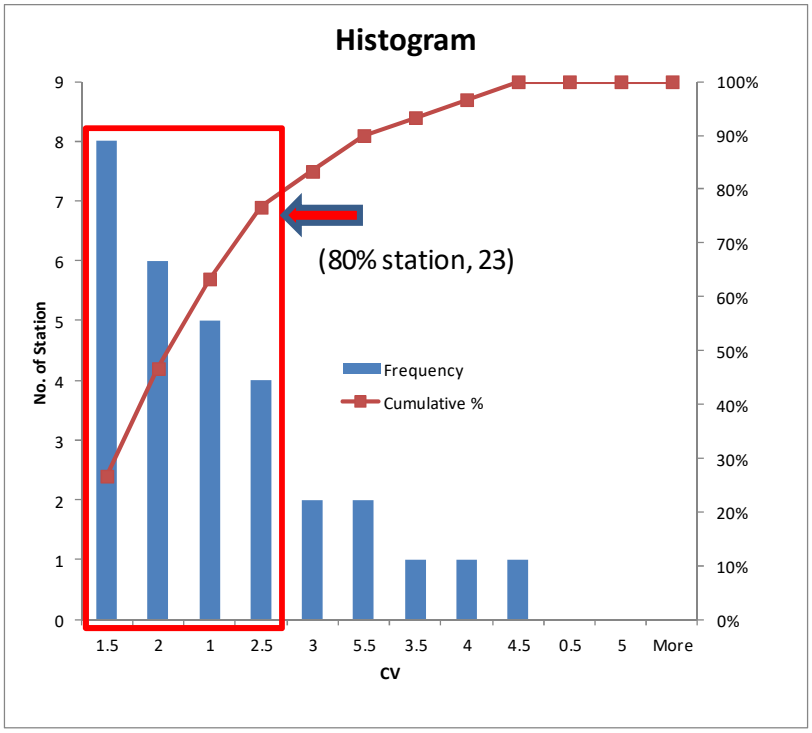

Figure 6.1: Histogram of CV of GD stations.

Thus in the final data set, annual flood peak of 23 stations are used for identify homogeneous cluster using Hierarchical clustering technique. The dendrogram of the Hierarchical clustering analysis is shown in Figure 6.2. The dendrogram illustrates the results of hierarchical clustering performed on GD stations in South Bihar using Euclidean distance (threshold ≈ 2.0) and complete linkage, based on parameters such as mean annual flow per unit area (QDA) and coefficient of variation (CV). The analysis categorizes the stations into three distinct clusters, reflecting varying degrees of hydrological similarity. Cluster 3 (green) constitutes the largest group, comprising 19 stations, indicating a high level of homogeneity in their characteristics and suggesting that these stations may belong to a similar hydrological regime. Cluster 2 (blue) includes stations such as Baksoti (Kuturichak), Baigudar, Kadirganj, and Gaya (CWC), which exhibit close similarity among them but are distinct from the larger group. Cluster 1 (red), consisting of stations like Telwa, Garhi, and Kharaura (Harnaut), forms a smaller, separate cluster, indicating comparatively different hydrological behaviour. The dendrogram shows that clusters merge at higher distance levels (approximately 2.5–3.0), confirming the presence of well-defined groupings.

Overall, the clustering provides a meaningful classification of stations into homogeneous regions, which is valuable for regional flood frequency analysis, model regionalization, and hydrological assessment.

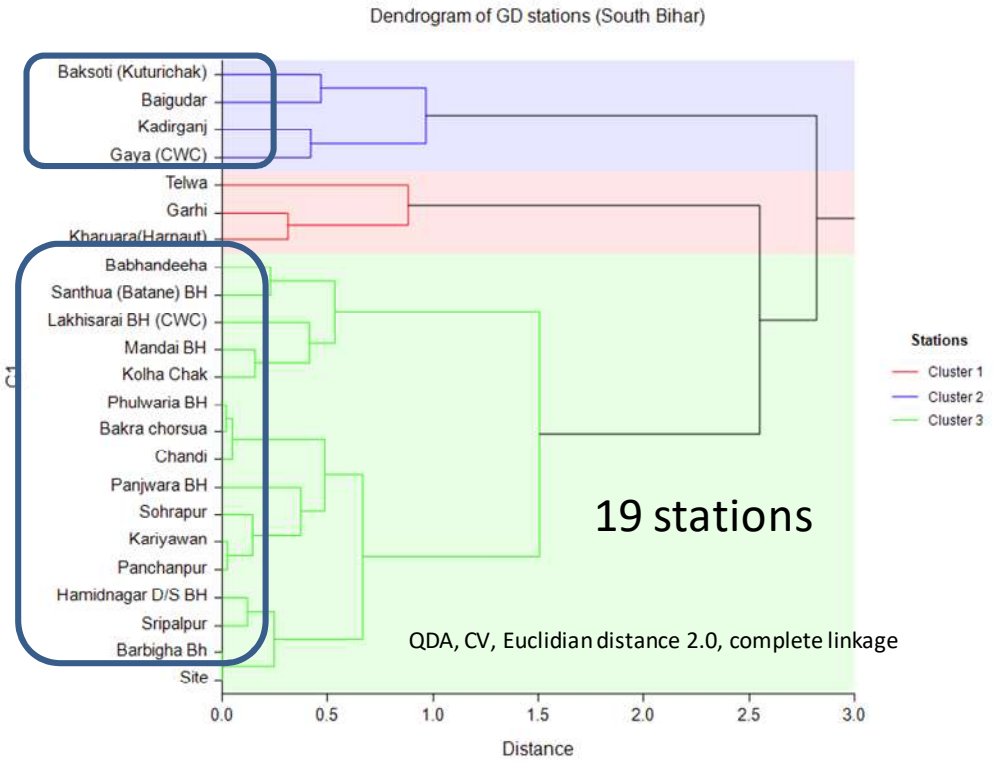

Figure 6.2: Dendrogram of GD stations

6.2 L-moment based Heterogeneity Measures

L-moment based regional frequency analysis for 14 stations were carried out to compute the site-specific statistics, regional averages, fitted distribution parameters, and heterogeneity measures.

For testing the regional homogeneity, a test statistic H, termed as heterogeneity measure was proposed by Hosking and Wallis (1993). It compares the inter-site variations in sample L-moments for the group of sites with what would be expected of a homogeneous region. The inter-site variation of L-moment ratio is measured as the standard deviation (V) of the at-site L-CV's weighted proportionally to the record length at each site. To establish what would be expected of a homogeneous region,

simulations are used. A number of, say 1000, data regions are generated based on the regional weighted average statistics using a four parameter distribution e.g. Kappa distribution. The inter-site variation of each generated region is computed and the mean (μ_v) and standard deviation (σ_v) of the computed inter-site variation is obtained. Then, heterogeneity measure H is computed as:

$$H = \frac{V-\mu_v}{\sigma_v} \quad (6.1)$$

The criteria established by Hosking and Wallis (1993) for assessing heterogeneity of a region is as follows:

If H < 1 Region is acceptably homogeneous.

If 1≤H < 2 Region is possibly heterogeneous.

If H ≥ 2 Region is definitely heterogeneous.

The heterogeneity measure for South Bihar river GD stations using the data of 15 sites was computed and the same was found to be greater than 1.0. Based on the statistical properties one sites of the region was excluded so that H value less than 1.0 was obtained. Thus, the region comprising of 14 sites was identified as the homogenous region. The values of heterogeneity measure computed by carrying out 1000 simulations using the Kappa distribution based on the data of 14 sites are given in Table 6.1.

Table 6.1: Heterogeneity measures for South Bihar GD Stations.

OBSERVED S.D. OF GROUP L-CV	0.117
SIM. MEAN OF S.D. OF GROUP L-CV	0.0578
SIM. S.D. OF S.D. OF GROUP L-CV	0.116
STANDARDIZED TEST VALUE H(1)	0.51
OBSERVED AVE. OF L-CV / L-SKEW DISTANCE	0.1966
SIM. MEAN OF AVE. L-CV / L-SKEW DISTANCE	0.1013
SIM. S.D. OF AVE. L-CV / L-SKEW DISTANCE	0.191
STANDARDIZED TEST VALUE H(2)	0.50
OBSERVED AVE. OF L-SKEW/L-KURT DISTANCE	0.1833
SIM. MEAN OF AVE. L-SKEW/L-KURT DISTANCE	0.1265
SIM. S.D. OF AVE. L-SKEW/L-KURT DISTANCE	0.23
STANDARDIZED TEST VALUE H(3)	0.25

The details of catchment area, sample size and sample statistics for the 14 sites which form the homogeneous region are given in Table 6.2 along with the Discordancy measure (Di) values. It is observed from Table that the Di values for the 14 sites vary from 0.30 to 2.07 and the same are less than the critical Di value of 2.140 (Hosking and Wallis, 1997). Hence, data of these 14 sites have been used for development of regional flood frequency relationships for the region.

Table 6.2: Catchment area and sample statistics and sample size for the 14 stations

Site Name	Data points	Area (sqkm)	mean flow (cumec)	L-CV	L-SKEW	L-KURT	Di
Barbigha	12	2984.8	233	0.1784	-0.0958	0.1207	1.63
Sripalpur	61	5380.1	576.164	0.2137	0.0046	0.1434	0.87
Panchanpur	9	2011.4	283.222	0.2537	0.2153	0.1245	2.07
Santhua	15	455.5	51.133	0.3974	0.3542	0.3352	0.95
Lakhisarai	61	2607.6	576.672	0.3925	0.2741	0.1477	0.31
Gaya	61	3131.1	723.082	0.4869	0.3523	0.1937	0.45
Kadirganj	31	1500	520.032	0.4926	0.4787	0.3338	0.56
Baigudar	6	1560.5	204.167	0.5372	0.6716	0.4337	1.49
Baksoti	7	1427.4	428.714	0.5568	0.549	0.3429	0.88
Kariyawan	7	348.7	47	0.2608	0.0204	0.047	0.42
Babhandeeha	6	122.2	18.5	0.4989	0.3072	0.0955	1.39
Sohrapur	7	3390.4	303.857	0.2783	0.0926	0.0979	0.3
Telwa	7	95.1	63.571	0.3446	-0.0033	-0.0023	1.11
Panjwara	8	552.3	119.375	0.3107	-0.0171	0.0374	0.76

6.3 Identification of Regional Frequency Distribution

The choice of an appropriate frequency distribution for a homogeneous region is made by comparing the moments of the distributions to the average moments statistics from regional data. The objective is to identify a distribution that best fits the observed data. The best fit is determined by how well the L-skewness and L-kurtosis of the fitted distribution match the regional average L-skewness and L-kurtosis of the observed data (Hosking, 1991). In this study, the L-moment ratio diagram and $|Z^{i}_{dist}|$ – statistic are used as the best fit criteria for identifying the regional distribution. L-moment ratio diagrams compare sample estimates of the dimensionless L-moment ratios with their theoretical counterparts (ZafirakouKoulouris et al., 1998).

The goodness-of-fit measure for a distribution, Z^i_{dist}–statistic defined by Hosking and Wallis (1993), is expressed as:

$$Z^i_{dist} = \frac{\left(\bar{\tau}_i^R - \tau_i^{dist}\right)}{\sigma_i^{dist}} \tag{6.2}$$

where $\bar{\tau}_i^R$is the weighted average of L-moment statistic i, τ_i^{dist} and σ_i^{dist} are the simulated average and standard deviation of L-moment statistics i for a given distribution. The distribution giving the minimum Z^i_{dist} value is considered as the best fit distribution. When all the three L-moment ratios are considered in the goodness-of-fit test, the distribution that gives the best overall fit is selected as the underlying frequency distribution. According to Hosking (1990), distribution is considered to give good fit if Z^i_{dist} is sufficiently close to zero, a reasonable criteria being $Z^i_{dist} <$ 1.64. As even for heterogeneous regions, it is important to use a distribution that is robust to moderate heterogeneity in the at-site frequency distribution. It is therefore preferred to use Wakeby distribution for heterogeneous regions. Further, the Wakeby distribution, which has five parameters, more than most of the common distributions can attain a wider range of distributional shapes than can the common distributions. This makes the Wakeby distribution particularly useful for simulating artificial data for use in studying the robustness, under changes in distributional form of methods of data analysis (Hosking and Wallis, 1997)

6.4 RFF relationship for the gauged catchments.

The L-moment ratio diagram of the RFFA is shown in Figure 6.3**Error! Reference source not found.**, the point defined by the regional average values of L-skewness i.e. $\tau 3 = 0.2342$ and L-kurtosis i.e. $\tau 4 = 0.1825$, lies closest to the GEV distribution. The $|Z^i_{dist}|$–statistic for the various three parameter distributions is given in Table 6.3. It is observed that the $|Z^i_{dist}|$–statistic values are lower than 1.64 for the four distributions viz. GEV, GLO, GNO and PE3. Further, the $|Z^i_{dist}|$–statistic is found to be the lowest for GEV distribution i.e. 0.41, which is very close to 0.0. Thus, based on the L-moment ratio diagram as well as $|Z^i_{dist}|$–statistic criteria, the GEV distribution is identified as the robust distribution for the study area.

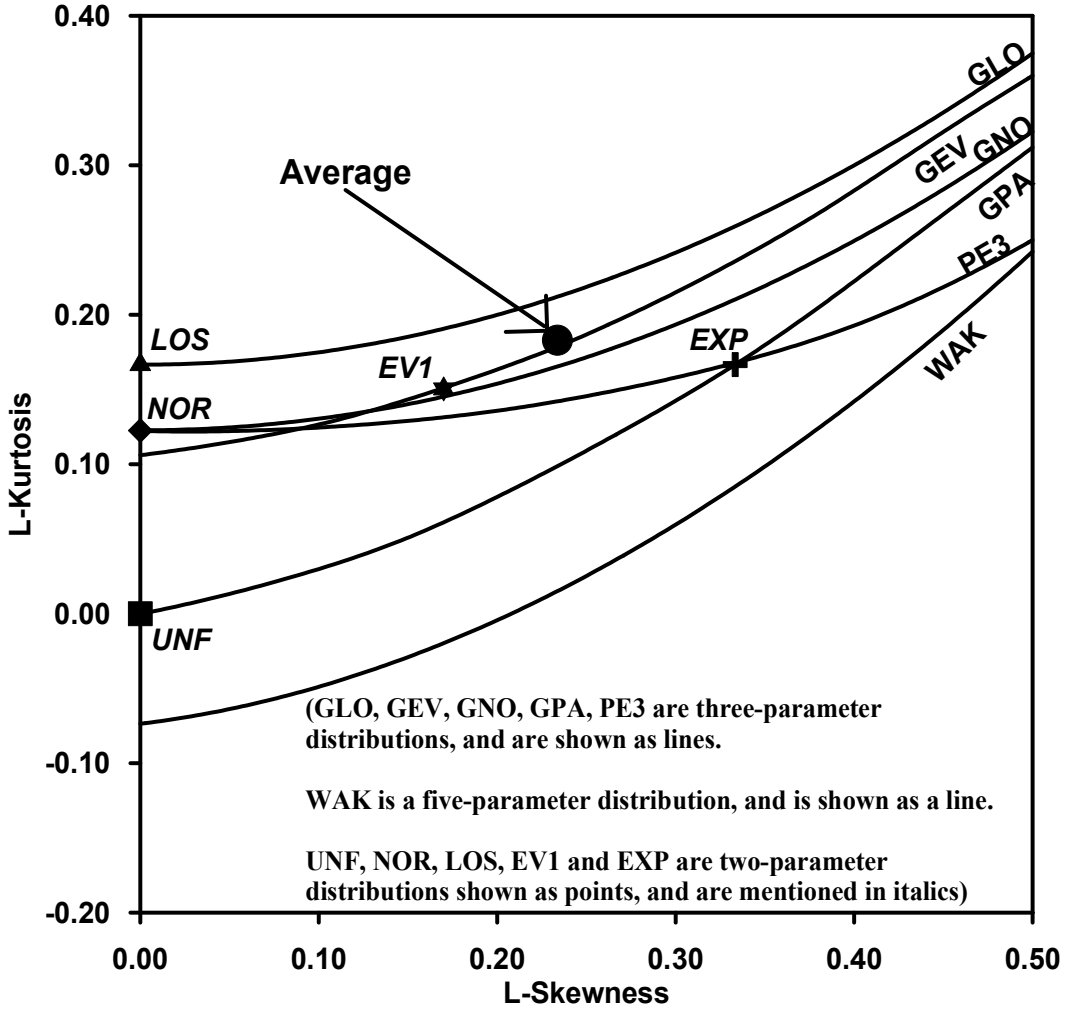

Figure 6.3: L moment ratio diagram for the region

Table 6.3: The first five lowest Z^i_{dist} –statistic for various distribution.

SN	Distribution	Z^i_{dist}–statistic
1	GEN. LOGISTIC	0.79
2	GEN. EXTREME VALUE	**-0.41**
3	GEN. NORMAL	-0.88
4	PEARSON TYPE III	-1.73
5	GEN. PARETO	-3.31

Quantile estimate, $Q_eT = Q_I^i * q(T)$ (6.3)

Where Q_I^i is Index flood for station i, $q(T)$ is the regional growth factor.

For Generalized Extreme Value, GEV distribution,

$$\text{Quantile estimate, } Q_e T = Q_I^i * q(T) \tag{6.4}$$

Where Q_I^i is Index flood for station i, $q(T)$ is the regional growth factor.

For Generalized Extreme Value, GEV distribution, (eq. 6.4)

$$q(T) = \varepsilon + \alpha * y_T \tag{6.5}$$

Where,

$$y_T = \frac{\left[1 - \left\{-ln\left(1 - \frac{1}{T}\right)\right\}^k\right]}{k} \tag{6.6}$$

The parameters for GEV distribution are; ε = 0.667, α = 0.487, κ = -0.098

Putting these values in **Error! Reference source not found.** and **Error! Reference source not found.**,

$$q(T) = 0.667 - 4.9693\left[1 - \left\{-ln\left(1 - \frac{1}{T}\right)\right\}^{-0.098}\right] \tag{6.7}$$

Table 6.4: Values of growth factors (QT /$\bar{Q}$) for various distributions

Distribution	Return period									
	2	10	20	25	50	100	200	500	1000	10000
	Growth factor									
GEN. LOGISTIC	0.86	1.838	2.302	2.465	3.021	3.668	4.425	5.63	6.728	11.962
GEN. EXTREME VALUE	0.849	1.894	2.346	2.497	2.982	3.497	4.047	4.831	5.472	7.941
GEN. NORMAL	0.846	1.91	2.354	2.498	2.956	3.429	3.921	4.602	5.143	7.118
WAKEBY	0.843	1.916	2.381	2.532	3.001	3.473	3.948	4.581	5.063	6.687

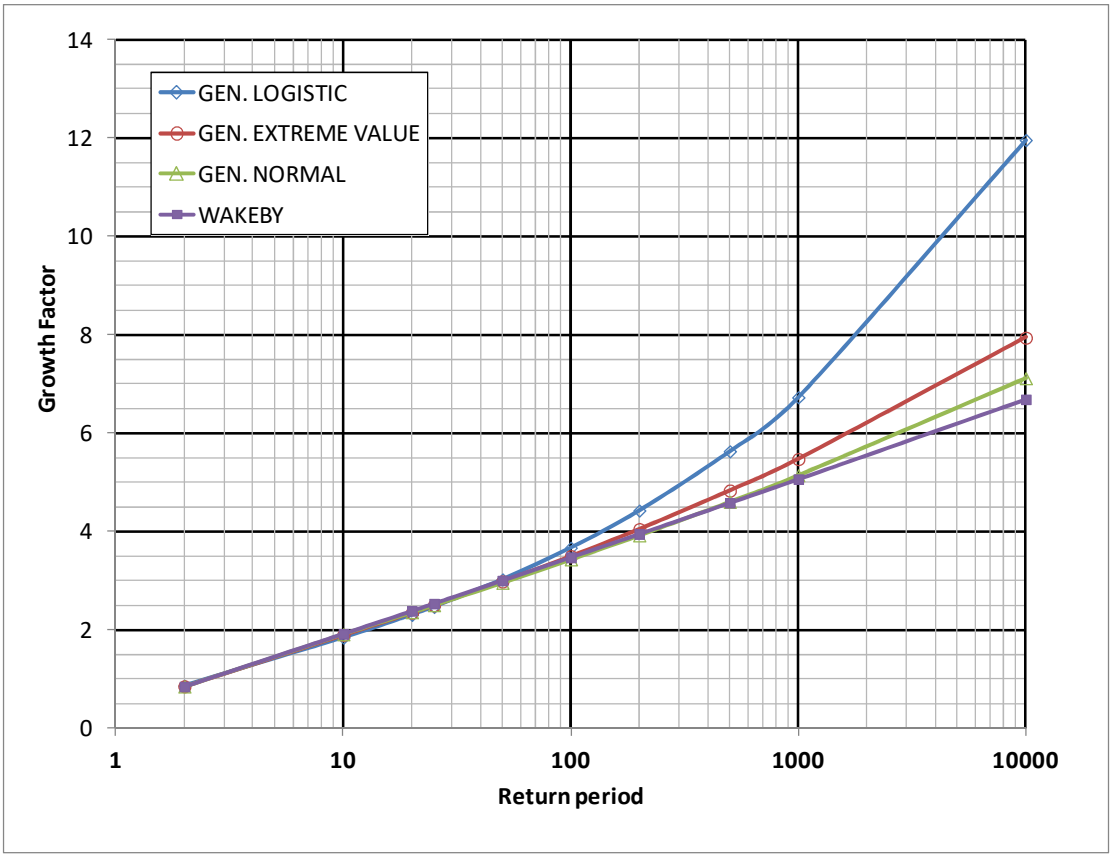

Figure 6.4: Development of regional growth factor for the region

Hence, the regional flood frequency relationship for estimation of floods of various return periods for the gauged catchments of south Bihar Rivers is expressed as:

$$Q(T) = \left[0.667 - 4.9693\left[1 - \left\{-ln\left(1 - \frac{1}{T}\right)\right\}^{-0.098}\right]\right]\bar{Q} \tag{6.8}$$

Where $\bar{Q}$ mean annual peak flood of the catchment

6.5 Developing relationship between peak flood and catchment physiographic characteristics.

The relationship development between the peak flood and catchment physiographic characteristics; namely catchment area, main stream length, slope, rainfall etc. have attempted in this section. In the analysis, Q/ Q_{peak} (m^3/s) is the dependent variable while the (i) A-Catchment Area (km^2), (ii) L-Main Stream Length (km), (iii) S-Slope (m/m), (iv) R-Rainfall (m) and (v) F-Forest cover (km^2) are considered as

independent variables. The data for 14 catchments (used in RFF analysis) have been used in this exercise.

6.6 Multivariate multiple regression

Regression technique has been attempted for three types of relationships namely; linear, logarithmic and exponential among dependent and independent variables. Further, the variables are selected using Forward selection and backward elimination techniques. For forward selection approach, the correlations of the dependent variable with independent variables have been used. The variable showing highest correlation has been analysed first and in successive trial the additional variables is added which shows next higher value. In each trial, the *Significance F* of the regression and *P-value* of individual variables smaller than 0.05 (for 95% confidence interval) have been used as selection criteria.

6.6.1 Multivariate Linear regression

Once, the forward selection is done with inclusion of all the variables one by one, the backward elimination technique is also tried to remove redundant variables. For this process again *Significance F* of the regression and *P-value* of individual variables smaller than 0.05 have been used. The following is the list of variables used in the analysis.

The correlation matrix for the variables is shown below:

Variables	***Q***	*A (km2)*	*F(KM2)*	*L (km)*	*S*	*R (m)*
Q	1					
A (km2)	0.733837	1				
F(KM2)	0.702137	0.592556	1			
L (km)	0.548523	0.927196	0.613834	1		
S	0.44626	0.571062	0.083464	0.386516	1	
R (m)	0.441268	-0.07104	0.438468	-0.21888	-0.02787	1

Inference: The correlation of 'Q' with 'A' 'F', 'L', 'S', and 'R' is estimated as 0.733837, 0.702137, 0.548523, 0.44626 and 0.441268, respectively. Thus the highest correlated variable is A', followed by 'F', 'L', and 'S'.

6.6.1.1 Forward Selection

(a) Set-1 (Q vs A)

Regression Statistics	
Multiple R	0.733837

R Square	0.538516
Adjusted R Square	0.500059
Standard Error	164.5802
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	379296.2	379296.2	14.00308	0.002811
Residual	12	325039.6	27086.63		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	93.27949	69.8459	1.335504	0.206491	-58.9017	245.4606
A (km2)	0.111174	0.029709	3.742069	0.002811	0.046443	0.175905

Linear relationship: Qp=93.2795+0.1112*A

Inference: For relationship between 'Q' and 'A'; R^2 is estimated as 0.54 with significance F 0.0028

(b) Set-2 (Q vs A and F)

Regression Statistics	
Multiple R	0.805373
R Square	0.648626
Adjusted R Square	0.58474
Standard Error	149.9955
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	456850.7	228425.4	10.15285	0.003175
Residual	11	247485.1	22498.64		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	63.61932	65.63031	0.969359	0.353198	-80.832	208.0707
A (km2)	0.074194	0.033613	2.207288	0.049452	0.000212	0.148177

F(KM2)	0.197985	0.106637	1.85663	0.090325	-0.03672	0.43269

Linear relationship: Qp=63.61932+0.074194*A+0.197985*F

Inference: For relationship between 'Q' and 'A'; 'F'; R^2 is estimated as 0.65 with significance F 0.0031

(c) Set-3 (Q vs A, F and L)

Regression Statistics	
Multiple R	0.914356
R Square	0.836047
Adjusted R Square	0.786861
Standard Error	107.4607
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	588857.7	196285.9	16.99767	0.000298
Residual	10	115478.1	11547.81		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	213.9387	64.71072	3.306078	0.007932	69.7542	358.1231
A (km2)	0.229829	0.05195	4.42401	0.001286	0.114076	0.345582
F(KM2)	0.254433	0.0782	3.253603	0.008667	0.080192	0.428674
L (km)	-4.10479	1.214065	-3.38103	0.006991	-6.80989	-1.39968

Linear relationship: Qp=213.9387+0.2298*A+0.2544*F-4.1048*L

Inference: For relationship between 'Q' and 'A'; 'F', 'L'; R^2 is estimated as 0.84 with significance F 0.0003

(d) Set-4 (Q vs A, F, L and S)

Regression Statistics	
Multiple R	0.914413
R Square	0.836151
Adjusted R Square	0.763329
Standard Error	113.2376
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	588930.9	147232.7	11.48214	0.001389
Residual	9	115404.9	12822.76		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	218.4834	90.92031	2.403021	0.039701	12.8074	424.1595
A (km2)	0.233359	0.071957	3.243045	0.010109	0.070581	0.396136
F(KM2)	0.252249	0.087325	2.888623	0.017924	0.054706	0.449791
L (km)	-4.15011	1.412952	-2.93719	0.016562	-7.34643	-0.95379
S	-2019.75	26726.43	-0.07557	0.941413	-62479.1	58439.65

Linear relationship: Qp=218.4834+0.2334*A+0.25225*F-4.15011*L-2019.74643*S

Inference: For relationship between 'Q' and 'A'; 'F', 'L' 'S'; R^2 is estimated as 0.84 with significance F 0.0014

(d) Set-5 (Q vs A, F, L, S and F)

Regression Statistics	
Multiple R	0.919923
R Square	0.846258
Adjusted R Square	0.750169
Standard Error	116.3433
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	596049.8	119210	8.807042	0.004142
Residual	8	108286	13535.75		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-515.403	1016.268	-0.50715	0.625729	-2858.92	1828.114
A (km2)	0.214778	0.078244	2.744996	0.025253	0.034348	0.395208
F(KM2)	0.176843	0.137335	1.287679	0.233856	-0.13985	0.493539
L (km)	-3.27744	1.885592	-1.73815	0.120382	-7.62563	1.07074
S	-3276.72	27514.07	-0.11909	0.908138	-66724.3	60170.85
R (m)	691.4764	953.4854	0.725209	0.488991	-1507.26	2890.218

Linear relationship: Qp=-515.403+0.2148*A+0.1768*F-3.2774*L-3276.72*S+691.4764*R

Inference: For relationship between 'Q' and 'A'; 'F', 'L' 'S', 'R'; R^2 is estimated as 0.846 with significance F 0.0041

6.6.1.2 Backward elimination

Considering, set-5, although the significance F is less than 0.05, the p values of variables *F, L, S* and *R* are higher than 0.05. The p values of only one variable *A* is less than 0.05. Hence, after elimination the variables *F, L, S* and *R*; the set becomes set-1 for which $R^2 = 0.538516$.

Similarly for set-2, the significance F is less than 0.05, the p value of variables *F* is higher than 0.05 while the p values of variable *A* is less than 0.05. Hence, after elimination the variables *F*; the set again becomes set-1 for which $R^2 = 0.538516$.

6.6.2 2. Multivariate nonlinear regression (logarithmic)

The correlation matrix with independent variables and dependent variables are estimated as shown below:

	Q	*LN(A)*	*LN(F)*	*LN(L)*	*LN(R)*	*LN(S)*
Q	1					
LN(A)	0.772304	1				
LN(F)	0.693142	0.837261	1			
LN(L)	0.609953	0.934942	0.697194	1		
LN(R)	0.438457	0.073767	0.379567	-0.17623	1	
LN(S)	0.353395	0.229218	0.308191	0.177713	0.026755	1

6.6.2.1 Forward Selection

(a) Set-1 (Q vs lnA)

Regression Statistics	
Multiple R	0.772304
R Square	0.596454
Adjusted R Square	0.562825
Standard Error	153.9025
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	420104	420104	17.73639	0.001207
Residual	12	284231.8	23685.99		

Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-699.874	240.0904	-2.91504	0.012959	-1222.99	-176.762
LN(A)	142.5715	33.85322	4.21146	0.001207	68.81164	216.3313

Nonlinear logarithmic relationship: Qp=-699.874+142.5715*lnA

Inference: For relationship between 'Q' and 'A'; R^2 is estimated as 0.596 with significance F 0.0012

(b) Set-2 (Q vs lnA and lnF)

Regression Statistics	
Multiple R	0.776977
R Square	0.603693
Adjusted R Square	0.531637
Standard Error	159.2977
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	425202.5	212601.2	8.378123	0.006154
Residual	11	279133.4	25375.76		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-571.317	379.4893	-1.50549	0.160362	-1406.57	263.9336
LN(A)	118.5221	64.0814	1.849556	0.091406	-22.5201	259.5643
LN(F)	9.368939	20.9016	0.44824	0.66268	-36.6352	55.37306

Nonlinear logarithmic relationship: Qp=-571.317+118.5221*lnA+9.3689*lnF

Inference: For relationship between 'Q' and 'A'and 'F'; R^2 is estimated as 0.604 with significance F 0.0062

(c) Set-3 (Q vs lnA, lnF and lnL)

Regression Statistics	
Multiple R	0.836635
R Square	0.699957
Adjusted R Square	0.609945
Standard Error	145.3722
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	493005.1	164335	7.7762	0.005699
Residual	10	211330.8	21133.08		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-577.706	346.3334	-1.66806	0.126265	-1349.39	193.9729
LN(A)	329.8259	131.6677	2.504987	0.031179	36.452	623.1997
LN(F)	-7.42853	21.25505	-0.34949	0.733966	-54.7877	39.93068
LN(L)	-309.831	172.9747	-1.79119	0.103526	-695.242	75.58097

Nonlinear logarithmic relationship: Qp=-577.706+329.8259*LN(A)-7.42853*LN(F)-309.831*LN(L)

Inference: For relationship between 'Q' and 'A', 'F' and 'L'; R^2 is estimated as 0.67 with significance F 0.0057

(c) Set-4 (Q vs lnA, lnF, lnL and lnR)

Regression Statistics	
Multiple R	0.882164
R Square	0.778213
Adjusted R Square	0.679641
Standard Error	131.7458
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	548123.3	137030.8	7.89487	0.00513
Residual	9	156212.5	17356.95		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-953.482	378.1285	-2.52158	0.032684	-1808.87	-98.0961
LN(A)	252.6755	126.9371	1.990557	0.077727	-34.4761	539.8271
LN(F)	-23.694	21.31584	-1.11157	0.295134	-71.9138	24.52576
LN(L)	-101.59	195.5238	-0.51958	0.615893	-543.895	340.7155
LN(R)	1765.763	990.8804	1.782014	0.108431	-475.765	4007.29

Nonlinear logarithmic relationship: Qp=-953.482+252.6755*lnA-23.694*lnF-101.59*lnL+1765.763*lnR

Inference: For relationship between 'Q' and 'A', 'F' and 'L', 'R'; R^2 is estimated as 0.78 with significance F 0.0051

(c) Set-5 (Q vs lnA, lnF, lnL, lnR and lnF)

Regression Statistics	
Multiple R	0.90948
R Square	0.827154
Adjusted R Square	0.719125
Standard Error	123.3602
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	582593.9	116518.8	7.656775	0.006445
Residual	8	121741.9	15217.74		
Total	13	704335.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-569.99	436.2161	-1.30667	0.227636	-1575.91	435.9267
LN(A)	243.2548	119.0223	2.043776	0.075232	-31.211	517.7207
LN(F)	-31.6414	20.64579	-1.53258	0.163918	-79.2506	15.96793
LN(L)	-66.8317	184.5296	-0.36217	0.726603	-492.358	358.6944
LN(R)	2039.632	945.4871	2.157229	0.063061	-140.665	4219.929
LN(S)	71.19629	47.30509	1.505045	0.170728	-37.8894	180.282

Nonlinear logarithmic relationship: Qp=-569.99+243.2548*lnA-31.6414*lnF-66.8317*lnL+2039.632*lnR+71.19629*lnS

Inference: For relationship between 'Q' and 'A', 'F' 'L' 'S', 'and R'; R^2 is estimated as 0.82 with significance F 0.0064

6.6.2.2 Backward elimination

With forward selection with 2 and higher variables (set-2 to set-5), the significance F is more than 0.05 and also the P-values of individual variables are also more than 0.05 so no backward elimination is applicable for these sets. Only logarithmic relationship exists with single variable i.e. lnA for which R^2 is 0.596454

6.6.3 3. Multivariate nonlinear regression (exponential)

Correlation matrix

	lnQ	*lnA*	*lnF*	*lnL*	*lnR*	*lnS*
lnQ	1					
lnA	0.884771	1				
lnF	0.872933	0.837261	1			
lnL	0.705712	0.934942	0.697194	1		
lnR	0.464249	0.073767	0.379567	-0.17623	1	
lnS	0.305955	0.229218	0.308191	0.177713	0.026755	1

6.6.3.1 Forward Selection

(a) Set-1 (lnQ vs lnA)

Regression Statistics	
Multiple R	0.884771
R Square	0.78282
Adjusted R Square	0.764722
Standard Error	0.552252
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	13.19163	13.19162615	43.25376	2.62E-05
Residual	12	3.659786	0.304982162		
Total	13	16.85141			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.34064	0.861522	-0.39539771	0.699488	-2.21774	1.536451
lnA	0.79892	0.121476	6.576759313	2.62E-05	0.534246	1.063594

Nonlinear exponential relationship: $Q=0.7113*A^{0.7989}$

Inference: For relationship between 'Q' and 'A'; R^2 is estimated as 0.78 with significance F 0.00026

(b) Set-2 (lnQ vs lnA and lnF)

Regression Statistics	
Multiple R	0.917184
R Square	0.841227
Adjusted R Square	0.812359
Standard Error	0.493185
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	14.17587	7.087933	29.14069	4.02E-05
Residual	11	2.675546	0.243231		
Total	13	16.85141			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.445538	1.174897	1.230353	0.244227	-1.14039	4.03147
lnA	0.464776	0.198396	2.342673	0.038988	0.02811	0.901442
lnF	0.130173	0.064711	2.011595	0.06941	-0.01226	0.272601

Nonlinear exponential relationship: $Q=4.2442*A^{0.4648}*F^{0.13017}$

Inference: For relationship between 'Q' and 'A' and 'F'; R^2 is estimated as 0.91 with significance F 0.0004

(c) Set-3 (lnQ vs lnA, lnF and lnL)

Regression Statistics	
Multiple R	0.954084
R Square	0.910275
Adjusted R Square	0.883358
Standard Error	0.388843
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	15.33943	5.113142	33.81741	1.51E-05
Residual	10	1.511985	0.151199		
Total	13	16.85141			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	1.41907	0.926375	1.531853	0.156561	-0.64502	3.483162
lnA	1.34012	0.352186	3.805153	0.003457	0.555402	2.124838
lnF	0.060588	0.056853	1.065692	0.311617	-0.06609	0.187265
lnL	-1.2835	0.462674	-2.77409	0.019649	-2.3144	-0.2526

Nonlinear exponential relationship: $Q=4.1333A^{1.3401}F^{0.06059}L^{-1.2835}$

Inference: For relationship between 'Q' and 'A' 'F' and 'L'; R^2 is estimated as 0.91 with significance F 0.000015

(d) Set-4 (lnQ vs lnA, lnF, lnL and lnR)

Regression Statistics	
Multiple R	0.975232
R Square	0.951078
Adjusted R Square	0.929335
Standard Error	0.302655
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	16.02701	4.006753	43.74189	6.69E-06
Residual	9	0.824399	0.0916		
Total	13	16.85141			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.091844	0.868661	0.10573	0.918115	-1.8732	2.056891
lnA	1.067628	0.291608	3.661176	0.005226	0.407965	1.727291
lnF	0.003139	0.048968	0.0641	0.950292	-0.10763	0.113912
lnL	-0.548	0.44917	-1.22003	0.253456	-1.56409	0.46809
lnR	6.236603	2.276314	2.739782	0.022856	1.087223	11.38598

Nonlinear exponential relationship: $Q=1.0962A^{1.0676}F^{0.00314}L^{-0.548}R^{6.2366}$

Inference: For relationship between 'Q' and 'A' 'F' 'L' & 'R'; R^2 is estimated as 0.95 with significance F 0.00006

(e) Set-5 (lnQ vs lnA, lnF, lnL, lnR and lnS)

Regression Statistics	
Multiple R	0.979583
R Square	0.959583
Adjusted R Square	0.934323
Standard Error	0.291779
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	16.17033	3.234067	37.98756	2.29E-05
Residual	8	0.681079	0.085135		
Total	13	16.85141			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.873807	1.031764	0.846905	0.421659	-1.50545	3.253059
lnA	1.048419	0.281519	3.724155	0.005836	0.399236	1.697602
lnF	-0.01307	0.048833	-0.26757	0.795799	-0.12567	0.099542
lnL	-0.47713	0.43646	-1.09318	0.306138	-1.48361	0.529352
lnR	6.795038	2.236322	3.038488	0.0161	1.63807	11.95201
lnS	0.145173	0.111889	1.297478	0.230628	-0.11284	0.403189

Nonlinear exponential relationship: $Q=1.096194A^{1.0676}F^{0.00314}L^{-0.548}R^{6.2366}S0^{0.1452}$

Inference: For relationship between 'Q' and 'A' 'F' 'L', 'R'; & 'S' R^2 is estimated as 0.96 with significance F 0.00003

6.6.3.2 Backward elimination

(f) Set-6 (lnQ vs lnA, and lnR)

Regression Statistics	
Multiple R	0.971019
R Square	0.942877
Adjusted R Square	0.932492
Standard Error	0.295819
Observations	14

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	15.88882	7.944408	90.78427	1.45E-07
Residual	11	0.962595	0.087509		
Total	13	16.85141			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.34961	0.461485	-0.75758	0.464622	-1.36533	0.666109
lnA	0.772199	0.065248	11.8349	1.34E-07	0.62859	0.915808
lnR	8.047613	1.449562	5.551756	0.000172	4.857149	11.23808

Nonlinear exponential relationship: $Q=0.7049*A^{0.7722*}R^{8.0476}$

Inference: For relationship between 'Q' and 'A' & 'R'; R^2 is estimated as 0.94 with significance F 0.000001

6.6.4 Summary of Multivariate multiple regression Analysis

The best fit relationship is developed with Q, A and R with the given data set is exponential in nature with R^2 =0.942877.

The relationship is given below:

$$\bar{Q} = 0.7049A^{0.7722}R^{8.0476} \tag{6.9}$$

6.7 RFF relationship for the ungauged catchments

In section 6.5 and 6.6, the relationship between mean annual flood and catchment characteristics has been developed which is shown in Eq. 6.9. Using the relationship and replacing the term $\bar{Q}$ in Eq. 6.7, the flood estimate for a given return period for unguaged catchment can be defined as flows:

$$Q(T) = \left[0.667 - 4.9693\ \left[1 - \left\{-ln\left(1 - \frac{1}{T}\right)\right\}^{-0.098}\right]\right] 0.7049A^{0.7722}R^{8.0476} \tag{6.10}$$

which simplified to,

$$Q(T) = \left[0.4702 - 3.5029\ \left[1 - \left\{-ln\left(1 - \frac{1}{T}\right)\right\}^{-0.069}\right]\right] A^{0.7722}R^{8.0476} \tag{6.11}$$

Eq. 6.11 is reported tabular form (**Table 6.5** and **Table 6.6**). To compute the annual average flood in a catchment for a given catchment area (sq.km) and annual average rainfall (in m) Q_0, can be read directly from the table, **Table 6.5**. Further to estimate the flood of specific return period, Q_0 is to be multiplied by the growth factor given in **Table 6.6**.

Further, Q_0 for the given catchment area (in sqkm) and annual average rainfall (in m) can also be estimated using the chart shown in Figure 6.5 and Figure 6.6. To estimate the flood of specific return period, Q_0 is to be multiplied by the growth factor.

Table 6.5: Table showing annual average flood (Q_0) using catchment characteristics

C A (km^2)	Rainfall (m)																									
	0.95	**0.96**	**0.97**	**0.98**	**0.99**	**1**	**1.01**	**1.02**	**1.03**	**1.04**	**1.05**	**1.06**	**1.07**	**1.08**	**1.09**	**1.1**	**1.11**	**1.12**	**1.13**	**1.14**	**1.15**	**1.16**	**1.17**	**1.18**	**1.19**	**1.2**
50	14	15	16	17	19	21	22	24	26	28	30	33	35	38	41	44	47	51	55	59	63	68	73	78	83	89
90	21	23	25	27	30	32	35	38	41	44	48	52	56	60	65	70	75	80	86	93	99	107	114	122	131	140
100	23	25	27	30	32	35	38	41	44	48	52	56	60	65	70	75	81	87	94	101	108	116	124	133	142	152
500	80	87	95	103	112	121	131	142	154	166	180	194	209	225	243	261	281	302	325	348	374	401	429	460	492	526
1000	137	149	162	176	191	207	225	243	263	284	307	331	357	385	415	446	480	516	554	595	638	684	733	785	841	899
1500	188	204	222	241	261	284	307	332	360	389	420	453	489	527	567	611	657	706	758	814	873	936	1003	1074	1150	1230
2000	234	255	277	301	327	354	384	415	449	485	524	566	610	658	708	762	820	881	947	1016	1090	1169	1253	1341	1436	1536
2500	278	303	329	358	388	421	456	493	534	577	623	672	725	781	842	906	974	1047	1125	1207	1295	1389	1488	1594	1706	1824
3000	320	349	379	412	447	484	525	568	614	664	717	774	835	900	969	1043	1121	1205	1295	1390	1491	1599	1713	1834	1963	2100
3500	361	393	427	464	503	545	591	640	692	748	808	872	940	1013	1091	1174	1263	1358	1458	1566	1680	1801	1930	2066	2212	2366
4000	400	435	473	514	558	605	655	709	767	829	895	966	1042	1123	1210	1302	1400	1505	1617	1736	1862	1996	2139	2291	2452	2623
4500	438	477	518	563	611	662	717	777	840	908	981	1058	1142	1230	1325	1426	1534	1649	1771	1901	2039	2186	2343	2509	2685	2872
5000	475	517	562	611	663	718	778	842	911	985	1064	1148	1238	1335	1437	1547	1664	1788	1921	2062	2212	2372	2541	2722	2913	3116
5500	512	557	605	657	713	773	838	907	981	1060	1145	1236	1333	1436	1547	1665	1791	1925	2068	2220	2381	2553	2736	2929	3135	3354
6000	547	595	647	703	763	827	896	970	1049	1134	1225	1322	1425	1536	1655	1781	1915	2059	2211	2374	2547	2730	2926	3133	3353	3587
6500	582	633	688	748	811	880	953	1032	1116	1206	1303	1406	1516	1634	1760	1894	2037	2190	2352	2525	2709	2904	3112	3333	3567	3815
7000	616	671	729	792	859	932	1009	1092	1182	1277	1379	1489	1606	1730	1864	2006	2157	2319	2491	2674	2869	3076	3295	3529	3777	4040
7500	650	707	769	835	906	982	1064	1152	1246	1347	1455	1570	1694	1825	1966	2116	2275	2446	2627	2820	3025	3244	3476	3722	3984	4261
8000	683	744	808	878	952	1033	1119	1211	1310	1416	1529	1651	1780	1918	2066	2224	2392	2571	2761	2964	3180	3410	3653	3912	4187	4479
8500	716	779	847	920	998	1082	1172	1269	1373	1484	1603	1730	1865	2010	2165	2330	2506	2694	2894	3106	3332	3573	3829	4100	4388	4694
9000	749	814	885	961	1043	1131	1225	1326	1435	1551	1675	1808	1950	2101	2263	2435	2619	2815	3024	3246	3483	3734	4001	4285	4586	4906

Table 6.6: Growth factor to estimate flood of various return period (Q_0xGF).

Return period (yr)	2	5	10	20	25	50	100	500	1000
Growth factor (GF)	0.5599	0.8522	1.0586	1.2670	1.3352	1.5525	1.7788	2.3454	2.6090

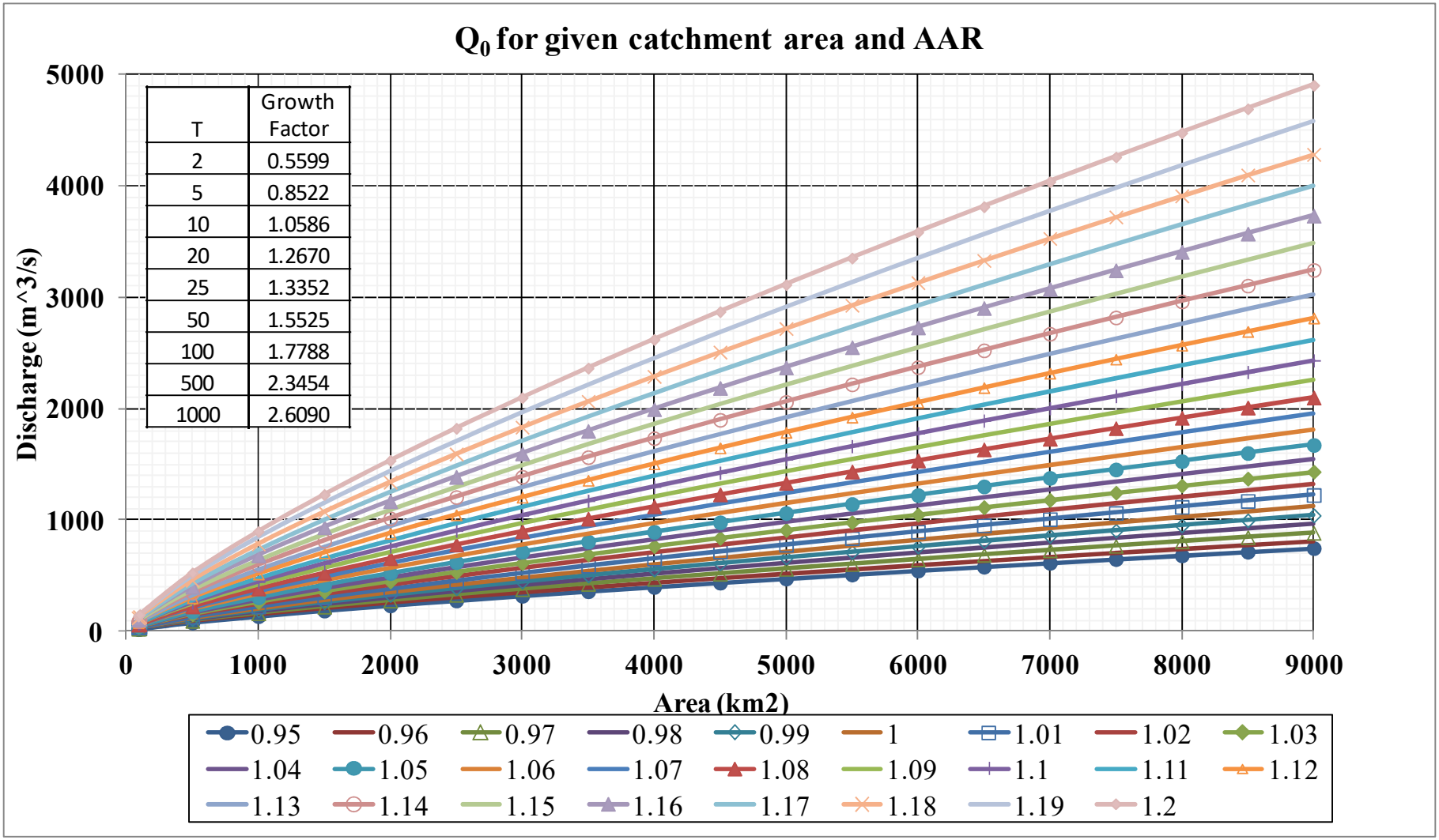

Figure 6.5: Average annual flood (Q_0) estimation chart for large catchments.

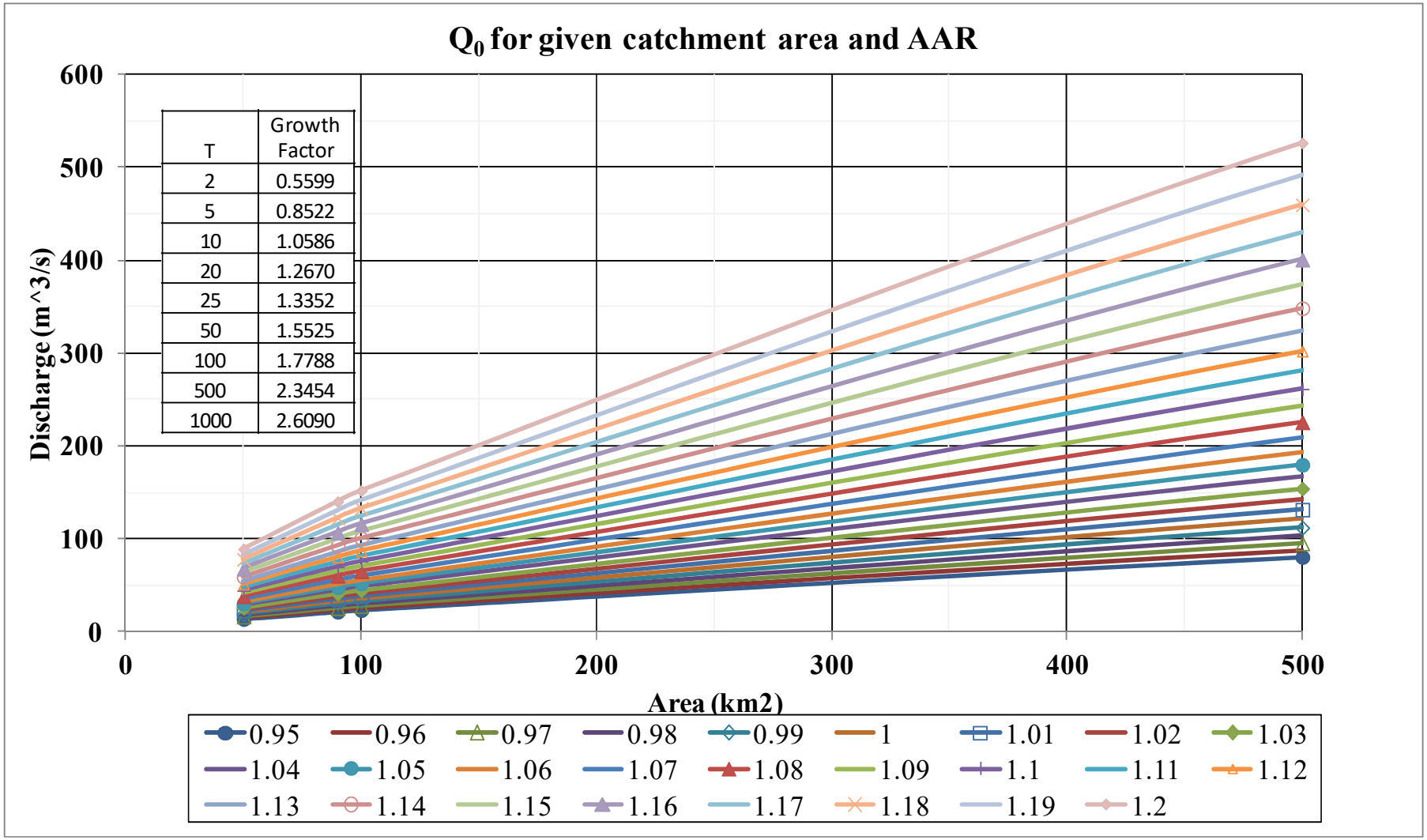

Figure 6.6: Average annual flood (Q_0) estimation chart for small catchments.

7 Summary and Conclusions

The important components of the work carried out in the study includes estimation of catchment characteristics (catchment area, main stream length, stream slope etc.) using online DEM and Google Earth images. The hydrological data for various stream/ sites (30 GD stations) in South Bihar region were collected from Central Water Commission and Water Resources Department, Government of Bihar. The annual peak flood data are screened for their consistency and long term availability. The data for short span (less than 6 year data) are ignored. Further, using the clustering technique, the homogeneous stations are identified. Further using L-moment RFFA, the data are further tested for homogeneity. RFFA is carried out to estimate flood of various return period for gauges and unguaged catchment of the region. The summary of the analysis and results are described as below:

7.1 Data Preparation & Clustering

- Annual peak flood data from 30 stations were collected; 23 reliable stations were selected based on consistency and sufficient records.
- Histogram analysis showed ~80% stations have low variability (low CV), indicating similar hydrological behaviour.
- Hierarchical clustering grouped stations into 3 clusters, with one large homogeneous cluster of 15 stations.

7.2 Regional Flood Frequency Analysis using L-moment approach

7.2.1 Homogeneity Testing (L-Moment Method)

- L-moment based heterogeneity measure (H) was used:
    - With 15 stations region was identified as heterogeneous.
    - After removing one site, 14 stations formed a homogeneous region (H < 1).
- Discordancy values (Di) for all 14 sites were within acceptable limits.

7.2.2 Selection of Best-Fit Distribution is carried out using L-moment ratio diagram and Z-statistic ($|Z^{i}_{dist}|$).

- Several distributions were tested (GEV, GLO, GNO, PE3).

- Generalized Extreme Value (GEV) distribution was identified as the best fit ($Z \approx 0.41$).

7.2.3 Regional Flood Frequency Relationship for gauged catchments

- Developed regional growth factors for different return periods.
- Flood estimation equation expressed as:
  - Flood = (Mean annual flood) × (Growth factor)

7.2.4 Regression Analysis for Flood Estimation

- Relationship between peak flood (Q) and catchment characteristics was analyzed using:
  - Linear regression
  - Logarithmic regression
  - Exponential regression

The Key findings of the regression analysis are:

- Catchment area (A) is the most influential variable; however, correlation can be further improved with additional catchment characteristics (rainfall).
  - The best identified model is the Exponential regression with catchment area (A) and rainfall (R) that achieved highest accuracy ($R^2 \approx 0.94$)**.**

7.2.5 Ungauged Catchment Flood Estimation

- Developed equations to estimate floods in ungauged basins using catchment area and Rainfall.
- Tables and charts are prepared from which the mean annual flood (Q_0) can be computed directly. To estimate the flood for any return period, the growth factor is to be used following relationship of, $Q_T = Q_0 \times$ Growth Factor

Based on the study following conclusions are made:

1. A homogeneous region of 14 stations was successfully identified in the South Bihar region to carry out the regional flood frequency analysis.
2. GEV distribution is identified as the most suitable PDF for regional flood frequency analysis in the region.
3. Exponential model with catchment area and annual average rainfall provides the best prediction for average annual flood in a catchment.
4. Practical tools (tables, charts, equations) were developed for both gauged and ungauged catchments. The developed charts and tables can directly be used by

field Engineers to estimate the annual average flood and floods of various return period in a given catchment.

References:

Acreman, M.C.; Wiltshire, S.E., 1989: The regions are dead. Long live the regions. Methods of identifying and dispensing with regions for flood frequency analysis. In: -FR1ðir'DS in Hydrology (ed. L. Roald, K. Nordseth and K.A Hassel), IAHS Publ. No- 187, r75-188

Ahmad, M. I., Sinclair, C. D. and Werritty, A. (1988). Log-logistic flood frequency analysis. Journal of Hydrology Vol (98), 205-224.

Bates, B.C.; Rahman, A.; Mein, R.G.; Weinmann, P.E. Climatic and physical factors that influence the homogeneity of regional floods in southeastern Australia. Water Resour. Res. 1998, 34, 3369–3381.

Benson, M.A. Evolution of methods for evaluating the occurrence of floods. US Geol. Surv. Water Supply Pap. 1962, 1580, 30.

Conleth Cunnane, Methods and merits of regional flood frequency analysis, Journal of Hydrology, Volume 100, Issues 1–3, 1988, Pages 269-290, ISSN 0022-1694, https://doi.org/10.1016/0022-1694(88)90188-6.

Cunnane, C. and Nash, J.E., 1971, Bayesian estimation of frequency of hydrological events, in: “Proceedings of the Warsaw Symposium. Mathematical Models in Hydrology, July 1971”, IAHS-AISH Publication No. 100, Vol. 1, pp. 47–55.

Cunnane, C., (1989). Statistical Distribution for Flood Frequency Analysis. Operational Hydrol. Rep. 33, World Meteorological Organisation, Geneva.

Dawdy, D. R. (1961), Variation of Flood Ratios with Size of Drainage Area, in Geological Survey Research 1961: U. S. G. S. Professional Paper 424-C, Article 160.

Dubey, Amit. (2019). Development of Regional Flood Frequency Relationship for Narmada Basin using Index Flood Procedure. International Journal of Innovative Technology and Exploring Engineering. 8. 2744-2752. 10.35940/ijitee.L2560.1081219.

Fill, H.D.; Stedinger, J.R. Homogeneity tests based upon Gumbel distribution and a critical appraisal of Dalrymple’s test. J. Hydrol. 1995, 166, 81–105.

Griffis, V. W., and Stedinger, J. R. (2007). “The log-Pearson type 3 distribution and its application in flood frequency analysis: 1. Distribution characteristics.” J. Hydrol. Eng., 12(5), 482–491.

Haddad K, Rahman A. 2007. Selection of the best fit flood frequency distribution and parameter estimation procedure—A case study for Tasmania in Australia. Stochastic Environmental Research and Risk Assessment DOI: 10.1007/s00477-010-0412-1.

Hailegeorgis, Teklu Tesfaye & Alfredsen, Knut. (2017). Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for mid-Norway, Journal of Hydrology: Regional Studies Volume 9, February 2017, Pages 104-126, https://doi.org/10.1016/j.ejrh.2016.11.004

Hewa, G.A., McMahon, T.A., Peel, M.C. and Nathan, R.J. (2003). Identification of the most appropriate regression procedure to regionalise extreme low flows, 28th Intl. Hydrology and Water Resour. Symp, 10-13 Nov., 2003

Hosking JR, M.; Wallis, J.R. Some statistics useful in regional frequency analysis. Water Resour. Res. 1993, 29, 271–281.

Hosking, J. R. M., and J. R. Wallis. 1986. Paleoflood hydrology and flood frequency analysis. Water Resources Research 22(4):543-550.

Hosking, J. R. M., and J. R.Wallis (1986), The Value of Historical Data in Flood Frequency Analysis, Water Resour. Res., 22(11), 1606–1612, doi:10.1029/WR022i011p01606.

Hosking, J. R. M., and Wallis, J. R. (1988). “The effect of intersite dependence on regional flood frequency analysis.” Water Resour. Res., 29, 271–281.

Hosking, J. R. M., and Wallis, J. R. (1997). Regional frequency analysis: An approach based on L-moments, Cambridge University Press, New York.

Houghton, J.C., 1978. Birth of a parent: the Wakeby distribution for modelling flood flows. Water Resources Res., 14(6), 1105–1109.

IACWD (Interagency Advisory Committee on Water Data). 1982. Guidelines for Determining Flood Flow Frequency, Bulletin 17-B, U.S. Department of the

Interior, U.S. Geological Survey, Office of Water Data Coordination, Reston, Virginia. National Academies of Sciences, Engineering, and Medicine. 1999. Improving American River Flood Frequency Analyses. Washington, DC: The National Academies Press. https://doi.org/10.17226/6483.

India-WRIS, 2022, Hydrological Observation Sites under CWC, Patnahttps://indiawris.gov.in/wiki/doku.php?id=cwc_hydro-meteorological_sites (last accessed in December 2022)

Institution of Engineers Australia (I.E. Aust.) (1987/2001) Australian rainfall and runoff: a guide to flood estimation. In: Pilgrim DH (ed), vol 1. I. E. Aust, Canberra

IS 12094 (2000): Guidelines for Planning and Design of River Embankments (Levees), Bureau of Indian Standards, Manak Bhavan, 9 Bahadur Shah Zafar Marg, New Delhi 110002.

IS 14815 (2000), Design flood for river diversion works – guidelines, Bureau of Indian Standards, Manak Bhavan, 9 Bahadur Shah Zafar Marg, New Delhi 110002.

IS 7784-1 (1993): Design of cross drainage works- Code of practice, Part 1: General features [WRD 13: Canals and Cross Drainage Works.

K.W. Potter, D.P. Lettenmaier A comparison of regional flood frequency estimation methods using a resampling method Water Resour. Res., 26 (3) (1990), pp. 415-424

Klemes, V., 1987. Dilettantism in hydrology: transition or destiny. Water Resources Res., 22(9), 1775–1885.

Kroll, K., and J. R. Stedinger. 1999. Estimation of moments and quantiles with censored data. Water Resources Research 32(4): 1005-1012. National Academies of Sciences, Engineering, and Medicine. 1999. Improving American River Flood Frequency Analyses. Washington, DC: The National Academies Press. https://doi.org/10.17226/6483.

Kuczera, G. (1982), Robust flood frequency models, Water Resour. Res., 18(2), 315–324, doi:10.1029/WR018i002p00315.

Kuczera, G. (1983), A Bayesian surrogate for regional skew in flood frequency analysis, Water Resour. Res., 19(3), 821–832, doi:10.1029/WR019i003p00821.

Kumar Rakesh, Development of regional flood frequency relationships using L-moments for gauged and ungauged catchments of India, Water Utility Journal 23: 37-47, 2019.

Kumar, R., Chatterjee, C. (2011). Development of Regional Flood Frequency Relationships for Gauged and Ungauged Catchments Using L-Moments. In: Kropp, J., Schellnhuber, HJ. (eds) In Extremis. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14863-7_5

Kumar, Rakesh & Chatterjee, Chandranath & Kumar, Sanjay & Lohani, Anil Kumar & Singh, R.. (2003). Development of Regional Flood Frequency Relationships Using L-moments for Middle Ganga Plains Subzone 1(f) of India. Water Resources Management. 17. 243-257. 10.1023/A:1024770124523.

Lay, M. (1989). 'Design flood estimation for ungauged rural catchments in Victoria'. Road Construction Authority, Tech Bulletin No. 38, pp 1-17.

Lettenmaier, D. P., and T. Y. Gan. 1990. Hydrologic sensitivities of the Sacramento San Joaquin River basin, California, to global warming. Water Resources Research 26(10):69-86. National Academies of Sciences, Engineering, and Medicine. 1999. Improving American River Flood Frequency Analyses. Washington, DC: The National Academies Press. https://doi.org/10.17226/6483.

Lettenmaier, D.P. & K.W. Potter (1985) Testing flood frequency estimation methods using a regional flood generation model. Water Resour. Res 21(12), 1903–1914.

Lettenmaier, D.P. and Potter, K.W. (1985). Testing flood frequency estimation methods using a regional flood generation model, Water Resour. Res., 21(12): 1903-1914.

LGBO, 2022, Hydrological Observation Sites under LGBO, CWC, Patna, https://cwc.gov.in/lgbo/ho-ff-network/ho, (last accessed in December 2022)

Lu L.-H. and Stedinger J.R., 1992. Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test. J. Hydrol., 138, 223−245.

Marin, C. (1983). Uncertainty in water resources planning, Ph.D. thesis, Harvard Univ., Cambridge Mass.

National Research Council (1988). Committee on Techniques for Estimating Probabilities of Extreme Floods, "Estimating Probabilities of Extreme Flooods, Methods and Recommended Research", National Academy Press, Washington, D.C.

Ouarda T.B.M.J., Diaz-Delgado K.M. Bâ, C., Cârsteanu A., Chokmani K., Gingras H., Quentin E., Trujillo E. and Bobée B., Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study, Journal of Hydrology, Volume 348, Issues 1–2, 2008, Pages 40-58, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2007.09.031.

Pilgrim, D.H.; Cordery, I. 1993: Flood runoff. In Maidment, D.R. {ed.), Handbook of Hydrology, McGraw Hill, New York, 9.1.

Pilon, Paul & Adamowski, Kaz. (2011). The Value of Regional Information to Flood Frequency Analysis Using the Method of L-Moments. Canadian Journal of Civil Engineering. 19. 137-147. 10.1139/l92-014.

Potter, K.W. and Lettenmaier, D.P. (1990). 'A comparison of regional flood frequency estimation mean using a resampling method'. Water Resour. Res, vol 26, iss 3 , pp 424.

Potter, K.W. and Lettenmaier, D.P. (1990). 'A comparison of regional flood frequency estimation mean using a resampling method'. Water Resour. Res, vol 26, iss 3 , pp 424.

Rahman, A. (1997). Flood Estimation for ungauged catchments: A regional approach using flood and catchment characteristics, PhD thesis, Department of Civil Engineering, Monash University.

Rahman, A. and Hollerbach, D. (2003). Study of Runoff Coefficients Associated with the Probabilistic Rational Method for Flood Estimation in South-east

Australia, 28th Hydrology and Water Resources Symposium, Wollongong, 10-13 November 2003, pp. 199-203.

Rahman, A., Weinmann, P.E. and Mein, R.G. (1999). At-site flood frequency analysis: LP3-product moment, GEV-L moment and GEV-LH moment procedures compared. In: Proceeding Hydrology and Water Resource Symposium, Brisbane, 6–8 July, 2, 715–720.

Requena, Ana & Ouarda, Taha & Chebana, Fateh. (2017). Flood Frequency Analysis at Ungauged Sites Based on Regionally Estimated Streamflows. Journal of Hydrometeorology. 18. 10.1175/JHM-D-16-0143.1.

Riggs, H.C., 1973. Regional analysis of streamflow characteristics. US Geol. Surv. Water Resources Invest. Tech., Book 4.

Rijal, N.; Rahman, A. Design flood estimation in ungauged catchments: Quantile regression technique and Probabilistic Rational Method compared. In Proceedings of the Modsim05: International Congress on Modelling and Simulation: Advances and Applications for Management and Decision Making, Melbourne, Australia, 12–15 December 2005.

Rossi, F., Fiorentino, M., and Versace, P. (1984). "Two-component estreme value distribution for flood frequency analysis." Water Resour. Res., 20(7), 847–856.

Stedinger J.R. & G.D. Tasker (1985) Regional hydrologic analysis—1. Ordinary, weighted and generalized least squares compared. Water Resour. Res. 21(9), 1421–1432.

Stedinger, J. R. (1983), Estimating a regional flood frequency distribution, Water Resour. Res., 19(2), 503–510, doi:10.1029/WR019i002p00503.

Stedinger, J.R. (1993) Estimating a regional flood frequency distribution. Water Resour. Res. 19(2), 503–510.

Tasker, G. D. (1980). "Hydrologic regression with weighted least squares." Water Resour. Res., 16(6), 1107–1113.

Tasker, G. D., and Stedinger, J. R. (1986). "Regional skew with weighted LS regression." J. Water Resour. Plann. Manage., 112(2), 225–237.

Tasker, G. D., and Stedinger, J. R. (1989). “An operational GLS model for hydrologic regression.” J. Hydrol., 111(1–4), 361–375.

Tasker, G.D and Stedinger, J.R. (1987). Regional regression offlood characteristics employing historical information. In: W.H. Kirby, S.Q. Hua and L.R. Beard (ed), Analysis of Extra-ordinary flood events. J. Hydrology., 96:255-264.

Teklu T. Hailegeorgis, Knut Alfredsen, Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for mid-Norway, Journal of Hydrology: Regional Studies, Volume 9, 2017, Pages 104-126, ISSN 2214-5818, https://doi.org/10.1016/j.ejrh.2016.11.004.

Vogel, R., Fennessey, N., 1993. L moment diagrams should replace product moment diagrams. Water Resources Research 29(6), 1745-1752.

Vogel, R.M., Kroll, C.N., 1990. Generalized Low-Flow Frequency Relationships for Ungaged Sites in Massachusetts1. JAWRA Journal of the American Water Resources Association 26, 241–253. https://doi.org/10.1111/j.1752-1688.1990.tb01367.x

Wallis, J.R. & E.F. Wood (1985) Relative accuracy of log Pearson III procedures. J. Hydraul. Eng. 111, 1043–1056.

Wiltshire, S. E. (1986). Regional flood frequency analysis I: Homogeneity statistics. Hydrological Sciences Journal, 31(3), 321–333. https://doi.org/10.1080/02626668609491051

Xiao Pan, Ataur Rahman, Khaled Haddad, Taha B.M.J. Ouarda, Ashish Sharma, Regional flood frequency analysis based on peaks-over-threshold approach: A case study for South-Eastern Australia, Journal of Hydrology: Regional Studies, Volume 47, 2023, 101407, ISSN 2214-5818, https://doi.org/10.1016/j.ejrh.2023.101407.