POLITECNICO DI TORINO

Ph.D. in Engineering for Natural and Built Environment

Parametric and non-parametric approaches for runoff and rainfall regionalization

Muhammad Uzair Qamar

matr. 189981

Advisors:

Prof. Pierluigi Claps

Dr. Daniele Ganora

Ph.D. Dissertation

February 2015

To my family

Abstract

The information on river flows is important for a number of reasons including; the construction of hydraulic structures for water management, for equitable distribution of water and for a number of environmental issues. The flow measurement devices are generally installed across the workspace at various locations to get data on river flows but due to a number of technical and accessibility issues, it is not always possible to get continuous data. The amount rainfall in a basin area also contributes towards the river flows and intense rainfall can cause flooding. The extended rainfall maps for the study areas to analyze these extreme events can be of great practical and theoretical interest.

This thesis can be generally regarded as a work on catchment hydrology and mapping rainfall extremes to estimate certain hydrological variables that are not only useful for future research but also for practical designing and management issues. We analyzed a number of existing techniques available in literature to extend the hydrological information from gauged basin to ungauged basin; and suggested improvements. The three main frontiers of our work are: Monthly runoff regime regionalization, Flow duration curves (FDCs) regionalization and preparing rainfall hazardous maps.

The proposed methods of regionalization for runoff regime and FDCs are tested for the basins located in northern Italy; whereas for rainfall extremes, the procedure is applied to the data points located in northern part of Pakistan.

I would like to extend my sincere thanks to my supervisors Prof. Pierluigi Claps and Dr. Daniele Ganora for their consistent guidance and support. I appreciate the way my lab mates and colleagues helped me out during my stay in Italy.

Contents

Chapter 1. Introduction

Water is, without any doubt, the important and essential natural resource. Since it thoroughly affects the life of earth, in every aspect, it is therefore important to understand the mechanism of water availability not only for scientific reasons but also for an efficient management of available water resources. The science that studies the movement, distribution pattern and occurrence of water within each phase of water cycle is called hydrology.

The understanding on water cycle is extremely limited mainly because of involvement of complex physical processes and also because these physical processes take place over a wide range of both spatial and temporal scales; for example Siberian winds bring chilling winter in Pakistan. Apart from a broader prospective, the study on hydrological issues is divided into different areas of expertise e.g. water resources management, meterogical science etc.

This thesis can be generally regarded as a work on catchment hydrology and mapping rainfall extremes to estimate certain hydrological variables that are not only useful for future research but also for practical designing and management issues.

A fundamental landscape unit that physically relates hydrological cycle with ecological, climatological, morphological, geochemical and other processes of an area is termed as basin [Sivapalan et al., 2003]. The aim of studying this interaction is to mainstream the concept of fluxes through the basin boundaries, particularly from and towards the atmosphere and groundwater (e.g. rainfall converted to surface runoff and a fraction of it percolates down to recharge groundwater level). Among all other variables calculated during this interaction, surface runoff or discharge is the one that stands out due to its importance in the study of flood estimation, water management and also in the designing of hydraulic structures (i.e. dams, reservoirs, barrage etc.). Since all the geomorphological and climatological processes converge towards the magnitude of surface runoff, discharges, in some way, summaries or at least are representative of catchment processes.

Due to ever changing enviromental, climatological and geomorphological parameters within a complex system of basin; it is only possible through macro-characteristics like magnitude, frequency and duration of hydrological events to replicate basin behavior. The statistical laws and theorems can be used to estimate these macro-characteristics by trying to interpret the hydrological patterns without having a prior knowledge about the physical processes.

The entire efforts, in this thesis, are made to estimate the magnitude of discharge at an instantaneous time interval and magnitude of peak flow, along with its timing by simulating monthly flow regime for an ungauged basin. Availability of flow, exceeding or equaling a certain value for a certain percentage of time, is elaborated through the regionalization of flow duration curve (FDC). Moreover, the primary cause of extreme flow, at least for Pakistan is extreme rainfall; therefore rainfall hazardous map is generated for a certain return period to broaden the scope of our research.

The hydrological processes occurring in a basin play a pivotal role in shaping the life style of human societies. For example, in early history a strong influence of river Nile on the lives of early Egyptian civilization, for instance, is documented in Karnak temple complex on the northern side [*Lauro*, 2009]. The propitiatory values of flood level along the length of the river at different locations are represented by hieroglyphics.

In the recent decades, due to the increase in urbanization and change in land use patterns have contributed towards the rapid exploitation of water resources. This emphasized the need for a reliable classification of surface water flows based on their magnitude and time of occurrence in order to manage this natural resource efficiently (reservoir operations). The demand of water on the downstream for different, sometimes conflicting, necessities (e.g energy production plant, agriculture, industry etc.) can make water management issues exponentially complex. Moreover the environmental issues related to water quantity (water logging and wetland etc) and water quality (industrial waste water, chemical industries and lather industries) need to be addressed properly. Apart from meeting the water demands of commuters, it is also important to save the community from the adversities related to water availability (drought and extreme floods in case of extreme rainfalls). To address all the issues regarding water management, it is desired to have efficient long term information about certain hydrological variables.

The methods for studying the catchment behavior can be either direct or indirect. The direct methods are more straight forward and reliable, and the implementation involves comprehensive study of streamflow time series and the parameter related to it e.g. vegetation, soil characteristics and precipitation etc. The application of this method requires the discharge data to be known at the point of interest. The at-site availability at data is entirely dependent, whether or not the measurement instrument are installed and working correctly at the site of interest. Both instrument installation and operation are unrealistic incase of remote sites and site is termed as an ungauged one. For hydrological characterization of an ungauged basin indirect methods are generally called in for service.

The fundamental concept in the implementation of indirect procedure is to transfer the hydrological information from the gauged stations to an ungauged station based on the developed physical and statistical laws. The US national Research Council [1988] for hydro-meteorological modeling proposed a principle "substitute time for space" which summarized this procedure. The lack or absolutely no availability of hydrological data are compensated by extrapolating the hydrological records from the neighboring gauged basins. This topic of transferring data from gauged basin to an ungauged one is the back bone in the field of catchment hydrology and sensing its importance a whole decade was explicitly dedicated to prediction of ungauged basin (PUB) initiative [Sivapalan et al., 2003].

Our work covers up the following dimensions of PUB:

- the current procedure used for the regionalization of flow regime is revisited and special attention is given to estimate the instantaneous flow magnitude and peak flow occurring at data-scarce stations.

- the uncertainty in the results generated by application of the procedure were analyzed.

- the large-scale statistical models developed by using local information are used for the correction of estimates.

- the non-convential descriptors data and procedures are used to make hydrological estimates.

More precisely, in chapter 2 a detailed discussion on regionalisation procedure of monthly flow regimes is being eleborated. The aim of the study is to overcome some of the limitations posed by classical regionalization approaches while specifically giving attention to the position of peak flow w.r.t time. In this regard the at-site descriptors data and hydrological data are related through regression models. Each regression model is passed through various statistical tests (VIF and mantel test) to check its stability. To transfer the information from gauged basin to ungauged basin, unlike classical regression approach, this approach allows one to introduce even the complex descriptors in the regression model. The regionalisation procedure assumed that any variation occuring on descriptors space will be responded, in the same manner, on discharge space. The best model based on comprehensive MRM between distance matrices of discharge and descriptors data and cross validation based on delta (Δ) factor is selected to estimate flow regime at an ungauged basin.

Streamflows are either constituted by runoff from rainfall or from melting of snow moving downstream as surface or subsurface flow. When large quantity of water in the form of runoff flows quickly into streams and rivers; floods occur. A number of factors affect magnitude of floods like: (1) Intensity of rainfall including its duration. (2) Amount of snowmelt under the effect of temperature. (3) The geology, vegetation cover, topography of the basins. (4) The hydrological characteristics effecting rainfall extremes and snowmelt events. In third world countries like Pakistan, extreme rainfall events in the recent past have caused natural hazard because they are a source of degradation processes like flash floods, landslide triggering and erosion which cause a severe damage to the land and properties. In 2011, massive flooding as a result of extreme rainfall in Pakistan affected over 6 million people. According to some rough estimates made by the government of Pakistan, it destroyed over a million houses and standing crops over 4.5 million acres of land. There was a serious need to study the extreme rainfall events over the entire area of the country. Mapping the hazard of extreme rainfall is important as it allows us to assess the spatial distribution of this climatic feature even at locations where no climatic record exists. In the fields of regional planning and environmental management, rainfall hazard maps, in general, can also be helpful as a part of decision making systems. The main objective of our work on rainfall extremes is to describe a method to obtain extreme precipitation hazard maps. We also developed a probabilistic model for downscaling monthly rainfall data into daily extremes. The probabilistic models used here are based on fitting GPD (Gumbel and pareto) to the monthly values of precipitation. The procedure was applied to precipitation data from 15 stations concentrated in northern part of the country. we use the extreme value theory, to describe the occurrence of extreme rainfalls in a region, which provides a complete analysis of the statistical distribution of extreme precipitation events, and allowing the construction of magnitude–frequency curves by fitting the distribution on rainfall data.

For the case when hydrological modeling is to be done for the estimation of FDC, a new regionalization procedure is discussed in chapter 3. Although the procedure is still classified as "regional", the underlying idea is very different from what is described in chapter 2. The dissimilarity between the flow regimes is function not only of magnitudinal comparison but also of lateral and vertical separation of peaks; but in case of FDCs, lateral and vertical dissimilarity functions can be ignored due to its functional nature. On the contrary to parametric representation of FDCs, this approach represents FDCs as a non-parametric entity. The dissimilarity between all the FDCs are executed and transformed into a distance matrix.

Within the vicinity of regional model, the workspace is divided into different clusters and a separate regional model is found for each cluster. The regional models of each cluster use the concept of dissimilarity to make estimates about hydrological parameters at an ungauged basin. The estimations of hydrological parameters for remotely located basins are improved by swapping models and bringing the remotely located basin into an area with better coverage of its neighbors around it (60^0 degree pruning).

Chapter 2 Monthly runoff regime regionalization through dissimilarity-based methods

2.1. Introduction

The topic of estimation of flow regimes in an ungauged basin has received extensive research efforts over the last two decades [*Blöschl et al.*, 2013]. The practical purposes for which prediction of flow regimes is important involve design and management of hydraulic structures, irrigation and hydropower systems, etc. In particular, the hydrological monthly flow regime is generally defined as the curve obtained with the 12 average monthly flows in a year. The shape and magnitude of flow regime curves depend on hydroclimatic processes and basin characteristics [*Bower & Hannah, 2002*] in a complex way. *Bower and Hannah* [2002] noticed that the basins associated with major aquifers within U.K. are characterized by more stable regimes and the variability in regime shape is a function of seasonal variability and amount of precipitation. They further stated that the double peaks are commonly observed in basins associated with large aquifers, whereas climatological extremes may result in single regime shape dominating across the entire area.

A number of methods can be cited from literature about flow regime estimation at ungauged sites [e.g. *Hrachowitz et al., 2013; Parajka et. al., 2013; Shoaib et al., 2013*]. These methods can be theoretically divided according to *Parajka et al.*, [2013] into: 1) Process-based methods [e.g., *Carrillo*, 2011] and 2) statistical methods [e.g., *Gallart et al., 2008; Samaniego et al., 2010; Girolamo, 2011; Renner and Bernhofer, 2011; Archfield et al., 2013*]. The former are fundamentally based on established physical laws which can capture the underlying dynamics of the watershed. However, they are not suitable to the case of ungauged basins, which is the main goal of the present approach, because they generally require the calibration of the parameters of the model. A number of statistical methods are also available in literature for the prediction of hydrological data at an ungauged basin. *Olden and Poff* [2003] provided a statistical frame work, called index method, for the characterization of hydrologic regimes by focusing on the inter-relationships among the hydrologic indices. Similarly a number of methods have been worked out for extrapolating flow regimes from gauged basins to ungauged ones using geostatistical and proximity methods using basin descriptors as predictor variables (e.g., Sauquet et al., 2000, 2008). *Laaha and Bloschl* [2006] compared a number of clustering methods for the calculation of low magnitude of flow from short stream flow records. They found that the method of clustering based on seasonality regions is the best one though all methods tend to underestimate in very wet catchments. The basic idea underlying statistical methods is to bring hydrological information from gauged basins to an ungauged basin using some basin characteristics, known as descriptors, as proxy of the hydrological information. This process refers to as "regionalization".

Classic regionalization approaches work either on each single monthly value or on a smaller set of representative parameters *[Krasovskaia et al., 1994]*. In case of classic regionalization approach, one regional model is to be defined for each month with an advantage of doing nothing on the hydrological data. Whereas, the parametric has an advantage of requiring fewer models (i.e. one for each parameter) but a fitting procedure makes it complicated. Moreover, a distance-based method is also worked out which requires only one regional model, defined by a suitable dissimilarity measure, with no fitting requirement. Alternative methods are non-parameteric and tend to consider the estimation of the entire curve (*see Ganora et al., 2009*) as a whole unique variable.

Another relevant application of the dissimilarity framework is reported by *Samaniego et. al.,* [2010] which incorporates copulas to find dissimilarity measures on daily streamflow time series by using three (dis)similarity measure. One of the similarity-measure considers the symmetry of the empirical copula density while the other two merge the degree of symmetry with correlation coefficient between time series for a pair of two catchments. By using a local variance reducing technique a transformation matrix was defined to relate m-dimensional space into k-dimensional transformed space measured with coordinate of meoscale hydrological model.

Ganora et al., [2009] used regression method to predict flow duration curves by linking descriptors data with hydrological data. To our knowledge no such technique has ever been tried for flow regimes. The dissimilarity-based method (sometimes also referred to as distance-based method) proposed by Ganora et al. (2009), considers the dissimilarity between the hydrological features of two basins measured by using a predefined metric in the hydrologic space. The application of the dissimilarity measure to all the possible combinations of basins, ultimately generates a distance matrix. The distance matrix of hydrological regimes can then be related to analogous distance matrices computed between basin characteristics for any couple of basins, with the final aim of using close basins in the space of characteristics to predict the hydrological behavior at an ungauged catchment. This procedure is delineated in the following sections 1 and 2. To our knowledge there is no other specific distance measurement technique available in the literature for the estimation of flow regimes (non monotonic functions) at ungauged basins.

Regardless of any research done in this regard, the magnitude and timing of occurrence of flow regime peaks is never discussed explicitly.

2.2. Dissimilarity between regimes

The dissimilarity-based method we propose starts from the comparison of the flow regimes of a pair of stations. For any two flow regimes belonging to the two gauged basin S and R, constituted by 12 elements each, $\{q_{1,S}, q_{2,S}, \dots q_{12,S}\}$ and $\{q_{2,R}, q_{2,R}, \dots q_{12,R}\}$, (i.e. the mean flows of each month) a dissimilarity measure can be defined in different ways. For instance, a function of point to point (magnitudinal) distance between monthly value can be used. A more complex definition of distance, accounting for the number and position of local maxima (peaks) and their position can be considered for flow regimes.

The magnitudinal dissimilarity used by *Ganora el. al.,* [2009] reads

$$\boldsymbol{D_{PtP} = \sum_{i=1}^{12} \left| q_{i,S} - q_{i,R} \right|,} \qquad \textbf{(2.1)}$$

where q_i is the monthly mean of the aforementioned stations S or R, D_{PtP} is the point to point difference and i is the index related to the monthly value.

Although eq (2.1) can be applied to flow regimes, it does not account for the possible shifting of peak positioning which is an important feature of flow regimes. We thus propose to add to the point-to-point difference D_{PtP} a "lateral distance measure" (L_{sp}), which considers the time difference between the occurence of peaks in the two regimes and a "vertical distance measure" (V_{sp}), which is the quantitative difference between these peaks. The two measures are then combined in a unique metric to account for all the main features of the regime, i.e. the total distance between two curves is the combination of these three modules (D_{PtP}, L_{sp}, V_{sp}):

$$D_T = D_{PtP} + L_{sp} + V_{sp}. \qquad (2.2)$$

The Point to point difference, lateral and vertical separations are sketched in fig. 2.1

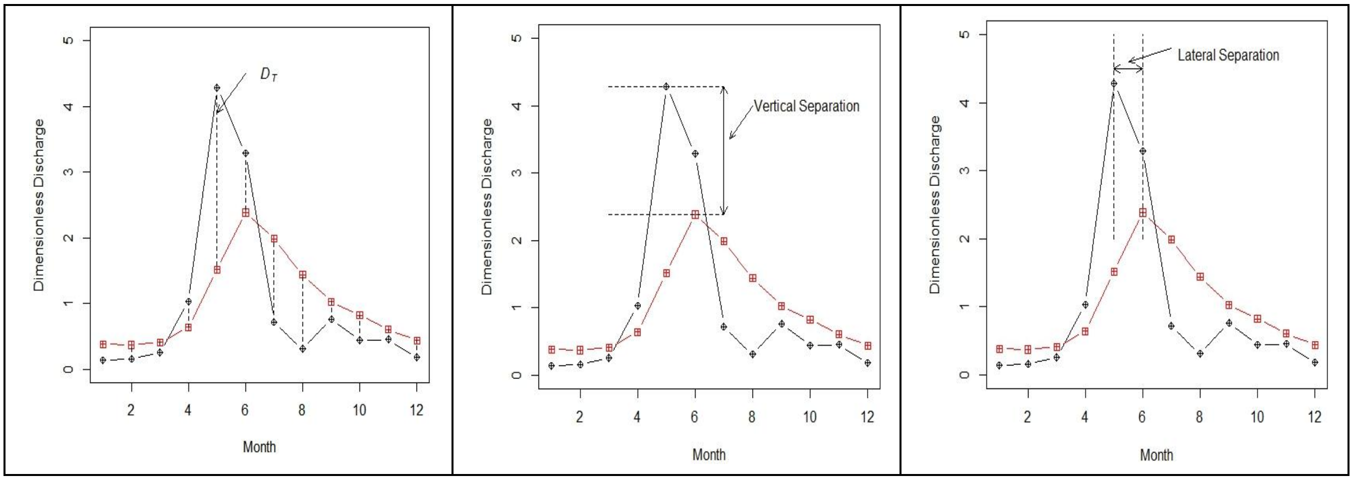


Figure 2.1. Distance between flow regimes in the month of May a) point-to-point distance, b) Vertical seperation of peaks and c) Lateral seperation of peaks.

The vertical distance is estimated as

$$\boldsymbol{V_{sp} = \left|q_{max,S} - q_{max,R}\right|}\ , \qquad (\mathbf{2.3})$$

where q_{max} is the magnitude of the highest peak discharge at stations S or R.

V_{sp} gives more importance to peak. If the compared peaks occur in the same month both

$V_{sp}, L_{sp} = 0$ ∀ regimes (*S*,*R*), since D_{PtP} takes into account the effect of both these dissimilarities. For estimating the lateral separation, we first need to define the number of peaks in flow regimes. Studying the following we will consider all the values greater or equal to $0.80 \cdot q_{max}$ as peaks. The lateral separation is a circular variable. Therefore, once the peaks have been defined we need to shift the regimes towards each other over the shortest possible span. For each time step of the movement of peaks, we calculate the change in PtP difference. The process of moving the peak stops, when the moving peak overshadows the peak of reference station. The peak being shifted is referred to be in shifted state and the state of stationary peak is termed as actual.

As an example, we compute the lateral separation measure L_{sp} considering station # 9 with a peak discharge (S_0) in April and station # 79 having a maximum discharge (R_0) in July as described in figure 2.2. The actual state (μ) of flow regimes at these respective stations is depicted in solid lines. Being μ the actual state, and σ the shifted configuration, the lateral separation reads:

$$L_{sp} = \sum_i \left| D_{PtP,\mu} - D_{PtP,\sigma_i} \right|, \qquad (4)$$

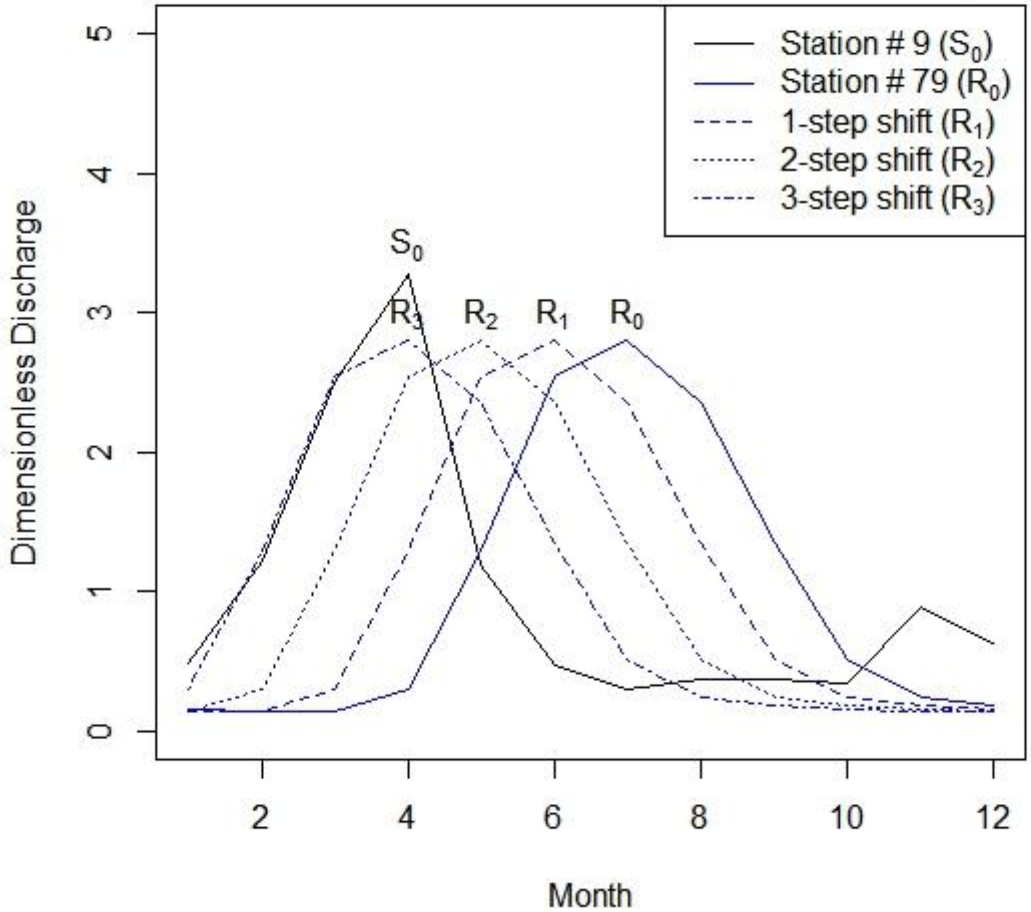


Figure 2.2. Compared peaks in actual state ($\mathrm{S}_0, \mathrm{R}_0$) and moving R_0as $\mathrm{R}_{1,}\mathrm{R}_2, \mathrm{R}_3$ towards S_0.

with i the index of the shifted stated. By definition, we have to move any peak (S_0 or R_0) towards the other over shortest possible span of time. Therefore, we move through these months backward (July→ June→ May→ April) instead of moving forward (July→ August→ Setember→

October→ November→ December→ January→ Feburary→ March→ April). The process of moving peaks towards each other stops once they are exactly underneath (S_0 or R_3) (see figure 2.2). L_{sp} is then equal to the sum of point-to-point distance computed for the actual configuration ($\mu_0 = D_{PtP}[S_0, R_0]$) and the shifted configuration comprised of 1st-step shift state ($\mu_1 = D_{PtP}[S_0, R_1]$) and 2nd-step shift state ($\mu_2 = D_{PtP}[S_0, R_2]$). For general case, the shift state can be defined as $\mu_i = D_{PtP}[S_0, R_i]$ with R_i as i^{th}-shift state of regime R.

The total Lateral Separation L_{sp} for the exemplified and general case is then defined as:

$$L_{sp} = \sum_i |\mu_{i-1} - \mu_i|, \tag{2.5}$$

where i is the difference in peak location in shortest possible path. To understand the difference between using the simple D_{PtP} distance and the comprehensive distance D_T of equation (2.2), we compare the two definitions of distances in figure (2.3) based on a set of 118 stations records used in our work. Their quantitative comparison is done in Table 2.1, where regimes from four regions (A,B,C and D) are put in evidence

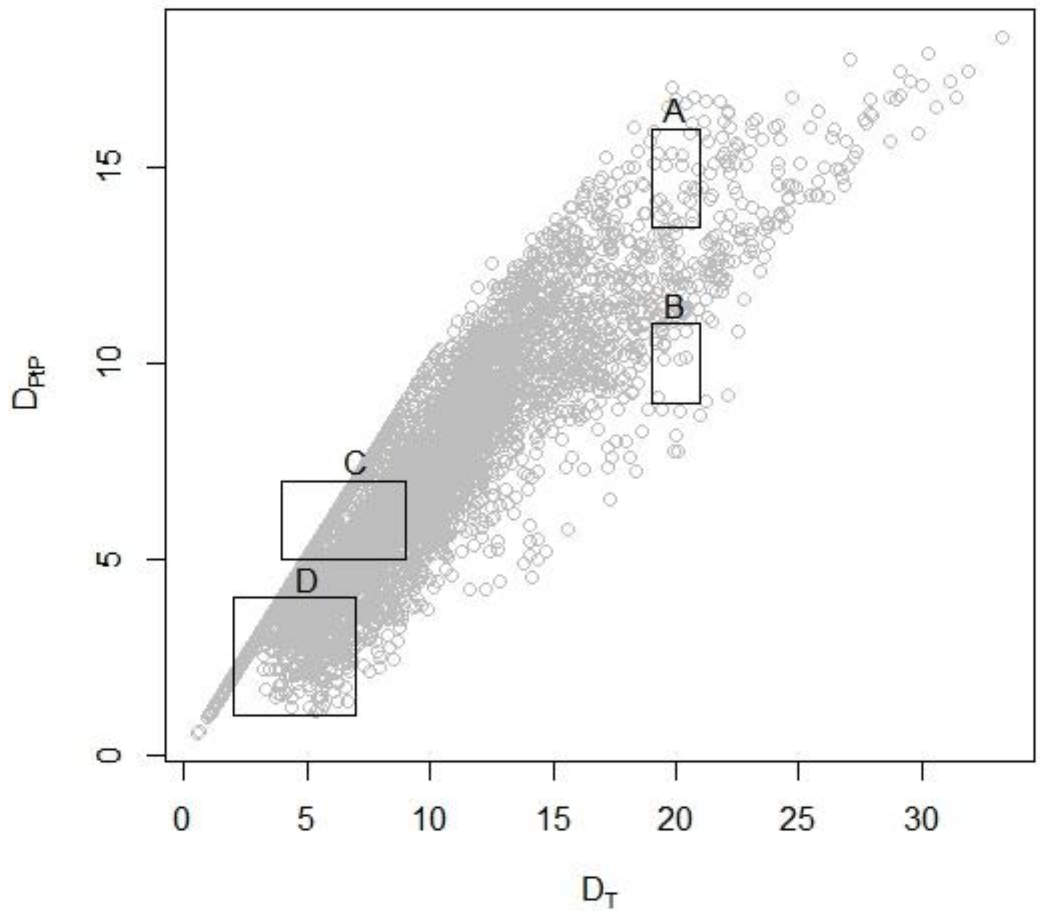


Figure 2.3. Comparison between Magnitudinal distance method and Newly developed method.

The regimes in those four regions are highlighted in the figure (2.4), whereas the points on the bisector line in figure 2.3 are representative of peaks of compared stations that are occuring in the same month (hence no lateral and vertical separation were to be considered). Let us compare a set of regimes in blocks A and B of the figure 2.3, to understand the difference between D_T and D_{PtP} (or PtP). In figure 2.4, the regimes have been actually drawn to further eleborate the difference. We will not only take into account, the trend (occurence of flow magnitude w.r.t time) of the regimes but also the time of occurence of peaks. In figure 2.4, the regimes in block B are similar to those in A (station # 84 ~ station # 89 and station # 15 ~ station # 27); the reason being small time-scale difference between the occurence of peaks and almost similar trends of regimes being compared in both blocks. By the definition of dissimilarity, the distance of both these blocks should somehow be similar. On the contrary, D_{PtP} distance changes dramatically from A to B but D_T remains consistent. Similarly, in block C(a), the regimes are more alike in trend and peak-occurence than those in block C(b) but the dissimlarity measures are otherwise for D_{PtP} while replicates the similar behavior for D_T. A more simpler case is described in block D, where besides being more similar in D(b) than in D(a), D_{PtP} counts larger difference between regimes in former and less in latter case. Whereas, D_T reproduces seemingly more meaninful translation of the results as shown in Table 2.1.

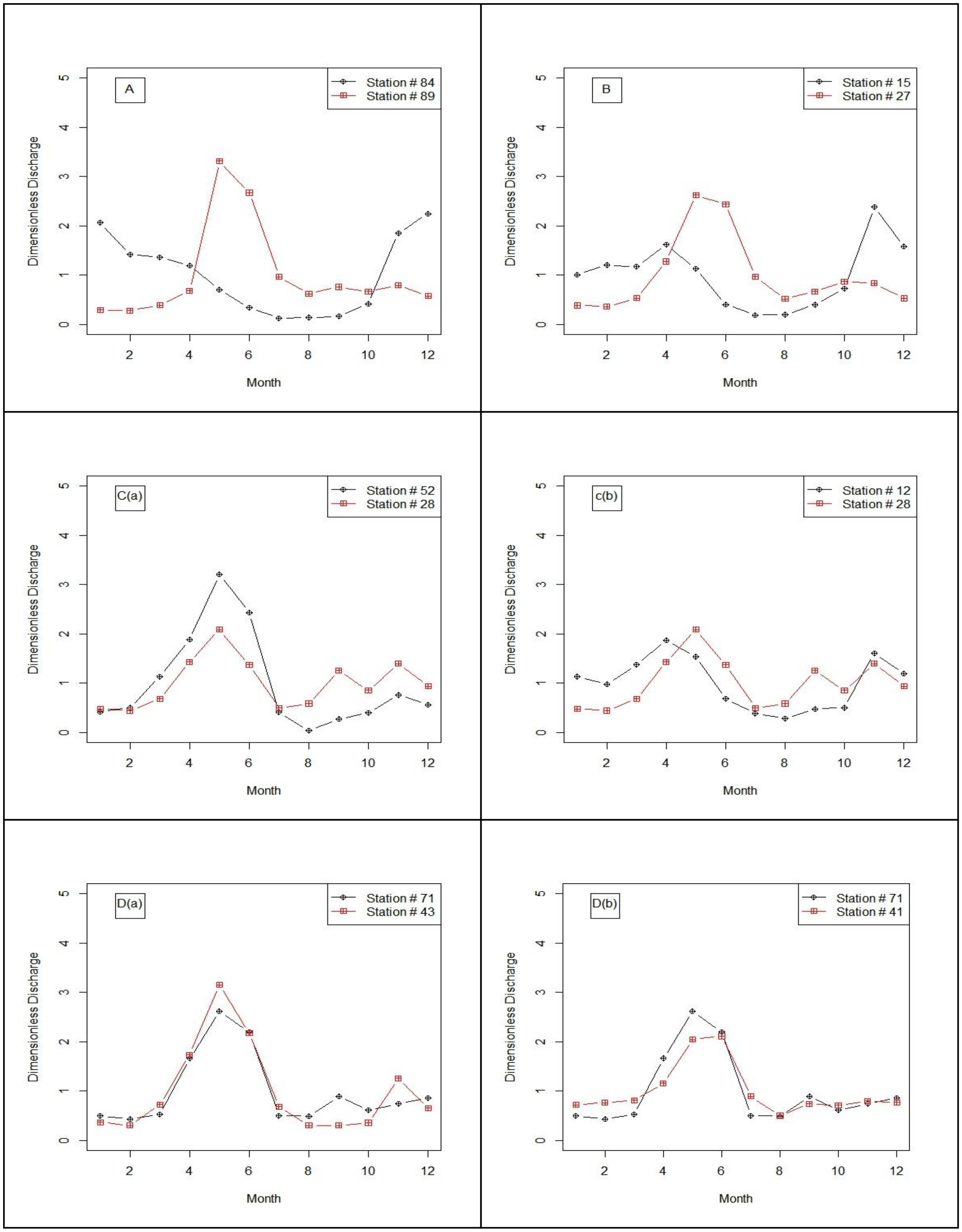


Figure 2.4. Sensitivity check for dissimilarity methods at various stations.

Table 2.1. Qaulitative comparison of Absolute distance method and New method

Stations	Region	D_{PtP}	D_T
84, 89	A	20.37	14.23
15, 27	B	20.20	10.09
52, 28	C (a)	6.28	6.28
7, 28	C (b)	7.56	5.58
71, 43	D (a)	3.01	3.01
71, 41	D (b)	6.39	2.80

2.3 Regional Model

The time series dataset of 118 stations in Northwestern Italy was considered for the application with variable length from a minimum of 5 to a maximum of 52 years, with a mean value of 12 years; the runoff data was extracted from the publications of the former Italian Hydrographic Service extended with the more recent measurements provided by the Regional Environmental Agency (ARPA) of the Piemonte Region. Basic measurements are at the daily scale and have been aggregated at the monthly scale for the purpose of this study. For effective application of the model, the data was made dimensionless by normalizing with the average monthly value for that site.

A number of geomorphological variables, referred to as descriptors, relative to the considered basins are extracted from the database developed by [*Ganora et al., 2013*]for the region of interest and based on the former CUBIST database [*CUBIST Team, 2007]* which contains data for more than 500 basins all over Italy. The catchment area of the considered basins ranges between 22 and 7983 km^2, and their average elevation ranges from 494 to 2694 m a.s.l. Geomorphological characteristics of each basin were obtained from the NASA SRTM [Farr et al., 2007] digital terrain model (pre-processed to a 100 m cell grid) using automatic GIS procedures under the GRASS GIS environment. Climatic, vegetation and land use descriptors were obtained by properly clipping thematic maps available for the area of interest.

The implementation of the regional procedure is based on the idea that similar hydrological behavior is related to basin similarity in a subset of descriptors. Similar basins are then usually pooled together by proximity in the descriptors space [*Samaniego et. al.,* 2010] and the average of hydrological properties, e.g. flow regime is taken as valid for the whole group.

In the context of the definition of the proposed regional procedure, Section 1 provides different ways to compute dissimilarities between streamflow regimes, with particular attention to the location of peak discharge of regime. An analogous procedure should be applied to compute dissimilarities between descriptors of two basins in order to implement the regional procedure. In fact, based on the dissimilarity of descriptors, one is expected to find low dissimilarity values for the basins with “similar” hydrological properties (Utopically). The way to compute the descriptors dissimilarity changes depending on the type of descriptors.

The simplest descriptors are basin elevation, basin area etc. and the dissimilarity can be computed simply as the absolute difference of the values. When the descriptor is represented by a monotonic function (as the hypsographic curve) the dissimilarity can be computed as the point-to-point distance as in equation (2.1). For more complex descriptors (in this case the rainfall regimes) the D_T dissimilarity is appropriate.

After the definition of descriptors dissimilarity matrices it is necessary to relate hydrological data to basin characteristics. This step is fundamental as only a small subset of descriptors is expected to be useful to represent the hydrological variability. As there is no prior information about this subset, it is defined through a statistical procedure which seeks for the descriptor distance matrices more correlated with the distance matrix of the hydrological regime.

The correlation between distance matrices is investigated through the Mantel test [*Mantel and Valand,* 1970]. In its simple version, it is used to evaluate the significance of the linear correlation between two distance matrices. This test is performed by computing a statistic (usually the Pearson correlation coefficient) between all the pair wise elements of the two matrices. Its significance is tested by repeatedly permuting the objects in one of the matrices, and recomputing the correlation coefficient each time; Permutations are performed simultaneously exchanging two rows (randomly) and their corresponding columns of the matrices (see *Legendre et al.,* 1994). The significance of the statistic is assessed by comparing its original value to the distribution of values obtained from the permutations, which are considered as many realizations of the null hypothesis of no correlation.

The relation between the discharge distance matrix, defined as $\mathcal{M}_H$, and various combinations of the distance matrices of descriptors ($\mathcal{M}_D$) is in general more interesting than the relationship with one single descriptor. To evaluate this kind of multiple relationship, a linear multiregressive approach has been adopted. We started considering a simple linear model,

$$\boldsymbol{\mathcal{M}_H = \beta_0 + \beta_1(\mathcal{M}_D)_1 + \cdots \beta_n(\mathcal{M}_D)_p + \varepsilon,} \tag{2.6}$$

with p as number of descriptors selected among the whole set of available characteristics, β_i as the generic regression coefficient and ε is the residual element of matrices, “unpacked” to vectors as described by *Lichstein* [2007]. The simple Mantel test can be extended to multiple linear regression models as described by equation (2.6) with the aid of an extension introduced by *Smouse et al.* [1986] and later on deliberated and improved by *Legendre et al.* [1994] and recently practiced by *Lichstein* [2007] in the ecological field. As described by *Lichstein* [2007], the redundant values of each distance matrix are eliminated and matrix is transformed into a vector of distance and regression is performed in a classical way. Then, the elements in a distance matrix of descriptors are permuted to construct a null distribution. The rows and the columns of the matrix $DiDM$ are permuted simultaneously and each regression coefficient is tested individually, similarly to what described for the simple Mantel test.

Several combinations of models were investigated using linear regression. They were built using different combination of (1) regimes distances, considering the three representations described before (point to point, lateral and vertical). As per the descriptors distance matrices $\mathcal{M}_D$, all possible combination from one to three descriptor matrices have been taken into account. The regressions were first tested for significance with the multiple Mantel test, with a significance level of 0.05. Models passing the Mantel test were then ranked according to the adjusted coefficient of determination defined as (e.g., *Kottegoda and Rosso*, 1997):

$$\boldsymbol{R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}}, \tag{2.7}$$

In the above equation (2.7), p stands for the number of descriptors, n is the total number of basins and R^2 defines the standard coefficient of determination, which alongside regression coefficients was computed in a standard way, defined by *Legendre et al.*, [1994]. As the distances inside a distance matrix are not mutually independent, it is advisable to use all the n values instead of classical $n(n-1)/_2$ values. Furthermore, a test against multicollinearity has been performed in order to exclude variables with redundant information in the descriptors.

The R^2_{adj} values observed with distance matrices of regression models are very low(always jaunted between 0.20 and 0.55), although the results are significant, statistically. Which is to say that regressions are only used to select dominating descriptors and not for any direct estimation.

This statistic is used to rank the models, but cannot be used to quantify the variance explained by the linear model as in classic regresions due to the mutual correlation of the values in the distance matrices. Besides R_{adj}^2 it is of great importance to investigate also the behavior of the residuals along the regression line and its development with time [*HP Training Module, 2002*], which is very difficult to interpret.

To check the quality of model output, we need to device a cross validation procedure. Generally, one station, in the entire dataset, is considered ungauged and its data (hydrological and descriptors) are removed from the database. Afterwards, the models are recalibrated and the unknown flow regime is estimated.

We used predictive Leave-one-out cross validation approach, to check the validity of regression models, for its convenience and fast computation. The full scale model validation is often extremely time consuming and sometimes computationally impossible due to large size of dataset and the complexity of model due to increasing number of descriptors. In our work, to reduce the computational burden, the regression models having good R^2 values, filtered through mantel test and VIF test are used to execute the regional regimes and that executed regime is then compared with the empirical regime in D_T space. The D_T space is defined as $\zeta_{D_{T\,Empirical}\quad ,Estimated}$ with ζ as an error magnitude for a single station. The model producing least overall error (Δ), between actual and regional regimes was selected, defined as $\Delta = \sum \frac{\zeta}{n}$ where n is the number of stations.

The proposed methodology of distance-based measurement was carried out in the R statistical environment [*R development core team, 2007*], desegregated for Mantel test and Multivariate Regression Analysis in nsRFA package [*viglione, 2007*].

Once the distance-based model is estimated, we find the distance matrices of descriptors in the selected model according to the type of the descriptors (Scalar or monotonic). After normalizing them by average distance and then summing them up to find the single representative distance matrix for finding the nearest neighbors of ungauged basin; considering the minimum value of the distance relative to the stations from the distance matrix of descriptors. The beauty of this technique lies in the ease with which a non monotonic function (complex descriptor) like rainfall was introduced with a scalar descriptor to define an appropriate space for the neighbor selection.

Another important step is to determine the optimum number of neighbors of an ungauged basin. Since too few neighbors resulted in over simplication of the results and in some cases even counter intruitive; whereas, too many neighbors may cause considerable error in the final results. In the present work we used cross-validation procedure to set the number of neighbors and after scrutinizing from 1 to 9, we finally found reasonable results with 5 neighbors.

The best models obtained by one, two and three descriptors were only considered. The model selection results is the combinations of descriptors which generated the lower values of Δ and reasonable values of R_{adj}^{2} ., are enlisted in Table 2.2.

Table 2.2. Models with 1, 2 and 3 descriptors enlisted in the order of R^2_{adj}

Model	Descriptors	overall error (Δ)	R^2_{adj}
1	Annual NDVI	3.539	0.484
1	Hypsographic Curve	3.862	0.424
1	Mean Basin Elevation	4.067	0.374
1	Max Basin Elevation	3.884	0.216
1	Rainfall Regime	4.044	0.014
2	Fourier Coefficient, Annual NDVI	3.149	0.517
2	Annual NDVI, Rainfall Regime	2.720	0.494
2	Hypsographic Curve, Rainfall Regime	3.018	0.437
2	Mean Basin Elevation , Rainfall Regime	2.940	0.391
2	Land use Index (Non-vegetated area), Rainfall Regime	2.960	0.314
3	Precipitation Intensity Coefficient, Annual NDVI, Rainfall Regime	2.759	0.531
3	Land use Index (Non-vegetated area), Annual NDVI, Rainfall Regime	2.798	0.515
3	Land use Index (Wetlands), Annual NDVI, Rainfall Regime	2.658	0.500
3	Basin Area, Annual NDVI, Rainfall Regime	2.759	0.495
3	Rainfall intensity Duration Curve, Annual NDVI, Rainfall Regime	2.736	0.494

Table 2.2 shows the five best models for each combination with one, two and three descriptors, where all the models have been tested for significance of regression coefficients with the Mantel test with a level of significance of 0.05. It appears that, considering together the three representations of combination of descriptors, the most significant descriptors are the rainfall regime and hypsographic curve.

The adoption of these two descriptors is coherent with the typology of investigated basins. In fact, since we are considering mainly mountain basins, the annual NDVI descriptor is expected to be relevant because of its strong relation to snow accumulation and snowmelt mechanisms.

Similarly, the rainfall regime provides a synthetic description of flow pattern. The ranges of some dominating descriptors are enlisted in Table 2.3.

Table 2.3. Range of variation of descriptors used by the distance-based model.

Descriptors	Maximum	Mean	Minimum
Land use Index (Wetlands)	7.890	0.190	0
Rainfall intensity Duration Curve	37.88	23.40	11.88
Basin Area	25640	1330.11	22
Maximum Basin Elevation	4743	2750	368
Rainfall Regime	Regime	Regime	Regime
Mean Basin Elevation	2682	1323.17	244
Hypsographic Curve	Curve	Curve	Curve
Y-Coordinate	5129050	4977667	4886350
Land use Index (Non-vegetated area)	78.68	16.03	0
Annual NDVI	0.644	0.447	0.082
Fourier Coefficient	49.563	-8.161	-56.554

The methodology can be summed up in the following steps

1) – Calculate the monthly mean discharge at each station.

2)-Identify the variable needed to calculate dissimilarities.

3)-Execute dissimilarities between stations by using specified techniques (point-to-point, lateral and vertical).

4)-Select best descriptor models by observing least Δ values and Multivariate regression analysis.

5)-On the Descriptors space find the nearest neighbors of missing data station and by using those NN execute a regime for that station.

2.4. Alternative Regional Models

2.4.1. Parametric representation of the regime

The dissimilarity-based approach was compared with a more traditional regional model based on the parametric representation of the regime curve, which were calibrated on the same set of basins. In contrast to the dissimilarity-based approach which aims at considering the regime as a whole element, here the shape of monthly averaged hydrological regimes is represented by using a certain of number of parameters. This parameterization is based on the fourier harmonic, and its form reads:

$$f(t) = A_0 + A_1 cos\left(\frac{2\pi t}{\tau} + \varphi_1\right) + A_2 cos\left(\frac{4\pi t}{\tau} + \varphi_2\right), \tag{2.8}$$

where the harmonics represent the 1-year-scale and the 6-months-scale fluctuations of the hydrologic regime. This analytical model to represent the regime has 5 parameters, among which A_0 can be neglected as the mean values is not considered in this work. Phase shifts φ_1 and φ_2 are circular variables so large values may be very close to small values, which on transformation can be sparse apart (e.g. $1^0 * \frac{\pi}{180}$ and $364^0 * \frac{\pi}{180}$). Therefore, in order to estimate them with a regional procedure, it is better to resort to a different representation

$$f(t) = A_0 + A_1 cos\left(\frac{2\pi}{\tau} t\right).cos(\varphi_1) - A_1 sin\left(\frac{2\pi}{\tau} t\right).sin(\varphi_1) + A_2 cos\left(\frac{4\pi}{\tau} t\right).cos(\varphi_2) - A_2 sin\left(\frac{4\pi}{\tau} t\right).sin(\varphi_2), \tag{2.9}$$

by separating the variables that don't depend on time t

$\theta_1 = A_1 cos(\varphi_1);$ $\theta_2 = A_2 cos(\varphi_2);$

$\theta_3 = -A_1 sin(\varphi_1);$ $\theta_4 = A_2 sin(\varphi_2);$

and those which depend on t

$X_1(t) = cos\left(\frac{2\pi}{\tau} t\right);$ $X_2(t) = cos\left(\frac{4\pi}{\tau} t\right);$

$Y_1(t) = sin\left(\frac{2\pi}{\tau}t\right);$ $Y_2(t) = sin\left(\frac{4\pi}{\tau}t\right);$

Equation (2.9) now reads (neglecting A_0):

$$f(t) = \theta_1.X_1(t) + \theta_2.Y_1(t) + \theta_3.X_2(t) + \theta_4.Y_2(t), \quad (2.10)$$

whose parameters can be easily fitted to a real dimensionless regime $f(t)$ made of 12 observations by the least squares method (see figure 2.5), where the vectors X_1, X_2, Y_1 and Y_2 are calculated using $t = 1,2,3 \dots \dots ,12$ and $\tau = 12$

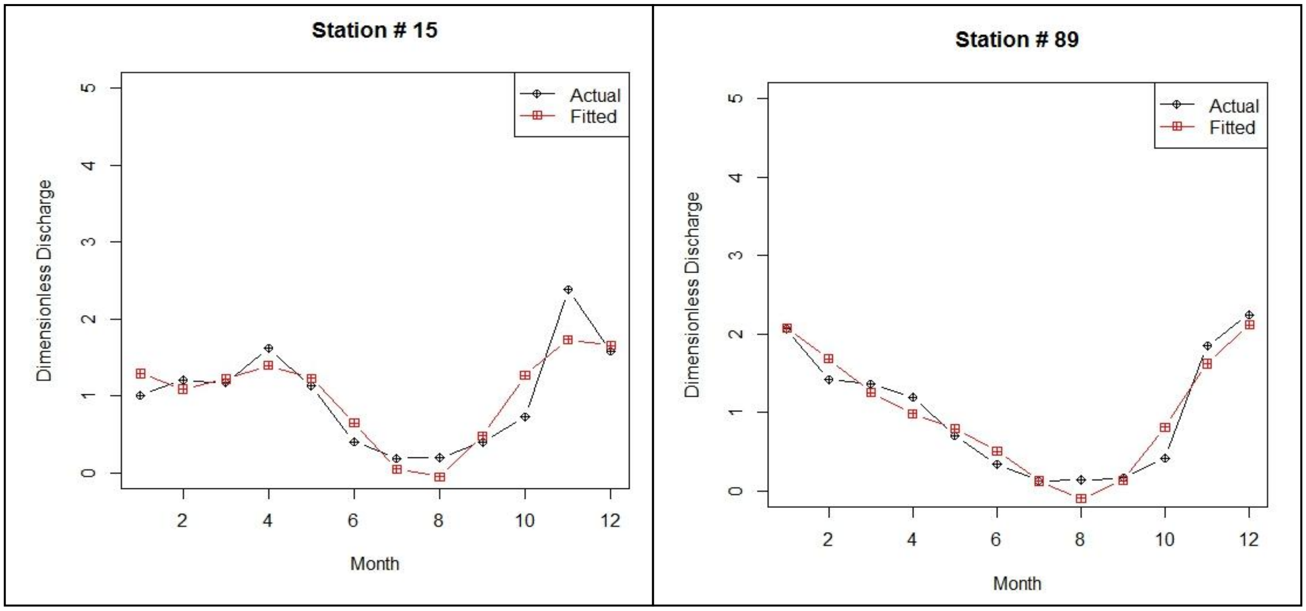


Figure 2.5. Fitted regimes over original regimes with parametric models.

After the fitting procedure of the θ parameters has been extended to all the 118 observed regimes, we proceeded to the regionalization phase. Each parameter θ_j is related to the catchments' descriptors d by a linear model of the form

$$\theta_j = a_0 + a_1.d_1 + a_2.d_2 + \cdots + a_n.d_n + \varepsilon, \quad (2.11)$$

where a_i are regression coefficients and $\boldsymbol{\varepsilon}$ is residual vector. The choice of a suitable regional model is an important step in the estimation of generic parameters at an ungauged basin. Many linear models of the form of equation (2.11) were considered and validated with a Student t test with a significance level of 0.05 followed by a multicollinearity (VIF>5) test and subsequently ordered by their values of R^2_{adj} [e.g., *Montgomery et al.*, 2001].

The leave-one-out validation scheme was used for evaluating the amplitudes and phases of the harmonics and reconstructing the regime. The predicted regime in an ungauged basin is evaluated by combining the basis (X_1, X_2, Y_1 and Y_2) to the estimated θ_j obtained by using the related descriptors. The best models for each θ are;

$$\theta_1 = 4.069 * 10^{-1} - 6.961 * 10^{-5}(Hypsographic\ Curve) + 8.795 * 10^{-4}(Average\ Aspect), \quad (2.12)$$

$$\theta_2 = 1.298 * 10^{1} - 1.073 * 10^{-2}(clc_4) + 2.528 * 10^{-6}(Basin\ Latitude), \quad (2.13)$$

$$\theta_3 = -1.025 + 2.779 * 10^{-5}(Hypsographic\ Curve) + 1.206 * 10^{-2}(cn_3), \quad (2.14)$$

$$\theta_4 = 3.5917 + 0.1473(cn_2) - 0.1684(cn_3), \quad (2.15)$$

where clc and cn are corine land cover and soil curve number respectively (for details see Ganora et al., 2013). The error measurement between predicted and actual regimes was obtained by comparing RMSE and NSE values.

2.4.2 Regionalization by geographical proximity

The dissimilarity-based approach was also tested against the geoghraphical distance norm which is used to measure the closeness (or dissimilarity) of basins in geoghraphical space. For the sake of simplicity, Euclidean norm was used to find the nearest neighbors of an ungauged basin. The efficiency of output was tested within a leave-one-out cross-validation scheme.

2.5. Results and Comparison

The three regional procedures presented in section 2 provide three different ways to estimate the dimensionless montly regime at ungauged sites. All the methods have been extensively applied to the 118 basin dataset of Italian catchments described above and are compared in the present section.

Among all the possible models ranked by the distance-based approach, the model containing two descriptors, namely annual NDVI and rainfall regime, was selected for its good global performance in cross validation. More descriptors can be used as well to obtain an enhanced estimator, however increasing the number of descriptor might make the model less robust. For the purposes of this work, the use of only two descriptors is shown to be effective, with performances overtaking those of other regional approaches based on two or more descriptors.

A proper metric to quantify the quality of fitting is not trivial to find, for the purpose of comparing the different models. Generally, the metrics are used to compare estimated and observed values (single value comparison); whereas, we need to compare a non monotonic function with a special emphasis on the peak discharge position. It's better to use different metrics to see the goodness of fit of each model by observing the fitting quality of models at each station and ultimately globally. We decided to use RMSE, which is one of the most commonly used error index statistics, and D_T since we are also interested in determining peak flow position.

On average the distance-based model (DBM) has smaller error (ζ) than parametric (PM) and geographical proximity (GM) as shown in table (2.5). Although performances quantified with the D_T metric are expected to favor the distance-based approach, due to peak-shift consideration, the distance-based approach prevails over other models even when RMSE was used for its evaluation.

The newly developed non parametric distance based approach executed, by far, good results compared to those of parametric and geographic proximity models as shown in figure (2.6). The

Table 2.4 illustrates a comparison of RSME and NSE values among parametric model, Geographic proximity and distance-based approach. It was observed that each parametric model was able to execute good results for a certain subset of basins, but not at all, when tested on whole of the dataset. The graphical representation of errors (ζ) obtained in different environments (D_T and RMSE) are shown in figure (2.7). The total magnitude of error over the entire sample and standard deviation of errors (δ) are enlisted in Table (2.5).

Table 2.4. RMSE, NSE and D_T obtained in the various basins using three types of methods

Basin	Area (Km^2)	New Method			Euclidean			Parametric Model		
Stations Codes		RMSE	NSE	D_T	RMSE	NSE	D_T	RMSE	NSE	D_T
3	41	0.265	0.918	2.752	0.528	0.675	5.665	0.439	0.775	4.503
6	262	0.382	0.713	4.408	0.483	0.542	8.559	0.462	0.582	5.254
14	127	0.238	0.861	2.301	0.389	0.630	4.144	0.394	0.619	4.692
21	75	0.353	0.820	4.087	0.461	0.692	5.501	0.678	0.334	8.433
28	152	0.145	0.912	1.642	0.388	0.369	7.017	0.373	0.417	6.945
37	106	0.317	0.719	3.246	0.483	0.350	6.323	0.429	0.485	6.864
45	212	0.216	0.900	2.590	0.243	0.873	5.613	0.778	-0.305	13.216
47	102	0.081	0.991	0.814	0.497	0.660	6.185	1.207	-1.008	17.114
48	160	0.251	0.854	4.430	0.445	0.541	5.620	0.770	-0.376	10.089
54	333	0.205	0.605	2.392	0.412	-0.595	4.370	0.448	-0.885	5.409
55	131	0.210	0.729	2.119	0.289	0.486	3.079	0.718	-2.182	9.254
63	838	0.211	0.925	6.855	0.329	0.818	7.378	0.854	-0.228	12.103
70	38	0.157	0.932	1.537	0.390	0.581	4.807	0.532	0.221	7.532

Table 2.5. Comparison of magnitudes of different errors (ζ) with corresponding Standard deviations (δ).

Model	ζ(RMSE,δ)	ζ(NSE,δ)	ζ(D_T,δ)
New Method	0.230(0.091)	0.812(0.293)	0.812(0.293)
Geographical Method	0.280(0.11)	0.735(0.346)	0.735(0.346)
Parametric Method	0.500(0.221)	0.273(0.595)	0.273(0.595)

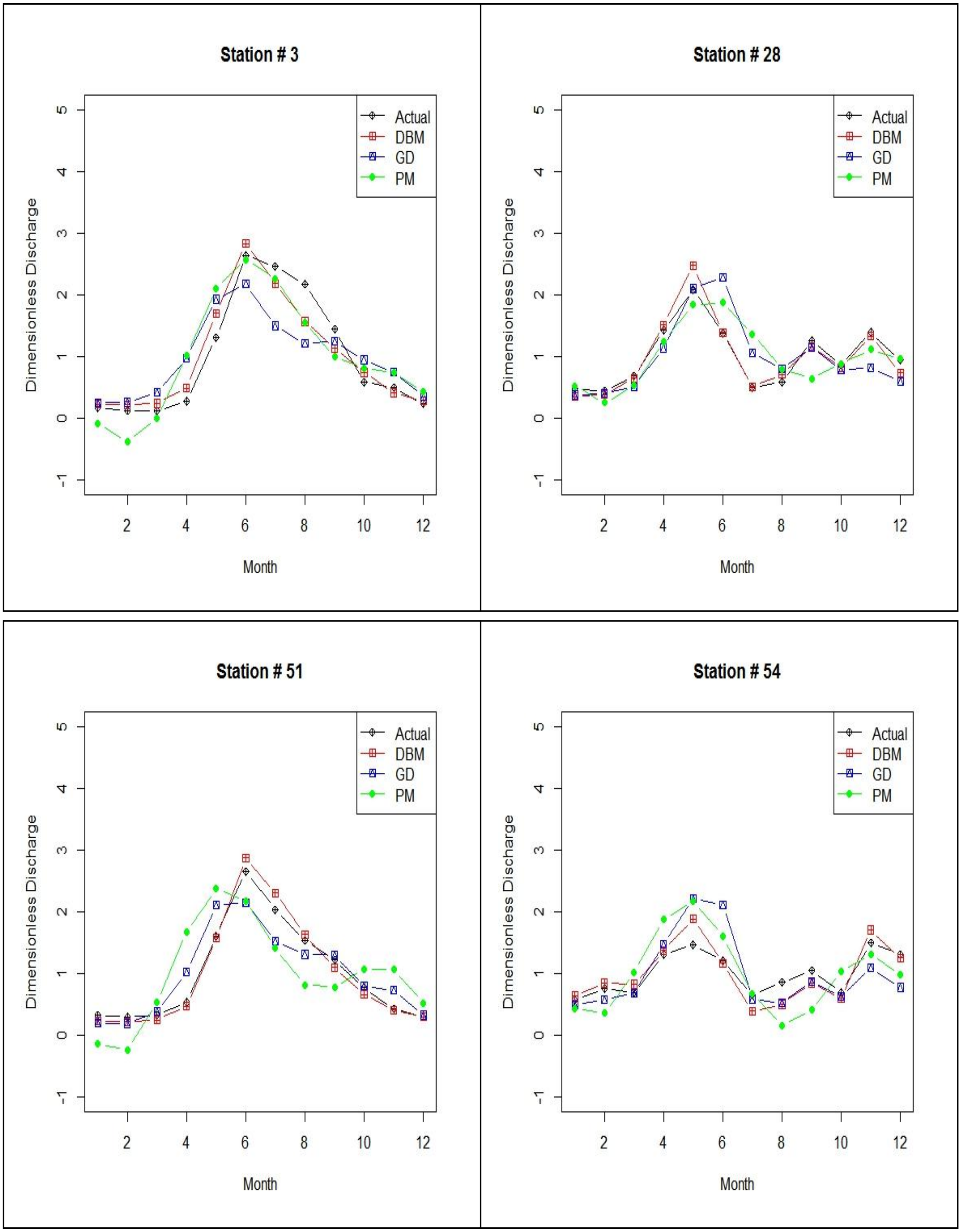
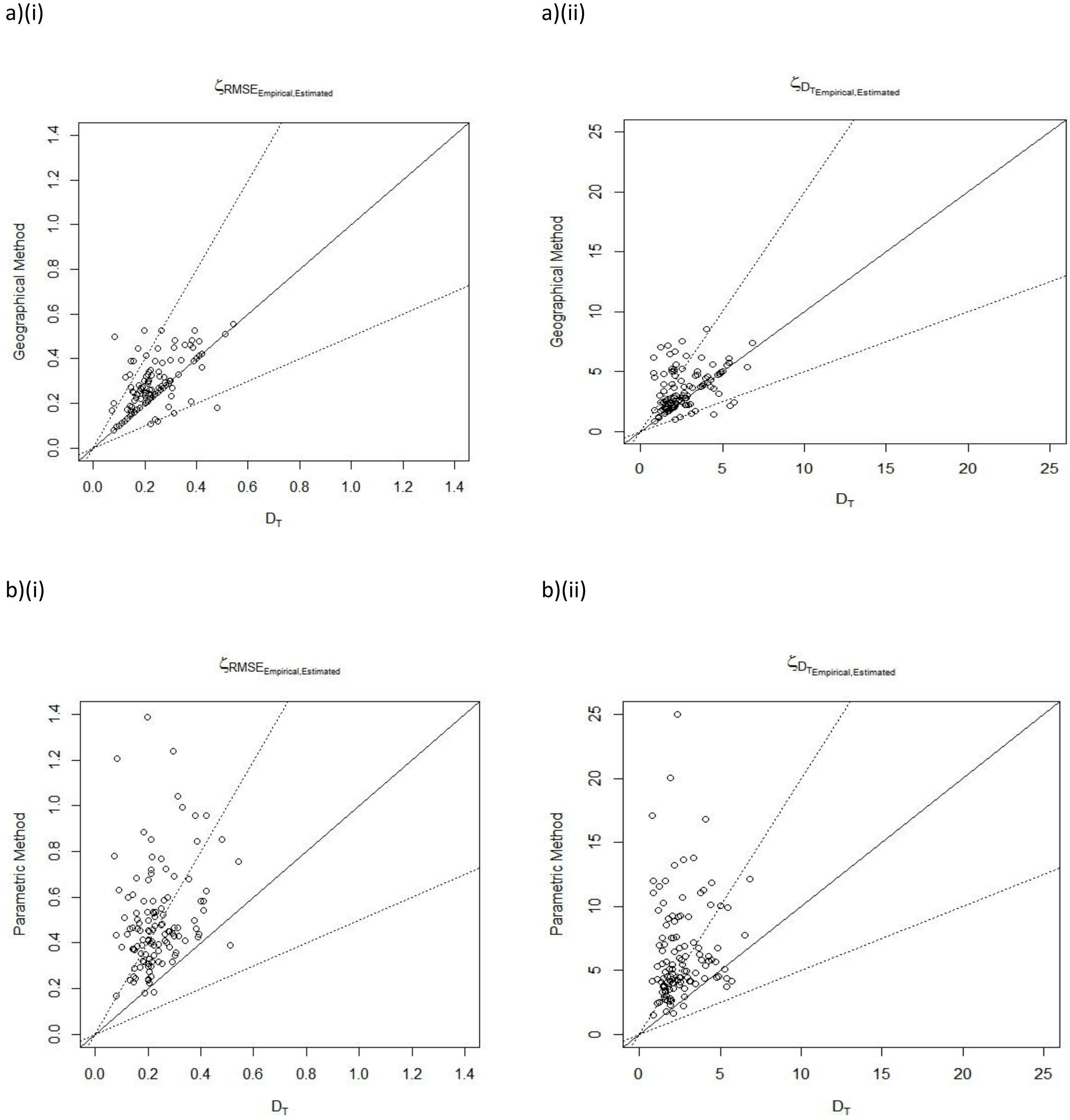

Station # 3
Station # 28
Station # 51
Station # 54
Actual
DBM
GD
PM
Dimensionless Discharge
Month
-1
0
1
2
3
4
5
2
4
6
8
10
12

Figure 2.6. Comparison between original and simulated regimes at selected stations.



a)(i)

a)(ii)

b)(i)

b)(ii)

Figure 2.7. Error comparison of distance based model in $RMSE$ and D_T enviroments compared with (a) the Geograhical method and (b) the Parametric method. The distance between the empirical regime and the estimated one, is reported in the scatterplot for each considered basin.

The solid line represents the ratio 1:1 between the errors, while dashed lines delimit the areas where errors for the distance-based model are twice the parametric ones and vice versa. Points above the solid line represent regimes better estimated by the distance-based method; points above the top dashed line represent regimes much better estimated by the distance-based method. From these results it can be concluded that the present method led to the most suitable results for flow regimes prediction in most basins with respect to RMSE and D_T. Though the new model performed generally well in all types of catchments, it presented some slight issues of magnitudinal differences between observed and simulated flow regimes for basins with extremely large (≥1000 km^2) or small areas ($< 100km^2$). The model predicted peaks of each regime correctly with slight variation in flat peaks but even in those cases the magnitude of discharge is very close to that of original peak discharge (fig. 2.8).

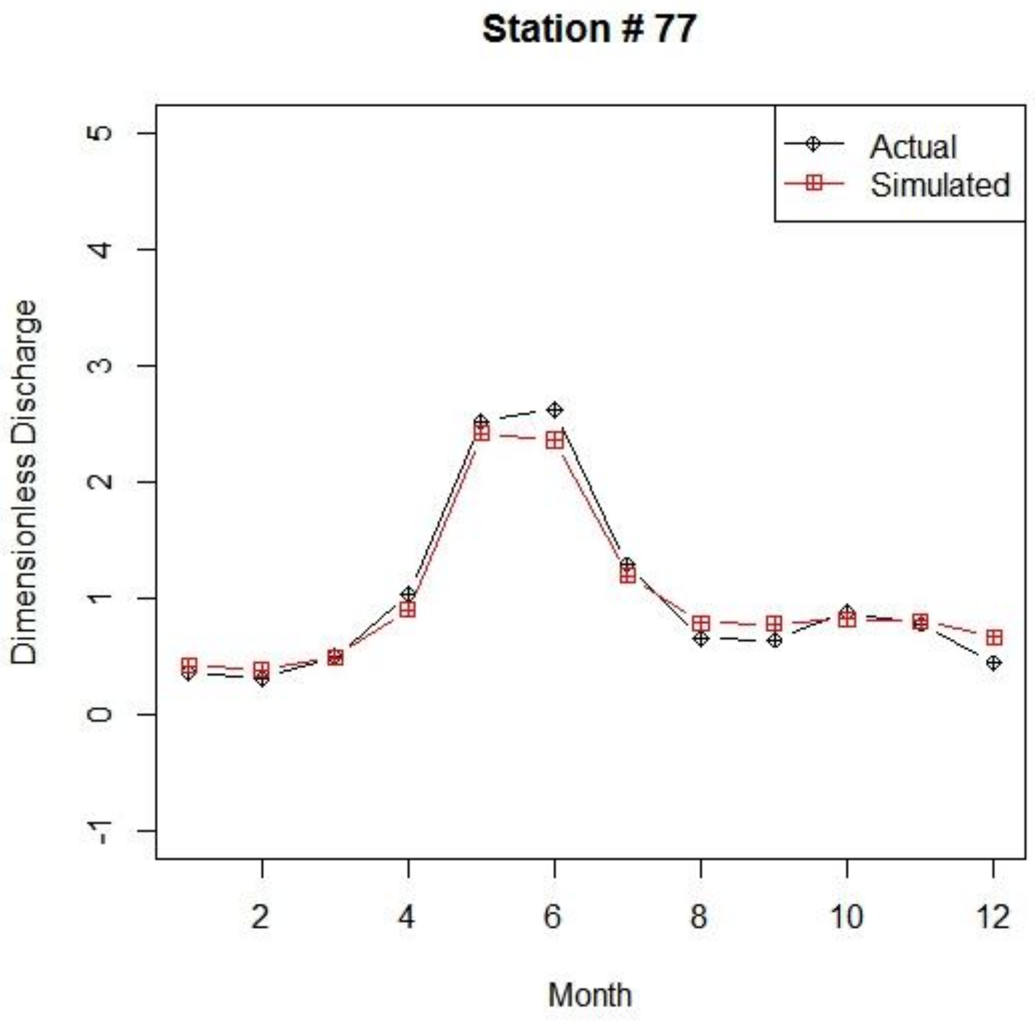


Figure 2.8. Estimation of regime in case of flat peak.

2.6. Conclusions

The dissimilarity technique between the flow regimes has been revisited in this paper. It has been shown that a good amount of information can be lost by considering, only, magnitude differences (e.g. the monthly-difference of streamflow data) between the flow regimes. While serveral authors contributed on the identification of the main parameters affecting the shapes of flow regimes, to our knowledge this is the first study which actually tries to integrate all those parameters into a dissimilarity measurement. This measure between regimes is used to account for both the magnitude and the position of the peaks, thus allowing one to quantitatively compare any couple of regimes. This concept is extended to the basin descriptors, so that a dissimilarity index between two sets of basin characteristics can be computed as well.

Information on both flow regime and basin descriptors have been combined to calibrate a regional model: the value of a vegetation index and the average rainfall regime of an ungauged basin are used to identify a set of gauged basins similar to the ungauged one. These are grouped together, and their streamflow records are used to predict the regime at the ungauged site.

The results made available by our distance-based model are comparable and are reasonably better than what we obtained by using other traditional approaches. Moreover, the ability of our model in prediction of complicated annual regimes can be achieved by using only two descriptors.

This approach demonstrates also that is possible to exploit the information of “complex” descriptors, in this case the average rainfall regime, without requiring any kind of parameterization and thus making the prediction procedure easily applicable.

Chapter 3. Regionalization of FDCs

3.1. Introduction

The flow data in river, particularly those of lower magnitude, are of great importance to meet the requirements of developmental projects for the management of water resources. The problem of estimating hydrological variables in ungauged basins located in difficult terrain has been the object of intense research activity in recent years. There are different methods which have been used to perform such estimation with the central idea of either extending or transferring the hydrological data from gauged to ungauged sites.

A flow duration curve is a cumulative-frequency curve which defines the relationship between magnitude of stream-flows of a certain time resolution (hourly, daily or monthly) and frequency of occurrence in any basin by translating the percentage of time for which a certain magnitude of flow equals or exceeds a certain flow value.

The most recent efforts in this field involved the construction of a distance-based regionalization model for the execution of flow duration curves (FDC) at sites with no or limited available data [*Razavi and Coulibaly*, 2013]. Classically, FDC at an ungauged basin are obtained by simple regression models. This allows establishing a link between flow quantiles or distribution parameters to the known characteristics of basins. A distance based technique has been introduced by *Ganora et al.*, [2009], which utilized a cluster analysis approach to group the similar basins by using non-parametric approach. In that method, the dissimilarities between FDCs were quantified as distances measured by comparing magnitudes of flows on FDCs occurring at the same time and then giving the computed dissimilarities among curves, a matrix form (i.e. distance matrix). The distance matrix was then co-related, by means of linear regression models to the distance matrices of each descriptor (mantel test). A strong co-relation value identified significant descriptor. Finally, cluster analysis was applied to group basins of similar characteristics; a suitable number of clusters were selected in order to provide adequately homogeneous (in statistical sense) pooling groups for which a single dimensionless flow duration curve was assumed as representation of the whole cluster. The study was conducted on 95 basins of Switzerland and northwestern Italy.

The distance between FDCs of any two stations was calculated by the use of following simple equation

$$D_a = \sum_{i=1}^{12} \left| Q_{i,s} - Q_{i,r} \right|; \qquad (3.1)$$

where the value D_a is total magnitude of dissimilarity between FDCs of stations "s" and "r" having flow magnitude of $Q_{i,s}$ and $Q_{i,r}$ respectively. The practical implementation of eq (3.1) is also exemplified in Figure 3.1.

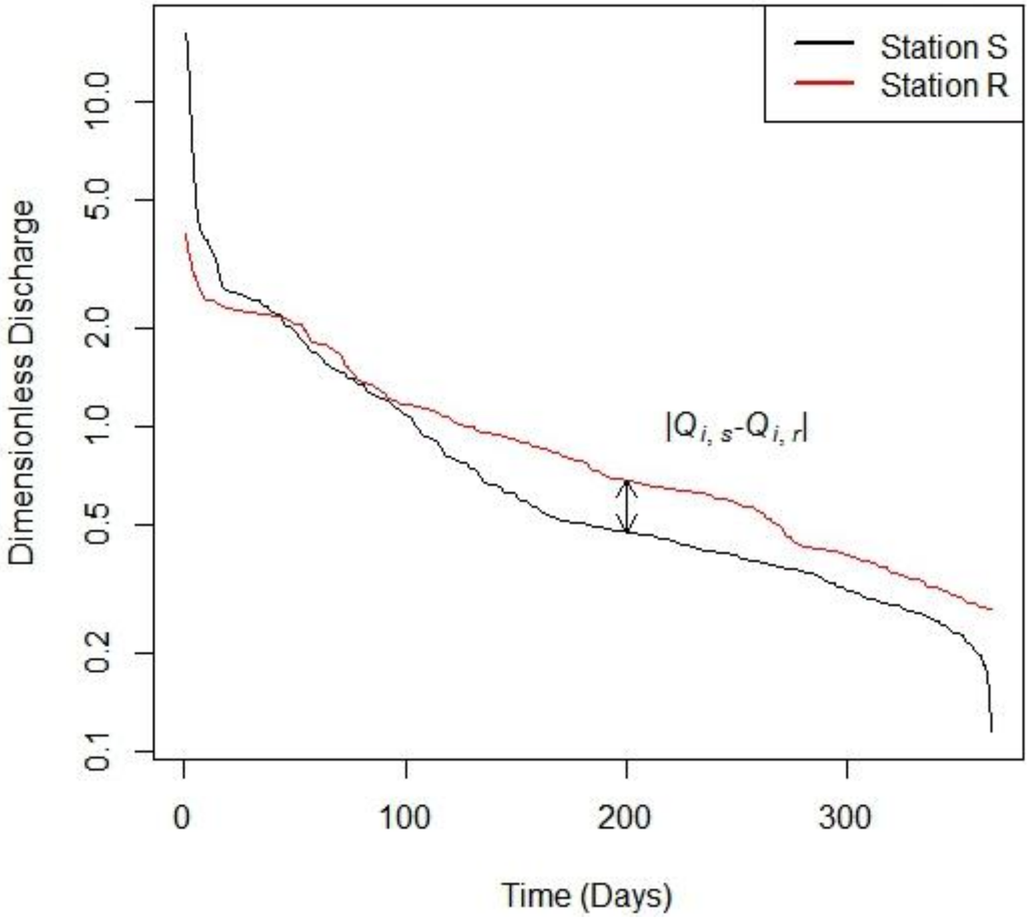


Figure 3.1. Distance calculation between two FDCs by using equation (1).

There were certain combinations of descriptors (2 to 5) which were tested against distance matrix of FDCs. The selected model was used for regionalization. The complete information regarding regionalization procedure is presented in the form of flow chart in Fig. 3.2.

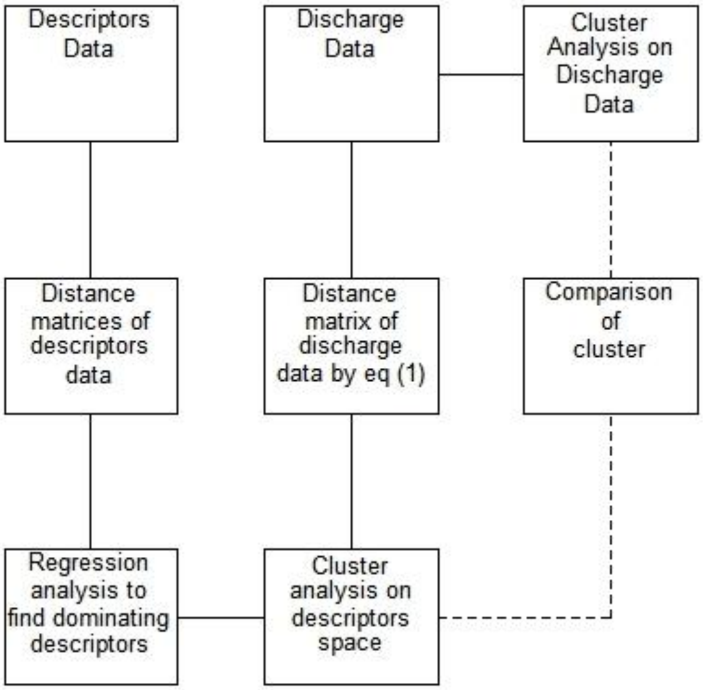

Descriptors Data
Discharge Data
Cluster Analysis on Discharge Data
Distance matrices of descriptors data
Distance matrix of discharge data by eq (1)
Comparison of cluster
Regression analysis to find dominating descriptors
Cluster analysis on descriptors space

Figure 3.2. Schematic diagram representing the steps involved in regionalization procedure followed by *Ganora et al.* [2009].

It can be noted that regression analysis used to estimate flow duration curve at an ungauged site is generally comprised of regression models between hydrological and geomorphological characteristics at gauged site to indentify dominating descriptors. A drawback of this approach is that the selected model deteriorates as it extends over the entire workspace [*Laaha and Bloschl*, 2006]. The shape of unknown flow duration curve thus obtained may be far from correct. In the present work, this problem was dealt with by a two pronged approach, 1)- obtaining best operational model for the whole work space and dividing the work into predefined number of clusters 2)- reselecting the best model for each cluster intending to avoid or minimize the deterioration in model output due to its extension over the entire workspace.

3.2. Methodology

3.2.1. Descriptors analysis

The first step of the procedure is to define the representative descriptors (dominating descriptors). To start with, we will first determine the distance matrix for each descriptor D_{Y_i} by absolute distance measurement method (eq. 3.2) and for FDCs (D_Q) by a predefined metric, in a classical way.

The dominating descriptors are bracketed by their relationship with FDCs. The multiregressive approach was used to assess the relationship between distance matrix of discharge and descriptors. The statistical model can be written as

$$D_Q = \beta_1 D_{Y_1} + \beta_2 D_{Y_2} + \beta_3 D_{Y_3} + \cdots + \beta_p D_{Y_p} + C_0, \tag{3.2}$$

where D_Q and D_Y are distance matrices of discharge and descriptors respectively unfolded to be represent able as vectors; P is the number of descriptors involved; β is regression coeeficient and C_0 is residual matrix. The strength of regression is determined by

$$R_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \tag{3.3}$$

with n as the number of basins and R as standard coefficient of determination [e.g., *Kottegoda and Rosso*, 1997].

Due to the large number of regressors, there is every chance of finding models with a non negligible correlation between descriptors. In these cases, the variance inflation factor (VIF) [e.g., *Montgomery et al.*, 2001], which in terms of quantity, measures the undesirability of multicollinearity in a least square regression analysis become unignorable. It quantifies, through an index estimation, the inflation occurred in variance of an estimated regression coefficient. A cut-off value of 5 was used beyond which the selected model was dropped.

Mantel Test [*Mantel and Valand*, 1970], was also applied to check the significance of the correlation between the distance matrices. Initially, mantel test was proposed to correlate two distance matrices however, modified version of mantel test called Partial Mantel Test made it possible to correlate three distance matrices by correlating two distance matrices while controlling the third matrix. The correlation process was done by unfolding the distance matrix into a vector. For more complex cases *Lichstein et al.*, [2007] provide a method for multiple regression on distance matrices (MRM) to correlate a number of distance matrices. MRM, an extension of partial Mantel analysis, is mathematically simple and can work on all data types. They further deliberated that this method can define any type of relationship such as linear, nonlinear, or nonparametric between a response distance matrix and any number of descriptor distance matrices.

The regression models having good R^2_{adj} values (previously filtered through mantel test and VIF test) are used to execute the FDCs at an ungauged basin. The simulated FDC is then compared with the empirical FDC in D_a space. The models producing least D_a between actual and regional FDCs were selected by using the statistics ζ as an error measure, where ζ is defined as the dissimilarity between the empirical and the estimated FDC.

The proposed methodology of distance-based measurement was carried out in the R statistical environment [*R development core team, 2013*], desegregated for Mantel test and Multivariate Regression Analysis in nsRFA package [*Viglione, 2007*].

The best results is the combinations of descriptors which generated the lowest values of error by Leave-one-out cross validation (LOOCV) procedure ($\Delta = \sum_{1}^{n} \zeta$) and R_{adj}^{2} . The R_{adj}^{2} values obtained with regression models with distance matrices are low, although the descriptors result are statistically significant. Lower R_{adj}^{2} values arise from simpler models with only two descriptors, as in Table 2.

3.2.2. Cluster Analysis

The procedure of the estimation of FDCs in an ungauged basin is based on the basins located nearest to it. For every ungauged basin we want to locate the basins around it that have geomorphological and climatical characteristics similar to that of ungauged basin. The FCDs of neighboring basins will be used to execute the FDCs for ungauged basin. There are different procedures available in the literature to choose the neighbouring basins, for example the formation of fixed regions through cluster analysis [*Hosking and Wallis, 1997*; *Viglione et al., 2007b*] or based on region of influence (*ROI*) [*Burn, 1990*]. Unlike *Ganora et al.* [2009], we will use the combination of classification techniques (e.g. fixed regions and *ROI*) having straight forward application. We do cluster analysis on dominating descriptors selected in the previous step; then by *ROI* technique in each cluster we assess FDC of an ungauged basin. We used ward hierarchical algorithm [*ward*, 1963] as it was able to generate compact clusters with evenly distributed basins in each of them. The wards algorithm starts by considering each basin in a single cluster and then progressively merge basins closet to each other in terms of descriptors magnitude. *Ganora et al.* [2010] also used a reallocation procedure to bring every element closer to center of each cluster. A controversial point in the reallocation procedure is the complication which may arise in case of many clusters. In our work we cluster on the basis of dominating descriptors space and treat each cluster as a separate entity; this means that no reallocation or homogeneity test for cluster independence is required is required.

Moreover, *Ganora et al.*, [2009] defined regional curve as representative of a whole cluster (unlike the present procedure where each station is treated as a separate entity), therefore defining the number of clusters to account for the variation in FDCs in the working area is extremely important. Yet, no specific procedure was defined in the procedure to access the optimal number of clusters and the variation of regimes within the cluster is ignored.

Generally, the aim is to get minimum number of clusters, so that each cluster has large enough number of elements in it. In our work, the number of clusters is selected by using NbClust package in R statistical environment which provides best clustering scheme by observing results obtained by varying number of clusters, distance measurement and clustering technique [*Charrad et al.,* 2014]. The noticeable point in the whole procedure is that the cluster analysis should not be taken in its usual sense of homogeneity. The aim of doing cluster analysis is to select a model for each region executed by dividing the entire workspace purely on the descriptors values to reduce the error magnitude resulting from extension of single model over the whole work space [*Laaha and Bloschl,* 2006].

3.3. Remotely located basins in Descriptors space (Model swapping)

When we talk about assessing hydrological data at the remote catchments, there will always be a considerable amount of error in the final calculation due to scarcity of data, which prevents the usage of standard models [*Pellicciotti et al., 2012*]. Generally, in any work space constituted by the selected descriptors (models), the stations located away from rest of the basins are termed as remote stations (see figure 3.3). In the space of dominating descriptors there can be stations having entirely different descriptors values from the rest of the sample; as the values of descriptors directly affect the hydrological properties of basins therefore the assessment made on its hydrological properties from its neighboring basins can introduce a reasonable amount of error [*Ganora et al.*, 2009; *Pechlivanidis et al.,* 2014].

From the previous discussion, it can be interpreted that the selection criteria for the model of each cluster are R_{adj}^2 and delta values. It is assumed that any two models having almost same R_{adj}^2 and delta (Δ) values can act as a proxy for each other. The concept of *model swapping* can be used to cope with the problem of remotely located basins; intending that model which passes statistical tests defined earlier and has almost similar R_{adj}^2 and delta values (Δ) will have smaller error values for the stations arranged in the middle of cluster of stations (see Figure 3.3). This would bring the remotely located basin to a location where it is crowded by other stations around it. In our work we call this region *Comfortable zone*. The definition of comfort zone came from *Korn et al,*. [2000], who were the first to study Reverse k Nearest Neighbors (R*k*NN) queries. They answered R*k*NN query by drawing a circle of predefined diameter around each data point (say *P*) such that NN of *P* lies on the perimeter of the circle. Later *Stanoi et al.* [2000], solve R*k*NN queries by partitioning the whole work-space around the data point into six equal regions (each of 60^0). We use the same concept in defining the confidence zone of an unknown data station in our work with a slight modification, since too few stations which could result from continuous filter, and continuous refinement phases, may result in over simplification of final results. The reason of locating the unknown data station in the middle of other stations is due to the fact that descriptors vary spatially and temporally [*Hessami et al.,* 2007; *Wilby and Dawson,* 2007]. We try to observe this variation of the descriptors along the defined six sections by selecting such an orientation; where unknown data station is surrounded by its neighbors from every direction.

Principally, the stations having different descriptor values than rest of basins in the selected workspace are classified as "remote stations" [*Pellicciotti et al.*, 2012]. To our knowledge, there is no mathematical definition present in literature for the definition of remotely located basins. In our work, we use following procedure to define remotely located basins;

1) A comparison of station-neighbors distance for any selected station, say X, with station-neighbors distances of rest of the stations.

If S_{nn}^{X} is the sum of station-neighbors distances for the basin X; then for more general case of n number of basins, we can write

$$TD_n = \frac{\sum_{i=1}^{n}(S_{nn}^{i})}{n}; \qquad (3.4)$$

where TD_n is average station-neighbors distances for the entire basins in the workspace. S_{nn}^{X} can be a remotely located basin if $\frac{S_{nn}^{X}}{TD_n} > 1.5$;

2) Observing the neighbors in six-regions around the station.

Generally, due to unique position of remote basin in the workspace, its nearest neighbors are either concentrated in one of the six regions around it or basin is covered from all sides [see Figure, 3]. The swapped model should increase the covering of basin by its neighbors.

Practically speaking, the orientation of NN for station # 3 (in red filled point) in x_i, y_i is more desirable than that of x_j, y_j.

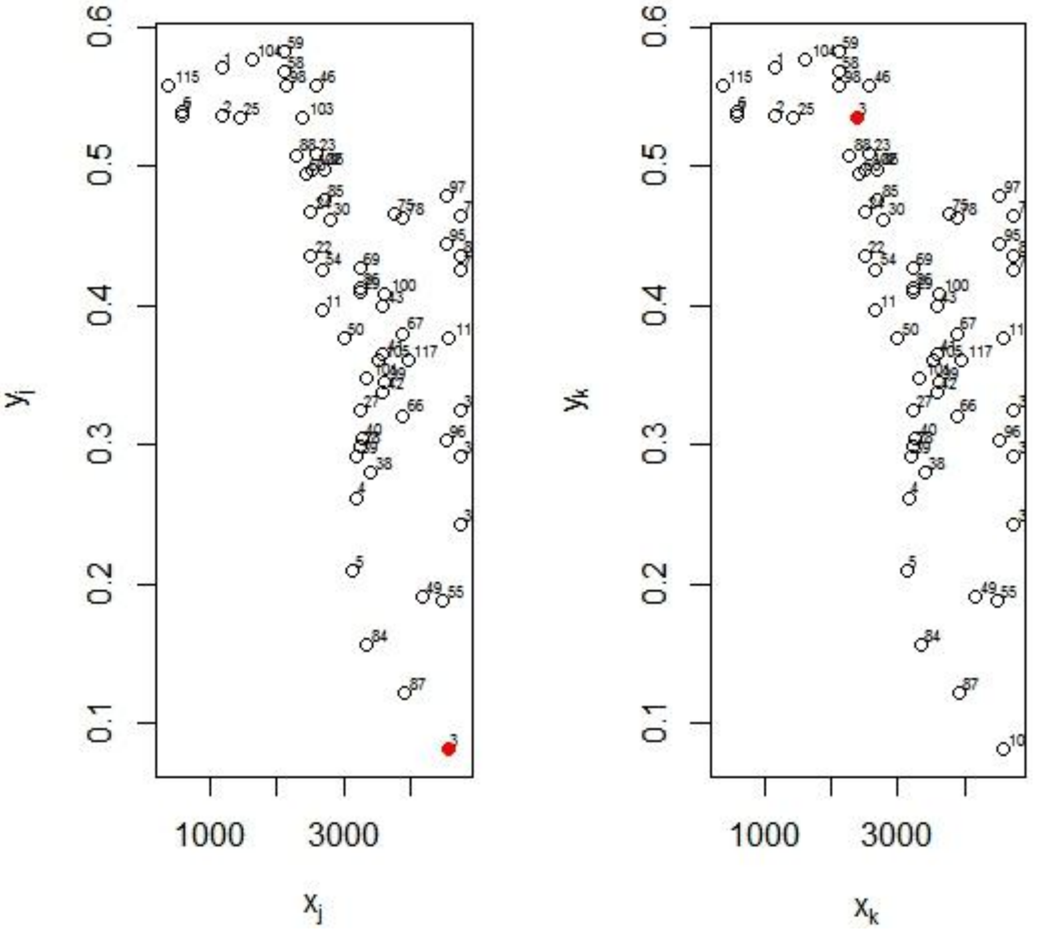


Figure 3.3. (left) Station located remotely in space of any selected model (j). (right) Remoteness eliminated by changing model descriptors (k).

The strength or efficiency ($\mathbb{E}$) of this procedure depends on the number of stations surrounding, in the six regions, the ungauged basin ($\mathbb{N}$) and their distance ($\mathbb{D}$) from the ungauged basin ($\mathbb{E} = f(\mathbb{N}, \mathbb{D})$).

Once the models are enlisted on the basis of their R^2_{adj} and delta values, it is advisable to select a sufficient number of models for each cluster so that each basin is well surrounded by other basins from all directions in case of remotely located basins.

After performing statistical tests, we selected models with 2 descriptors for clusters and overall workspace.

Ideally, each descriptor value of each station should be uniformly scattered over the entire workspace, which is measured by density plots of each descriptors by considering its shape.

3.3.1 Example of model swapping procedure

Let us dissect the concept of changing operational model to have better spatial coverage around unknown data point. We compare the output of nearest neighborhood analysis, for station # 4 (represented with red filled dot in Fig 3.4) in cluster 1 and station # 45 (represented with red filled dot in Fig 3.5) in cluster 2, before and after improving the spatial coverage of neighbors (represented with green filled dot). The selected 5 models for overall workspace and selected clusters are enlisted in Table 1 which carries Δ values and R_{adj}^2 of the models. The outputs of originally selected and swapped models (in terms of RMSE, NSE and MAE) for the considered station are represented in Table (2).

Table 3.1. Descriptor models for overall workspace and clusters enlisted in the order of R_{adj}^2

Model	Descriptors	R_{adj}^2	VIF	Delta Factor
Overall				
	clc_5[1], $quota\_massima$[2]	*0.024*	*<5*	*59.00*
	$y - baricentro$[3], $Hypsographic\ Curve$[4]	*0.041*	*<5*	*59.66*
	$quota\_massima$[2], $NDVIanno$[5]	*0.030*	*<5*	*60.26*
	$NDVIanno$[5], $delta\_z$[6]	*0.023*	*<5*	*60.48*
	$y - baricentro$[3], $quota\_media$[7]	*0.033*	*<5*	*60.82*
Cluster-1				
	MAP_std[8], $NDVIanno$[5]	*0.035*	*<5*	*57.33*
	clc_4[9], cv_rp[10]	*0.048*	*<5*	*58.85*
	$quota\_massima$[2], $NDVIanno$[5]	*0.038*	*<5*	*59.06*
	clc_4[9], clc_5[11]	*0.047*	*<5*	*59.74*
	$quota\_massima$[2], $Hypsographic\ curve$[12]	*0.049*	*<5*	*59.84*
Cluster-2				
	$quota\_massima$[2], $delta\_mese$[13]	*0.066*	*<5*	*51.46*
	clc_5[11], $quota\_massima$[2]	*0.069*	*<5*	*53.66*
	$quota\_massima$[2], $NDVIanno$[5]	*0.065*	*<5*	*56.17*
	MAP[14], C_int[15]	*0.029*	*<5*	*56.24*
	$quota\_massima$[2], C_int[15]	*0.066*	*<5*	*57.64*

[1] percentage area of the basin as wetlands
[2] maximum elevation of the basin (m)
[3] latitude of basin (m)
[4] hypsographic Curve (m m.s.l)
[5] annual Normalized Difference Vegetation Index (NDVI)
[6] interquartile distance between basin elevation at 25% and 75% of area dominated by hypsographic curve (delta_z)
[7] average basin Evelation (m) (quota_media)
[8] annual Normalized Difference Vegetation Index (NDVI)
[9] percentage area of the basin which is not vegetated (e.g mining areas, landfills and construction sites, industrial, trade and communication networks)
[10] coeff. of variation in rainfall patterns
[11] percentage area of the basin as wetlands
[12] hypsographic Curve (m m.s.l)
[13] time Interval Between Maximum and Minimum Monthly Averages of Rains (delta mese)
[14] average total annual rainfall [mm] (MAP)
[15] coeff. of precipitation intensity (C_int)

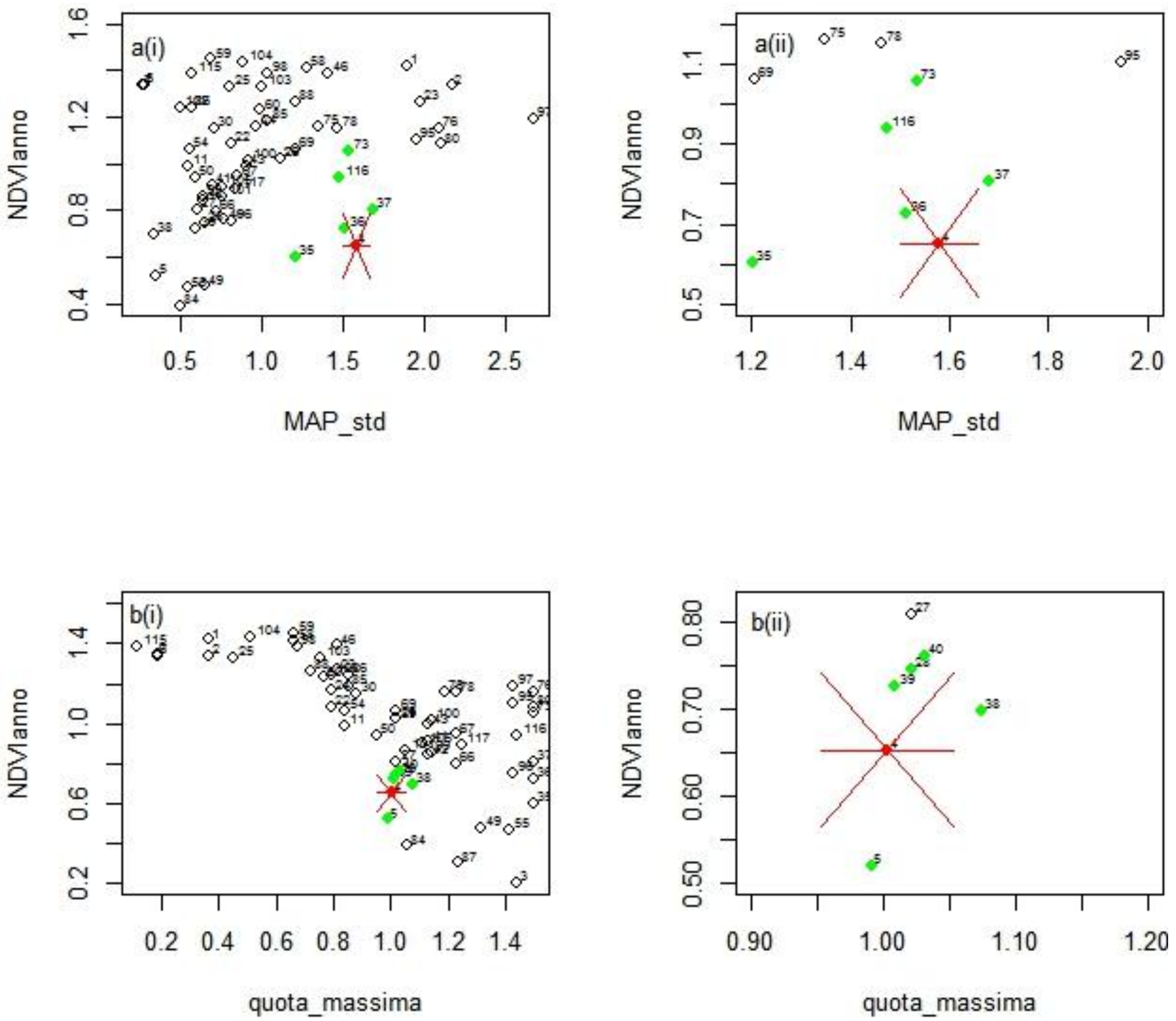


Figure 3.4. a(i) Basins arrangement in selected model for overall cluster 1; a(ii) detailed view of selected basin and its neighbors; b(i) and b(ii) swapping model to give better neighbor coverage and detailed view respectively.

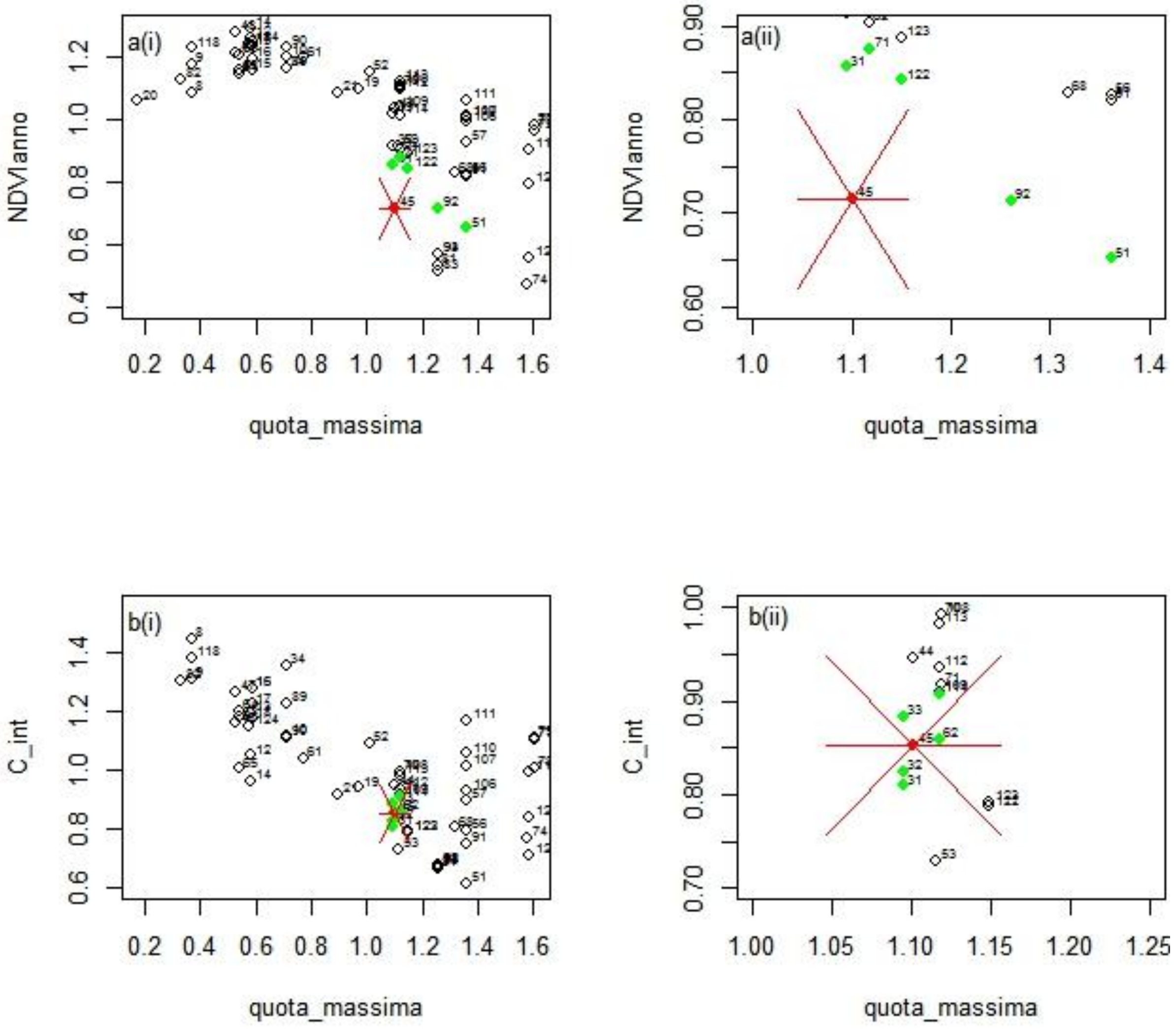


Figure 3.5. a(i) Basins arrangement in selected model for overall cluster 2; a(ii) detailed view of selected basin and its neighbors; b(i) and b(ii) swapping model to give better neighbor coverage and detailed view respectively.

Table 3.2. Models with 1, 2 and 3 descriptors enlisted in the order of R^2_{adj}

Basins	Cluster Number	Original Model	RMSE	NSE	D_T	Swapped Model	RMSE	NSE	D_T
4	1	MAP_std, NDVIanno	0.144	0.960	34.071	Quota_massima, NDVIanno	0.075	0.989	21.492
45	2	Quota_massima,	0.214	0.949	59.600	Quota_massima,	0.155	0.973	41.972

The complete procedure of execution of FDCs would require: (1) For each station, we first found the dissimilarity index by a predefined metric, (2) The distance matrix of each descriptor was found by comparing their magnitudinal values for each basin, (3) The delta factor and MRM will give us operational models for overall workspace, (4) In the vicinity of selected model for overall workspace, the workspace is divided into smaller regions, (5) Based on the previously defined procedure of model selection, model is selected for each smaller region, (6) Remotely located basins were given better spatial coverage by swapping model technique, (7) The regional dimensionless FDCs are estimated by nearest neighborhood NN method.

3.4. Alternative Procedures

3.4.1. Parametric Model

Since the flow duration curve represents the number of days in a year during which flow is available to a certain extent. It is immediately evident that the duration can be expressed in terms of frequency or percentage of time in which a certain level of flow is equaled or exceeded. In the context of *frequency*, it is equivalent to the frequency of exceedance of the flow over a designed discharge. It is therefore natural to interpret the FDCs as the frequency curves and represent them in an analytical way by the means of probability distributions.

In literature there are different functions of cumulative probability which represent the FDCs, such as the Generalized Pareto distribution with three parameters [*Fennessey*, 1994], the Gumbel distribution [*Kottegoda and Red*, 1997], the distribution normal [Singh et al., 2001] and, usually two or three parameters log-normal distribution [*Fennessey and Vogel*, 1990, *Claps and Fiorentino*, 1997]. In more recent times other distributions have also been used, for example, the Kappa [*Castellarin et al.*, 2007] or the EtaBeta [*Iacobellis*, 2008]. The choice of the distribution depends on the ability to adapt to the observed data and the possibility to estimate parameters in a robust way.

The difficulties encountered in choosing an appropriate distribution to represent FDCs, among those commonly used in the field hydrological, substantially led to the introduction of other types of distribution. Among them, particularly convenient is Burr distribution (also known as Burr type XII) introduced by *Burr* [1942] [see also *Rodriguez*, 1977] and is used in different scientific fields, but little known in the field of hydrology [*Shao et al.*, 2004; *Nadarajah and Kotz*, 2006].

The cumulative distribution function of Burr with its 3-parameters can be written as

$$\boldsymbol{P(x) = 1 - \left(1 + b(\frac{x}{a})^{c}\right)^{-1/b}} \tag{3.5}$$

where a is the scale parameter, b and c are the two shape parameters. The presence of two shape parameters allows us to represent adequately, the various forms of FDCs.

The analytical form of (3.5) allows to derive simple expressions for the probability density

$$\boldsymbol{p(x) = \frac{c}{a}\left(\frac{x}{a}\right)^{-1+c}\left(1 + b\left(\frac{x}{a}\right)^{c}\right)^{-1-\frac{1}{b}}} \qquad (3.6)$$

and the quantile function

$$x(P) = a\left(\frac{(1-P)^{-b}-1}{b}\right)^{\frac{1}{c}} \qquad (3.7)$$

Furthermore, the Burr distribution has a limitation when P equals to 0 (for $P \rightarrow 0$); a condition for which x becomes undefined, and therefore the values of flow become negative.

There are two limiting cases for Burr distributions, in particular:

- the lower limit corresponds to the case where $b \rightarrow 0$ and the distribution becomes a two-parameter Weibull;
- the upper limit corresponds to the case in which $c \rightarrow \infty$ and the distribution becomes a two-parameter Pareto.

To decide which distribution we are going to use, we used L- moments as a descriptive statistical indexes of FDC which, like moments contain information on the average value of the variability, asymmetry, etc., of distribution.

We define following applications to select the distributions:

- L_1 called L-moment of order 1, which represents the average of a distribution;
- LCV is defined as the dimensionless ratio between L-moment of order 2 and L_1, which represents the variability of the distribution (analogous to the coefficient of variation of theory of moments);
- LCA is defined as the dimensionless ratio between L - moment of order 3 and L-moment of order 2, representing the asymmetry of the distribution (similar to the coefficient of asymmetry or skewness of theory of moments);
- $Lkurtosis$, the dimensionless ratio between L-order moment 4 and L-moment of order 2 represents, the flattening of the distribution (similar to the coefficient of flatness or kurtosis of theory of moments).

In the space of the variables LCV and LCA, the Burr distribution has a spindle-shaped domain [*see Ganora et al.*, 2014], delimited by a lower and an upper limit, respectively, by the following equations

$$LCA^{inf} = \frac{1}{LCV}\left(-2 + 2.3^{\frac{\log(1-LCV)}{\log(2)}} + 3LCV\right) \tag{3.8}$$

$$LCA^{sup} = \frac{1+3LCV}{3+LCV} \tag{3.9}$$

The estimation of the parameters a, b, and c is also made using the method of L-moments.

$$L_1 = \frac{a}{b^{1/c}}\frac{\Gamma\left[\frac{1}{b}-\frac{1}{c}\right].\Gamma[1-\frac{1}{c}]}{\Gamma[\frac{1}{b}]} \tag{3.10}$$

$$LCV = \frac{1-\Gamma\left[\frac{1}{b}\right].\Gamma[\frac{2}{b}-\frac{1}{c}]}{\Gamma\left[\frac{2}{b}\right].\Gamma[\frac{1}{b}-\frac{1}{c}]} \tag{3.11}$$

$$LCA = \frac{\frac{\Gamma[\frac{1}{b}-\frac{1}{c}]}{\Gamma[\frac{1}{b}]}-3\frac{\Gamma[\frac{2}{b}-\frac{1}{c}]}{\Gamma[\frac{2}{b}]}+2\frac{\Gamma[\frac{3}{b}-\frac{1}{c}]}{\Gamma[\frac{3}{b}]}}{\frac{\Gamma[\frac{1}{b}-\frac{1}{c}]}{\Gamma[\frac{1}{b}]}-3\frac{\Gamma[\frac{2}{b}-\frac{1}{c}]}{\Gamma[\frac{2}{b}]}} \tag{3.12}$$

Where $\Gamma[.]$ is the gamma function.

It is necessary, in some cases, to estimate the FDCs by the Weibull distribution or the Pareto. If we fall in the limiting case of the distribution Weibull parameters to 2, the shape of cumulative probability function becomes:

$$P(x) = 1 - \exp[-\left(\frac{x}{a_W}\right)^{c_W}] \tag{3.13}$$

where the subscript W indicates that the parameters a and c refer to the distribution of Weibull. The quantile function becomes:

$$x(P) = a_W(-\log(1-P))^{1/c_W} \tag{3.14}$$

whose parameters can be easily estimated from the L-moments as:

$$c_W = -\frac{\log(2)}{\log(1-LCV)} \tag{3.15}$$

$$a_W = \frac{L_1.c_W}{\Gamma[\frac{1}{c_W}]} \tag{3.16}$$

Similarly, when you fall in the limiting case of the Pareto distribution with 2 parameters, the cumulative function is:

$$P(x) = 1 - \left(\frac{x}{a_P}\right)^{c_P} \tag{3.17}$$

where the subscript *P* indicates that the parameters a and c refer to the Pareto distribution. The quantile function becomes:

$$x(P) = a_P(1 - P)^{1/c_P} \tag{3.18}$$

whose parameters are estimated from the L-moments as:

$$c_P = \frac{-LCV+1}{2LCV} \tag{3.19}$$

$$a_P = \frac{L_1(1-c_P)}{c_P} \tag{3.20}$$

3.4.2. Geographical distance method

A more common and straight forward way for the selection of NN is to use geographical distance method. In the vicinity of geographical space it is assumed that the stations having similar hydrological properties are located closer to each other and hence it is reasonable to asses hydrological properties of ungauged catchments based on spatial proximity [*Bloschl*, 2005]. The Euclidean distance norm is generally used to calculate distance between a pair of catchments.

In newly developed distance based method described in section 2, we decided to adopt models with two descriptors because of their higher robustness and primarily for an ease of comparison with Euclidean distance, which is also a combination of two descriptors (Latitude and Longitude).

3.5. Result

The results generated by our model, geographical distance method and parametric model are tested by using cross-validation procedure. It was done by considering one station as ungauged, removing it from the whole database and estimating FDC for that station. The process was repeated for all the stations and error was measured between estimated FDCs and empirical FDCs. Generally, the agreement between actual and predicted FDCs is more qualitative.

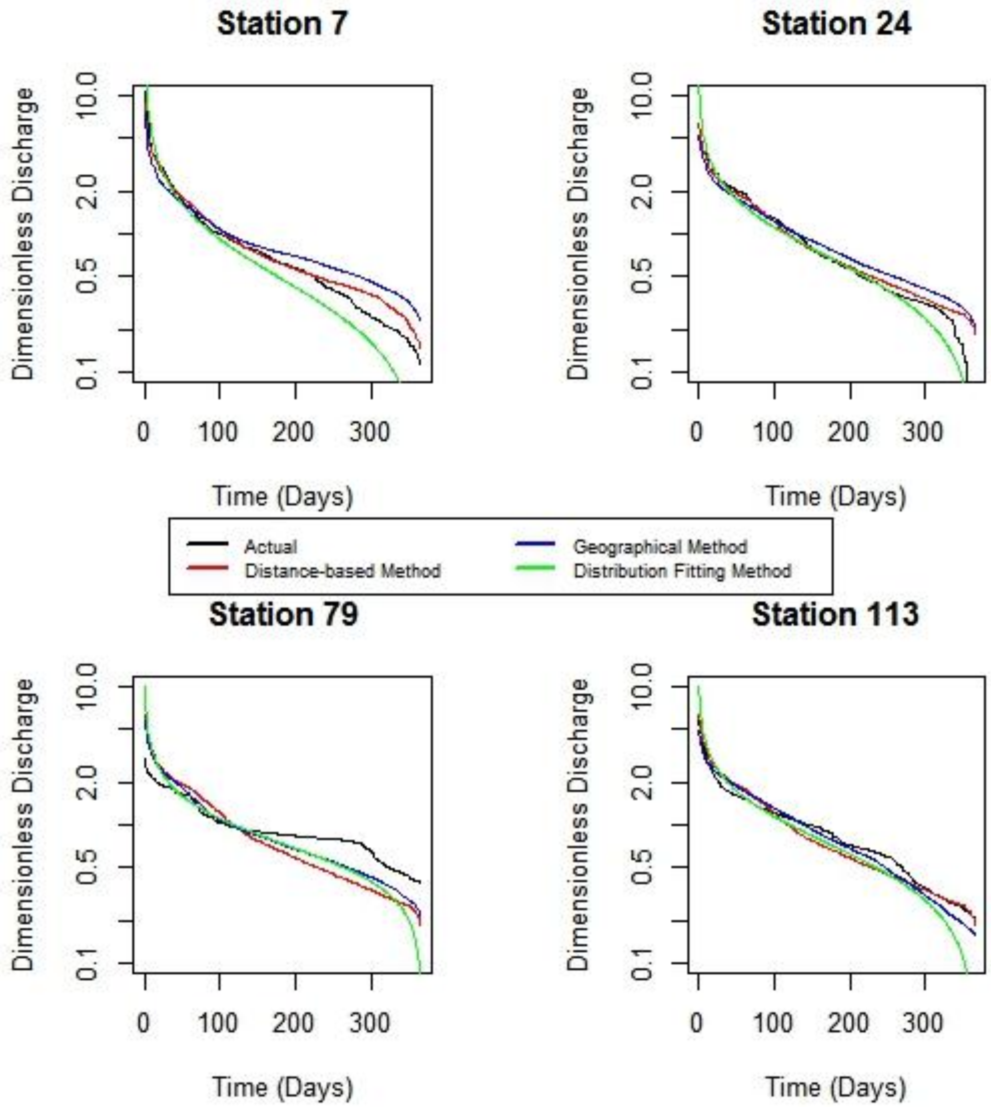


Figure 3.6. Comparison of simulated FDCs with actual FDCs at selected stations.

As performance indexes, the root mean square error (RMSE), D_a and Nash-Sutcliffe efficiency (NSE) have been evaluated. These performance indexes for the 3 considered procedures are listed in table (1) for some selected basins, while a complete comparison of 124 stations is shown in figure (3.7). The newly developed method performed better for most of the stations then its other counterparts.

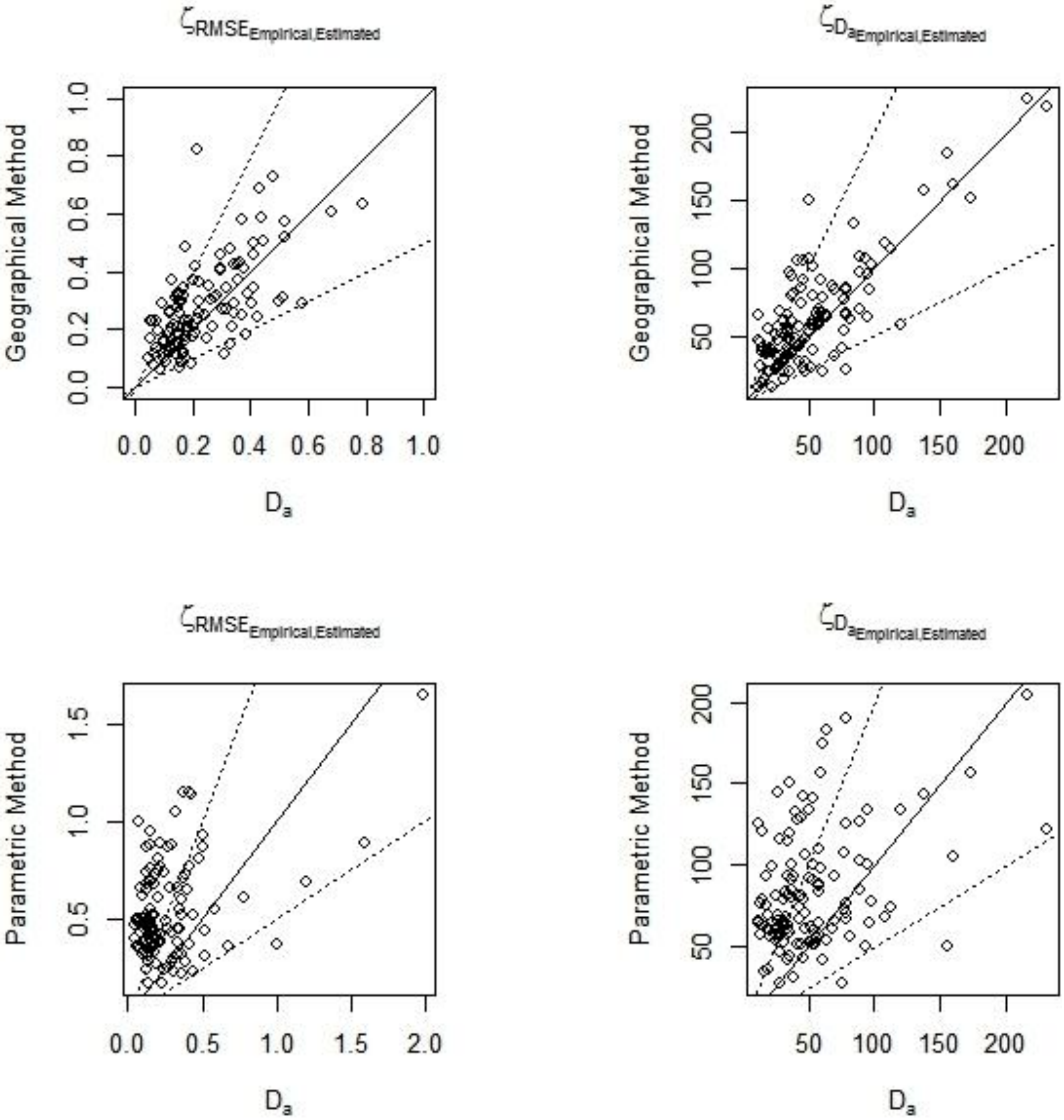


Figure 3.7. Comparison of RMSE and D_a of distance-based model, geoghraphical method and parametric model.

Table 3.3. RMSE, NSE and D_a obtained in the various basins using three types of methods

Basin	Area (Km^2)	New Method			Euclidean			Parametric Model		
Stations Codes		RMSE	NSE	D_a	RMSE	NSE	D_a	RMSE	NSE	D_a
3	41	0.166	0.970	42.028	0.342	0.874	104.835	0.516	0.713	127.837
5	43	0.223	0.950	56.305	0.258	0.934	66.931	0.385	0.852	54.362
7	350	0.296	0.947	36.794	0.460	0.873	79.695	0.884	0.530	80.096
24	133	0.122	0.982	25.610	0.187	0.960	51.476	0.520	0.691	53.362
38	207	0.297	0.726	77.000	0.413	0.471	77.132	0.432	0.421	72.378
46	256	0.185	0.963	28.915	0.230	0.943	35.408	0.266	0.924	27.031
65	74	0.145	0.970	26.745	0.320	0.848	56.499	0.878	-0.149	144.930
70	360	0.151	0.916	45.943	0.316	0.635	91.332	0.548	-0.101	79.154
72	3956	0.141	0.928	31.355	0.307	0.662	61.000	0.502	0.096	62.735
76	25640	0.174	0.867	36.706	0.486	-0.030	93.631	0.526	-0.203	90.795

Table 3.4. Comparison of magnitudes of different errors (ζ) with corresponding Standard deviations (δ).

Model	ζ(RMSE,δ)	ζ(NSE,δ)	ζ(D_a,δ)
New Method	0.271(0.264)	0.863(0.210)	53.721(37.698)
Geographical Method	0.310(0.266)	0.783(0.470)	62.735(38.750)
Parametric Method	0.535(0.240)	0.327(0.240)	83.465(35.490)

3.6. Conclusion

The procedure is applied to 124 basins in Northern Italy and Switzerland. The basins used in our analysis present different hydrological behavior and cover a wide range of descriptors (area, elevation etc). The distance-based model proposed here is able to reproduce the unknown FDCs in an efficient way if compared to geographical distance method and parametric model. Unlike classical parametric approach; the present approach deals with FDCs as a whole function. The discharge distance matrix is linked to basin descriptors' distance matrices through regression. By using the R^2_{adj} and delta values, models were selected for overall work space. Later the workspace, comprised of descriptors in best model, was divided into different clusters and the best models are found for each cluster. The simplicity of the proposed procedure makes it a valuable tool for FDCs assessment in an ungauged basin. The results obtained by our model are comparable with and better at many basins than other models.

The present work also covered some of the short falls in previous work done by *Ganora et al.,* [2009]: (1) each cluster was characterized by a single dimensionless FDC, hence a reasonable error might be introduced in case of remotely located basin. On the contrary, in our work regime of each ungauged basin is assessed through predefined number of neighbors and incase of remotely gauged basins the model was changed to eliminate remoteness. (2) Since the basins are scattered over a wide range of descriptors values, therefore using only a single model for the whole work space is over simplification. In the new model this issue is addressed by dividing the whole workspace into smaller clusters and a separate model for each cluster was found, (3) reallocation procedure applied previously by Ganora et al., [2009] might be complicated in case of many clusters whereas no reallocation procedure is required in the proposed methodology, (4) Finally cluster analysis is simultaneously done on descriptors and hydrological data spaces consequently clusters may not necessarily overlap in the exact same manner and this causes magnitudinal error in final regime. In the present procedure, cluster analysis is only done on descriptor space.

Chapter 4. Rainfall Extremes Analysis

4.1. Introduction

The topic of analysing rainfall extreme events is currently one of the leading topics, in the field of climatology because of its adverse impacts on human lives and properties [*Loo et al.*, 2014]. The researches have shown quite an interest in this field recently and almost the entire community of researchers within the climatic change paradigm agrees on a hypothesis that there is going be an increase in the magnitudes of extreme events due to the increase in climatic variability [*see Neumayer and Barthel*, 2011; *Bouwer*, 2011a; *Barthel and Neumayer*, 2012, and references therein]. A clear increment in magnitude of extreme events has been documented by *BEH* [2012], which resulted in adverse impacts on community and the environment [*IPCC*, 2007]. In third world countries like Pakistan, extreme rainfall events cause natural hazards because they are a source of degradation processes like flash floods, landslide triggering and erosion; which can cause a severe damage to the land and properties.

Mapping the hazards of extreme rainfall allows us to assess the spatial distribution of this climatic feature even at locations where no climatic record exists. In the fields of regional planning and environmental management; Climatic hazard maps, in general, can also be helpful as a part of decision making systems. The main objective of our work on rainfall extremes is to describe a method to obtain extreme precipitation hazard maps.

The rainfall data provides point information which needs to be translated to a spatially continuous variable. Over the course of research done in this regard, different variables have been proposed to describe the rainfall extremes. For example, *Prudhomme* [1999] and *Prudhomme and Reed* [1999] used the median of the annual maximum daily precipitation; *Lorente and Beguería* [2002] used the median of the annual maximum precipitation accumulated in 1, 3, 5, and 7 days. According to some authors a median value of extreme precipitation is not the most adequate variable to express extreme events [*Beguería and Vicente-Serrano,* 2006]. Attempts have also been made by using absolute maxima with a very little success [*García Ruiz et al.*, 2000].

To describe the occurrence of extreme rainfalls in a region, the extreme value theory can be used which provides a complete analysis of the statistical distribution of extreme precipitation events and allows the construction of magnitude–frequency curves by fitting the selected distribution to rainfall data. The degree of hazard related to extreme precipitation at a given location can be expressed by using driven statistical laws like quantile estimates in which average magnitude of an extreme event for a given return period is simulated. The spatial distribution of the hazard of extreme rainfalls can be mapped by combining quantile estimates with spatial interpolation techniques. For example, *Gajic-Capka* [1991] and *Lana et al.*, [1995] used local interpolation methods to map quantile estimations obtained by fitting a Gumbel model to series of annual maxima. *Beguería and Lorente* [1999] provided the 100-yr daily maximum rainfall estimates by the fitting of Gumbel model on rainfall data at several points in the study area, by using ordinary regression against relief parameters. *Weisse and Bois* [2001] modelled 10- and 100-yr rainfall estimates for rainfall duration of 1–24 h by comparing kriging and ordinary regression against topography.

The main limitations of existing techniques are: (1) calculation of the extreme quantiles by the extreme value theory in these examples was reduced to at-site estimations of the model parameters; the existence of a spatial structure was not addressed. (2) If extreme rainfall hazards are to be mapped for a different return period or hazard level, new at-site quantile estimations and interpolation are needed.

In our work we explore the possibility to build a probability model over a spatially continuous space by allowing the parameters of fitted distribution vary spatially. Many probability distributions have been considered, for the probabilistic modeling of extreme precipitation events like the extreme value distributions (Generalized Extreme Value (GEV), Gumbel and Log-normal distributions), the distributions of the transformed normal or gamma families [e.g., *Kottegoda and Rosso*, 1997]. The parameters of fitted distributions are measured by fitting it on at-site rainfall data and then with spatial interpolation techniques, the executed parameters are distributed spatially. The estimation of spatial model from distribution parameters helps in estimating hazardous rainfall maps for different return periods without having a need to apply new spatial interpolations. The advantage of using analysis of the spatial distribution of the model parameters, relative to using a set of unrelated at-site probability models is that it results in a much more robust regional probability model.

4.2. Study area and elaboration of the database

We have tested our methodology in Pakistan in which extreme precipitation is frequent and causes important social, economic, and environmental damage [*White et al.,* 1997; *García-Ruiz et al.,* 2000; *Lasanta* 2003]. For example, flood events of 2011, which levelled off the whole sindh province of the country resulting in the loss of human lives and properties.

Pakistan has a very high seasonal interannual variability because of dominating atmospheric patterns in different parts of country. The variation in annual precipitation oscillates between 131mm in gilgit situated in extreme north to 1761mm centrally-north in balakot. Due to these extreme fluctuations the precipitation magnitude, in some years, exceeds the mean value while in other years, the country faces long droughts periods which are particularly frequent in areas of Baluchistan and Sindh.

Globally, raingauges are always used to measure the depths and rates of rainfall events. The original database consisted of 21 series of daily precipitation with different lengths mostly concentrated in the nothern part of the country. There were raingauge stations in our analysis e.g. Balakot and Kotli, where within the same locality; position of the observatories were shifted. For these cases by merging the data of the observatories located in the same location, we created new series. At some sites we faced a problem of missing rainfall values. To overcome this challenge we selected a data series with less than 15% missing values [*Karl et al.,* 1995] for a common record period of 50 years (1961-2010). The reason of adopting this strict criterion is to ensure that all of the time series data are sufficiently long so that they not only provide reliable estimates of the extreme events probability [*Jones,* 1997], but also cover the same record period to avoid variability in the estimation of the parameter as a result of inter annual climatic cycles. This led to a final database of 15 observatories.

The actual rainfall data was obtained for the raingauges which are concentrated in the northern part of the country (see Fig. 4.7). This caused a serious problem of very small spatial coverage. The reduction of the spatial coverage of the weather stations can introduce some limitations to the analysis of climatic variables because to assess certain rainfall coefficients for extreme events analysis, we need to have a high resolution rain data [*Austin and Houze*, 1972, *Rodrıguez-Iturbe and Mejia*, 1974]. The spacing between raingauges located across any research area are generally very large and do not correspond to the variability of rainfall spatially [*Felgate and Read*, 1975].

The problem is even more magnified on small scales. In addition, the density of rainguages is extremely low over the difficult terrain. Satellite precipitation data can be used due to its high resolution. Satellite missions generate data of various temporal and spatial scales. To get this data the satellite missions are mounted with either infrared or more recently microwave devices. The former although have an ability to cover large spatial and temporal scale but they are unable to map rainfall events efficiently because of their lower penetration into the dense clouds. To overcome the lack of penetration power, infrared radiations have recently been replaced by microwave radiations as in case of TRMM mission. The TRMM mission takes data from multiple satellites and by using simple possible way, the values are averaged to produce a best estimate of the mean rainfall rate over the selected interval [*Huffman*, 2007]. The TRMM data was downloaded on 327 grid points shown in figure 4.7, located across the whole country.

The length of the dataset is an important aspect in the analysis of climatological variables, but in the case of extreme value analysis as the samples are reduced to only the highest values in the range of the variable [*Jones*, 1997], the length of dataset becomes critical. Pakistan can be characterized by high interannual variability of climate and since in the recent decades, significant differences in the annual averages have been found therefore the need of long time series to provide robust estimate is seriously needed. Significant temporal variability and trends in extreme events have also been found in different areas of the Pakistan.

The adequacy of the length of series required to obtain reliable estimations about the frequency of extreme preceipitation events is subject to debate. Some authors indicate that for the return interval estimations of 50-yr, by fitting gumbel distribution on annual maxima with 25% error, require 39 years of data [*Benson*, 1960]. While the others analyzed that 20 years of data provide the estimates for extreme events against return periods with a 20% rate of error; data for more than 25 years or more reduces this error magnitude to less than 20% [*Porth et al.*, 2001].

Considering the previous discussion, despite the reduction of spatial coverage due to concentration of rainguages in north, we decided to maintain a large temporal extent of the database (50 yr). We used station-year method for the construction of growth curve and to analyze the probability of the extreme precipitation events which do not require a complete data time series: instead, the entire data was arranged in a single vector.

The checking of data quality and homogeneity of daily climatological records, which is required for the construction of growth curve, is a very complex procedure. There are a number standard tools in the scientific literature to test the homogeneity of climatic dataset [see *Buishand*, 1981 for details], because different authors have addressed the problem in a different way [*Manton et al.*, 2001; *Brunetti et al.*, 2002]. *Hosking and Wallis*, 1997 used L-moments to find the homogenity of a climatic region. This has been the method followed in this study. For the preparation of extreme rainfall hazardous map for a certain return period, a single growth curve was prepared for statistically homogenous regions.

4.3. Downscaling

Annual maxima of daily rainfall for the years 1961–2010 are modelled for fifteen locations in Pakistan (chosen to give a good geographical representation of the country). The gumbel distribution is fitted to data from each location to describe the extremes of daily rainfall and to predict its future behavior from monthly data. We find evidence to suggest that the Gumbel distribution provides the most reasonable model for four of the fifteen locations considered. We explore the possibility of trends in the data but find no evidence suggesting of any trend. We derive estimates of 2, 50, 100, 500 year return levels for daily rainfall.

The satellite data consist of a total monthly rainfall and number of monthly wet days data for the years from 1961 to 2010 for the 327 grid points pointed out in figure 4.7. The data were extracted from the website http://badc.nerc.ac.uk of the British Atmospheric Data Centre, which lists the monthly rainfall figures for 327 grid points in Pakistan. But there are only 15 sites that have data going back to 1961 (the earliest year for which data are available). Our analyses are limited to 15 of these grid points. These fifteen raingauges, located in northern part of Pakistan, are chosen because of the availability of actual data.

In this section we focus on the quantitative assessment of extreme precipitation events. To do this, it is necessary to have data on daily temporal scale; unlike the evaluation of only the behavior of extreme events, for which one could also refer to time series with monthly temporal scale. In this regard, we will make use of only the CRU database due to limited spatial coverage of actual data from Pakistan. The traditional hydrological analyses are based for example on historical daily rainfall values measured on a single station. It is then possible to adapt a probability distribution of extreme values and thus build a curve that shows the entity of the event as a function of the probability of non-exceedance. Most often, instead of the probability of non-exceedance, the so-called return period that represents the average number of years that one must wait for an event, is used. The return time is usually preferred because the probability interpretation is more intuitive. The CRU precipitation data are available but only at a monthly scale. Therefore, for its temporal downscaling we must look for other information attributable to the daily scale which is already contained in the database that reports the number of rainy days per month.

4.3.1 Distributions of extreme values

One can derive the distribution of the extreme values of a certain event (P_{max}) from the probability distribution of a single event (P_E) through the following expression

$$P_{max}(x) = e^{-\lambda[1-P_E(x)]} \tag{4.1}$$

where the term λ is the average number of events per year.

Initially we assume that the depth of rain (h) follows a probability distribution of a single parameter, in particular an exponential law of the type

$$p_H(h) = \frac{1}{\alpha_G} e^{-\frac{h}{\alpha_G}} \tag{4.2}$$

where α_G represents the intensity of the average annual precipitation. The Probability density function turns out to be:

$$P_H(h) = 1 - e^{-\frac{h}{\alpha_G}}. \tag{4.3}$$

The distribution of the maximum, which is obtained by replacing the function (5.3) in (5.1), can be traced back to the known distribution of Gumbel

$$P(h_g) = e^{-\lambda e^{\frac{h_g}{\alpha}}} \tag{4.4}$$

which is a function of daily rainfall depth (h_g) expressed by two parameters. One can easily notice that in the equations from (4.2) to (4.4) the number of parameters of a distribution increase from one to two. To estimate this additional parameter we will make reference to the number of rainy days.

The expression just obtained can be extrapolated for h_g,

$$h_{g,P} = -\alpha_G ln\left[-\frac{1}{\lambda}\ln(P)\right] \tag{4.5}$$

it can be rewritten as a function of the return time

$$h_{g,P} = -\alpha_G ln\left[-\frac{1}{\lambda}\ln(1-\frac{1}{T})\right]. \tag{4.6}$$

now by repeating the same procedure, but considering a function that describes the rainfall depth as pareto distribution of two parameters (instead of the exponential of a single parameter). This distribution can be written as

$$P_H(h) = 1 - \left[1 - k\frac{h}{\alpha_P}\right]^{\frac{1}{k}} \tag{4.7}$$

where k is the shape parameter. The mean and variance of the Pareto can be expressed as:

$$\mu = \frac{\alpha_P}{1+k} \tag{4.8}$$

$$\sigma^2 = \frac{\alpha_P^2}{(1+k)^2(1+2k)} \tag{4.9}$$

now the distribution of extreme values is obtained with the distribution of three parameters; the depth of daily rainfall can then be expressed as the function of return period with the expression in the form

$$h_{g,T} = \frac{\alpha_P}{k}\left[1 - \left(\frac{1}{\lambda T}\right)^k\right]. \tag{4.10}$$

4.3.2 Application of the distributions

In order to apply the newly derived distributions we must decide whether to use the Gumbel distribution (4.6), or the Pareto (4.10). However, for both distributions the parameter of number of rainy days per month is very important to make some considerations. The database containing the number of rainy days (NWD) was built by considering a minimum threshold of rain depth equal to 0.1 mm. But since extreme events are intense in nature, this threshold is extremely low to be representative. It is, for this reason, considered necessary to increase the threshold to have more meaningful representation on number of wet days.

The selection of an appropriate threshold value to extract the exceedance series for every station of work space constitutes as the most critical parameter selection in PD series modeling because the level of threshold defines the size of the selected data. A low threshold is generally preferred over a high one for its ability to bracket maximum amount of rainfall data. Too low threshold value should however be avoided as distribution model would not correctly fit on the selected data. To set an appropriate threshold one of the advised procedure in literature is the mean excess plot, complemented by plotting the average mean excess over a certain threshold against the value of the threshold itself. Mean excess plots were constituted by using different threshold values to evaluate the fit of the selected distribution model to the data as shown in figure 4.1. It can be noticed that graph in figure 4.1 follows a straight line and the best line regression model, allows the selection of threshold equal to 10mm. The threshold values are counter checked by fitting the same distribution model (GEV) on the actual daily data at known stations.

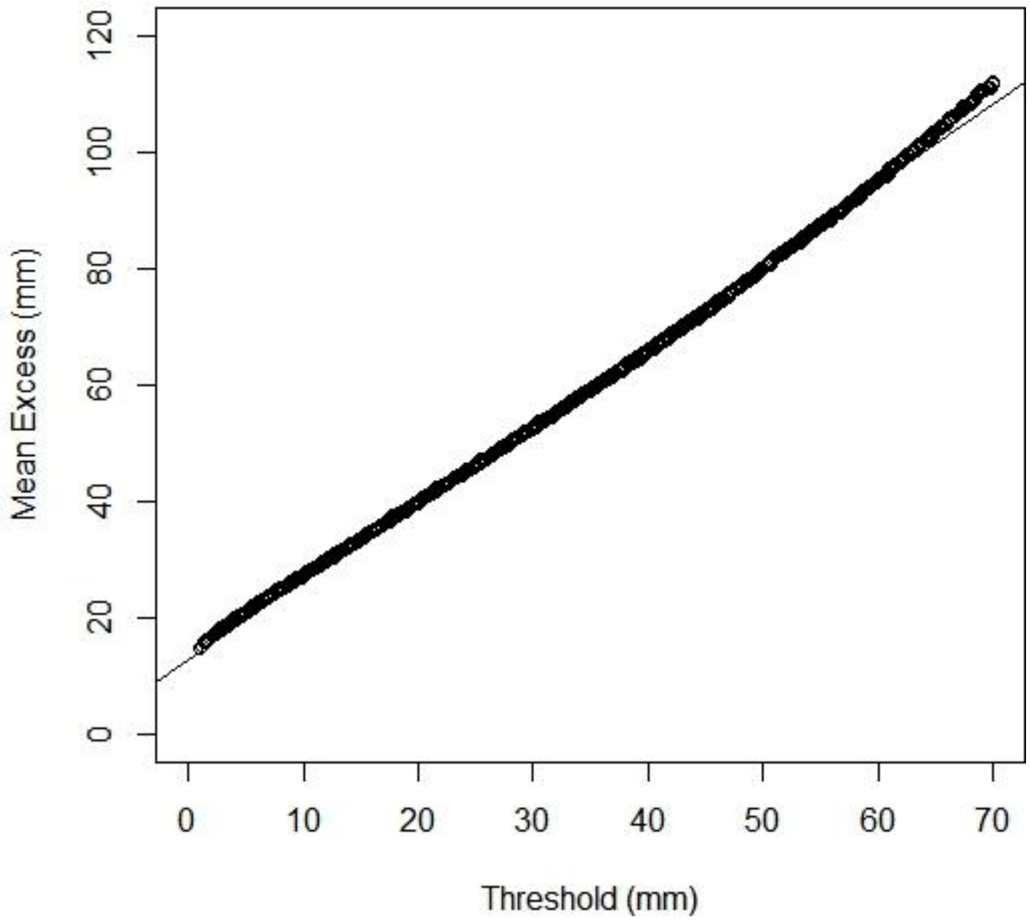


Figure 4.1. Mean excess plot for the selection of threshold.

The transition for NWD for monthly data (from 0.1mm threshold) to a higher threshold is made by considering the actual data, for which a best fit line between the number of events corresponding to the threshold 0.1mm and the number of events corresponding to the threshold 10mm was fitted through the data. This allowed us to define a multiplication factor of 0.4096 (almost half) to be multiplied with the number of events corresponding to a threshold of 0.1mm to be able to use them for a higher threshold of 10 mm.

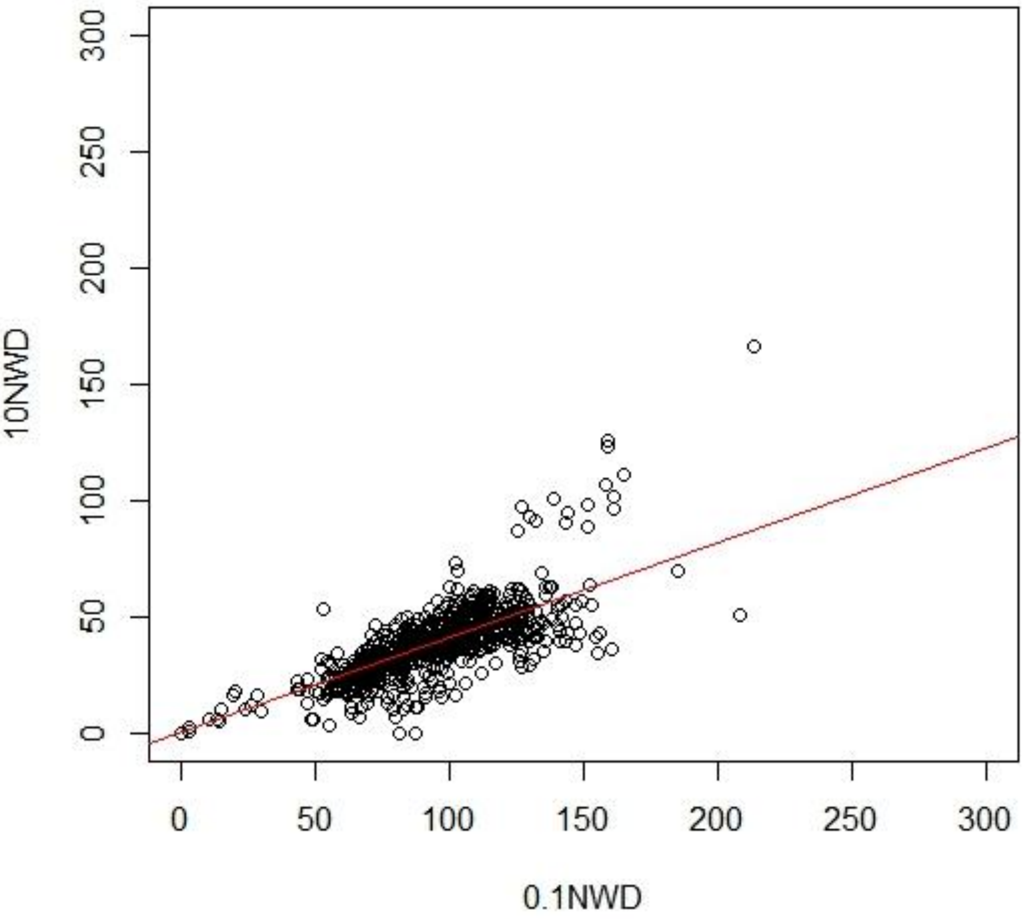


Figure 4.2. Comparison between NWDs for thresholds of 0.1mm and 10mm ($R^2_{adj} = 0.937, slope = 0.4096$).

Focusing on a single grid point, we define:

- A vector containing the individual values of average annual precipitation, indicated with P_a;
- A vector containing the number of rainy days of any year, indicated with nwd;

- An additional vector represented by I_a contains average annual intensity obtained by dividing each element of P_a with nwd.

If gumbel distribution is used, we can estimate the parametric distribution with the following expressions:

$$\alpha_G = E(I_a) \tag{4.11}$$

$$\lambda = E(nwd). \tag{4.12}$$

When Pareto distribution is being refered, the estimate becomes a little more complicated by the fact that an additional parameter is present. Consider a variable y representing average annual rainfall and the variable x indicating precipitation corresponding to a single event (daily scale). The calculation is done in the following manner:

$$\mu_y = E(P_a) \tag{4.13}$$

$$\sigma_y^2 = var(P_a). \tag{4.14}$$

From which coefficient of variation becomes

$$CV_y = \frac{\sigma_y}{\mu_y} \tag{4.15}$$

considering the rules of the compound Poisson process, for which

$$E(y) = E(nwd)E(x) \tag{4.16}$$

$$var(y) = E(nwd)var(x) + E(x)^2 var(nwd) \tag{4.17}$$

and using the definition of coefficient of variation, the following expressions are obtained which allow to calculate the mean and coefficient of variation of the variable x:

$$\mu_x = \lambda\mu_y \tag{4.18}$$

$$CV_x = \sqrt{\lambda CV_y^2 - \frac{\sigma_{nwd}^2}{\lambda}} \tag{4.19}$$

The expression (5.16) can simplify, if the number of rainy days per year can be considered as a variable of poisson distribution, in the following way

$$CV_x^2 = \lambda CV_y^2 - 1. \tag{4.20}$$

At this point, the mean and the coefficient of variation of the pareto distribution obtained from (4.8) and (4.9) are equal to the ones obtained with the expressions (4.18) and (4.19), we can estimate the parameters of the distribution in the following way:

$$k = \frac{1}{2CV_x^2} - \frac{1}{2} \tag{4.21}$$

$$\alpha_P = \mu_x(1 + k). \tag{4.22}$$

4.4. Results obtained with two different distributions

The depth of daily rainfall is estimated for a given return period by applying the gumbel and Pareto distributions, and results are compared with the values provided by the fitting the same distribution on actual data provided by Pakistan Meteorological Department (PMD). The analysis for a number of stations in the northern areas of Pakistan, for the return periods of 2, 5, 10, 20, 50, 100, 500 years was done. The values obtained by fitting the distributions on actual data were taken as a reference to evaluate the goodness of the estimates made by the procedures described in the previous paragraph. The depth of rainfall we obtained reasonably agreed with what we obtained from actual data. The estimates made by the Pareto distribution overestimated the reference values. This overestimation may be caused by the extra parameter in Pareto distribution which can introduce more uncertainty during the estimation. The comparison of the values for the only northern Pakistan (Figure 4.3) shows that the values estimated by the Gumbel distribution are actually closer to the value obtained by the actual data.

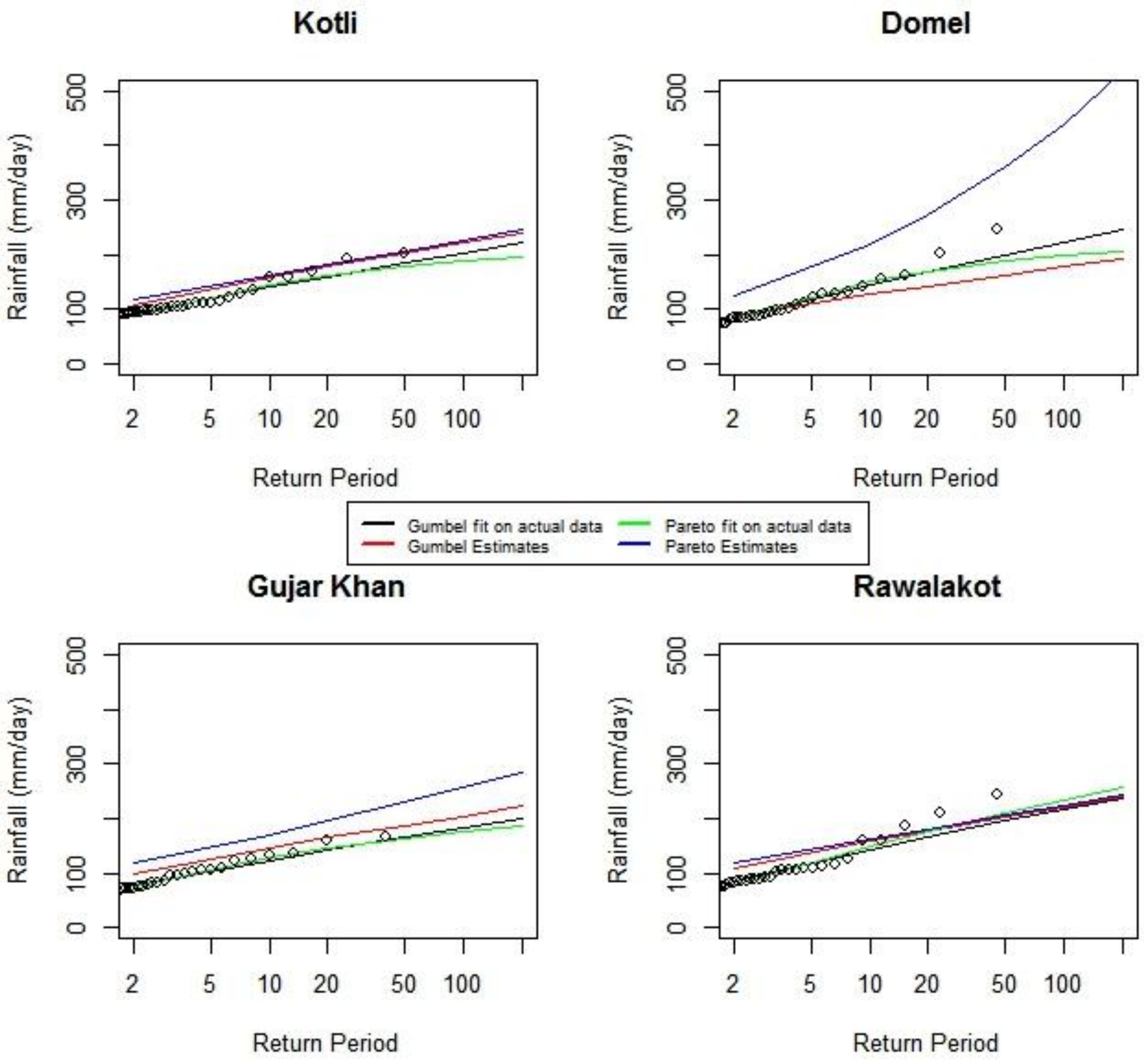


Figure 4.3. Comparison of magnitude of an event after a certain return period predicted by actual data and downscaled data.

Due to high intensity of rainfall in Pakistan very large rainfall depths in short durations occur every now and then especially in monsoon season. As a result, massive flooding can cause a very serious damage to hydraulic structures, property, disrupting communication system and ultimately causing a loss of human lives. To estimate the probability of occurrence of an extreme event for a certain year we prepare the growth curve for the region.

For a long return period (50, 100, 200, 500 etc), the growth curves are extended by using classical methods for grouping extreme precipitation values and specially treated annual maximas of each raingauge stations located across the study area. By doing this, an additional parameter called growth factor is defined as a ratio of T-year extreme values and an index extreme value generally taken as mean of at-site annual maximas. *Reed et al.*, 1999 suggested that a data record of 10 or more years is sufficient to obtain a reliable value of index variable.

4.5. Regional Analysis of Annual Maxima

The reliable climatological extremes can be estimated for a long return period by combining information from several sites [e.g. *NERC*, 1975; *Buishand*, 1989; *Cunnane*, 1989]. Jenkinson summed up the results of a long-standing regional procedure for precipitation frequency estimation in the Flood Studies Report [*NERC*, 1975]. The defined technique adopted an index variable approach: in this approach an extreme precipitation event of a certain duration D-hour for a specific T-year return period is synthesized as the product of 5-year values (represented by M5) and growth factor which replicates the ratio of T-year and 5-year value. Initially, they divided the whole work space into two regions and did broad regionalization by preparing growth curve. They further introduced an additional regionalization by pooling data according to M5 for growth curve estimation. The application of this complex regionalization scheme is difficult and the properties are too complicated to analyze. The raingauge stations are usually pooled according to geographical location with an intention of pooling stations having similar variable values, long-term annual rainfall, maximum annual rainfall, etc.

By adopting fixed geographical regions *Dales and Reed* [1989] derived a regional rainfall frequency model by analysing grouped annual maximas. They fit a Generalized Extreme Value (GEV) distribution on annual maximas by the method of regional Probability-Weighted Moments (forerunner of regional L-moment methods). Following iterative approach *Schaefer* [1990] formed regions by dividing the whole workspace and pooling together the sites having similar mean annual precipitation. He fits the very flexible 5-parameter Wakeby distribution by using regional PWM values, to make a choice between competing 3-parameter distributions fitted on the data. Schaefer adopts the GEV distribution, after the exploratory analysis, and applies it for the regionalization procedure by decreasing the variability and skewness in annual maxima with a typical increase in mean annual precipitation. *Buishand* [1991] again adopting a GEV distribution, introduces a regional analysis based on maximizing a joint likelihood function across all rainguage stations in a region. He further concluded that due to unbaised nature of PWM towards the assumed distribution, it can provide better estimates for regional parameter.

Hosking and Wallis [1997] used L-moment methods to introduce a comprehensive approach to regional frequency analysis based. L-moment ratios are derived for each station in the workspace and then regionalized by averaging and these averages are weighted depending on recorded data length. The process can be visualized by superimposing the regionally-fitted distribution and individual extreme-value plots for each station in one diagram. The precipitation data for each site is standardized by an index value which is generally the mean of the annual maxima. Additional smoothing that is helpful in stabilizing estimates for very long return periods, can be induced by pooling data from a larger number of stations, by selecting a larger region. However, the grouping of large number of sites and using regional growth curve for regionalization procedure can incur biasness through excessive generalization.

The selection of appropriate regions is therefore a problem. *Guttman* [1993] based on an exploratory data analysis adopts fixed regions approach. Using fixed region approach might be convenient but it can induce discontinuities in growth factor estimates on the boundaries of the regions except when the boundaries of regions are marked by major topographical divides. The problem of discontinuity on the boundaries can be resolved by using flexible regions approach but the approach is less convenient since the growth curve cannot be summarized in a map or table [*Reed and Stewart,* 1989; *Burn,* 1990]. *Rossi and Villani* [1994] use regional methods for flood estimation by exploiting physical mean to transfer information from rainfall to flood by adopting fixed regions. Their approach in principle can be applied to flexible regions.

Another widely accepted approach for the preparation of Growth Curve is the station-year approach. This approach is used to estimate long return period extreme events by completely avoiding or atleast reducing the need for extrapolation of a fitted distribution on data of daily annual maxima. Instead, the approach combines yearly rainfall records from individual sites, to form a single vector containing record equal in length to the sum of record lengths at an individual site in the workspace. Once the annual maxima have been standardized by division using an index variable (e.g., average of at site annual maxima) the approach, which is similar to other methods of pooling, assumes that the distribution of extremes is identical at each site. The station-year method, further, assumes that the at-sites rainfall records are mutually independent. This assumption only holds if the sites are very widely scattered or if the periods of data record are mutually exclusive. To ensure mutual exclusiveness among the stations using station-year method, a typical procedure is to avoid pooling data from nearest neighbor so that excessive dependence in the pooled extremes can be avoided. *Reed and Stewart* [1989] noticed that arranging the different data from different sites into one vector whose length is equal to sum of individual at-site record length introduces less bias than expected initially. The reason is that the implementation of the method provides an auxiliary to include extreme events occurring at more than one site to contribute to the analysis more than one station-year point.

4.5.1. Homogeneity Test

To assess whether a proposed group of sites intending to be pooled together are homogeneous or not; an examination of homogeneity is applied. The tests to check the homogeneity of pooling groups involve a concept generally based on a statistic that cognates to the formulation of a frequency distribution model, e.g. their L-moment parameters [*Dewar and Wallis,* 1990; *Hosking and Wallis,* 1997] or of dimensionless quantiles against a certain return period such as the 10-year event [*Mimikov and Gordios,* 1989; *Kuczera,* 1982], the coefficient of variation [*Lettenmaier,* 1985; *Viglione,* 2010] and/or skew coefficient.

Hosking and Wallis [1993, 1997] proposed homogeneity tests based on L-moment ratios. They defined two indexes which are widely used in flood frequency analysis such as 1)- H1 based on L-CV alone and it is recommended by these authors for having better power to discriminate between homogeneous and heterogeneous regions; 2)- H2 founded on L-CV & L-skewness jointly. The underlying idea *Hosking and Wallis* [1993] heterogeneity statistics is to compare the variation that would be expected in a homogeneous region to the measured sample variability of the L-moment ratios. The former is simulated from a kappa distribution of four parameters through repeated simulations of homogeneous regions by drawing samples from the fitted distribution [see e.g., *Hosking and Wallis,* 1997,pp. 202-204]. Very recently, *Viglione et al.,* [2007] reviewed several homogeneity tests and stated that H1 test is ahead of all others when the L-skewness is lower than 0.23. They further concluded that the H2 as a homogeneity test lacks power. These findings certainly indicate that the heterogeneity among the sites in a group is mainly due to variations in the sample L-CVs.

4.5.2. Choice of distribution

The log-normal distribution and generalized extreme value distribution are commonly used in the field of hydrology as reported by *Sevruk and Geiger* [1981]. They further advised that the research should be carried out to find out the physical reason to choose the distribution to be fitted on the data. *Pegram and Adamson* [1988] came up with an analysis called two-component approach for annual maximum rainfall in subtropical climates [e.g., *Arnell and Gabriele*, 1988]. The approach is based on the assumption that annual maxima of rainfall come from a mix of populations and of its two components, one represents normal extreme events and the other represents exceptional extreme events. Ideally, based on physical reasoning and concurrent weather observations (e.g. rainfall), annual maxima are defined prior to the analysis. The authors (Pegram and Adamson) were unable to define physical reasons for the selection of distribution but they gave physical interpretation to the spatial variation occurring in annual maxima. In our work gumbel, GEV, Log-normal, general pareto and exponential distributions were fitted on normalized annual maxima data by the method of moment fitting [e.g., *Griffiths and Pearson*, 1993; *Smithers*, 1996; *Hosking*, 1990 or *Hosking and Wallis*, 1997]. The goodness of fit was tested by procedure defined by Laio [2004].

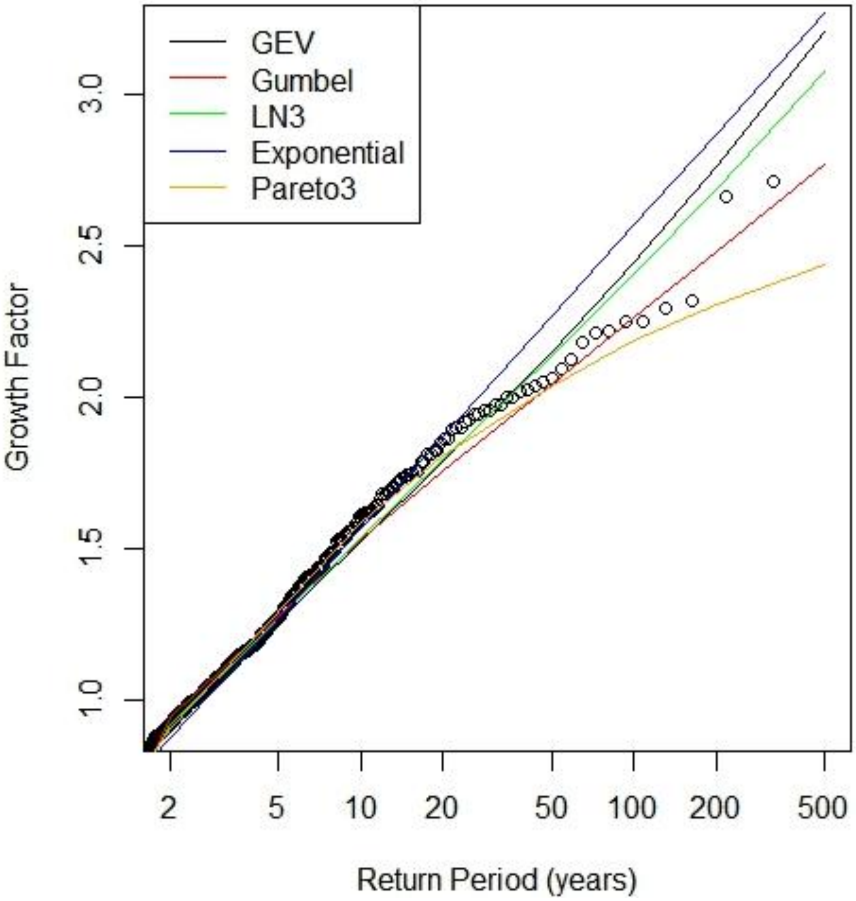


Figure 4.4. Distributions fitting on actual annual maxima.

4.5.3. Goodness of fit test

The use of goodness of fit tests based on Anderson-Darling statistics has classically been discussed based on the hypothesis that the sample of rainfall data belongs to a distribution, say F_H, whose parameters(shape, scale etc) are unspecified. The approach gets lethargic since for each hypothetical distribution F_H, the criticl region of the test is to be redetermined. Laio 2004 overcame this challenge by adopting a transformed procedure which generates a new test statistic, independent of F_H. The transformation procedure is carried out by calculating three coefficients (location, scale and shape) which are determined by fitting the empirical distribution function on the data using the asymptotic theory of tests. Thus standard coefficients summarized in a single table are sufficient for carrying out the fitness test of different hypothetical distributions.

In our work the set of probability models considered here, include 1)- extreme value 1 and 2; 2)-normal and lognormal; 3)-generalized extreme value; 4)- three-parameter gamma; and 5)-log-Pearson type 3. The parameters of the fitted distribution were calculated by the maximum likelihood method in every case. A series of Monte Carlo simulation experiments is used to assess the performance of these tests through calculation of predefined statistics A^2.

When the parameters of the hypothetical distribution are unknown, which is the most common case in hydrology; the case will be referred to as *case* p, while it will be referred to as *case* 0 when the parameters are fully specified a priori [*Stephens*, 1986].

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\left[(2i-1)ln[F(x_i,\theta)] + (2n+1-2i)ln[1-F(x_i,\theta)]\right] \qquad (4.23)$$

where F defines the family of distributions (such as normal or gamma) and θ is a vector of parameters. Here we define a variable ω, whose distribution is close to that of the case 0 of Anderson-Darling statistics.

$$\omega = \beta_0\left(\frac{Q_p^2-\xi_p}{\beta_p}\right)^{\frac{\eta p}{\eta 0}} + \xi_0, \text{ If } 1.2\xi_p \le Q_p^2 \text{ else,} \qquad (4.24)$$

$$\omega = \left[\beta_0\left(\frac{0.2\xi_p}{\beta_p}\right)^{\frac{\eta p}{\eta 0}} + \xi_0\right]\frac{Q_p^2-0.2\xi_p}{\xi_p}, Q_p^2 \qquad (4.25)$$

where ξ, β and η are location, scale and shape parameters, respectively. The second equation in equation (4.25) is necessary for keeping ω on the real positive axis also when $Q_p^2 < \xi_p$, and to improve the accuracy of the transformation in the very low part of the distribution. This correction is seldom required, since the probability of having $Q_p^2 < \xi_p$ is below 0.12 [*Laio*, 2004].

Adopting the described methods [see *Laio,* 2004], the case 0 coefficients were as follows:

$\xi_0 = 0.0403$ $\qquad \beta_0 = 0.116$ $\qquad \eta_0 = 0.851.$

While the parameters for case p are set to be

$\xi_p = 0.188$ $\qquad \beta_p = 0.281$ $\qquad \eta_p = 1.111.$

Once the transformation (11) is adopted, ω becomes the new test statistic and the null hypothesis is rejected if ω is greater than 0.347, 0.461, and 0.743 for significance levels α = 0.10, 0.05, and 0.01, respectively.

Table 4.1. Coefficients to be set in equation (4.24/4.25) for the Anderson-Darling Statistic

Distribution	ξ_p	β_p	η_p
EV1, EV2	0.169	0.169	0.169
NORM, LN	0.167	0.167	0.167
GEV	0.147(1+0.13k+0.21k^2+0.09 k^3)	0.147(1+0.13k+0.21k^2+0.09 k^3)	0.147(1+0.13k+0.21k^2+0.09 k^3)
GAM, LP3	0.145(1+0.17k^{-1}+0.33k^{-2})	0.145(1+0.17k^{-1}+0.33k^{-2})	0.145(1+0.17k^{-1}+0.33k^{-2})

Here k is an asymptotic efficient estimator (usually maximum likelihood) of the shape parameter of the distribution.

The Anderson-Darling test for the four distributions in Table 4.1 can proceed as follows: (1) Estimate the parameters from the sample data using maximum likelihood (ML) or *Smith's* [1985] estimators when necessary. (2) Sort the data in ascending order, find $F(x_i, \theta)$ for any distribution, and calculate A2 from equation (4.23). (3) Determine the case p coefficients using already defined ξ_p, β_p and $\eta_p,$ with the appropriate sample size and k values. (4) Find ω from equations (4.24 or 4.25), using the case 0 coefficients. (5) Compare ω to the appropriate percentage points for the selected significance level (e.g., 0.461 for α = 0.05).

Based on defined Goodness of fit procedure, GEV distribution found to be well fitted on the data having $\omega = 0.321$ for significance level $\alpha = 0.10$.

Once the average values of the regionalized maximum annual rainfall for durations of 1, 3, 6, 12 and 24 hours is determined, it is now possible to calculate the probability curve of average rainfall, for each point of the domain of study. Furthermore, in order to statistically characterize the rains extreme, it is also necessary to determine the indexes which allowed associating a certain event to a return period depending on the probabilistic formulations adopted.

The probability curve allows you to evaluate the maximum depth of rain for an assigned duration that can occur in a certain area and is usually expressed by means of the formula

$$h(d) = ad^n \tag{4.26}$$

The depth of rainfall h is expressed as a function of the duration d of the event than two parameters:

- The coefficient of rainfall duration “a”, which is the average depth of rain fall in a time interval equal to one hour;

- The exponent of scale invariance “n”, which governs the shape of the curve and the extent to which it depends on the duration of the precipitation.

By using equation of $h(d)$ and the growth curve, a depth-duration frequency model is developed which allows for the estimation of point rainfall frequencies for a range of durations for any location in Pakistan. The model consists of an index (mean of maximum) rainfall and a log-logistic growth curve which provides a multiplier of the index rainfall. Since the original data is not spatially adequate, therefore the parameters "a" and "n" are found by TRMM data (see Figure 4.5 and 4.6). The data is downloaded across the entire length of the country on 0.5*0.5 degree grids (see Figure 4.7).

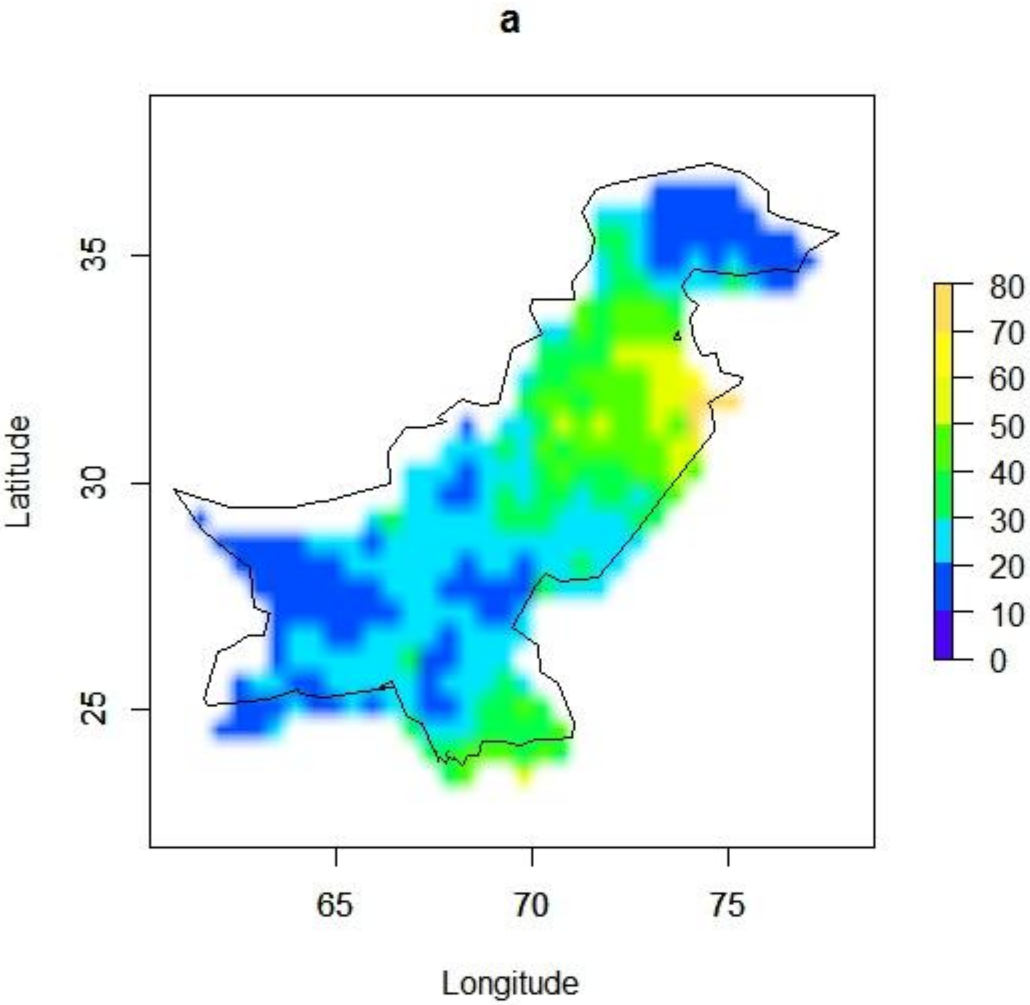


Figure 4.5. Coefficient "a" for the maximum annual precipitation for durations of 1, 3, 6, 12 and 24 hours

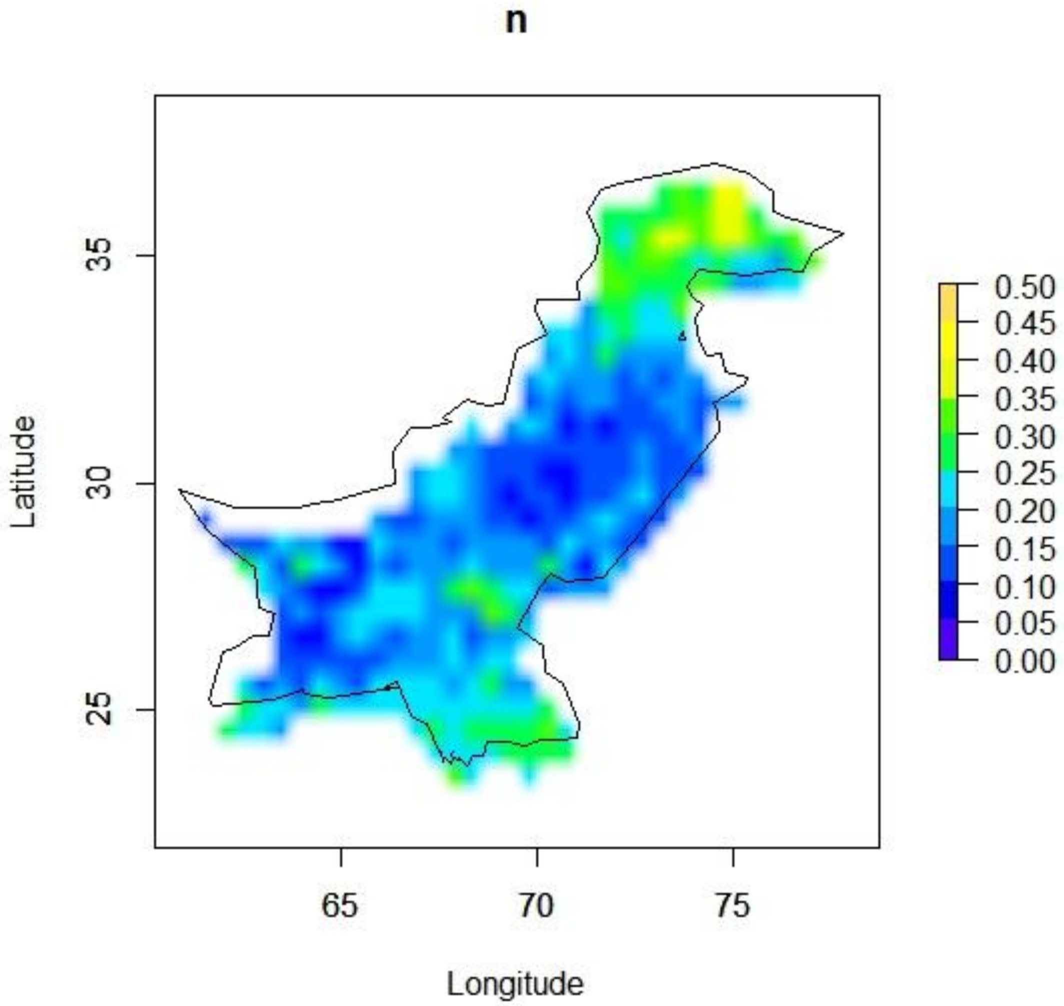


Figure 4.6. Coefficient "*n*" for the maximum annual precipitation for durations of 1, 3, 6, 12 and 24 hours

4.6. Results of mapping rainfall extremes for a certain return period

The values of rain depths for the stations located in northern areas of Pakistan were compared against the values obtained by the procedure previously described, for a return period of 200 years. The distribution fit on original data calculated the depth values of rain at the specific station for a certain return period and duration. These values were taken as a reference to evaluate the goodness of estimates made by the procedures described previously and a comparison in shown the table 4.2.

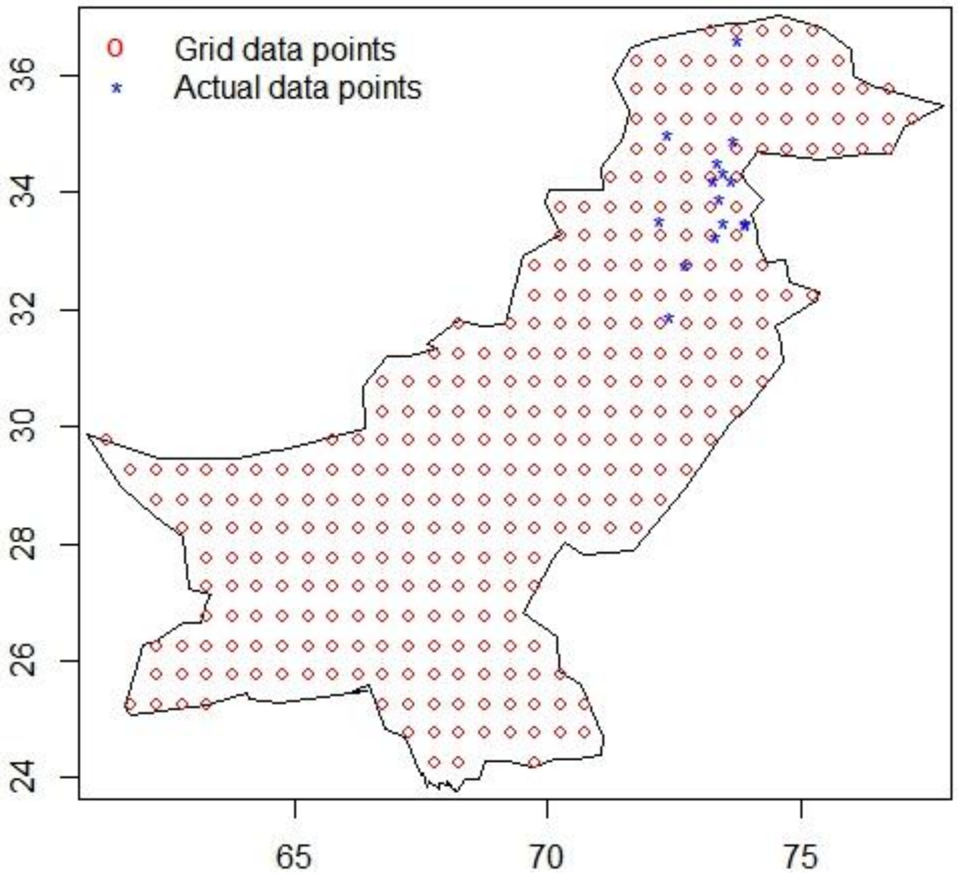


Figure 4.7. TRMM and raingauge data points identification across Pakistan

The rainfall estimates are made only at a certain point in the workspace whereas the satellite data is executed by averaging rainfall amount over a certain area. The translation of area-averaged information to point-information does not come without the introduction of unavoidable error. It is therefore important to introduce a correctness factor to improve the estimates of TRMM data, overland. In our work, to improve the estimates of TRMM data a comparison of yearly mean values (μ) with the actual rainguage data from ground stations was done (ground truthing). The slope of regression line provided the "correctness-factor" (see figure 4.8).

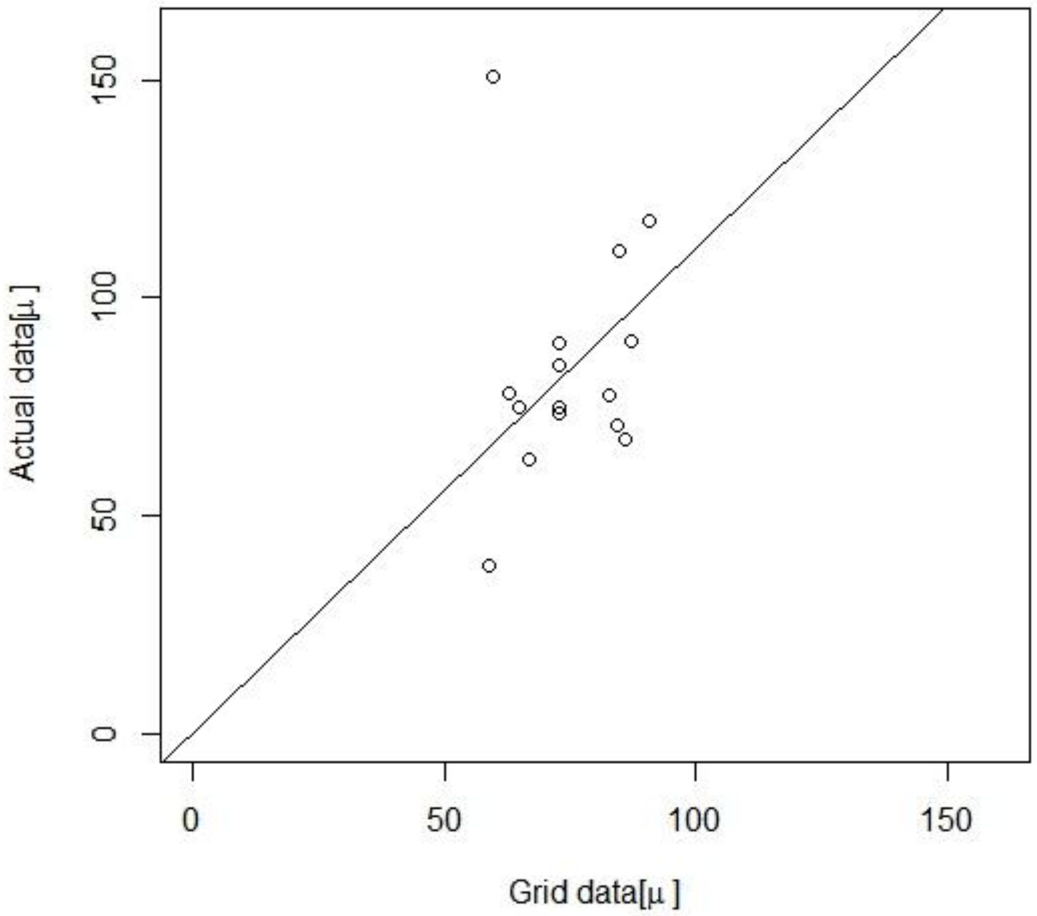


Figure 4.8. Estimating multiplication factor to correct area-averaged rainfall information to point-information.

Table 4.2. comparison of simulated values and values obtained by fitting GEV distribution in original data

Distribution	Distribution fit on actual data	Simulated values
Balakot	403.89	184.26
Gharidopatta	224.08	222.32
Kotli	264.14	258.57
Murree	378.90	263.27
Muzzafarabad	263.21	264.00
Bagh	217.63	222.32
Domel	269.97	265.57
Gujar Khan	221.67	251.16
Kallar	242.10	198.37
Mangla	206.50	203.51
Naran	187.70	182.81
Rawalakot	302.76	258.57
Rehman Br.	234.47	258.00
Khandar	317.86	233.40
Sehr. Kokata	231.95	222.32

The results evaluated for rainfall extremes for 200 years return period are summed up in Table 4.2 and a raster map (see figure 4.9). A certain level of fluctuations can be observed in some cases but overall the level of agreement between the results can be considered good. Note the difference between actual and simulated values at Balakot due to unexpected extreme rainfall events under the influence of monsoon spell in summer and winter precipitation caused by western circulations [*Tahir et al.,* 2015].

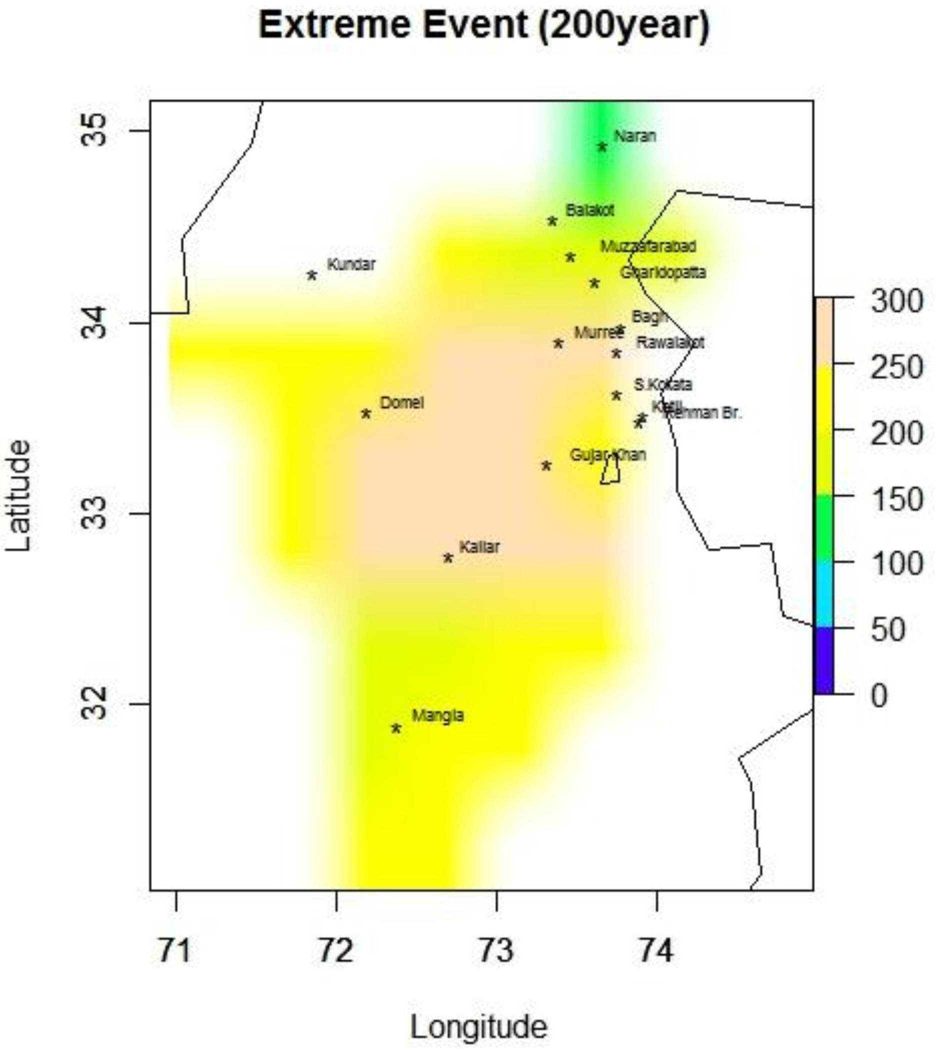


Figure 4.9. Raster map of extreme events for 200 years return period.

4.7. Conclusion

The rainfall extremes are tackled from two different prospective. Firstly, we developed a probabilistic model for downscaling monthly rainfall data into daily extremes. The probabilistic models used here are based on fitting GPD (Gumbel and Pareto) to the monthly values of precipitation. The procedure was applied to precipitation data from 15 stations located in northern part of the country. Before performing the fitting procedure a threshold value is selected to count the number of rainy events occurring in each year.

Secondly, we use the method of L-moments to estimate the magnitude of rainfall extremes for a certain period with a GEV model. The risk maps are drawn for the northern area of Pakistan. The parameters "a" and "n" are estimated using TRMM data extracted at 0.5*0.5 resolutions across the entire length of the country. Later, a growth curve for northern homogenous region, containing 15 raingauges, is executed. A multiplication factor was derived to account for the error that might have caused when transferring satellite data to point data of raingauge.

The CV of the results obtained for both cases showed that the estimated results are in good agreement when GEV distribution was fitted on actual daily data and actual daily annual maxima for a certain return period. The estimates are slightly less accurate for the areas where extreme rainfall events are unpredictable due to seasonal variations (e.g., Balakot).

References:

Archfield, S. A., Pugliese, A., Castellarin, A., Skøien, J. O., and Kiang, J. E.: Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach?, *Hydrol. Earth Syst. Sci.*, 17, 1575-1588, doi:10.5194/hess-17-1575-2013, 2013.

Arnell, N. W., and Gabriele, S. 1988. The performance of the two component extreme value distribution in regional flood frequency analysis. *Water Resour. Res*., 246, 879–887.

Austin, P.M.; Houze, R.A. Jr. 1972: Analysis of the structure of precipitation patterns in New England. *J. Appl. Meteor.* 11,926-935

Barthel F, Neumayer E (2012) A trend analysis of normalized insured damage from natural disasters. *Clim Change* 113:215–237

Beguería, S., and A. Lorente, 1999: Distribución espacial del riesgo de precipitaciones extremas en el Pirineo aragonés occidental. *Geographicalia*, 37, 17–36.

BEH (2012) World risk report 2012. United Nations University (EHS). *Bundnis Entwicklung Hilft*, Berlin

Benson, M. A., 1960: Characteristics of frequency curves based on a theoretical 1000-year record. Flood Frequency Analysis, T. Dalrymple, Ed., U.S. *Geological Survey Water Supply Paper* 1543-A, 51–77.

Blöschl, G.: Rainfall-Runoff Modeling of Ungauged Catchments, edited by: Anderson, M. G., *Encyclopedia of Hydrological Sciences*, 2005.

Bouwer LM (2013) Projections of future extreme weather losses under changes in climate and exposure. *Risk Anal.* 33(5):915–930

Bower, D., Hannah D.M. (2002) Spatial and temporal variability of UK river flow regimes, *FRIEND 2002- Regional Hydrology: Bridging the Gap between Research and Practice* (Proceedings of Cape Town Conference). IAHS Publication, 274, 457-464.

Brunetti, M., M. Mangueri, T. Nanni and A. Navarra, 2002: Droughts and extreme events in regional daily Italian precipitation series. *Int. J. Climatol.*, 22, 543–558.

Buishand T.A., 1989. Statistics of extremes in climatology. *Statistica Neerlandica* 43, 1-30.

Buishand T.A., 1991. Extreme rainfall estimation by combining data from several sites. *Hydrological Sciences Journal* 36, 345-365.

Buishand, T. A.: The analysis of homogeneity of long-term rainfall records in The Netherlands, R. Neth. *Meteorol. Inst. (K.N.M.I.), De Bilt, Sci. Rep. No. 81-7*, 77 pp., 1981.

Burn, D. H. (1990), Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resour. Res.*, 26(10), 2257 – 2265.

Burn, D. H. (1990), Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resour. Res.,* 26(10), 2257 – 2265.

Burn, D.H., Boorman, D.B., 1993. Estimation of hydrological parameters at ungauged catchments. *J. Hydrol.* 143, 429–454.

Burr, W., *Cumulative frequency functions.* The Annals of Mathematical Statistics, 13(2):215-232, Jun. 1942.

Carrillo, G., Troch, P. A., Sivapalan, M., Wagener, T., Harman, C., and Sawicz, K.: Catchment classification: hydrological analysis of catchment behavior through process-based modeling along a climate gradient, *Hydrol. Earth Syst. Sci.*, 15, 3411-3430, doi:10.5194/hess-15-3411-2011, 2011.

Castellarin, A., G. Camorani, and A. Brath (2007), Predicting annual and long-term flow-duration curves in ungauged basins, *Adv. Water Resour.*, 30(4), 937 – 953.

Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2014). "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.", *"Journal of Statistical Software, 61(6), 1-36."*, "URL http://www.jstatsoft.org/v61/i06/".

Claps, P., and M. Fiorentino (1997), Probabilistic flow duration curvers for use in environmental planning and management, in *Integrated Approach to Environmental Data Management Systems, NATO ASI Ser., Partnership Subser. 2,* vol. 31, edited by N. B. Harmancioglu et al., pp. 255-266, Kluwer, Dordrecht, Netherlands.

CUBIST Team (2007), CUBIST project: Characterisation of ungauged basins by integrated use of hydrological techniques, *Geophys. Res. Abstr.*, 10, 12048, sref:1607-7962/gra/EGU2008-A-12048.

Cunnane, C., 1989: Statistical distributions for flood frequency analysis. *World Meteorological Organization, Operational Hydrology Report No.33, WMO Publ. No.718, Geneva*.

Dales, M. Y. & Reed, D. W. (1989) *Regional flood and storm hazard assessment*. Institute of Hydrology Report no. 102, Institute of Hydrology, Wallingford, Oxfordshire, OX10 8BB, UK.

De Girolamo, A.M., Calabrese, A., Lo Porto, A., Oueslati, O., Pappagallo, G., Santese, G. (2011), Hydrologic regime characterization for a semi-arid watershed, 39-45. *In Bodenkultur* 62 (1-4).

Discovery. 44–53.

Farr, T. G., et al. (2007), The Shuttle Radar Topography Mission, *Rev. Geophys.*, 45, RG2004, doi:10.1029/2005RG000183.

Felgate, D. G., and D. G. Read, Correlation analysis of the cellular structure of storms observed by rain gauges, *J. Hydrol.*, 24, 191-200, 1975.

Fennessey, N. and Vogel, R. M.: Regional Flow-Duration Curves for Ungauged Sites in Massachusetts, *J. Water Resour. Plan. Manage.-ASCE*, 116, 530–549, 1990.

Fennessey, N. M. (1994), *A hydro-climatological model of daily streamflows for the northeast United States*, Ph.D. dissertation, Tufts Univ., Medford, Mass.

Gajic-Capka, M., 1991: Short-term precipitation maxima in different precipitation climate zones of Croatia, Yugoslavia. *Int. J. Climatol.*, 11, 677–687.

Gallart, F., Amaxidis, Y., Botti, P., Cane, G., Castillo, V., Chapman, P., Froebrich, J., Garcia-Pintado, J., Latron, J. & Llorens, P. (2008), Investigating hydrological regimes and processes in a set of catchments with temporary waters in Mediterranean Europe. *Hydrol. Sci. J.* 53, 618–628.

Ganora, D., Claps, P., Laio, F., Viglione, A., An approach to estimate non-parametric flow duration curves in ungauged basins. *Water Resources Research,* 45, OCT 15 2009. ISSN 0043-1397. doi: 10.1029/2008WR007472.

Ganora, D., Gallo, E., Laio, F., Masoero, A., and Claps, P., *Analisi idrologiche e valutazione del potenziale idroelettrico dei bacini piemontesi.* ISBN: 978-88-96046-07-4, Progetto RENERFOR, 2013.

Ganora, D., P. Claps, F. Laio, and A. Viglione (2009), An approach to estimate nonparametric flow duration curves in ungauged basins, *Water Resour. Res.*, 45, W10418, doi:10.1029/2008WR007472.

García-Ruiz, J. M., J. Arnáez, S. M. White, A. Lorente, and S. Beguería, 2000: Uncertainty assessment in the prediction of extreme rainfall events: An example from the central Spanish Pyrenees. *Hydrol. Process.*, 14, 887–898.

Griffiths, G.A.; Pearson, C.P. 1993. Distribution of high intensity rainfalls in metropolitan Christchurch, New Zealand. *Journal of Hydrology (NZ)* 31(1): 5-22.

Hessami, M., Gachon, P., ourda, T. B. M. J., and St-Hilaire, A. (2007) Automated regression based statistical downscaling tool. *Environmental Modeling and Software , Science Direct*.

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

Hosking, J., and J. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-Moments,* Cambridge Univ. Press, New York.

Hosking, J.R.M., Wallis, J.R., 1993. Some useful statistics in regional frequency analysis. *Water Resources Research* 29 (2), 271–281.

Hosking, J.R.M., Wallis, J.R., 1997. *Regional frequency analysis –an approach based on L-moments*. Cambridge University Press, New York, p. 224.

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J.E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., and Cudennec, C., 2013. A decade of Predictions in Ungauged Basins (PUB)—a review. *Hydrological Sciences Journal,* 58 (6), 1198–1255.

Huffman,G.J.,R.F. Adler, D.T. Bolvin, G. Gu, E.J. Nelkin, Y. Hong, and D.B. Wolff, 2007: The TRMM multisatellite precipitation analysis (TMPA): quasi-global, multiyear,combined-sensor precipitation estimates at fine scales. *J. Hydrometeor*, 8, 38-55.

Iacobellis, V. (2008), Probabilistic model for the estimation of T year flow duration curves, *Water Resour. Res.,* 44, W02413, doi:10.1029/ 2006WR005400.

In Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge
IPCC, 2007: Climate Change 2007: *Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change,* M.L. Parry, O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson, Eds., Cambridge University Press, Cambridge, UK, 976pp.

J. Parajka, V. Andréassian, S.A. Archfield, A. Bárdossy, G. Blöschl, F.H.S. Chiew, Q. Duan, A.N. Gelfan, K. Hlavčová, R. Merz, N. McIntyre, L. Oudin, C. Perrin, M Rogger, J. Salinas, H. Savenije, J. Skøien, T. Wagener, E. Zehe, Y. Zhang: "*Predictions of Runoff Hydrographs in Ungauged Basins*"; in: "*Runoff Prediction in Ungauged Basins - Synthesis across Processes, Places and Scales*", G. Blöschl, M. Sivapalan, T. Wagener, A. Viglione, H. Savenije (ed.); Cambridge University Press, 2013, (invited), ISBN: 978-1-107-02818-0, 227 - 360.

Jones, J. A. A., 1997, *Global hydrology: processes, resources and enviromental management.* Addison Wesley Longman, 399 pp

Karim, F., Petheram, C., Marvanek, S., Ticehurst, C., Wallace, J. and Gouweleeuw, B. (2011). *The use of hydrodynamic modelling and remote sensing to estimate floodplain inundation and flood discharge in a large tropical catchment*, in Chan, F., Marinova, D. and Andersson, R.S. (Eds.), *19th International Congress on Modelling and Simulation*, Perth, December 12- December 16, 2011, pp.3796-3802.

Karl, T. R., R. W. Knight and N. Plummer, 1995: Trends in high-frequency climate variability in the twentieth century. *Nature,* 377, 217–220.

Konz, M., Finger, D., Buergi, C., Normand, S., Immerzeel WW, Merz J, Giriraj A, Burlando P. 2010. Calibration of a distributed hydrological model for simulations of remote glacierized Himalayan catchments using MODIS snow cover data. *In: Global Change: Facing Risks and Threats to Water Resources.* Proceedings of the Sixth World FRIEND Conference, Fez, Morocco, 25–29 October 2010. IAHS [International Association of Hydrological Sciences] Publication 340. IAHS, pp 465–473. Available at: http://www.futurewater.nl/wp-content/uploads/2011/05/FRIEND2010_konzetal.pdf; accessed in September 2011.

Korn, F., Muthukrishnan, S. 2000. Influenced sets based on reverse nearest neighbor queries. *In SIGMOD.* 201–212.

Kottegoda, N. T., and R. Rosso (1997), *Statistics, Probability, and Reliability for Civil and Environmental Engineers,* McGraw-Hill, New York.

Kottegoda, N.T., and R. Rosso (1997), *Statistics, probability and reliability for civil and environmental engineers*, Par 6.2, Mc-Graw-Hill Publishing Company, New York.

Krasovskaia, I., Arnell, N. W. & Gottschalk, L. (1994) Flow regimes in northern and western Europe: development and application of procedures for classifying flow regimes. *In: FRIEND: Flow Regimes from International Experimental and Network Data* (ed. by P. Seuna. A. Gustard,

N. W. Arnell & G. A. Cole) (Proc. Braunschweig Conf., October 1993), 185–193. IAHS Publ. 221, IAHS Press, Wallingford, UK.

Kuczera, G., On the relationship of the reliability of parameter estimates and hydrologic time series data used in calibration, *Water Resour. Res.*, 18, 146-154, 1982.

Laaha, G. and Blöschl, G., 2006. Seasonality indices for regionalizing low flows, *Hydrological Processes*, 20, 3851–3878.

Laaha, G., and G. Bloschl (2006a), A comparison of low flow regionalisation methods: catchment grouping, *Journal of Hydrology*, 323, 1—4, 193—214.

Laio, F.: Cramer-von Mises and Anderson-Darling goodness of fit tests for extreme value distributions with unknown parameters, *Water Resour. Res.*, 40, W09308, doi:10.1029/2004WR003204, 2004

Lana, X., G. Fernandez-Mills, and A. Burgueño, 1995: Daily precipitation maxima in Catalonia (North-east Spain): Expected values and their spatial distribution. *Int. J. Climatol.*, 15, 341–354.

Lasanta, T., 2003: Gestión agrícola y erosión del suelo en la cuenca del Ebro: El estado de la cuestión. *Zubía*, 21, 76–96.

Lauro, M. Cartigli e faraoni d'egitto. Internet resource (in Italian), 2009.

Legendre, P. (1993), Spatial Autocorrelation: Trouble or New Paradigm?, *Ecology*, 74(6), 1659–73.

Lettenmaier, D.P., A Review of Regional Flood Frequency Estimation Methods, *Proceedings*, US/Peoples Republic of China Bilateral Symposium on the Analysis of Extraordinary Flood Events, Nanjing, China, 1985.

Lichstein, J. (2007), Multiple regression on distance matrices: A multivariate spatial analysis tool, *Plant Ecology* 188: 117-131.

Lichstein, J. (2007), Multiple regression on distance matrices: A multivariate spatial analysis tool, *Plant Ecol.*, 188(2), 117 – 131.

Loo, Y.Y., et al., Effect of climate change on seasonal monsoon in Asia and its impact on the variability of monsoon rainfall in Southeast Asia, *Geoscience Frontiers (2014)*, http://dx.doi.org/10.1016/j.gsf.2014.02.009

Lorente, A., and S. Beguería, 2002: Variation saisonniere de l'intensité des précipitations maximales dans les Pyreneées Centrales: Analyse spatiale et cartographique. *Publications de la Association Internationale de Climatologie, C. Kergomard, Ed., Vol. 14,* Association Internationale de Climatologie, 327–334.

Mantel, N. & Valand, R.S. (1970), A technique of nonparametric multivariate analysis, *Biometrics, 27*, 209 – 220.

Manton, M. J., and Coauthors, 2001: Trends in extreme daily rainfall and temperature in Southeast Asia and the South Pacific: 1961-1998. *Int. J. Climatol.*, 21, 269–284.

Mimikou, M. & Gordios, J. (1989) Predicting the mean annual flood and flood quantiles for ungauged catchments in Greece. *Hydrol. Sci. J.* 34 (2), 169-184

Montgomery, D., E. Peck, and G. Vining (2001), *Introduction to Linear Regression Analysis*, 3rd ed., John Wiley, New York.

Nadarajah, S., Kotz, S., Discussion of models for extremes using the extended three-parameter burr xii system with application to flood frequency analysis". *Hydrological Sciences Journal-Journal des Sciences Hydrologiques,* 51(6):1203-1204, 2006. ISSN 0262-6667.

Nathaniel B. Guttman, 1993: The Use of L-Moments in the Determination of Regional Precipitation Climates. *J. Climate*, 6, 2309–2325. doi: http://dx.doi.org/10.1175/1520-0442(1993)006<2309:TUOLMI>2.0.CO;2

NERC 1975. *Flood Studies Report.* Natural Environment Research Council, London (five volumes).

Neumayer E, Barthel F (2011) Normalizing economic loss from natural disaster: a global analysis. *Glob Environ Change* 21:13–24

Olden, J. D. and Poff, N. L. (2003), Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River Res. Applic.*, 19: 101–121. doi: 10.1002/rra.700

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies, *Hydrol. Earth Syst. Sci.*, 17, 1783-1795, doi:10.5194/hess-17-1783-2013, 2013.

Pechlivanidis I.G., Jackson B., McMillan H., Gupta H., 2014, 'Robust informational entropy-based descriptors of flow in catchment hydrology', *Hydrologic Sciences Journal,* doi: 10.1080/02626667.2014.983516

Pegram, G. G. S. & Adamson, P. T. (1988) Revised risk analysis for extreme storms and Hoods in Nalal/Kwazulu. *The Civ. Eng. in SA*, January, 15-20, and discussion July, 331—336.

Pellicciotti, F., C. Buergi, W. Immerzeel, M. Konz, and A. Shrestha (2012), Challenges and uncertainties in hydrological modelling of remote Hindu Kush–Karakoram–Himalayan (HKH) basins: Suggestions for calibration strategies, *Mt. Res. Dev.*, 32(1), 39–50, doi:10.1659/MRD-JOURNAL-D-11-00092.1.

Porth, L. S., D. C. Boes, R. A. Davis, C. A. Troendle, and R. M. King, 2001: Development of a technique to determine adequate sample size using subsampling and return interval estimation. *J. Hydrol.,* 251, 110–116.

Prudhomme, C., 1999: Mapping a statistic of extreme rainfall in a mountainous region. *Phys. Chem. Earth,* 24B, 79–84.

Prudhomme, C., and D. W. Reed, 1999: Mapping extreme rainfall in a mountainous region using geostatistical techniques: A case study in Scotland. *Int. J. Climatol.*, 19, 1337–1356.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Found. for Stat. Comput., Vienna

Razavi, T. and Coulibaly, P. (2013). ”Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods.” *J. Hydrol. Eng.*, 18(8), 958–975.

Reed, D. W. & Stewart, E. J. (1989) Focus on rainfall growth estimation. *Proc. 2nd National Hydrology Symposium*, Sheffield, September 1989, 3.57-3.65.

Renner, M. and Bernhofer, C.: Long term variability of the annual hydrological regime and sensitivity to temperature phase shifts in Saxony/Germany, *Hydrology and Earth System Sciences*, 15, 1819–1833, doi:10.5194/hess-15-1819-2011, 2011.

Rodriguez, R.N., A guide to the burr type xii distributions. *Biometrika*, 64 (1):129-134, Apr. 1977.

Rodriguez-Iturbe I, Mejía JM. 1974. On the transformation of point rainfall to areal rainfall. *Water Resources Research* 10: 729–735, DOI: 10.1029/WR010i004p00729.

Rossi, F., and P. Villani, A project for regional analysis of floods in Italy, in *Coping With Floods*, edited by G. Rossi, N. Harmancioglu, and V. Yevjevich, pp. 193 – 218, Kluwer Acad., Norwell, Mass., 1994.

Samaniego, L., A. Bardossy, and R. Kumar (2010), Streamflow prediction in ungauged catchments using copula-based dissimilarity measures, *Water Resour. Res.*, 46, W02506, doi:10.1029/2008WR007695.
Santiago Beguería and Sergio M. Vicente-Serrano, 2006: Mapping the Hazard of Extreme Rainfall by Peaks over Threshold Extreme Value Analysis and Spatial Regression Techniques. *J. Appl. Meteor. Climatol.*, 45, 108–124. doi: http://dx.doi.org/10.1175/JAM2324.1

Sauquet, E., Gottschalk, L. & Leblois, E., 2000. Mapping average annual runoff: A hierarchical approach applying a stochastic interpolation scheme. *Hydrological Sciences Journal*, 45 (6), 799-815.

Sauquet, E., M.-H. Ramos, L. Chapel, and P. Bernardara (2008), Streamflow scaling properties: investigating characteristic scales from different statistical approaches, *Hydrol. Processes*, 22, 3462–3475, doi:10.1002/hyp.6952.

Sevruk, B. & Geiger, H. (1981) Selection of distribution types for extremes of precipitation. *Operational Hydrology Report no 15* (WMO no. 560). WMO Geneva.

Shao, Q.X., *Notes on maximum likelihood estimation for the three-parameter burr xii distribution. Computational Statistics & Data Analysis*, 45(3):675-687, 2004. ISSN 0167-9473.

Singh, R., S. Mishra, and H. Chowdhary (2001), Regional flow-duration models for large number of ungauged Himalayan catchments for planning microhydro projects, *J. Hydrol. Eng.*, 6(4), 310 – 316.

Smith, R. L. (1985), Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, 72(1), 67 – 90.

Smithers, .J. C. (1996) Short-duration rainfall frequency model selection in southern Africa. *Water SA* 22(3), 211-217.

Smouse, P.E., J.C. Long, R.R. Sokal, Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence, *Systematic Zoology*, 35(4), 627-632, 1986.

Stanoi, I., Agrawal, D., Abbadi, A. E., 2000. Reverse nearest neighbor queries for dynamic databases.
Stephens, M. A. (1986), Tests based on EDF statistics, in *Goodness-of-Fit Techniques,* edited by R. B. D'Agostino and A. M. Stephens, pp. 97 – 194, Marcel Dekker, New York.
Syed Abu Shoaib, András Bárdossy, Thorsten Wagener, Yingchun Huang, Nahid Sultana, 2013. A different light in Predicting Ungauged Basins: Regionalization approach based on Eastern USA Catchments, *Journal of Civil Engineering and Architecture,* USA (ISSN1934-7359) Issue 3. Vol.7, March 2013, pp 364-378

Tahir, A. A., Chevallier, P., Arnaud, Y., Ashraf, M., Bhatti, M. T., (2015) Snow cover trend and hydrological characteristics of the Astore River basin (Western Himalayas) and its comparison to the Hunza basin (Karakoram region). Science of The Total Environment 505, 748-761. Online publication date: 1-Feb-2015.

Training module # SWDP – 37 (2002), *How to do hydrological data validation using regression,* [Online].Available:http://www.cwc.gov.in/main/HP/download/37%20How%20to%20do%20hydrological%20data%20validation%20using%20regression.pdf.

Viglione, A. (2007), nsRFA: Non-supervised regional frequency analysis, R package version 0.4-5, (Available at http://www.r-project.org/), R Found. for Stat. Comput., Vienna.

Viglione, A., F. Laio, and P. Claps (2007b), A comparison of homogeneity tests for regional frequency analysis, Water Resour. Res., 43, W03428, doi:10.1029/2006WR005095.

Viglione, A., Laio, F., and Claps, P.: A comparison of homogeneity tests for regional frequency analysis, *Water Resour. Res.,* 43(3), W03428, doi:10.1029/2006WR005095, 2007

Viglione, A.: nsRFA: Non-supervised Regional Frequency Analysis, R package, http://cran.r-project.org/web/packages/nsRFA/, 2010.

Ward, J. (1963), Hierarchical grouping to optimize an objective function, *J. Am. Stat. Assoc.*, 58, 236 – 244.

Weisse, A. K., and P. Bois, 2001: Topographic effects on statistical characteristics of heavy rainfall and mapping in the French Alps. *J. Appl. Meteor.,* 40, 720–740.

White, S., J. M. Garcìa-Ruiz, C. Martì, B. Valero, M. P. Errea, and A. Gûmez Villar, 1997: The 1996 Biescas campsite disaster in the Central Spanish Pyrenees, and its temporal and spatial context. *Hydrol. Process.,* 11, 1797–1812.

Wiesner, C. J., (1970): *Hydrometeorology.* Chapman and Hall, Ltd. London.

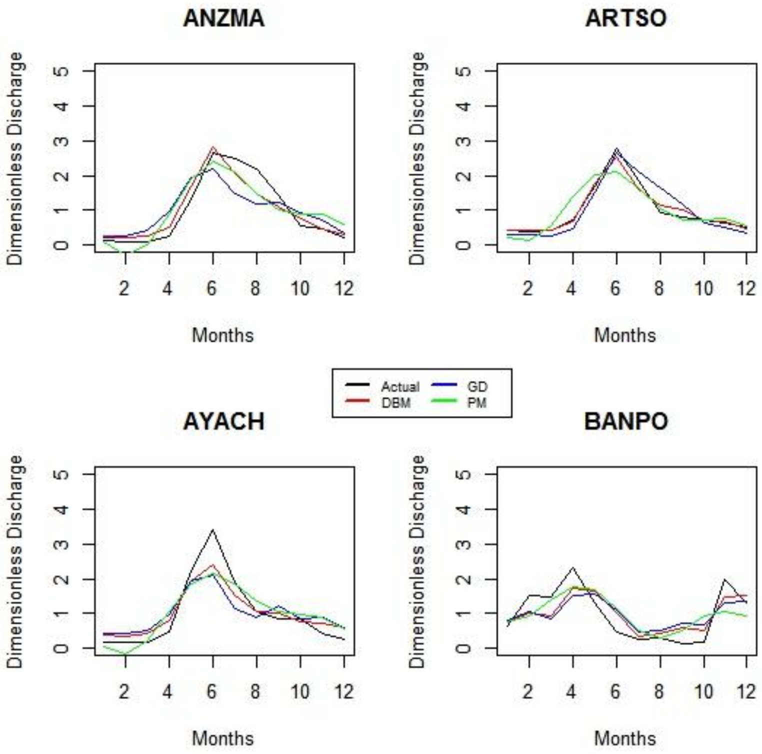
Wilby, R. L., and Dawson, C. W. (2007) SDSM 4.2 — *A decision support tool for the assessment of regional climate change impacts,*User Manual.

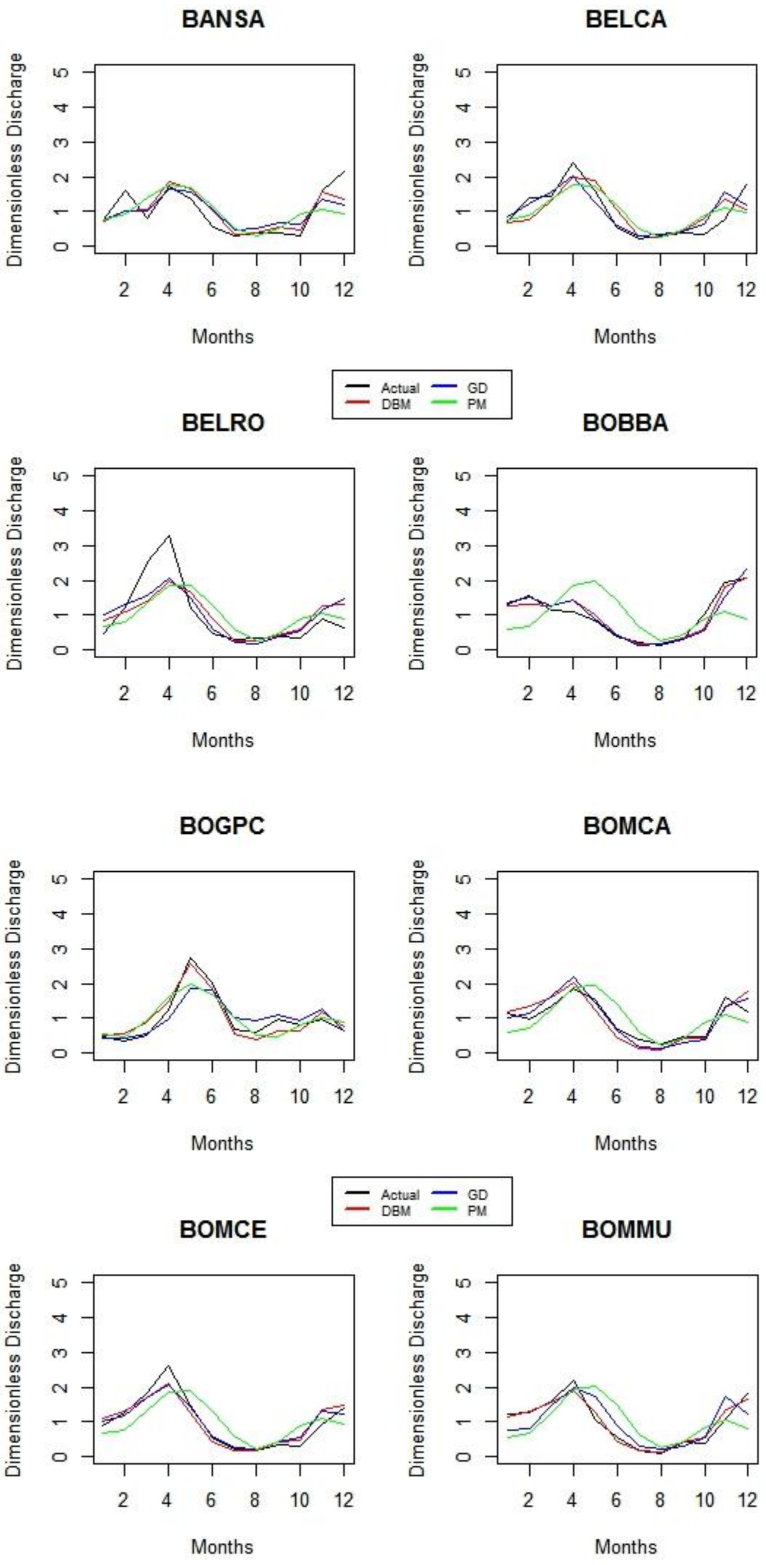
Appendix A.

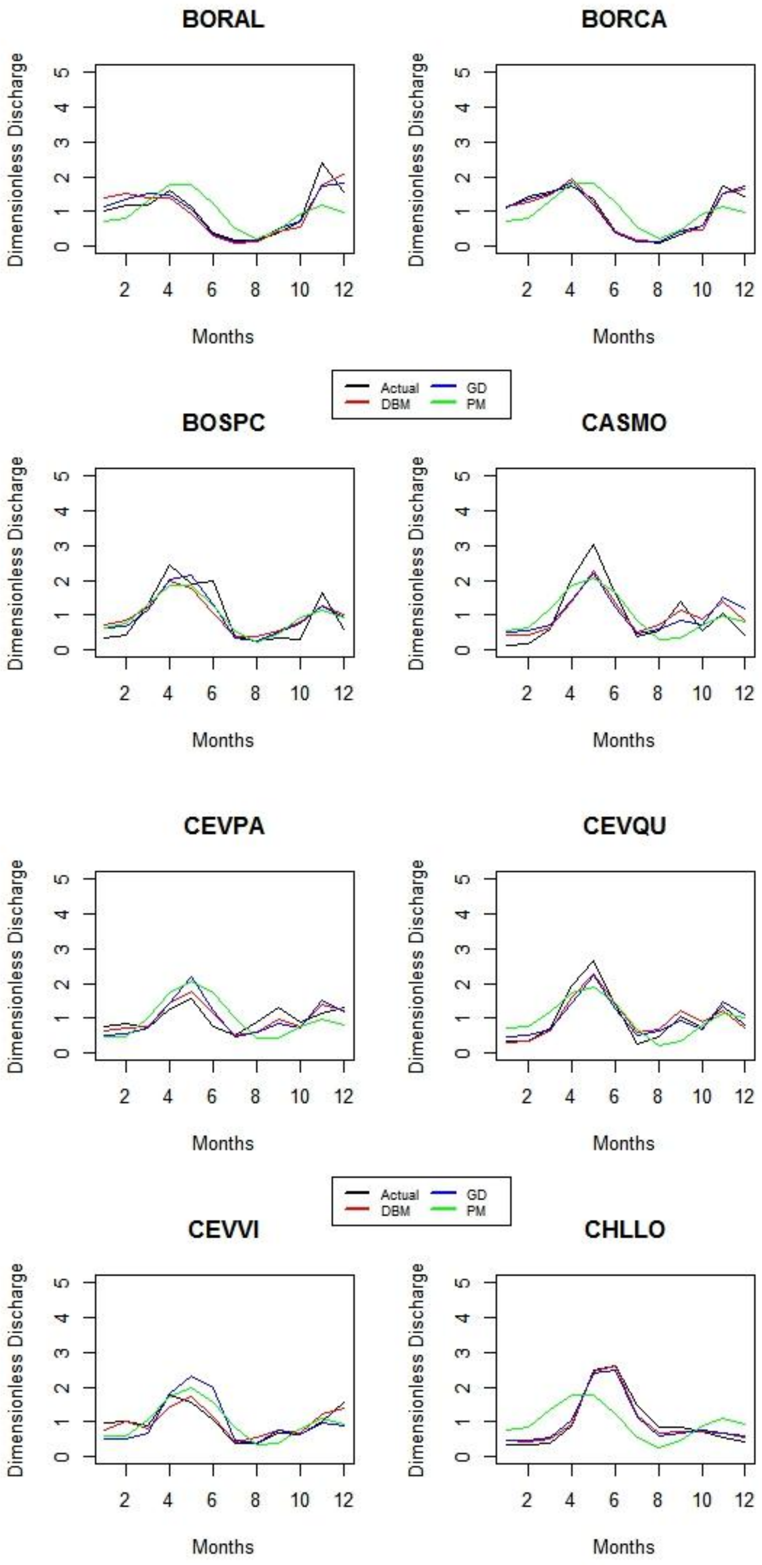
Summary of comparisons between Simulated and Actual flow regimes and FDCs by Distance-based, Geographical and Parametric method

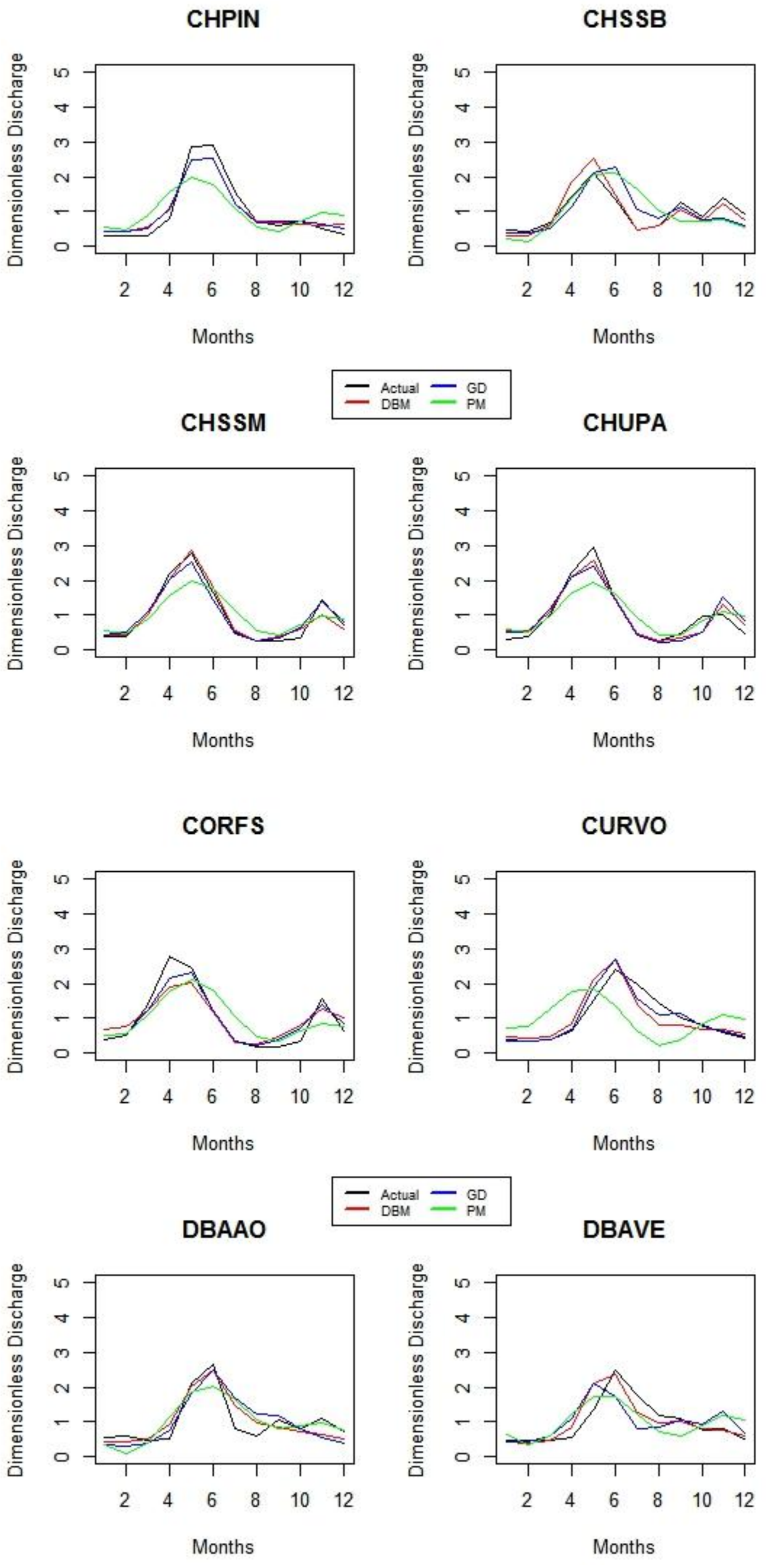
The analyses of the flow regime and FDCs in chapter 2 and 3 respectively, are based on a set of basins located in northern part of Italy. The following diagrams provide a short summary of results obtained at some selected stations by using Distance-based method, Geographical distance method and parametric method.

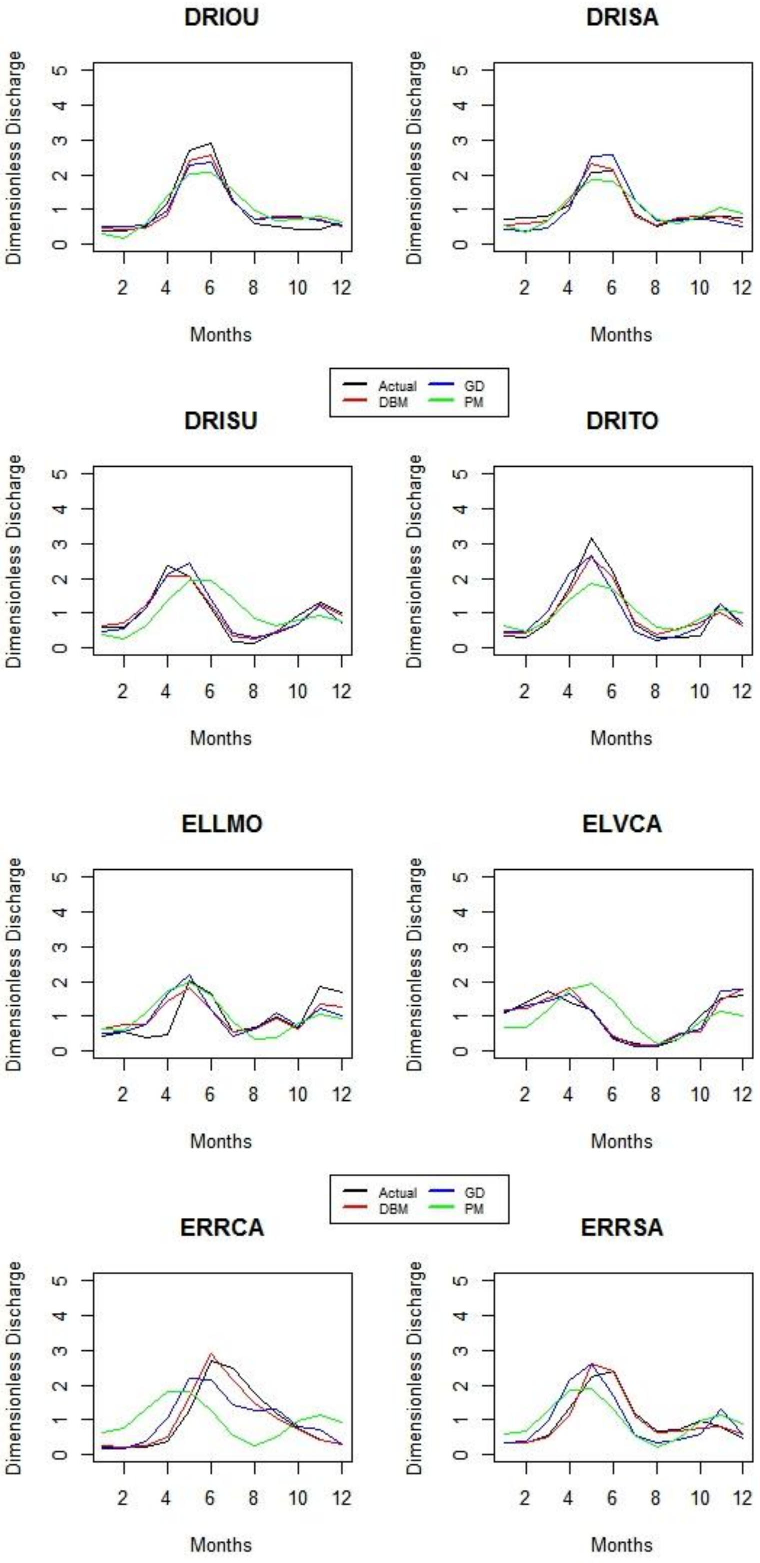
A.1. Flow regimes modeling

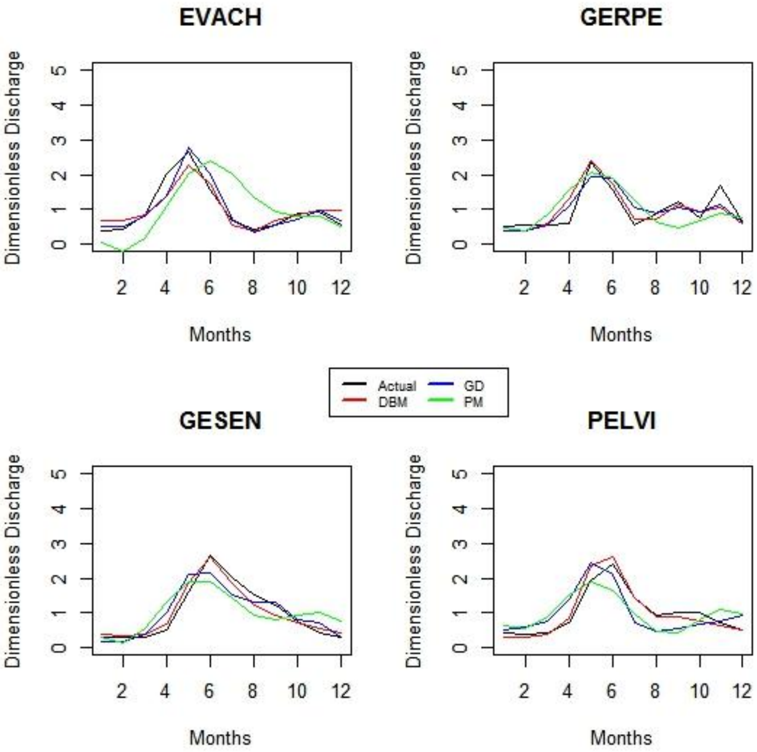



BANSA
BELCA
BELRO
BOBBA
BOGPC
BOMCA
BOMCE
BOMMU
Dimensionless Discharge
0
1
2
3
4
5
Months
2
4
6
8
10
12
Actual
GD
DBM
PM

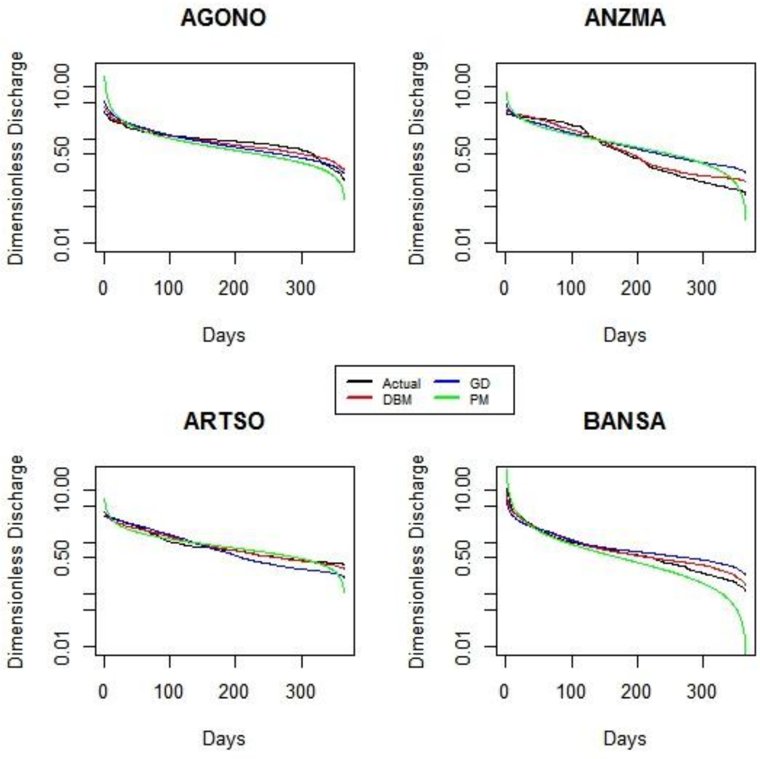

BORAL
BORCA
BOSPC
CASMO
CEVPA
CEVQU
CEVVI
CHLLO
Dimensionless Discharge
0
1
2
3
4
5
Months
2
4
6
8
10
12
Actual
GD
DBM
PM

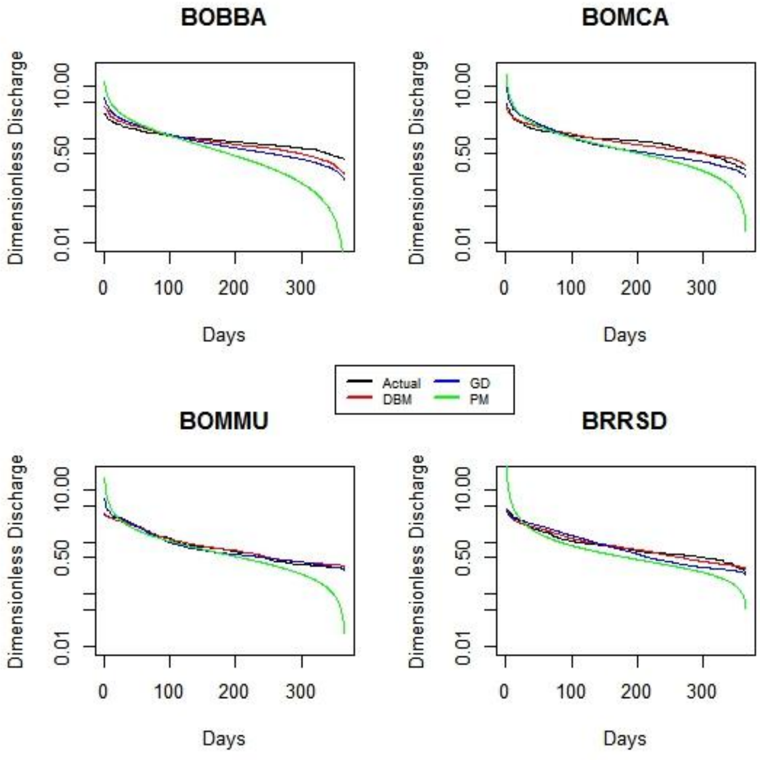

CHPIN
CHSSB
CHSSM
CHUPA
CORFS
CURVO
DBAAO
DBAVE
Dimensionless Discharge
Months
0
1
2
3
4
5
2
4
6
8
10
12
Actual
GD
DBM
PM

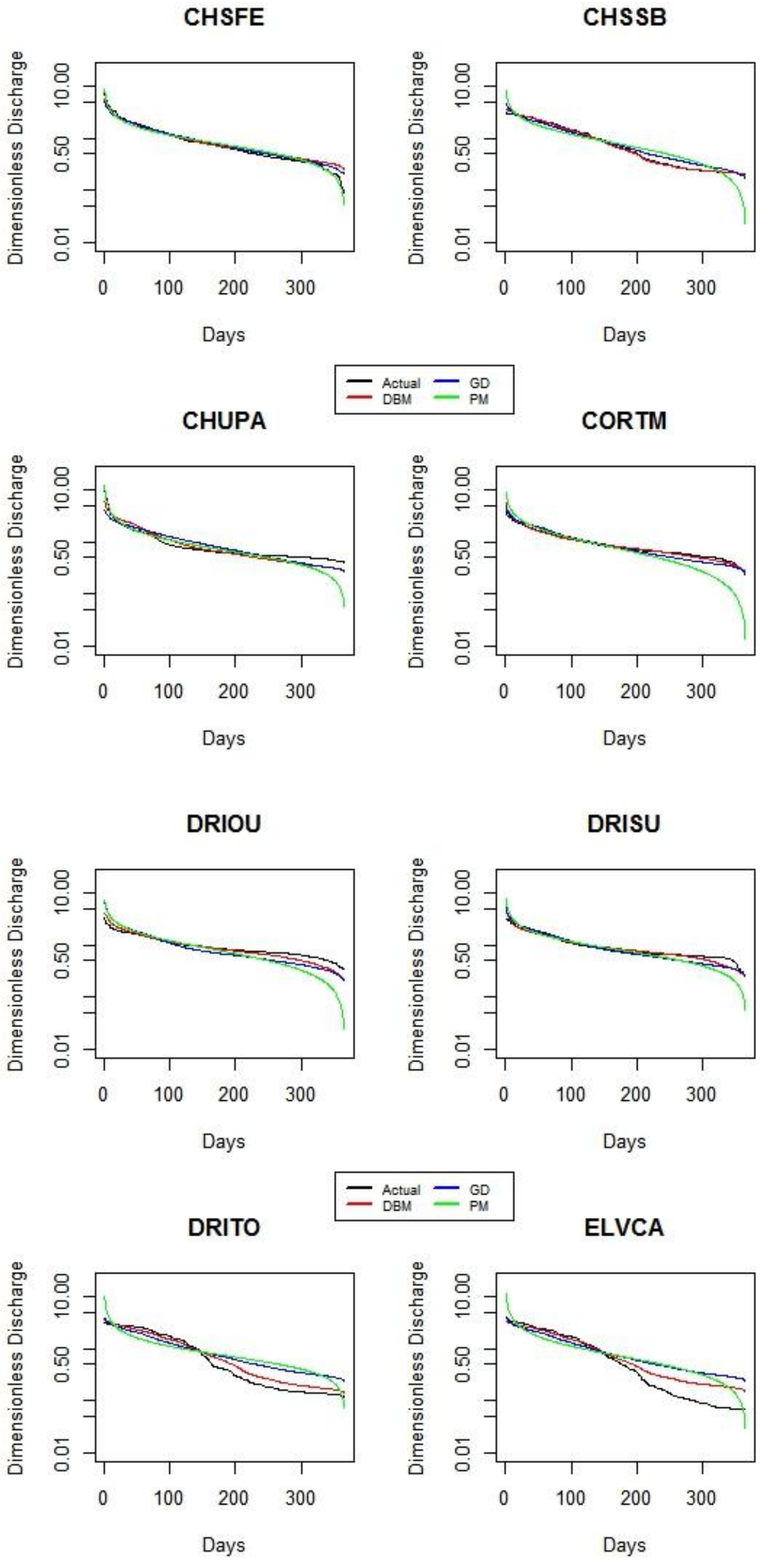

DRIOU
DRISA
DRISU
DRITO
ELLMO
ELVCA
ERRCA
ERRSA
Dimensionless Discharge
0
1
2
3
4
5
Months
2
4
6
8
10
12
Actual
GD
DBM
PM

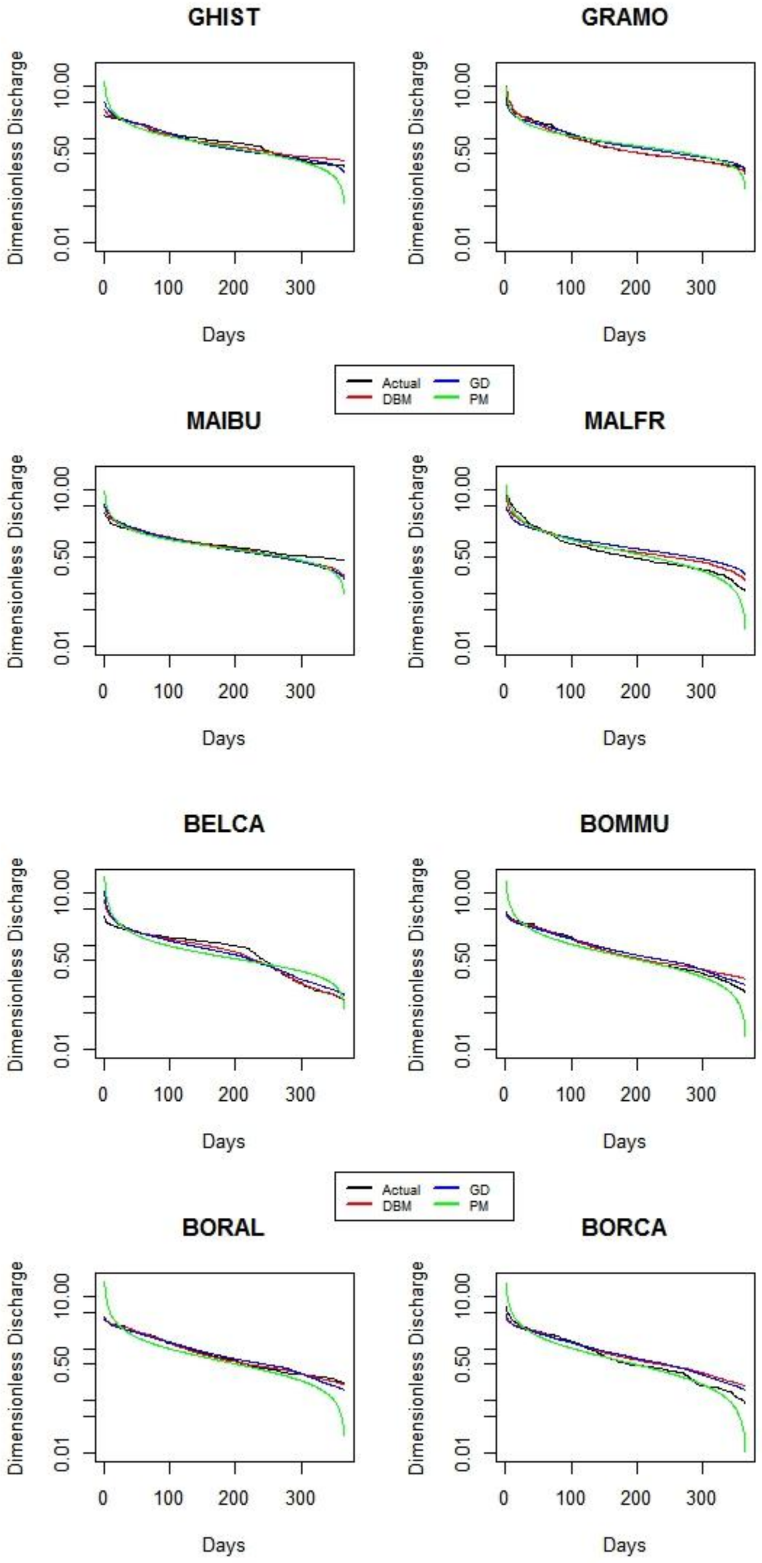

EVACH
GERPE
GESEN
PELVI
Dimensionless Discharge
0
1
2
3
4
5
Months
2
4
6
8
10
12
Actual
GD
DBM
PM

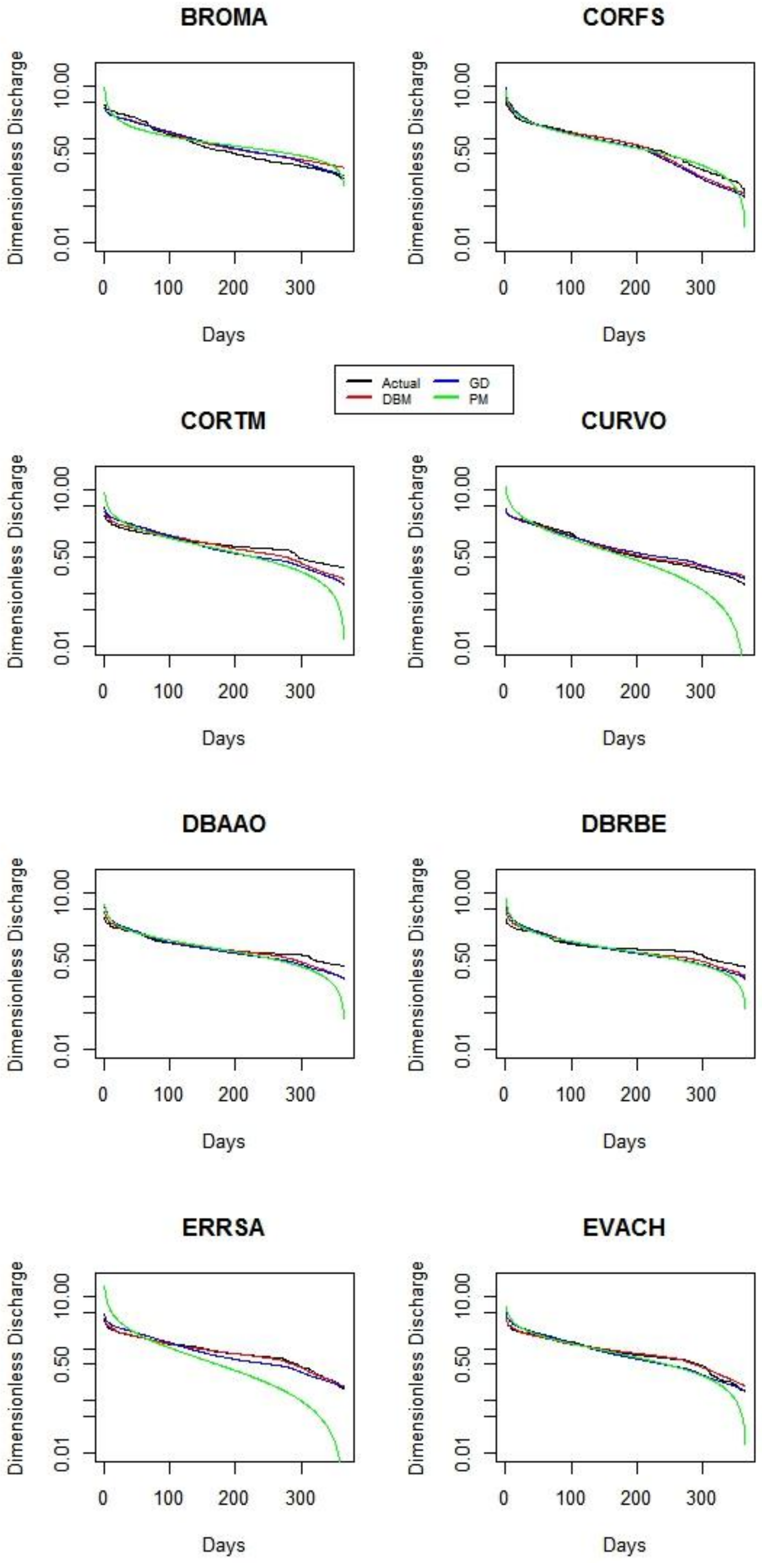
A.2. Flow duration curves modeling

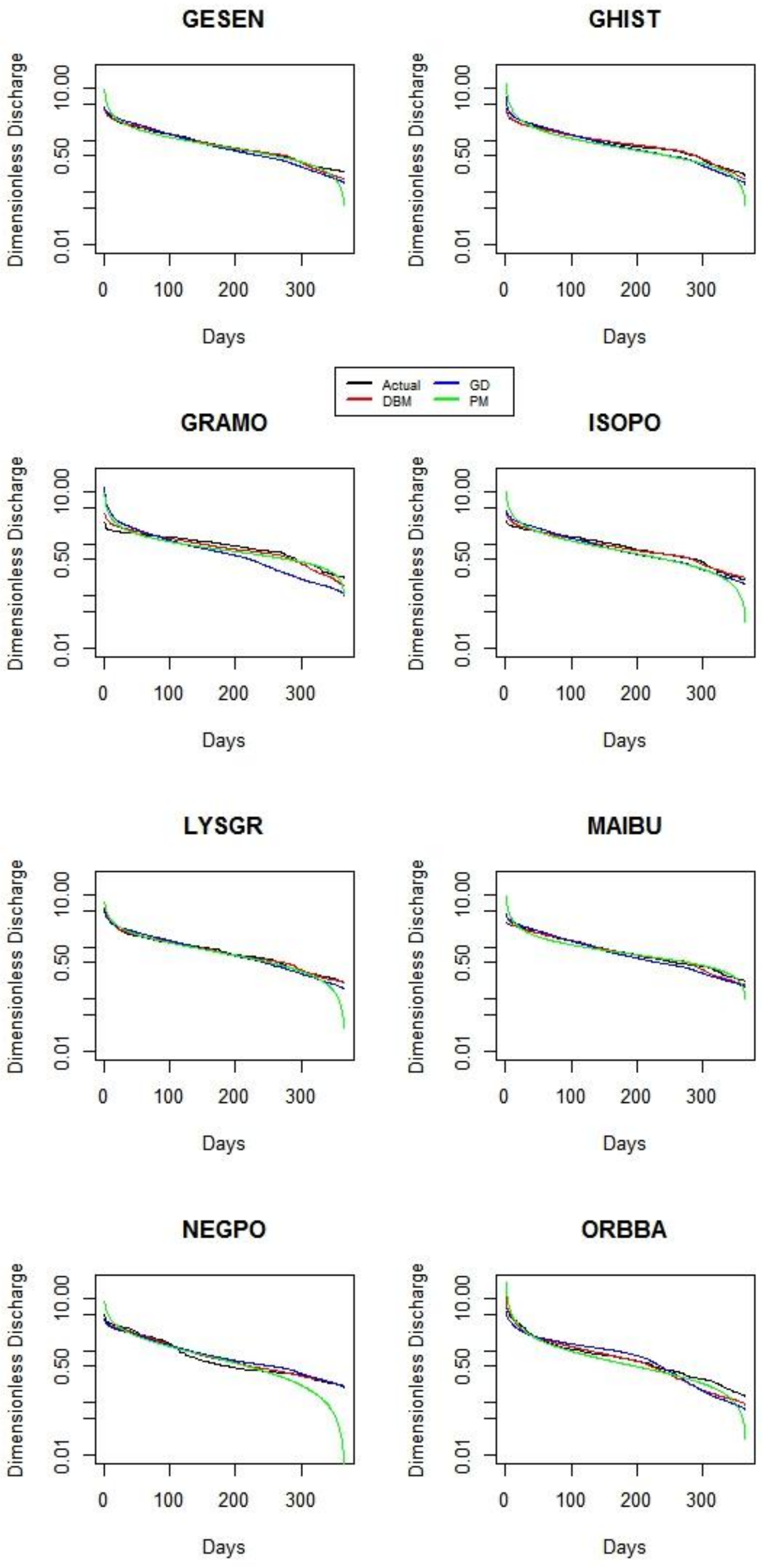





CHSFE
CHSSB
Dimensionless Discharge
10.00
0.50
0.01
0
100
200
300
Days
Actual
DBM
GD
PM
CHUPA
CORTM
DRIOU
DRISU
Actual
DBM
GD
PM
DRITO
ELVCA


GHIST
GRAMO
MAIBU
MALFR
BELCA
BOMMU
BORAL
BORCA
Dimensionless Discharge
0.01
0.50
10.00
Days
0
100
200
300
Actual
DBM
GD
PM

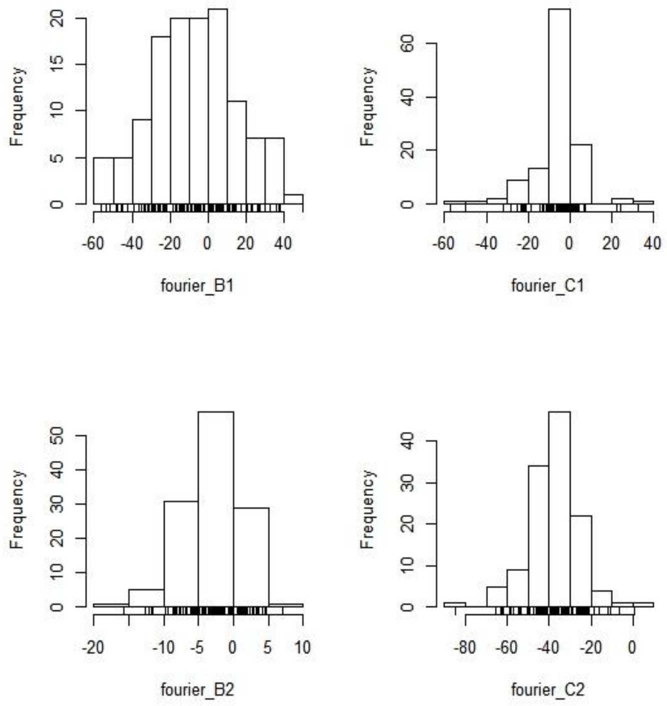

BROMA
CORFS
CORTM
CURVO
DBAAO
DBRBE
ERRSA
EVACH
Dimensionless Discharge
0.01
0.50
10.00
Days
0
100
200
300
Actual
GD
DBM
PM

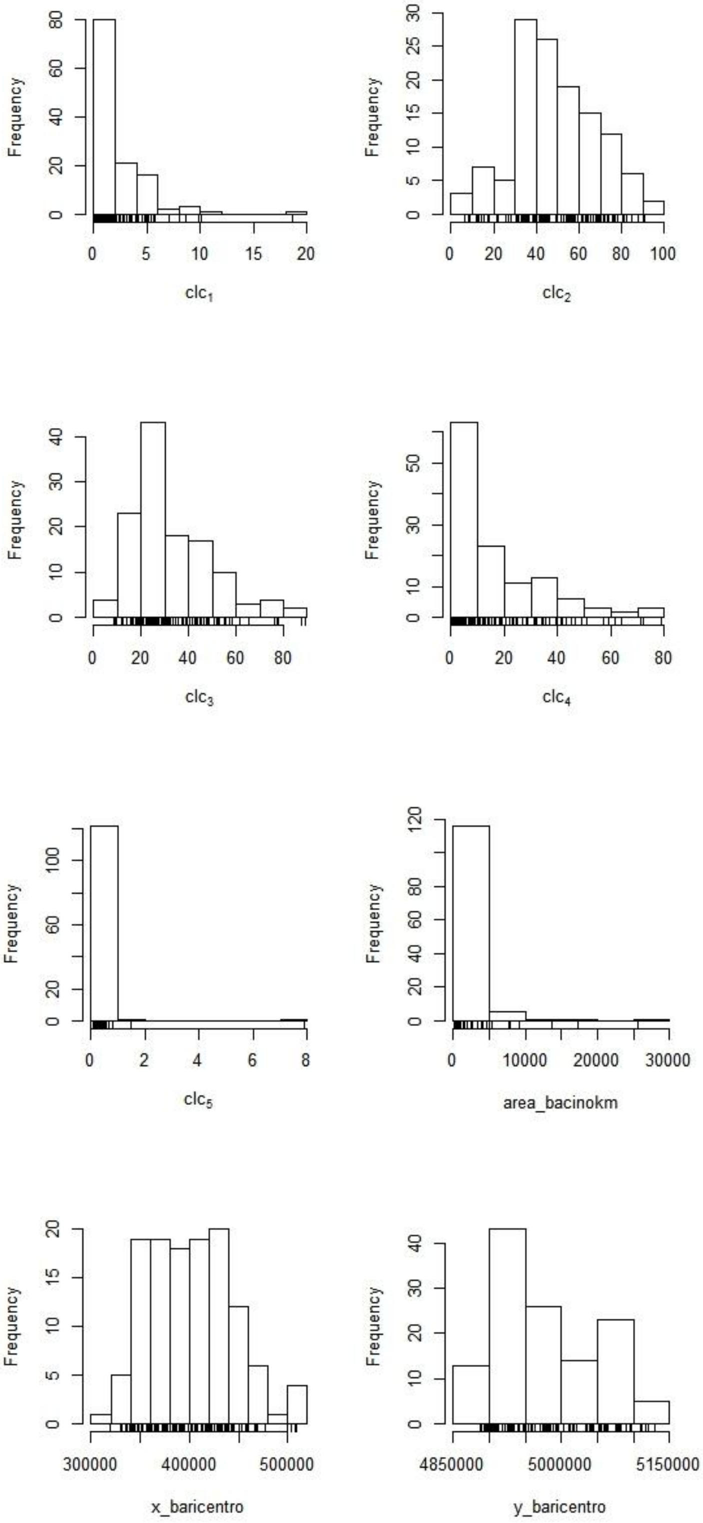

GESEN
GHIST
GRAMO
ISOPO
LYSGR
MAIBU
NEGPO
ORBBA
Dimensionless Discharge
0.01
0.50
10.00
Days
0
100
200
300
Actual
DBM
GD
PM

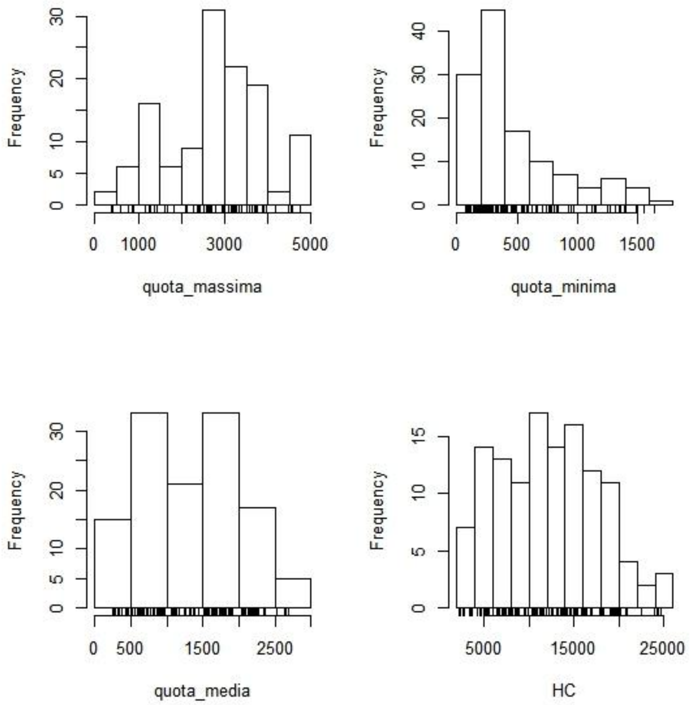
Appendix B.

The descriptors used for the regionalization of flow regimes and FDCs.

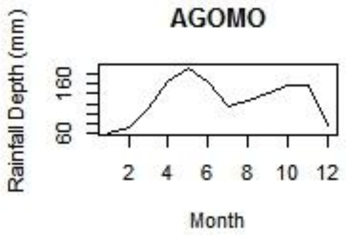
B.1. The histograms of some selected descriptors

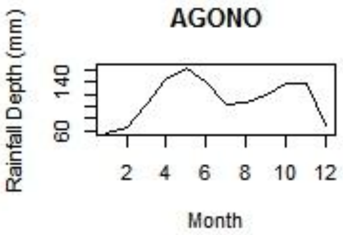


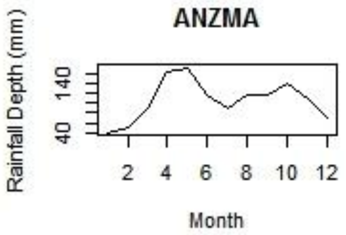

Frequency
0
20
40
60
80
0
5
10
15
20
clc$_1$
Frequency
0
5
10
15
20
25
30
0
20
40
60
80
100
clc$_2$
Frequency
0
10
20
30
40
0
20
40
60
80
clc$_3$
Frequency
0
10
30
50
0
20
40
60
80
clc$_4$
Frequency
0
20
60
100
0
2
4
6
8
clc$_5$
Frequency
0
20
40
60
80
120
0
10000
20000
30000
area_bacinokm
Frequency
0
5
10
15
20
300000
400000
500000
x_baricentro
Frequency
0
10
20
30
40
4850000
5000000
5150000
y_baricentro

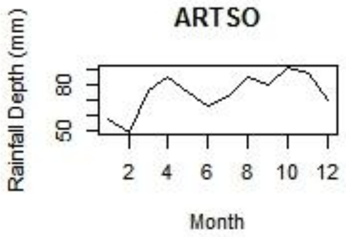

Frequency
0
5
10
20
30
0
1000
3000
5000
quota_massima
Frequency
0
10
20
30
40
0
500
1000
1500
quota_minima
Frequency
0
5
10
20
30
0
500
1500
2500
quota_media
Frequency
0
5
10
15
5000
15000
25000
HC

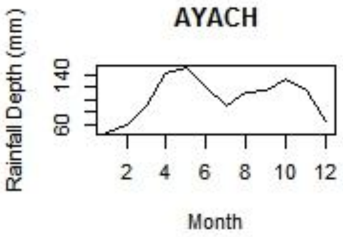
B.2. Rain Regimes at the selected stations

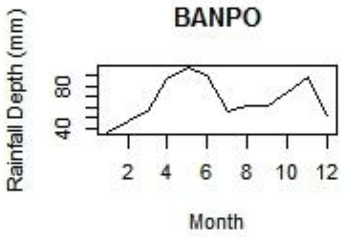

AGOMO

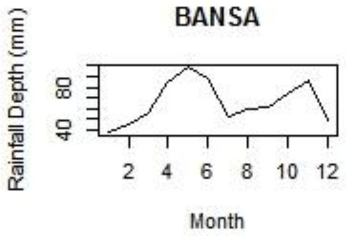

AGONO

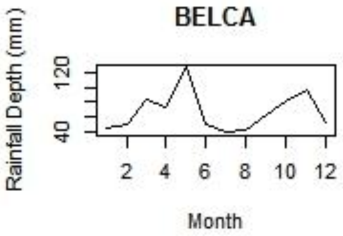

ANZMA

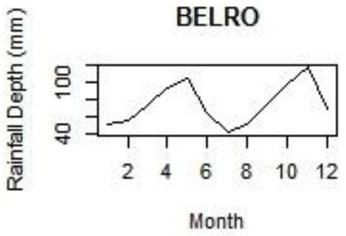

ARTSO

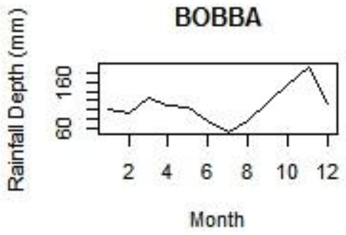

AYACH

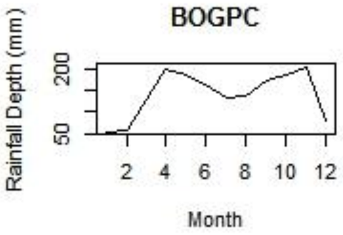

BANPO

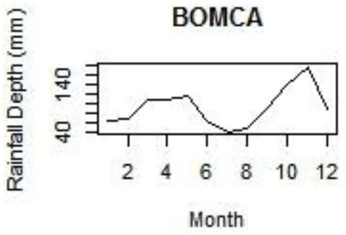

BANSA

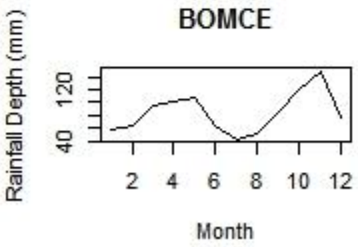

BELCA

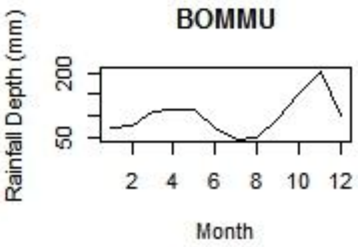

BELRO

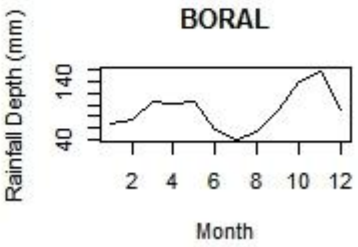

BOBBA

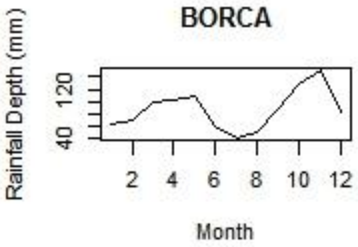

BOGPC

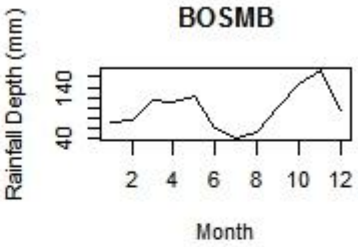

BOMCA

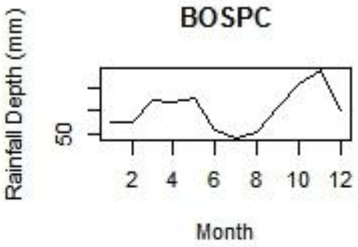

BOMCE
Rainfall Depth (mm)
40
120
2 4 6 8 10 12
Month

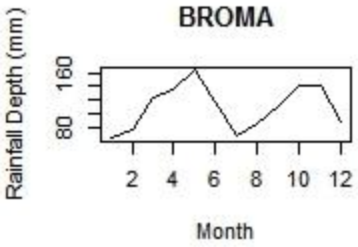

BOMMU
Rainfall Depth (mm)
50
200
2 4 6 8 10 12
Month

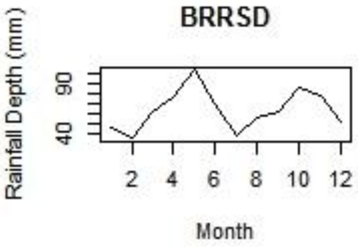

BORAL
Rainfall Depth (mm)
40
140
2 4 6 8 10 12
Month

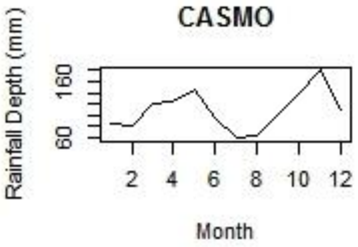

BORCA
Rainfall Depth (mm)
40
120
2 4 6 8 10 12
Month

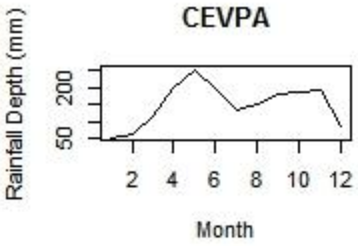

BOSMB
Rainfall Depth (mm)
40
140
2 4 6 8 10 12
Month

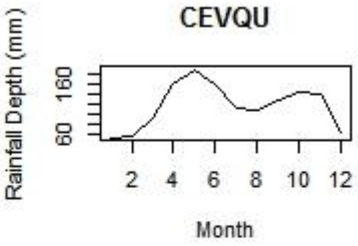

BOSPC
Rainfall Depth (mm)
50
2 4 6 8 10 12
Month

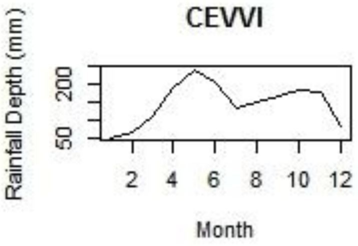

BROMA
Rainfall Depth (mm)
80
160
2 4 6 8 10 12
Month

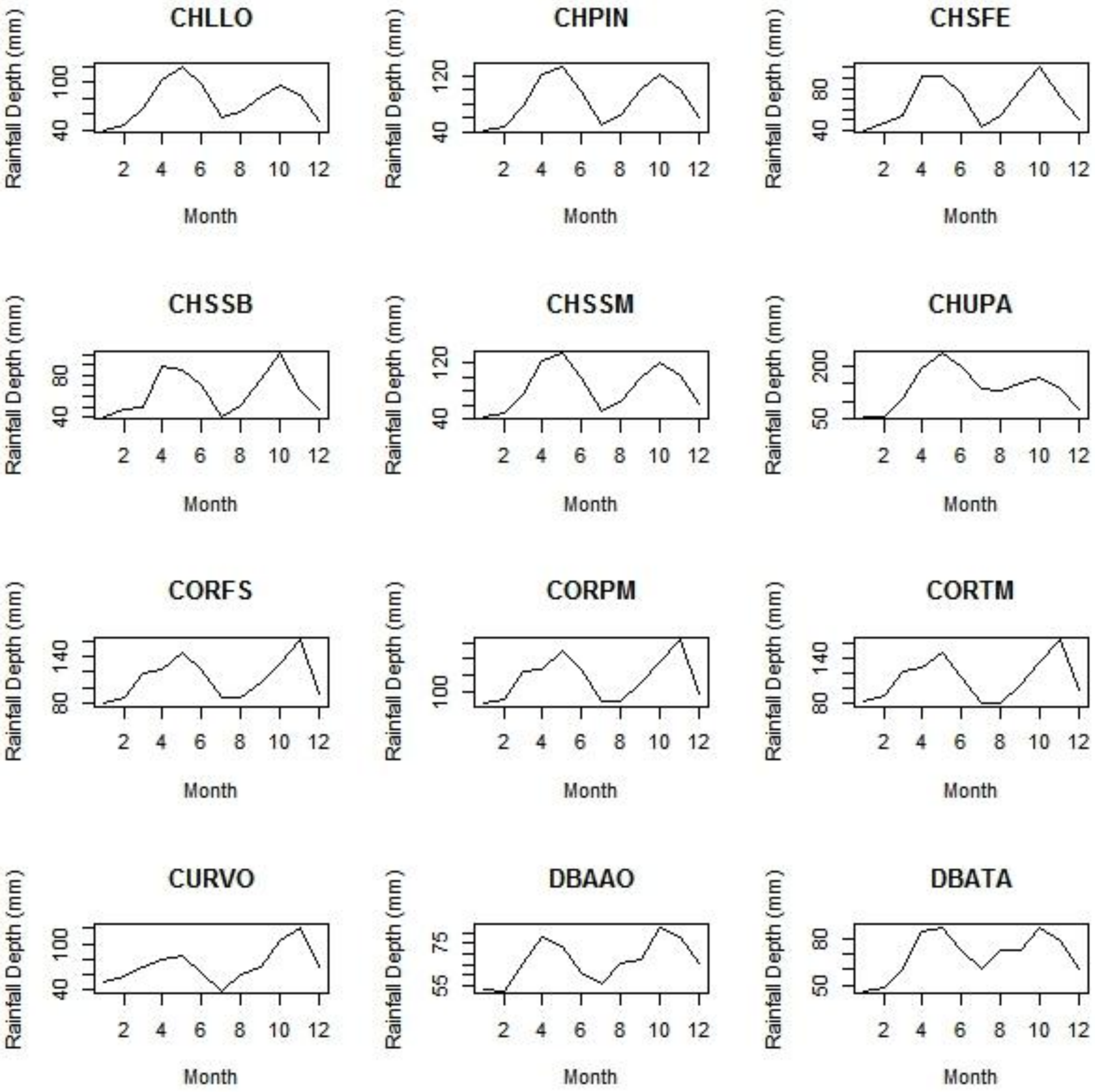

BRRSD
Rainfall Depth (mm)
40
90
2 4 6 8 10 12
Month

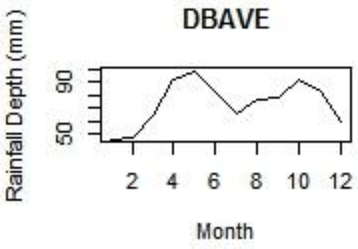

CASMO
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month

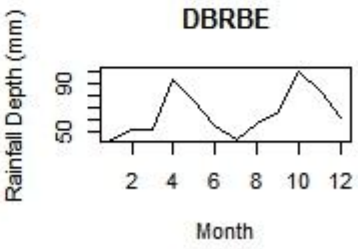

CEVPA
Rainfall Depth (mm)
50
200
2 4 6 8 10 12
Month

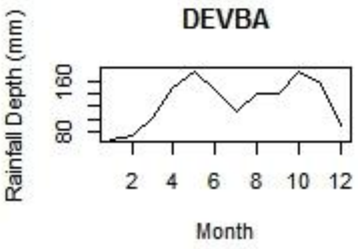

CEVQU
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month

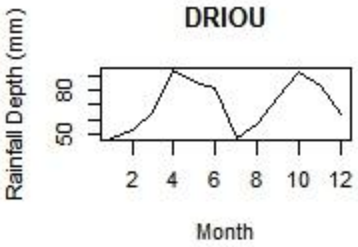

CEVVI
Rainfall Depth (mm)
50
200
2 4 6 8 10 12
Month

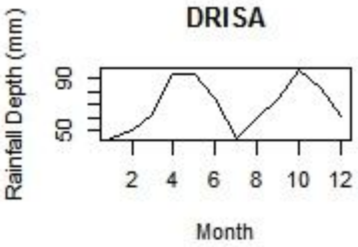

CHLLO
CHPIN
CHSFE
CHSSB
CHSSM
CHUPA
CORFS
CORPM
CORTM
CURVO
DBAAO
DBATA
Rainfall Depth (mm)
Month
2 4 6 8 10 12
40 100
40 120
40 80
40 80
40 120
50 200
80 140
100
80 140
40 100
55 75
50 80

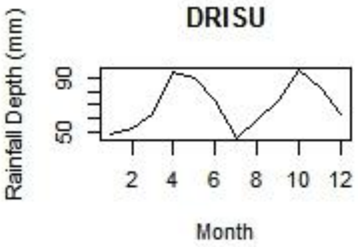

DBAVE
Rainfall Depth (mm)
50
90
2
4
6
8
10
12
Month

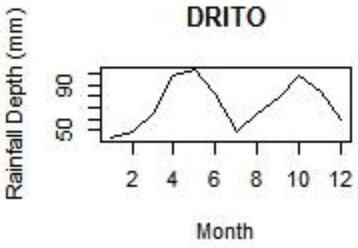

DBRBE
Rainfall Depth (mm)
50
90
2
4
6
8
10
12
Month

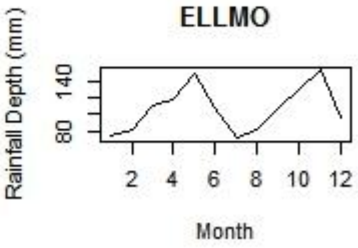

DEVBA
Rainfall Depth (mm)
80
160
2
4
6
8
10
12
Month

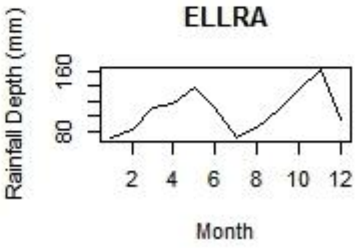

DRIOU
Rainfall Depth (mm)
50
80
2
4
6
8
10
12
Month

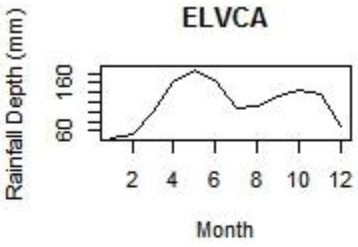

DRISA
Rainfall Depth (mm)
50
90
2
4
6
8
10
12
Month

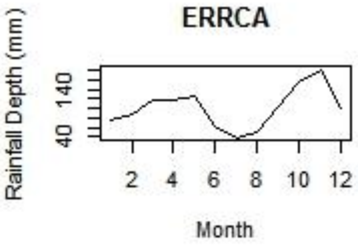

DRISU
Rainfall Depth (mm)
50
90
2
4
6
8
10
12
Month

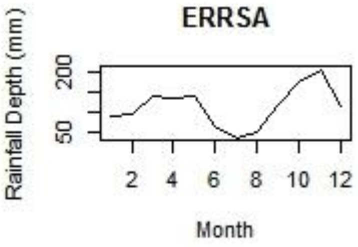

DRITO
Rainfall Depth (mm)
50
90
2
4
6
8
10
12
Month

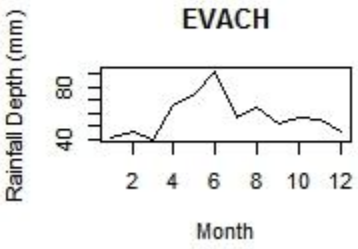

ELLMO
Rainfall Depth (mm)
80
140
2
4
6
8
10
12
Month

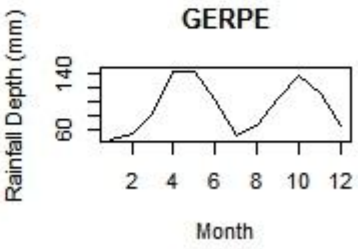

ELLRA
Rainfall Depth (mm)
80
160
2
4
6
8
10
12
Month

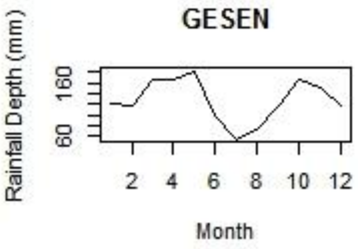

ELVCA
Rainfall Depth (mm)
60
160
2
4
6
8
10
12
Month

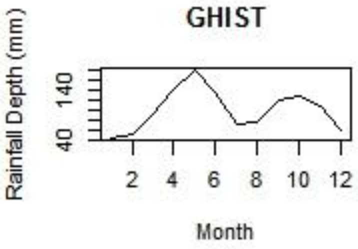

ERRCA
Rainfall Depth (mm)
40
140
2
4
6
8
10
12
Month

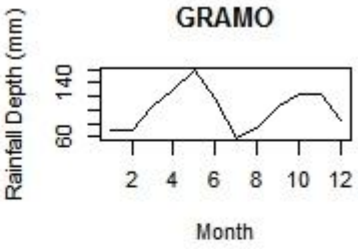

ERRSA
Rainfall Depth (mm)
50
200
2
4
6
8
10
12
Month

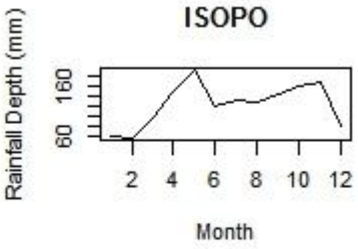

EVACH
Rainfall Depth (mm)
40
80
2 4 6 8 10 12
Month

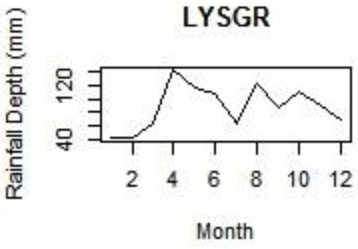

GERPE
Rainfall Depth (mm)
60
140
2 4 6 8 10 12
Month

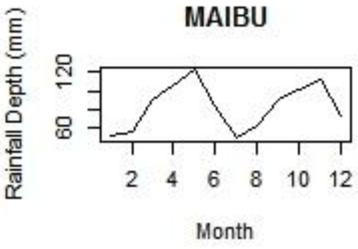

GESEN
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month

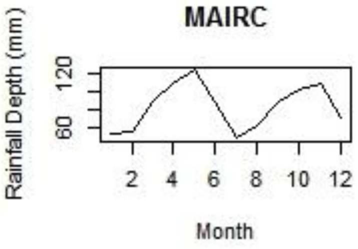

GHIST
Rainfall Depth (mm)
40
140
2 4 6 8 10 12
Month

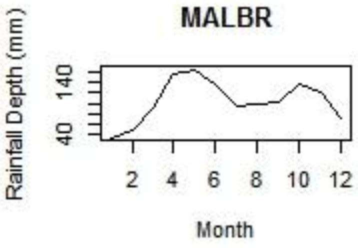

GRAMO
Rainfall Depth (mm)
60
140
2 4 6 8 10 12
Month

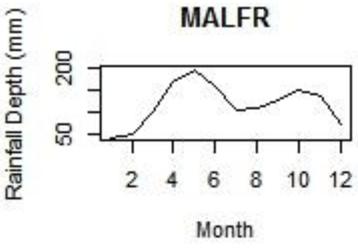

ISOPO
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month

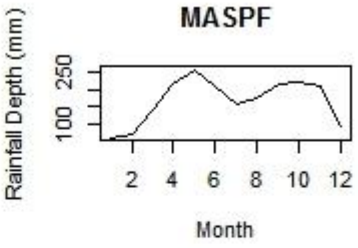

LYSGR
Rainfall Depth (mm)
40
120
2 4 6 8 10 12
Month

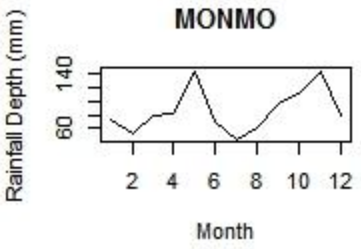

MAIBU
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month

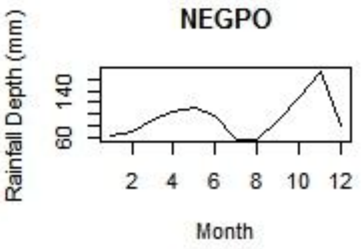

MAIRC
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month

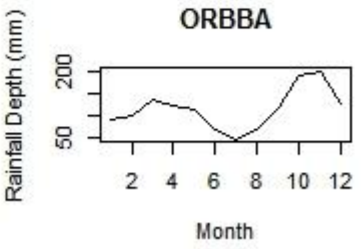

MALBR
Rainfall Depth (mm)
40
140
2 4 6 8 10 12
Month

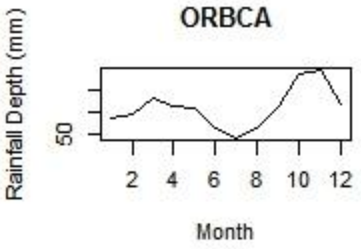

MALFR
Rainfall Depth (mm)
50
200
2 4 6 8 10 12
Month

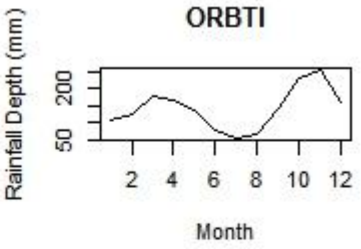

MASPF
Rainfall Depth (mm)
100
250
2 4 6 8 10 12
Month

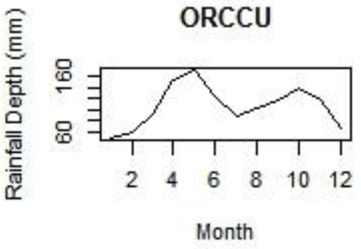

MONMO
Rainfall Depth (mm)
60
140
2 4 6 8 10 12
Month

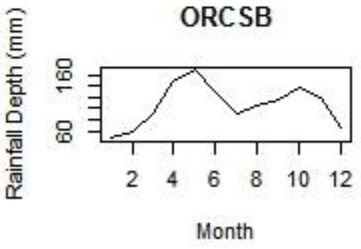

NEGPO
Rainfall Depth (mm)
60
140
2 4 6 8 10 12
Month

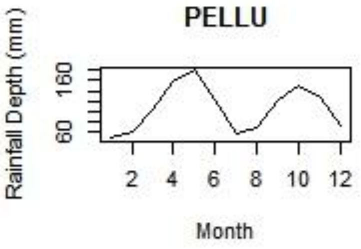

ORBBA
Rainfall Depth (mm)
50
200
2 4 6 8 10 12
Month

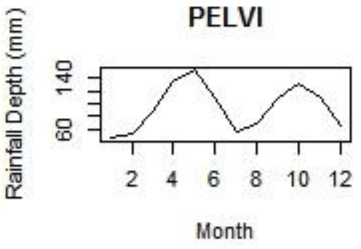

ORBCA
Rainfall Depth (mm)
50
2 4 6 8 10 12
Month

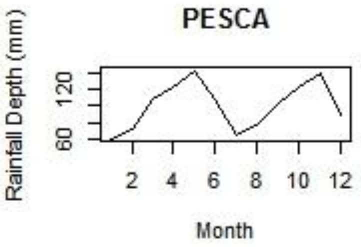

ORBTI
Rainfall Depth (mm)
50
200
2 4 6 8 10 12
Month

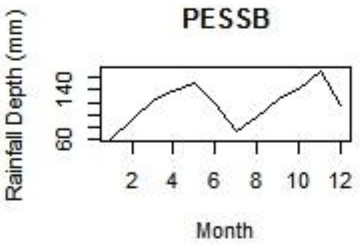

ORCCU
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month

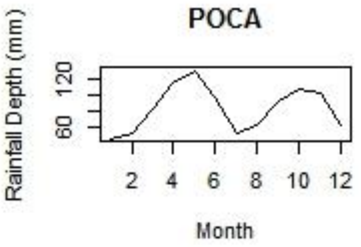

ORCSB
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month

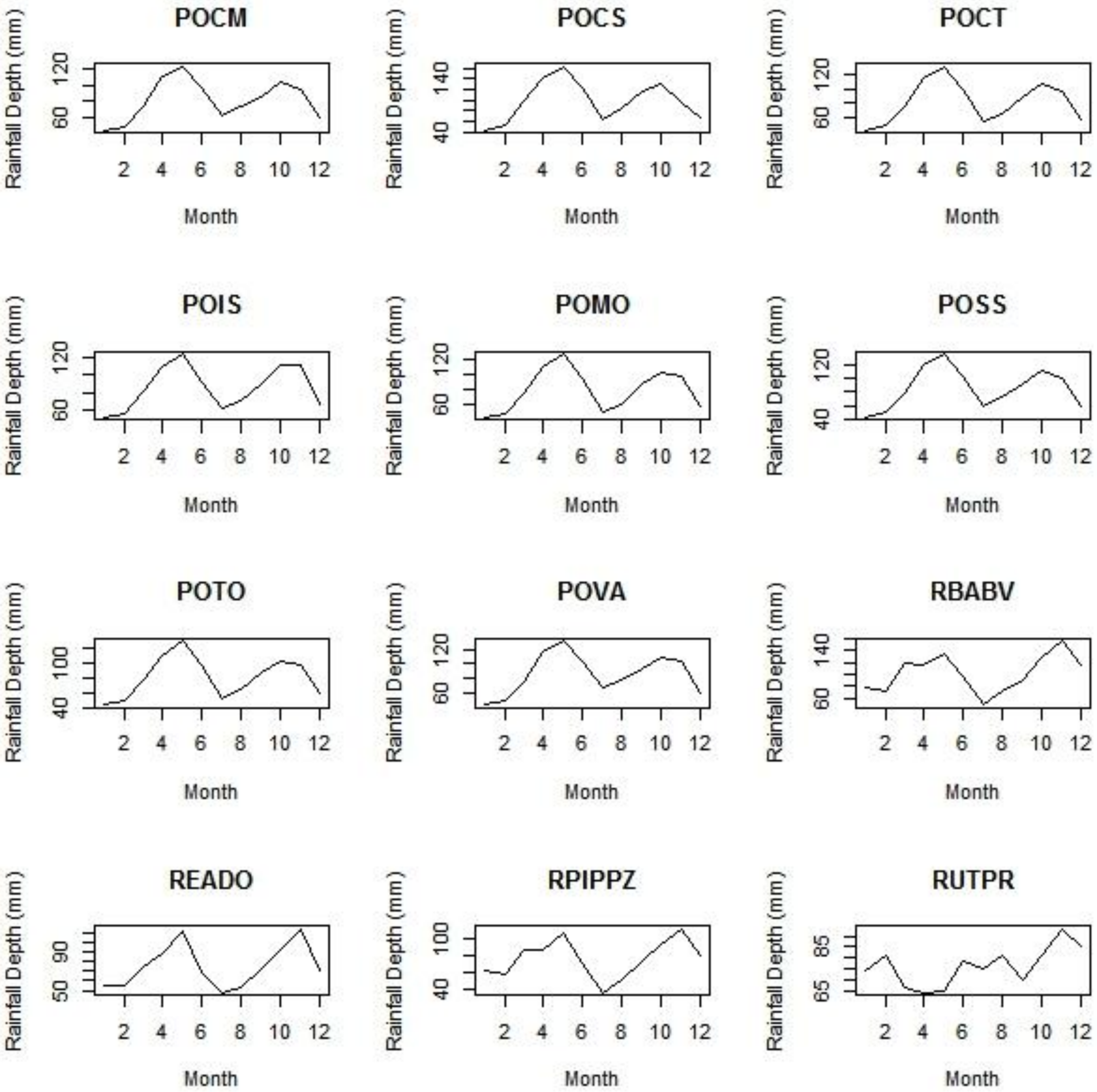

PELLU
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month

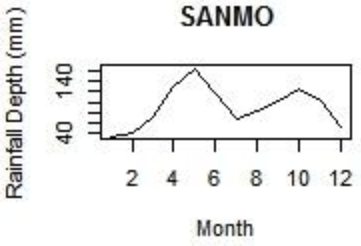

PELVI
Rainfall Depth (mm)
60
140
2 4 6 8 10 12
Month


PESCA
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month

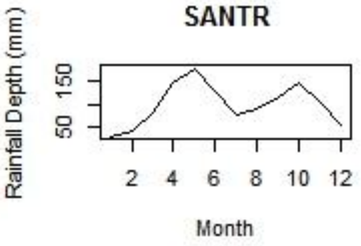

PESSB
Rainfall Depth (mm)
60
140
2 4 6 8 10 12
Month


POCA
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month

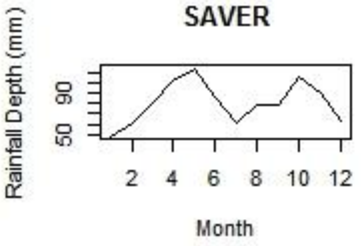

POCM
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month
POCS
Rainfall Depth (mm)
40
140
2 4 6 8 10 12
Month
POCT
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month
POIS
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month
POMO
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month
POSS
Rainfall Depth (mm)
40
120
2 4 6 8 10 12
Month
POTO
Rainfall Depth (mm)
40
100
2 4 6 8 10 12
Month
POVA
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month
RBABV
Rainfall Depth (mm)
60
140
2 4 6 8 10 12
Month
READO
Rainfall Depth (mm)
50
90
2 4 6 8 10 12
Month
RPIPPZ
Rainfall Depth (mm)
40
100
2 4 6 8 10 12
Month
RUTPR
Rainfall Depth (mm)
65
85
2 4 6 8 10 12
Month



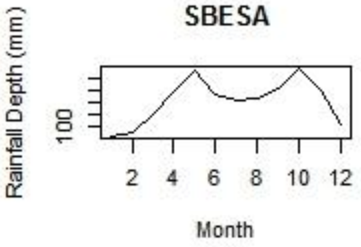
SANMO



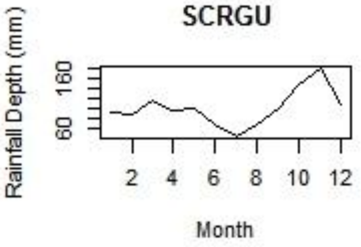
SANTR



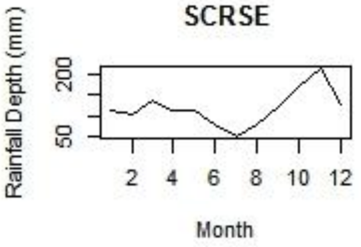
SAVER



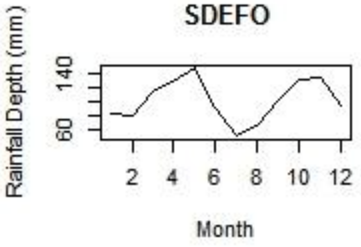
SBESA



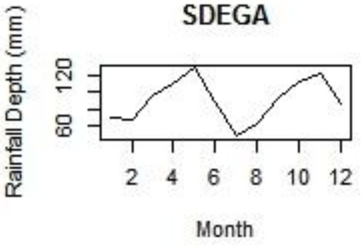
SCRGU



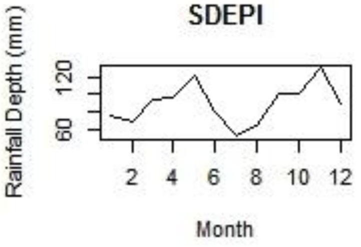
SCRSE



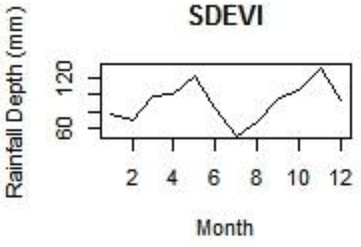
SDEFO



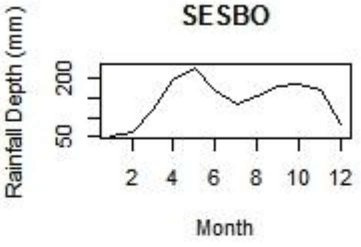
SDEGA



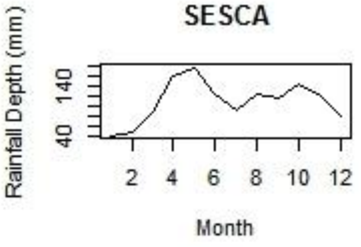
SDEPI



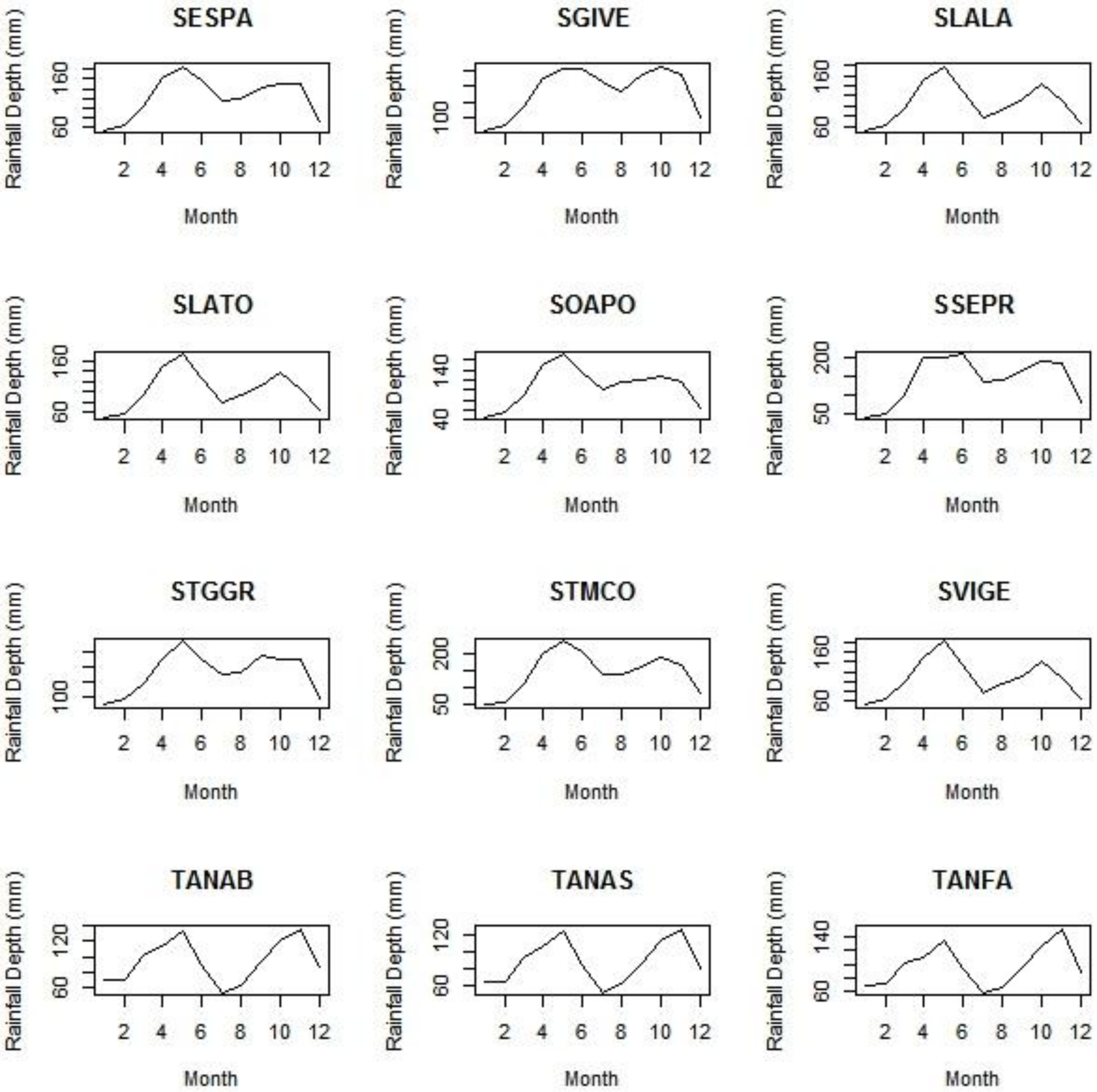
SDEVI

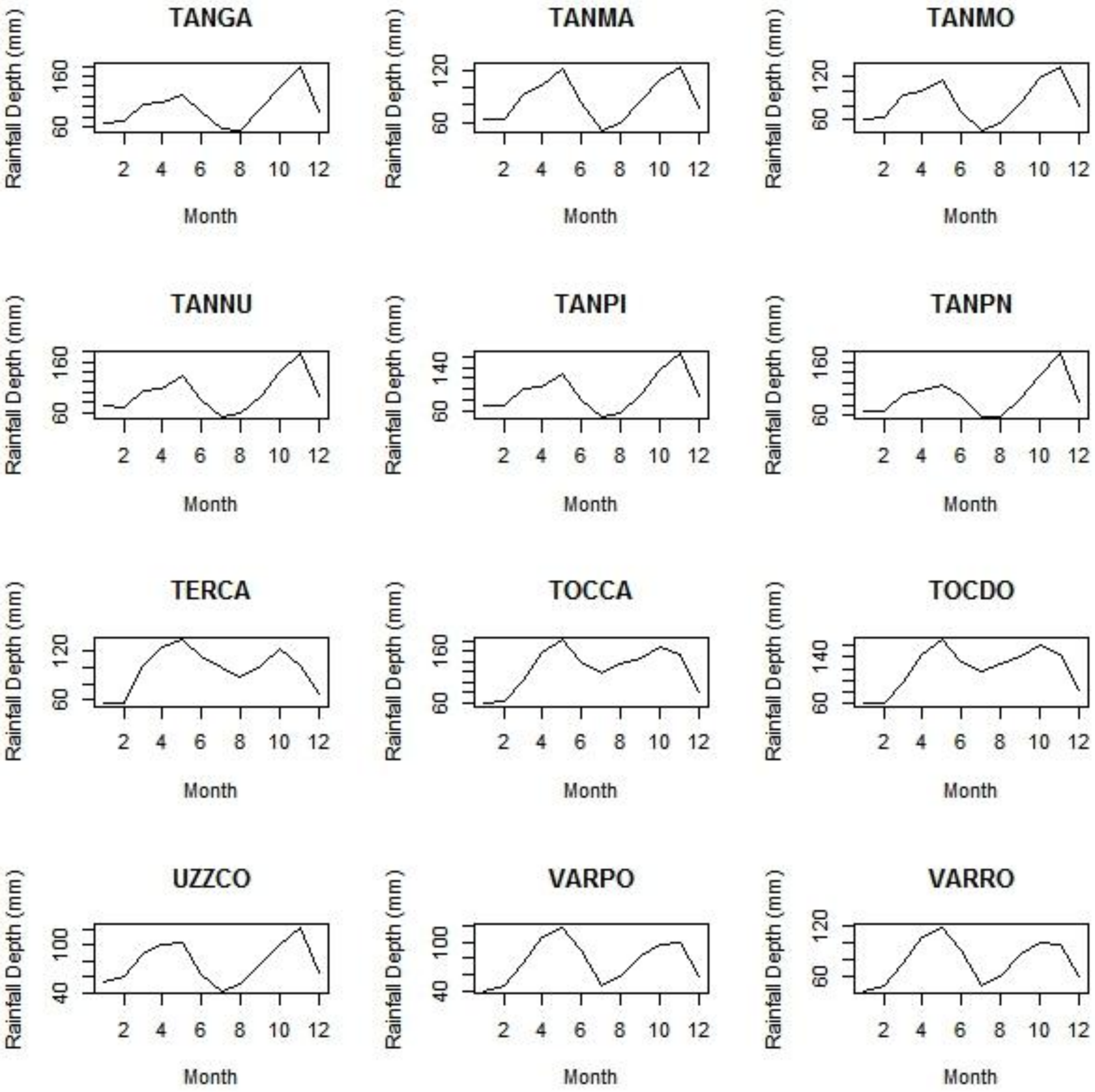


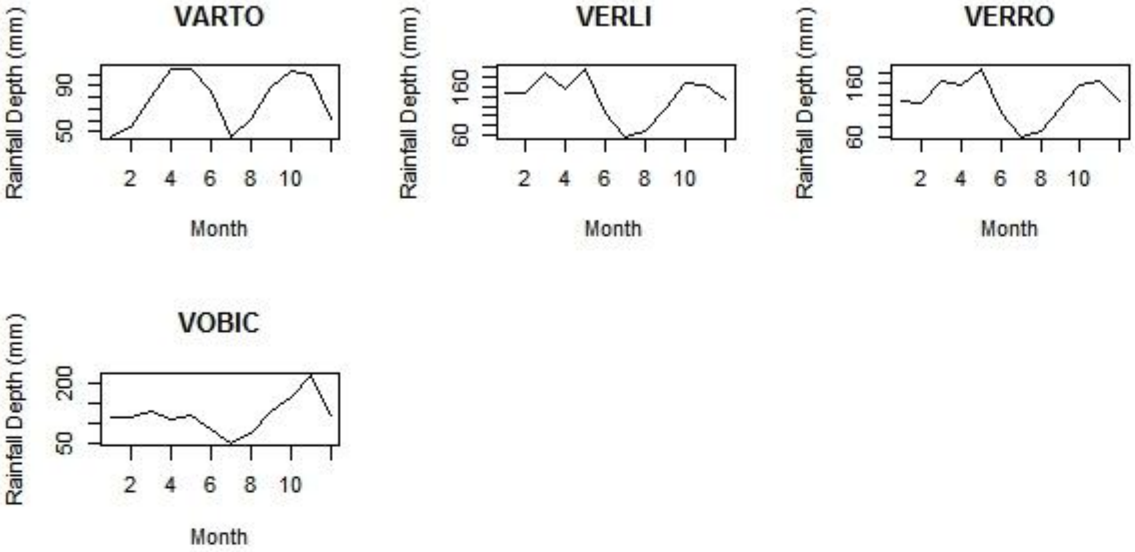
SESBO



SESCA


SESPA
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month
SGIVE
Rainfall Depth (mm)
100
2 4 6 8 10 12
Month
SLALA
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month
SLATO
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month
SOAPO
Rainfall Depth (mm)
40
140
2 4 6 8 10 12
Month
SSEPR
Rainfall Depth (mm)
50
200
2 4 6 8 10 12
Month
STGGR
Rainfall Depth (mm)
100
2 4 6 8 10 12
Month
STMCO
Rainfall Depth (mm)
50
200
2 4 6 8 10 12
Month
SVIGE
Rainfall Depth (mm)
60
160
2 4 6 8 10 12
Month
TANAB
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month
TANAS
Rainfall Depth (mm)
60
120
2 4 6 8 10 12
Month
TANFA
Rainfall Depth (mm)
60
140
2 4 6 8 10 12
Month


TANGA
Rainfall Depth (mm)
60
160
Month
2 4 6 8 10 12
TANMA
Rainfall Depth (mm)
60
120
Month
2 4 6 8 10 12
TANMO
Rainfall Depth (mm)
60
120
Month
2 4 6 8 10 12
TANNU
Rainfall Depth (mm)
60
160
Month
2 4 6 8 10 12
TANPI
Rainfall Depth (mm)
60
140
Month
2 4 6 8 10 12
TANPN
Rainfall Depth (mm)
60
160
Month
2 4 6 8 10 12
TERCA
Rainfall Depth (mm)
60
120
Month
2 4 6 8 10 12
TOCCA
Rainfall Depth (mm)
60
160
Month
2 4 6 8 10 12
TOCDO
Rainfall Depth (mm)
60
140
Month
2 4 6 8 10 12
UZZCO
Rainfall Depth (mm)
40
100
Month
2 4 6 8 10 12
VARPO
Rainfall Depth (mm)
40
100
Month
2 4 6 8 10 12
VARRO
Rainfall Depth (mm)
60
120
Month
2 4 6 8 10 12


VARTO
Rainfall Depth (mm)
50
90
2
4
6
8
10
Month
VERLI
Rainfall Depth (mm)
60
160
2
4
6
8
10
Month
VERRO
Rainfall Depth (mm)
60
160
2
4
6
8
10
Month
VOBIC
Rainfall Depth (mm)
50
200
2
4
6
8
10
Month

B.3. The notations used in the thesis for descriptors

- *fourier_B1*, *fourier_C1*, *fourier_B2*, *fourier_C2*
  Average values of the coefficients of Fourier series representing rainfall patterns
- clc_1
  Percentage area of the basin with continuous fabric of urbanized areas and urban areas discontinuous
- clc_2
  Percentage area of the basin under forests, arboreal vegetation, shrubs and bushy
- clc_3
  Percentage area of the basin containing herbaceous vegetation, grass-grazing, special crops, olive groves, vineyards, crops
- clc_4
  Percentage area of the basin not vegetated e.g., mining areas, landfills and construction sites, industrial, trade and communication networks
- clc_5
  Percentage area of the basin in the form of wetlands
- $area\_bacinokm$
  Basin Area (km^2)
- $x\_baricentro$
  Basin Longitude
- $y\_baricentro$
  Basin Latitude
- $qouta\_massima$
  Maximum Basin Elevation
- $qouta\_minima$
  Minimum Basin Elevation
- $qouta\_media$
  Average Basin Elevation
- HC
  Hypsographic Curve
- $lunghezza\_asta\_principale\_km$
  The longest sequence of segments that connect the source to a closing section of the basin (km)
- $longest\_drainage\_path\_length\_km$
  Longest Drainage Path Length in km
- $pendenza\_media\_LDP$
  Average slope of the Longest drainage path (LDP) (%)
- $lunghezza\_vett\_orient\_km$
  Length of the Orientation Vector represents the length of the segment joining the center of the basin to the section closing (km)
- $slope\_medio1$
  Average basin gradient which represents the average slope values associated with each pixel
- $slope\_medio2$

Average basin gradient calculated with respect to a basin of square shape equivalent to the real one, and does not take into account its actual form, which can be more or less elongated

- $aspect\_medio$
  Aspect Ratio which represents the angle of exposure of a cell in the horizontal plane, expressed in degrees
- R_al
  Elongation Ratio; which is the ratio between the diameter of the circle of the same area as a basin and basin length
- F_f
  Form Factor; which is the ratio between the area of the basin and the square of the length of the basin
- $media\_fa, varianza\_fa, skewness\_fa, kurtosis\_fa, fa5percento, fa10percento,$
  $fa15percento, fa30percento, fa40percento, fa50percento,$
  $fa60percento, fa70percento, fa85percento, fa95percento$
  The amplitude function is defined by counting the number of pixels having the same distance metric from the closing section; the distance measured by following the drainage directions. This function is calculated by first 4 statistical moments (mean, variance, skewness, kurtosis) and vector percentiles, or the distances from the closing section within which percentage of pixels equal to: (5%, 15%, 30%, 40%, 50%, 60%, 70%, 85%, 95%) are contained
- $lungh\_media\_vers$
  Average slope length: The average of the distances, measured following the drainage directions, of all pixels not belonging to the lattice, starting from the first pixel of the lattice in which they drain (km)
- $diam\_topol$
  Topological diameter representing number of segments (links) that form the main path. It indicates the number of confluences detected on the main path
- R_b
  Bifurcation ratio: The ratio of number of stream segments of one order to the number of the next higher order
- R_l
  The stream length ratio is defined as the ratio of average stream lengths of streams of order n and n+1, respectively
- $LCV24h, LCA24h$
  Average return period for short duration heavy rainfall events (LCV, LCA) for durations of 1, 3, 6, 12, 24 hours
- $IDFa$
  Coefficient of rainfall duration of rainfall intensity-duration curve in the form $h = ad^n$ (mm/hr)
- $IDFa\_std$
  Standard Deviation of $IDFa$ (mm/hr)
- $IDFn$
  Exponent of scale invariance of rainfall intensity-duration curve in the form $h = ad^n$ (mm/hr)
- $IDFn\_std$
  Standard Deviation of $IDFn$ (mm/hr)
- MAP

The total average annual rainfall in mm

- MAP_std
  Standard deviation of MAP
- $NDVIanno$
  Annual Normalized Difference Vegetation Index
- $NDVIanno\_std$
  Standard deviation of NDVI
- cf
  Coefficient of permeability
- cf_std
  Standard deviation of cf
- $cn1, cn2, cn3$
  Curve Number is an empirical parameter used in hydrology to define the part of rain that infiltrates into the ground. The curve number relative to the ground Cleaning is the $cn1$, one related to the moist soil is the $cn3$ and $cn2$ to describe the curves number of a soil with average soil moisture content
- $cn1\_std, cn2\_std, cn3\_std$
  Standard deviation of $cn1, cn2$ and $cn3$
- $delta\_z$
  Interquartile distance between basin elevation at 25% and 75% of area dominated by hypsographic curve (m m.s.l)
- c_int
  coeff. of precipitation intensity
- sd_rp
  Standard deviation of the rainfall regime (mm)
- cv_rp
  Coefficient of Variation in rainfall patterns
- $delta\_mese$
  Time interval between maximum and minimum monthly averages of rains