geosciences

MDPI

Article

Best-Fit Probability Models for Maximum Monthly Rainfall in Bangladesh Using Gaussian Mixture Distributions

Md Ashraful Alam, Craig Farnham and Kazuo Emura *

Department of Housing and Environmental Design, Graduate School of Human Life Science, Osaka City University, Osaka 558-8585, Japan; alam13ocu@gmail.com (M.A.A.); farnham@life.osaka-cu.ac.jp (C.F.)

* Correspondence: emura@life.osaka-cu.ac.jp; Tel.: +81-6-6605-2820

Received: 26 February 2018; Accepted: 17 April 2018; Published: 19 April 2018

Abstract: In this study, Gaussian/normal distributions (N) and mixtures of two normal (N2), three normal (N3), four normal (N4), or five normal (N5) distributions were applied to data with extreme values for precipitation for 35 weather stations in Bangladesh. For parameter estimation, maximum likelihood estimation was applied by using an expectation-maximization algorithm. For selecting the best-fit model, graphical inspection (probability density function (pdf), cumulative density function (cdf), quantile-quantile (Q-Q) plot) and numerical criteria (Akaike's information criterion (AIC), Bayesian information criterion (BIC), root mean square percentage error (RMSPE)) were used. In most of the cases, AIC and BIC gave the same best-fit results but their RMSPE results differed. The best-fit result of each station was chosen as the distribution with the lowest sum of the rank scores from each test statistic. The N distribution gave the best-fit result for 51% of the stations. N2 and N3 gave the best-fit for 20% and 14% of stations, respectively. N5 gave 11% of the best-fit results. This study also calculated the rainfall heights corresponding to 10-year, 25-year, 50-year, and 100-year return periods for each location by using the distributions to project more extreme values.

Keywords: Gaussian mixture distributions; maximum likelihood; expectation-maximization; extreme events; return period

1. Introduction

For analyzing the risk of rare events, extreme value analysis (EVA) is widely used in various disciplines, including environmental science [1], engineering [2], finance [3], and water resources engineering and management [4–6]. Typically, EVA is used for describing unusual or rare events, (e.g., the upper or lower tails of a distribution) [7]. In hydrology, the purpose of extreme event analysis, such as of floods or precipitation, is to estimate the risk to human beings and environments by extrapolating the observed range of sample data. Extreme precipitation analysis gives some basic information which can be used for the risk assessment of some natural disasters such as floods, droughts, landslides, and so on. The extreme events are expressed in terms of recurrence interval or "return period", the average recurrence interval between events. It can be derived from quantiles of a parametric probability distribution fitted to the extreme values [8].

In probability theory and statistics, the concept of mixture distributions is the combination of two or more probability distributions [9,10] to create a new probability distribution. Finite mixture densities have served as important models for complex processes [11]. The most frequently applied finite mixture distributions are Gaussian mixtures. Gaussian mixture distributions (GMDs) are formed by taking linear combinations of Gaussian distributions. It is a weighted sum of Gaussian component

densities. The applications of GMD can be found in various disciplines, such as biometric systems [12], astronomy [13], biology [14], finance [15], environment (such as water quality) [16], and floods [17,18]. However, in precipitation analysis, GMD is seldom used, whereas other mixture models—such as mixtures of gamma and generalized Pareto distributions (GPD)—were implemented [19–21].

The most commonly used probability distributions in hydrology include normal (N), log–normal (LN2), Pearson type 3 (P3), log–Pearson type 3 (LP3), generalized extreme value (GEV), and Gumbel (GUM) [22,23]. On the other hand, in empirical finance, there are many studies on the estimation of portfolio returns and value at risk (VaR) by using the class of Gaussian mixture distributions [24,25]. He [16] used the GMD model for environmental data, such as water quality data. The GMD model shows a great flexibility in capturing various density shapes. However, this same flexibility leads to some estimation problems. There are many methods that have been developed for solving the parameter estimation problems ranging from Pearson's method of moments, through the formal maximum likelihood method, to informal graphical techniques. Among these methods, maximum likelihood (ML) estimation is the most widely used method because it possesses desirable statistical properties. An ML estimate related to a sample of observations is a selection of parameters which maximizes the probability density function of the sample, called (in this context) the likelihood function (LF). LF plays an important role in statistical inference, especially in the method of parameter estimation from a set of statistics. The most commonly used and powerful method for solving the ML estimation problem is called the expectation-maximization algorithm, or EM algorithm [26,27]. The mixture-density parameter estimation problem is one of the most frequent applications of the EM algorithm in the computational pattern recognition discipline.

In water resources design and management, return period analysis is widely used in the management and communication of risk. Its use is especially common in determining hydrologic risk of failure. A common use of return period is to estimate the recurrence interval of an event such as a flood, drought, landslide, earthquake, and others. The return period of an event (e.g., precipitation, flood) is the interval between the events which exceeds a selected threshold [28,29]. In water resources engineering, the term "return period" can be defined as the average number of years to the first occurrence of an event of magnitude greater than a given level [30].

The precipitation pattern and its quantity during a specific duration—such as hourly, daily, monthly, and yearly—play a crucial role in water resources planning and management. For regional rainfall frequency analysis in the Zayandehrood Basin in Iran, Eslamian and Feizi [31] used maximum monthly rainfall, taken as the wettest month in each year, as the extreme event and found generalized extreme-value and Pearson type-3 distributions were the best-fit distributions for a specific station in that area.

The main objectives of this study were (1) to select the best-fit distributions of the GMD and (2) to estimate the highest rainfall values corresponding to the return period values equal to 10, 25, 50, and 100 years. The results of return period of best-fit distributions for the meteorological stations of Bangladesh can be used for risk policy and design purposes.

2. Data and Study Area

Bangladesh is in the Ganges-Brahmaputra-Meghna (GBM) river basin, which is the third largest freshwater outlet to the world's oceans. The country is between latitudes 20°30′ N and 26°45′ N and longitudes 88°0′ E and 92°45′ E (Figure 1). The total land area is 147,570 km^2. In the GBM basin there are many rivers, most of them originating from the Himalayas, north of Bangladesh, and passing through the country to the Bay of Bengal, south of the country. Bangladesh is a riverine country, with 79% of the country being a floodplain. The land was formed by the river delta process. This fertile floodplain land contributes to a significant agriculture-based economy. On the other hand, there are some hilly areas, 12% of total area, which are located in the southeast and northeast part of the country. Nine percent of the land area is occupied by four uplifted blocks, which are mainly located in the northwest and central parts of the country. In the floodplain area, the highest elevation is about 105 m

above sea level, which is in the north part of the country. Elevation decreases in the coastal south. In the hilly areas, the southeast part of the country, elevation varies from 600 to 900 m above sea level.

Bangladesh is an agricultural-based economy, where the role of precipitation is important. The Bay of Bengal lies to the southern part of the country, so much water vapor comes to the country and causes rainfall. Rivers which originate from the Himalayas to the north flow through Bangladesh and often cause flooding during this time. As the geographical conditions affect the precipitation patterns, these studies will play an important role on flood prevention and protection of natural assets.

The climate of Bangladesh is tropical monsoon-type, with a hot summer monsoon and a pronounced dry season in the winter. The effect of climate on hydrology in this tropical area has many facets. During the summer monsoon period, from June to October, excessive rainfall occurs—about 72% of annual rainfall occurs during this time period [32]. This excessive seasonal rainfall causes floods during this time. Temperatures throughout the country are almost uniform spatially, the month of July (28–29 °C) showing the highest and the month of January (17–19 °C) showing the lowest temperature, on average.

The daily rainfall data were collected from the Bangladesh Meteorological Department (BMD) from 35 different locations across the country (Figure 1). Rainfall stations are marked with a serial number from 1 to 35 in order of north to south on the map in Figure 1. The elevation of the locations of each station, the period of observation data, and percentage of missing values are presented in Table 1. The elevation of the stations was measured from "Google Earth" by using the coordinates of the locations. The geographical and climatological conditions are different, and the rainfall patterns also vary from station to station. The data was provided as the daily total rainfall in millimeters at each location. In this analysis, for most of the stations, 30 years of data (1984–2013) are used. However, there are some newer stations which were installed more recently that have less than 30 years of recorded data. These are Ambagan (15 years), Chuadanga (25 years), Mongla (23 years), Kutubdia (29 years), Sydpur (23 years), and Tangail (27 years). Firstly, the summation of daily rainfall of each calendar month was calculated. Then, the highest total in each year was taken as the maximum monthly rainfall for each location. This yields 30 maxima (1 for each year) for each station. This maximum monthly rainfall was used as the variable for analysis of extreme value (rainfall) estimation. Generally, the monsoon period, from June to October, has the maximum monthly rainfall each year all over the country. So, the maximum monthly rainfall came from the calendar month of July, August, or September in all 30 years studied. The best-fit probability distribution of these meteorological locations in Bangladesh was determined by using the GMD.

Geographical conditions play an important role in the precipitation pattern of a certain area. Geographic location, elevation, and adjacent environmental factors have a significant role on the rainfall pattern of a certain area. The compiled data varies from site to site. The southeastern part of Bangladesh has the highest amount of measured precipitation, mainly due to it being bounded by hills and the sea. For example, one station, named Sandwip, on the coast has recorded 3001 mm monthly maximum rainfall in the past 35 years. The northeastern part also has large amounts of precipitation. The main reason is that it is surrounded by the hilly areas of India, with the Tibetan plateau nearby. The Himalayan range and the Tibetan plateau are the source of many rivers in this area. Because of the unique geographical pattern of this area, with the combined influence of the Himalayan range and the Tibetan plateau, on the floodplain of the lower part of the Brahmaputra basin, with the addition of the monsoon driven with a distinct wet season from June to September, the total amount of precipitation and its frequency can produce particularly intense floods in this area.

The rest of the land is the part of the Ganges river basin. In the territory of Bangladesh, the basin is mostly floodplain and shows lower elevation than other parts of the country. The stations in the northwestern part of Bangladesh measured lower amounts of precipitation (such as Ishardi station, with 664 mm monthly maximum) than the southeastern and northeastern parts of the country.

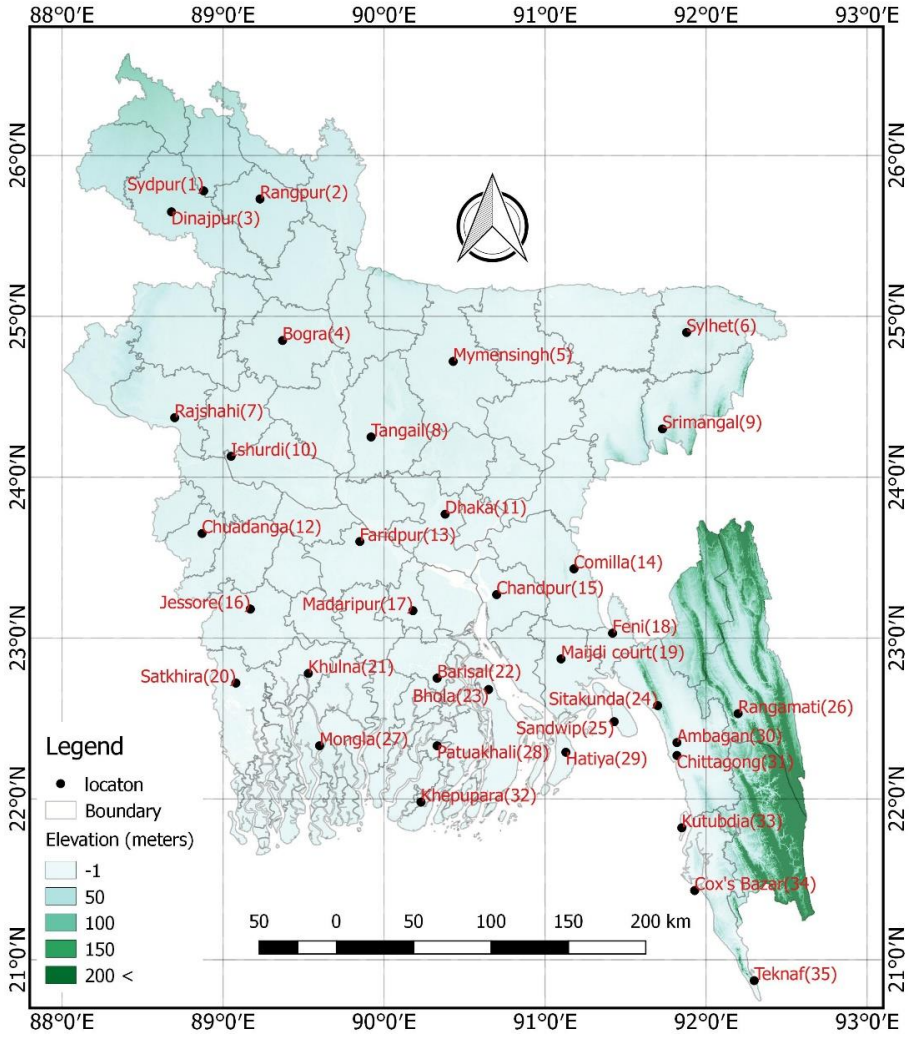

Figure 1. Meteorological stations of Bangladesh.

Table 1. Descriptions of data set of the Bangladesh Meteorological Department (BMD) stations.

St.No	Station Name	Elevation (m)	Missing Values (%)	Observed Period	St.No	Station Name	Elevation (m)	Missing Values (%)	Observed Period
1	Sydpur	45	0	1991–2013	19	Maijdi Court	9	0.41	1984–2013
2	Rangpur	34	0	1984–2013	20	Satkhira	6	0	1984–2013
3	Dinajpur	37	0.03	1984–2013	21	Khulna	4	0.28	1984–2013
4	Bogra	20	0.05	1984–2013	22	Barisal	4	0.05	1984–2013
5	Mymensingh	19	0	1984–2013	23	Bhola	5	0.05	1984–2013
6	Sylhet	35	0.07	1984–2013	24	Sitakunda	4	0.24	1984–2013
7	Rajshahi	20	0.14	1984–2013	25	Sandwip	6	1.58	1984–2013
8	Tangail	10	0.07	1987–2013	26	Rangamati	63	0.05	1984–2013
9	Srimangal	43	1.19	1984–2013	27	Mongla	4	0	1991–2013
10	Ishurdi	14	0.14	1984–2013	28	Patuakhali	3	0	1984–2013
11	Dhaka	9	0.03	1984–2013	29	Hatiya	4	15.18	1984–2013
12	Chuadanga	12	0.92	1989–2013	30	Ambagan	21	0	1999–2013
13	Faridpur	9	0	1984–2013	31	Chittagong	6	13.86	1984–2013
14	Comilla	12	0.06	1984–2013	32	Khepupara	3	0.34	1984–2013
15	Chandpur	7	0.14	1984–2013	33	Kutubdia	6	0.15	1985–2013
16	Jessore	7	0	1984–2013	34	Cox's Bazar	4	0.32	1984–2013
17	Madaripur	5	0.02	1984–2013	35	Teknaf	4	0.06	1984–2013
18	Feni	8	0.34	1984–2013					

3. Methodology

For selecting the best-fit model for a certain location, choice of the model definition, parameter estimation, and model selection tools are important. In this section, these are described. The method of parameter estimation of the distributions is presented in Section 3.1. In Section 3.2, the procedure of goodness-of-fit tests for model selection, both numerically and graphically, is discussed. In Section 3.3, the return period estimation procedure of extreme event is discussed.

3.1. Gaussian Mixture Distributions

GMD, the most popular mixture model, is a useful tool for density estimation. The Gaussian distribution is the most important and widespread distribution in the field of statistical modeling. The mixture of Gaussian distributions yielded a wide variety of curves that describe the statistical variability. One reason for this is that the univariate Gaussian distribution is simple and requires only two parameters, the mean μ and the variance σ^2. The Gaussian density is symmetric, unimodal, isotropic, and assumes the least prior knowledge. With a given mean and variance, it is easy to estimate an unknown probability density [33]. These characteristics and as its well-studied status provide Gaussian mixture density models more power and effectiveness than other mixture densities. For an independently and identically distributed (iid) random variable X drawn from K different normal distributions with weights p_k, the component probability density function of GMD can be written as [34,35]:

$$p(x) = \sum_{k=1}^{K} p_k N_k(x|\mu_k\sigma_k^2) \tag{1}$$

where x represents a one-dimensional random variable; k = 1, 2, . . . , K. The mixing coefficients p_k must satisfy the conditions $0 \leq p_k \leq 1$ and $\sum_{k=1}^{K} p_k = 1$ in order to be valid. The component Gaussian densities, $N_k(x|\mu_k\sigma_k^2)$, can be expressed as:

$$N_k(x|\mu_k\sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\} \tag{2}$$

where μ_k is the mean and ${\sigma_k}^2$ is the variance for the kth Gaussian distribution.

Maximum likelihood estimators, the well-known parameter estimators, have desirable asymptotic properties. Thus, it is a commonly used method for estimating the parameters in a mixture of Gaussian distributions. The likelihood function of the GMD can be defined as [34,35]:

$$l(x|\Theta) = \log\prod_{i=1}^{N}\sum_{k=1}^{K} p_k N_k(x|\mu_k\sigma_k^2) = \sum_{i=1}^{N}\log\sum_{k=1}^{K} p_k N_k(x|\mu_k\sigma_k^2) \tag{3}$$

where $\Theta = (\theta_1, \theta_2, \ldots . \theta_K)$ and $\theta_k = (p_k, \mu_k, {\sigma_k}^2)$. For K sets of Gaussian distributions, the same sets of parameters are needed to calculate the estimate.

In general, it is useless to obtain an analytical solution to maximize Equation (3) due to the composite operation of component-wise product and sum. The EM algorithm, the powerful method for finding maximum likelihood estimators, is applied to generate the unknown parameters in GMD. This algorithm is an iterative procedure for estimating the parameters of a certain distribution. There are two steps—the expectation (E-step) and the maximization (M-step)—for obtaining the maximum likelihood estimate [34,35].

E-step: calculate the responsibilities associated with data point x using the current parameter values:

$$E(p_k|x,\Theta) = \frac{p_k N_k(x|\Theta)}{\sum_{k=1}^{K} p_k N_k(x|\Theta)} \tag{4}$$

M-step: re-estimate and update the parameters using the current responsibilities:

$$\mu_k^{new} = \frac{1}{N_k}\sum_{i=1}^{N} E(p_k|x,\Theta)x \tag{5}$$

$$\sigma_k^{new} = \frac{1}{N_k}\sum_{i=1}^{N} E(p_k|x,\Theta)(x-\mu_k^{new})^2 \tag{6}$$

$$p_k^{new} = \frac{N_k}{N} \tag{7}$$

Firstly, some initial values are chosen for the means, variance, and weights. Then, these are used to get first estimates of $E(p_k \mid x, \Theta)$, which is inserted into Equations (5)–(7) to give revised parameter estimates. An alteration procedure between the above two steps is operated until some convergence criterion is reached. During each update of the parameters resulting from an E-step followed by an M-step, it is guaranteed to increase the log likelihood function. The algorithm is considered to have converged when the change in the log likelihood function, or alternatively, in the parameters, falls below some threshold [34,35].

In this study, the Gaussian distributions used were single normal distributions (N), mixtures of two normal distributions (N2), mixtures of three normal distributions (N3), mixtures of four normal distributions (N4), and mixtures of five normal distributions (N5). The N, N2, N3, N4, and N5 require 2, 5, 8, 11, and 14 parameters, respectively. The calculations were implemented with code written in the "R" programming language.

3.2. Goodness-of-Fit Tests

Goodness-of-fit test statistics are used for checking the validity and choosing the best-fit model among various distribution models for a specific data set. There are many procedures for testing the normality: graphical methods such as histograms with probability distributions, box plots, Q-Q plots, and the formal normality tests such as Akaike's information criterion (AIC), Bayesian information criterion (BIC), root mean square percentage error (RMSPE), and Kolmogorov–Smirnov (K–S). In the present study, AIC, BIC, and RMSPE were used.

According to the AIC and BIC criteria, the value of log-likelihood function is required to estimate the results of AIC and BIC. AIC is a different approach to model selection [36,37]. The AIC is an asymptotically unbiased estimator. For a given model, the AIC can be expressed as:

$$\mathrm{AIC} = -2l + 2K \tag{8}$$

where l denotes the maximum value of the likelihood function and K denotes the number of parameters. Given a set of candidate models for a data set, the best-fit model has the minimum value of the AIC.

The BIC is a criterion for model selection, closely related to the AIC, among a finite set of models. Like the AIC, for a given set of candidate models for a data set, the minimum value gives the best-fit model. The BIC was developed by Schwarz [38], where he explained a Bayesian argument for adopting it. The BIC is defined as:

$$\mathrm{BIC} = -2l + \ln(n)K \tag{9}$$

where n denotes the sample size.

The RMSPE is one of the most common methods to measure residuals—the differences between the observed and simulated values. The smallest RMSPE value gives the best-fit model for a given set of candidate models. It is also a good indicator for measuring errors of various models of particular variables. The RMSPE is expressed as the following equation:

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{x_i - X}{X} * 100\right)^2} \quad (10)$$

where x_i denotes the simulated value, X denotes the observed value, and n denotes the sample size.

Graphical display is one of the most simple and powerful techniques for selecting the best-fit model. The quantile-quantile (Q-Q) plot is implemented to visualize the fitness of model distributions. To calculate the plotting position of the non-exceedance probability $p_{i:n}$, Blom's plotting position formula, shown in Equation (11), is applied to yield approximately unbiased quantiles for a wide range of distributions. Blom's plotting position formula is expressed as:

$$p_{i:n} = \frac{n - (3/8)}{N + 0.25} \quad (11)$$

where N = total number of observed values, n = the rank of the observed value of X ($X_{(i)}$ = ascending order), n = 1, 2, 3, . . . , N. To construct the Q-Q plot, $X_{(i)}$ versus $x(F)$ is plotted, where F is the $p_{i:n}$ for the certain component of the Gaussian mixture distribution.

3.3. Return Period

The most important objective of extreme value frequency analysis is to calculate the recurrence interval or return period. In the mathematical definition, if the variable (X) equal to or greater than an event of magnitude x_T occurs once in T years, then the probability of occurrence $P(X \geq x)$ in a given year of the variable is expressed as:

$$P(X \geq x_T) = \frac{1}{T} \quad (12)$$

$$T = \frac{1}{1 - P(X \leq x_T)} \quad (13)$$

The precipitation amounts associated with the 50-year or 100-year average return periods cannot be directly calculated from the data set used here, but must be extrapolated from the 98th and 99th percentiles, respectively, of a fitted distribution (i.e., $[1 - 0.98^{-\text{year}}]^{-1}$ = 50 years; $[1 - 0.99^{-\text{year}}]^{-1}$ = 100 years) [8]. Statistical estimates are often presented with a range within which the true value can be expected to lie. One type is the confidence interval (CI). The range of the CI depends on the chosen confidence level. The upper and lower boundary levels of the CI are called confidence limits. In the return period estimations here, the 95% CI of each return period level was calculated.

4. Result and Discussion

Besides many parametric distributions, finite mixture densities have served as important models for complex processes. The main goal of this paper is to identify the best-fit Gaussian mixture distribution model for every station which yields the maximum monthly rainfall for return periods of 10, 25, 50, and 100 years.

4.1. Selecting the Best-Fit Results

Multiple distributions are usually tested against the real data to identify which distribution fits the data the best. Hence, the goal of distribution fitting is to anticipate the probability and frequency of occurrence of a phenomenon of a given magnitude within a certain interval. The selection of the best-fit mixture distribution depends in part on the presence or absence of symmetry of the real data with respect to the mean value. The visual technique of plotting data is one of the important methods for selecting a probability distribution. It is easy to look at the shape of the distribution and judge a best-fit of a given data set. This includes examining a histogram with the distribution overlaid and comparing the empirical model to the theoretical model.

Distributions can be expressed as probability density function (pdf) or cumulative distribution function (cdf). A pdf denotes a continuous probability distribution in terms of integrals. The pdf can be seen as a smoothed version of a probability histogram. The cdf is monotonically increasing between the limits from 0 to 1. Graphical comparisons of all five mixture distributions were created, where pdfs of all five distributions were overlaid onto the histograms of the observed data and cdfs of all five distributions were overlaid onto the empirical cdfs of the observed data sets. Some locations showed best-fit with a larger number of Gaussian distributions, whereas some were best-fit by only a single normal distribution. The fit depends on the pattern of the observed data set. As an example, two locations are illustrated in Figure 2, which shows the fitted pdf with observed data histogram (left side) and cdf with empirical cdf (right side).

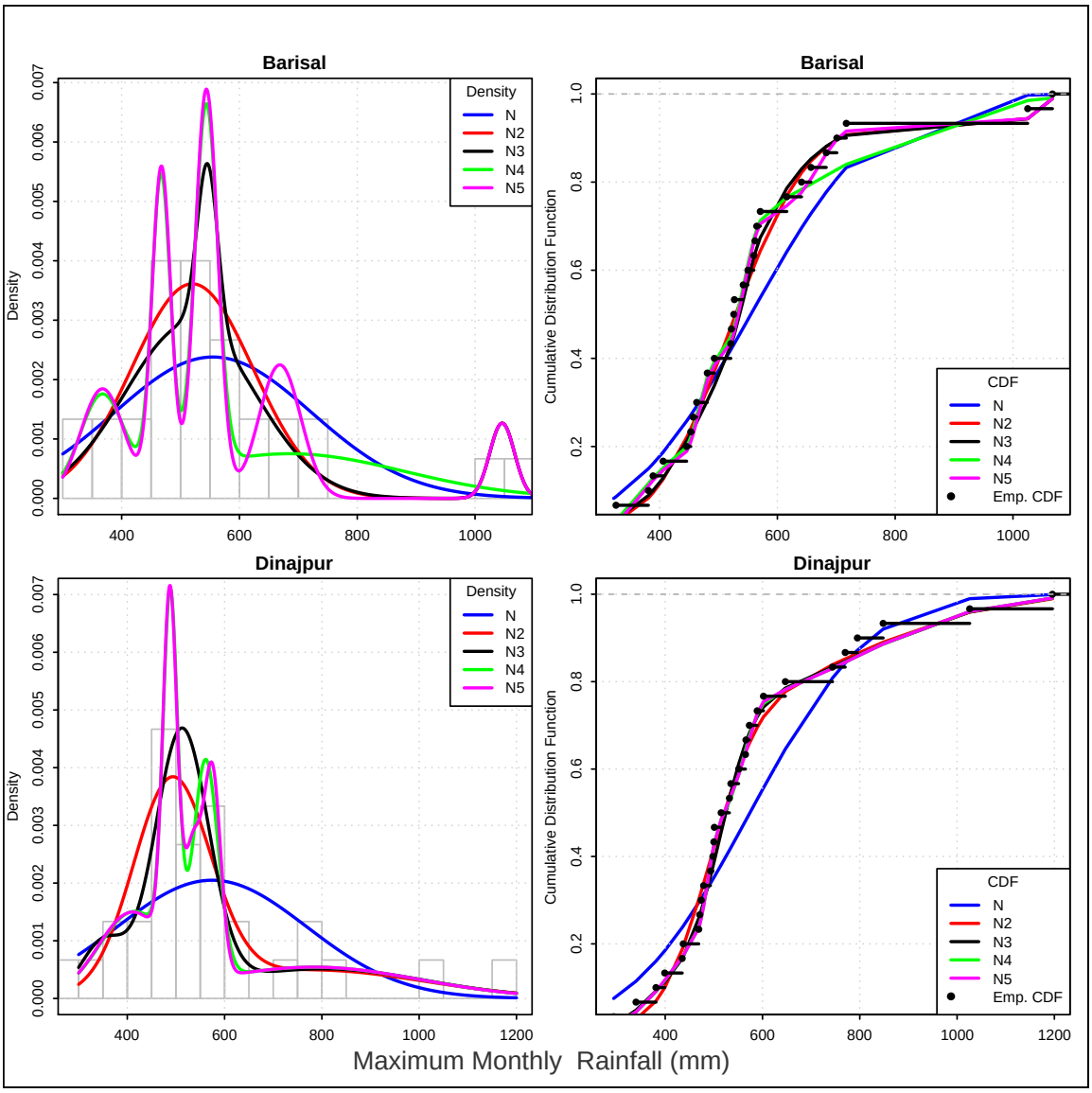

Figure 2. Probability distribution function (pdf) and cumulative distribution function (cdf) of Gaussian mixture distributions of two locations.

For the pdf and cdf plots, the horizontal axis is the range of maximum monthly rainfall data. For the pdf plots, the vertical axis shows the probability density, which varies between the lowest and highest possible values. For the cdf plots, the vertical axis shows the cumulative density function,

where the values increase from 0 to 1 as we go from left to right on the horizontal axis. These figures represent the fit distribution model for the given locations.

The term "probability plot" sometimes refers specifically to a Q-Q plot. This can allow an assessment of "goodness-of-fit" that is graphical, rather than reduction to a numerical summary. Thus, it is easier to judge where the curve best-fits or differs from the data. In general, the basic idea is to calculate the theoretically expected value for each data point based on the distribution in question. The Q-Q plots of the five distributions for each station were created. The distribution fit with observed data was found using RMSPE. By using Q-Q plots, the level of fit on the extreme right tail can be examined [39]. Any perfect data points would follow the [1:1] line. In Figure 3, examples of Q-Q plots for the same two stations of Figure 2 are shown.

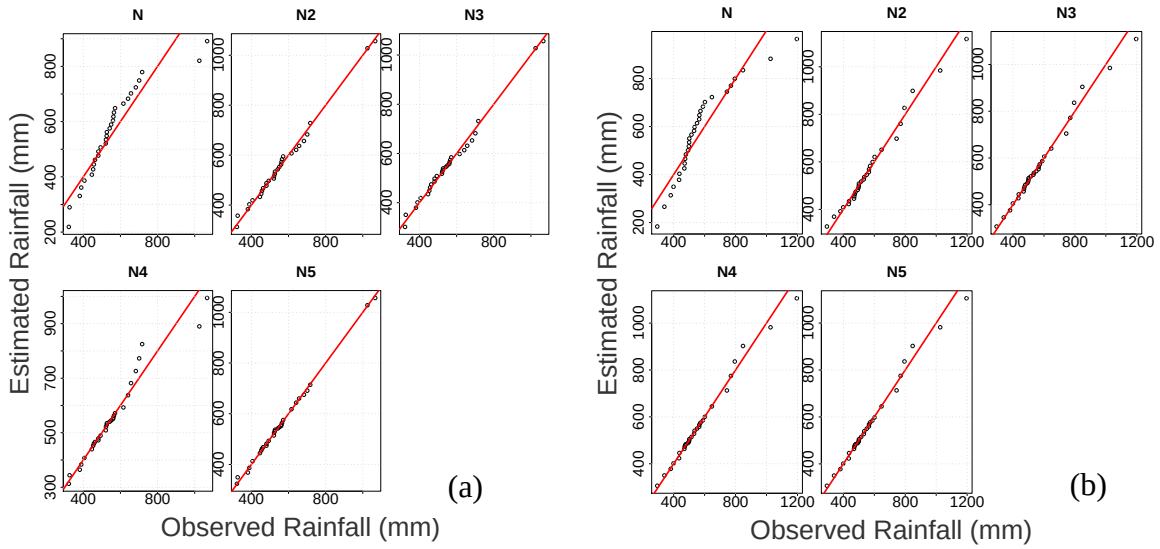

Figure 3. Quantile-quantile (Q-Q) plots for distributions as an example of two stations. (**a**) Q-Q plots of station Barisal; (**b**) Q-Q plots of station Dinajpur.

The horizontal axis shows the observed rainfall data in millimeters and the vertical axis shows the estimated rainfall of the five distributions of Gaussian mixtures. The right tail of the distributions' alignment with the [1:1] line is of interest here. In Figure 2, for the station at Barisal, the N2, N3, N5 distributions visually seem to have the best-fit among all distributions. In Figure 3, the N and N4 distributions deviate significantly from the [1:1] line, which shows the model does not match observed data. The main goal of this probability distribution fitting is to extrapolate the low-probability, extreme events on the extreme right tail. In the case of all other stations, there is no recognizable pattern of best-fit mixture distributions. Sometimes, the right tail can be found to be overestimated or underestimated. However, to determine the best-fit model from the Gaussian mixture distributions, the graphical observation alone is not enough; numerical tests are also needed.

Besides the visual comparison of the shape of the observed data histogram with the pdf, the empirical cdf with the theoretical cdf, and the Q-Q plot, the validity of the specified or assumed distribution models may be verified or disproved statistically by numerical fit tests. Table 2 shows the station names, the best-fit results of AIC, BIC, RMSPE, and best-scored results or highest ranked distribution results from the various components of the Gaussian mixture distributions.

Given a set of candidate models for a data set, the best-fit model is taken as the minimum value of the goodness-of-fit test statistic for every case of AIC, BIC, and RMSPE. In most of the cases, AIC and BIC give the same best-fit distribution for a certain station. The main reason for this is that the log-likelihood function and number of parameters are used for calculating the AIC and BIC. On the other hand, only the simulated value and observed value were used for calculating the RMSPE.

Table 2. Statistical and best-fit results of the BMD stations.

St.No	Station Name	Best-Fit Statistic Results			Highest Ranked Distribution (Sum of Ranks)
		AIC	BIC	RMSPE	
1	Sydpur	N (304.89)	N (307.16)	N5 (3.36)	N (7)
2	Rangpur	N5 (374.15)	N5 (393.76)	N5 (1.76)	N5 (3)
3	Dinajpur	N2 (395.76)	N2 (402.77)	N4 (2.64)	N2 (6)
4	Bogra	N (389.22)	N (392.02)	N5(3.34)	N (6)
5	Mymensingh	N3 (381.7)	N3 (392.91)	N5(1.52)	N3 (4)
6	Sylhet	N (407.99)	N (410.79)	N4(1.37)	N (7)
7	Rajshahi	N4 (375.9)	N (379.96)	N3 (2.61)	N (6)
8	Tangail	N (340.42)	N (343.01)	N2 (1.31)	N (4)
9	Srimangal	N5 (375.67)	N5 (395.29)	N5 (1.78)	N5 (3)
10	Ishurdi	N (382.11)	N (384.91)	N5(2.51)	N (5)
11	Dhaka	N3 (384.59)	N (390.43)	N5 (1.89)	N3 (6)
12	Chuadanga	N (323.73)	N (326.17)	N5(2.84)	N (6)
13	Faridpur	N3 (381.59)	N2 (392.48)	N4 (1.8)	N3 (5)
14	Comilla	N (389.59)	N (392.39)	N5(1.29)	N (5)
15	Chandpur	N5 (378.87)	N5 (398.48)	N5 (1.96)	N5 (3)
16	Jessore	N2 (392.14)	N2 (399.15)	N5 (1.87)	N2 (6)
17	Madaripur	N (390.4)	N (393.2)	N4(1.34)	N (5)
18	Feni	N5 (375.52)	N5 (395.14)	N5 (2.84)	N5 (3)
19	Maijdi Court	N5 (402.56)	N (409.5)	N5(0.91)	N (7)
20	Satkhira	N5 (371.38)	N (376.98)	N4(1.6)	N (7)
21	Khulna	N5 (389.96)	N (395.13)	N2 (2.1)	N (6)
22	Barisal	N2 (381.6)	N2 (388.6)	N2 (2.77)	N2 (3)
23	Bhola	N2 (394.57)	N (400.75)	N5 (2.07)	N2 (5)
24	Sitakunda	N (417.13)	N (419.93)	N5 (1.69)	N (7)
25	Sandwip	N3 (427.37)	N3 (438.31)	N2 (−3.35)	N2 (5)
26	Rangamati	N2 (402.14)	N2 (409.15)	N4 (2.85)	N2 (5)
27	Mongla	N2 (261.95)	N2 (267.63)	N2 (3.85)	N2 (3)
28	Patuakhali	N (398.24)	N (401.04)	N3(1.51)	N (5)
29	Hatiya	N (363.83)	N (366.34)	N5 (3.75)	N (7)
30	Ambagan	N4 (183.44)	N4 (191.23)	N5 (2.03)	N4 (4)
31	Chittagong	N3 (355.94)	N (363.86)	N5 (1.69)	N3 (6)
32	Khepupara	N (389.05)	N (391.86)	N4 (1.62)	N (6)
33	Kutubdia	N2 (409.68)	N2 (416.52)	N2(4.18)	N2 (3)
34	Cox's Bazar	N5 (368.89)	N3 (388.22)	N4 (1.07)	N3 (6)
35	Teknaf	N (416.61)	N (419.41)	N3 (1.28)	N (7)

All developed probability distributions were ranked for each selection tool (rank 1 is the best-fit). The three ranking results were summed to yield a ranking score. For each station, the distribution model with the smallest ranking score was selected as the best-fit and included in Table 2. For most of the stations, the selected best-fit model results match both the AIC and BIC results. In six stations, all three test statistic results are the same. Also, in the higher mixture distributions (N4 or N5), the differences of mixing proportions are very small. This is also shown in the pdf graphs in Figure 2. For the station at Dinajpur, the pdfs of the N4 and the N5 distributions almost overlap at the right tail of the distribution. The main reason is that here the proportion is very small. In the mixture distribution, the proportion among every single mode is an important parameter. In the probability distribution literature, sometimes a single distribution does not give a proper fit, so the mixture of distributions can give a better result. Though it must be kept in mind that increasing the number of parameters could result in overfitting—that is, the creation of a fit that matches the particular data set but has little or no general applicability or predictive power. In this study, a single Gaussian distribution was the most common best-fit, accounting for 51% of the best-fit results. N2 and N3 gave 20% and 14% of the best-fits, respectively. The five-component mixture distribution, N5, gave 11% of the best-fit results.

4.2. Return Period Results

The practical application part of this extreme value frequency analysis is the return period analysis, which yields risk estimations for a certain event. Figure 4 shows rainfall heights of 10-year, 25-year, 50-year, and 100-year return periods of best-fit distributions of all stations with 95% confidence intervals. The horizontal axis represents the station numbers, which are in the "St. No." column of Table 2. The vertical axis represents the expected maximum monthly rainfall. The type of distribution is indicated by the marker shape and color.

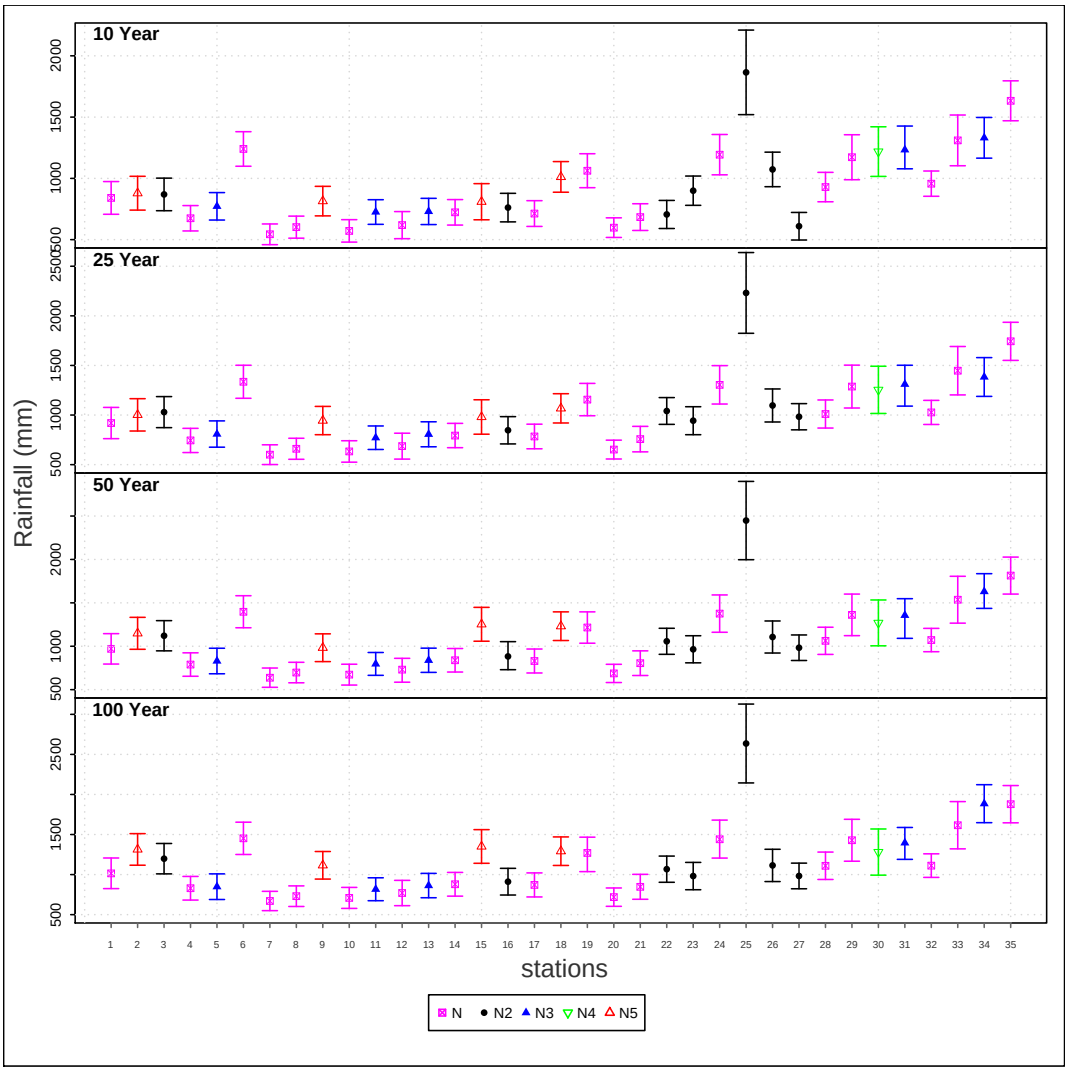

Figure 4. Maximum monthly rainfall heights (in mm, on the y-axis) estimated for each station (on x-axis, the station serial number of Table 1) and for various return period values (10, 25, 50, 100 years). For each station, the rainfall height was calculated by means of the best-fit distribution.

Figure 4 has four sections. From top to bottom, these indicate the rainfall heights corresponding to return period of 10-year, 25-year, 50-year, and 100-year, respectively. As an example, for the station Barisal (St. No. 22), which is best-fit by an N2 distribution, the rainfall amounts of 10-, 25-, 50-, and 100-year return periods are 706 mm, 1041 mm, 1057 mm, and 1068 mm, respectively. In the southeastern region—for example, near the stations including Kutubdia (St. No. 33), Cittagong (St. No. 31), Cox's Bazar (St. No. 34), and Teknaf (St. No. 35)—there is more intense rainfall than in the other regions. For statistical estimates, for expressing the uncertainty level, the CI is crucial in risk analysis as well as in the design purposes.

4.3. Spatial Variability of Extremes

Interpolation can be used to predict unidentified values for any geographic point of data, such as rainfall. It predicts values for cells in a raster from an inadequate number of sample data points. There are various interpolation techniques used to obtain gridded precipitation data based on gauge observations. Here, inverse distance weighting (IDW) with a distance coefficient of 2 was used by "QGIS" for calculating the spatial variability of extreme precipitation in Bangladesh. Spatial interpolation of 10-year, 25-year, 50-year, and 100-year return period of best-fit extreme value distribution are shown in Figure 5.

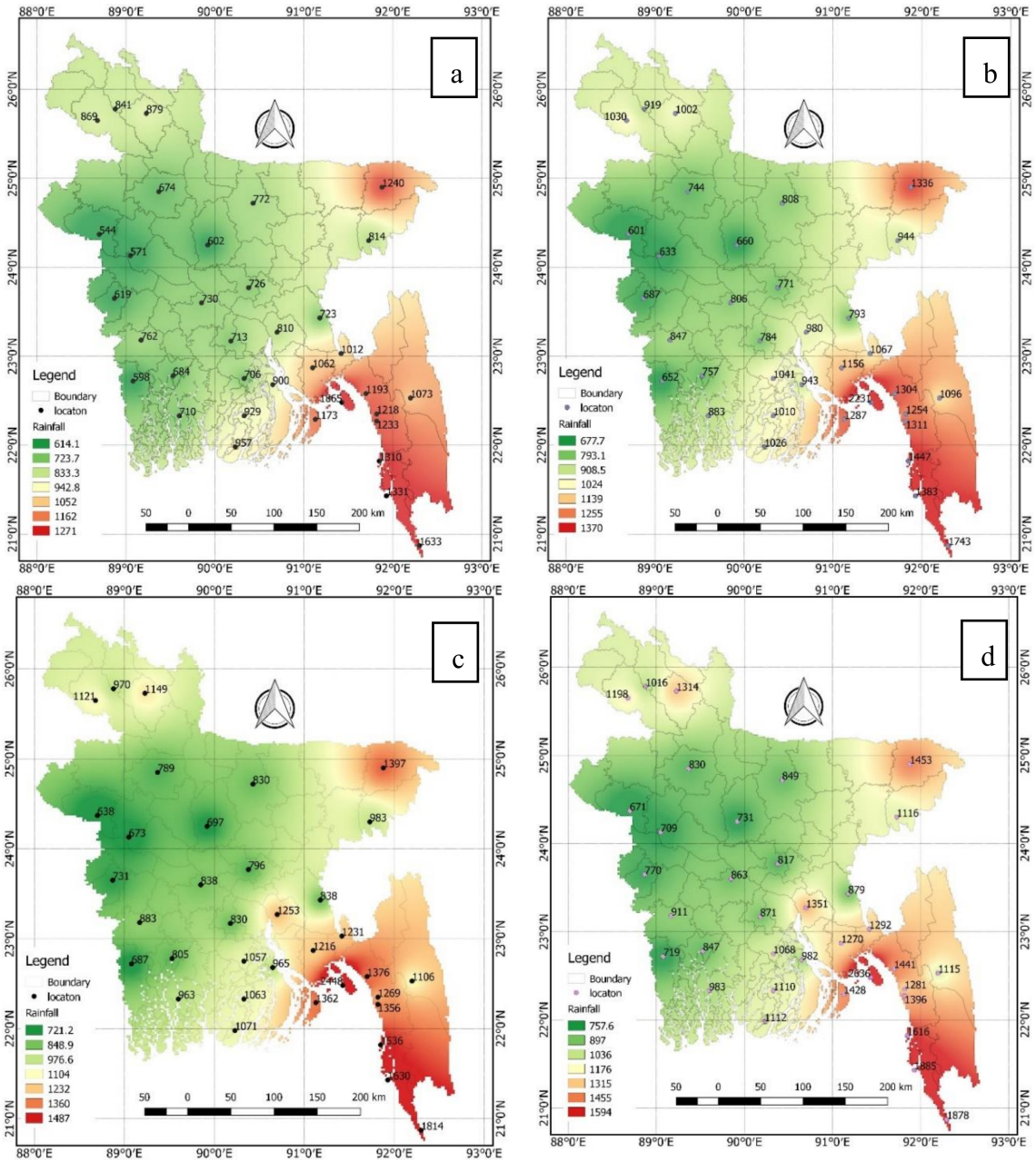

Figure 5. Spatial interpolation of maximum monthly rainfall heights calculated by means of the best-fit distribution for different return period values: 10 years (**a**), 25 years (**b**), 50 years (**c**), and 100 years (**d**).

The southeastern part of the country shows the highest amount of rainfall because of the presence of hills and it is near to the Bay of Bengal. The northeastern part also contains hills but is far from the ocean. The northeastern part is also near the Himalayas. The rest of the country is low elevation, floodplain areas. Sometimes, the western region faces drought because of less rainfall and water flow. Yen [40] claims that, for infrastructural flood design, a 100-year return period is useful. Overall, the use of return period duration depends on the purpose or intent of the policymakers.

5. Conclusions

Finite mixture distributions, especially the Gaussian mixture distribution, are widely used in various disciplines. This study applied from one to five components of univariate Gaussian mixtures to analyze the extreme values of precipitation. The rainfall pattern across the country differs. The geographical and physical condition varies. The southeastern part is bordered by both hills and the sea. The intensity of rainfall is higher than the other areas. During the monsoon season, this leads to more floods and landslides in this area, which cause deaths and damage of assets. Sarker and Rashid [41] also mentioned that excessive rainfall in the piedmont of hilly areas is the main source of flashfloods and the resultant landslides, specifically in the areas composed of unconsolidated rocks. Slope saturation by water is the main cause of these landslides. A number of graphical and numerical performance criteria were used to assess both the descriptive and predictive abilities of the models. More specifically, graphical inspection (pdf, cdf, Q-Q plot) and numerical criteria (AIC, BIC, RMSPE) were used to select the best-fit model for each of the 35 weather stations. In most of the cases, AIC and BIC give the same best-fit results, but differ from the results of RMSPE. This makes it complex to make a decision as to which is the best-fit. A scoring system was applied to choose the best-fit distribution for each location. The best-fit result of each station was chosen as the distribution with the lowest sum of the rank scores from each test statistic. The N (single distribution) gives the best-fit result for 51% of the stations. N2 and N3 gave best-fit for 20% and 14% of stations, respectively. The five-component mixture distribution, N5, gave 11% of the best-fit results.

This study also shows the return period calculation for each location by using the components of Gaussian mixture distributions. The rainfall heights corresponding to the 10-year, 25-year, 50-year, and 100-year return periods were calculated. The selection of return period levels depends on the decision-makers to choose the duration and risk level. This study can help policymakers to plan initiatives that could result in saving lives and assets.

Acknowledgments: The authors acknowledge the financial support of the general research funding of the Osaka City University School of Human Life Science.

Author Contributions: Md Ashraful Alam and Kazuo Emura conceived the research theme. Md Ashraful Alam performed the data analysis, the interpretation of results and wrote the paper. Craig Farnham contributed to writing the paper as well as evaluated the results, chose and advised on calculation methods with Kazuo Emura.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Smith, R.L. Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone. *Stat. Sci.* **1989**, *4*, 367–377. [CrossRef]
2. Castillo, E.; Hadi, A.; Balakrishnan, N.; Sarabia, J. *Extreme Value and Related Models with Applications in Engineering and Science*, 1st ed.; John Wiley & Sons: New York, NY, USA, 2004; pp. 3–18, ISBN 0-471-67172-X.
3. Longin, F. *Extreme Events in Finance: A Handbook of Extreme Value Theory and Its Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2016; pp. 1–10, ISBN 1118650190.
4. Katz, R.; Parlange, M.; Naveau, P. Statistics of extremes in hydrology. *Adv. Water Resour.* **2002**, *25*, 1287–1304. [CrossRef]
5. Smith, R.L. Extreme values. In *Environmental Statistics*; Department of Statistics, University of North Carolina: Chapel Hill, NC, USA, 2001; pp. 300–357. Available online: www.stat.unc.edu/postscript/rs/envnotes.pdf (accessed on 26 February 2018).

6. Coles, S. *An Introduction to Statistical Modeling of Extreme Values*, 1st ed.; Springer: London, UK, 2001; pp. 1–72, ISBN 978-1-84996-874-4.
7. MacDonald, A.; Scarrott, C.J.; Lee, D.; Darlow, B.; Reale, M.; Russell, G. A flexible extreme value mixture model. *Comput. Stat. Data Anal.* **2011**, *55*, 2137–2157. [CrossRef]
8. Wilks, D.S. Comparison of three-parameter probability distributions for representing annual extreme and partial duration precipitation series. *Water Resour. Res.* **1993**, *29*, 3543–3549. [CrossRef]
9. McLachlan, G.J.; Basford, K.E. *Mixture Modelss: Inference and Applications to Clustering*, 1st ed.; Marcel Dekker, Inc.: New York, NY, USA, 1987; pp. 1–8, ISBN 0-8247-7691-7.
10. McLachlan, G.J.; Peel, D. *Finite Mixture Models*, 1st ed.; John Wiley & Sons: New York, NY, USA, 2000; pp. 1–19, ISBN 0-471-00626-2.
11. Zhuang, X.; Huang, Y.; Palaniappan, K.; Zhao, Y. Gaussian Mixture Density Modeling, Decomposition, and Applications. *IEEE Trans. Image Process.* **1996**, *5*, 1293–1302. [CrossRef] [PubMed]
12. Reyonlds, D. Gaussian mixture models. In *Encyclopedia of Biometrics*, 2nd ed.; Li, S.Z., Jain, A.K., Eds.; Springer: New York, NY, USA, 2007; pp. 827–832, ISBN 978-1-4899-7487-7.
13. Lee, K.J.; Guillemot, L.; Yue, Y.L.; Kramer, M.; Champion, D.J. Application of the Gaussian mixture model in pulsar astronomy—Pulsar classification and candidates ranking for the Fermi 2FGL catalogue. *Mon. Not. R. Astron. Soc.* **2012**, *424*, 2832–2840. [CrossRef]
14. Gregor, J. An algorithm for the decomposition of a distribution into Gaussian components. *Biometrics* **1969**, *25*, 79–93. [CrossRef] [PubMed]
15. Wirjanto, T.S.; Xu, D. *The Applications of Mixtures of Normal Distributions in Empirical Finance: A Selected Survey*; Working Paper; University of Waterloo: Waterloo, ON, Canada, 2009.
16. He, J. Mixed model based multivariate statistical analysis of multiply censored environmental data. *Adv. Water Resour.* **2013**, *59*, 14–24. [CrossRef]
17. Diehli, T.; Potter, K.W. Mixed flood distribution in Wisconsin. In *Hydrologic Frequency Modeling*; Singh, V.P., Ed.; Springer: Dordrecht, The Netherlands, 1987; pp. 213–226.
18. Fan, Y.R.; Huang, W.W.; Huang, G.H.; Li, Y.P.; Huang, K.; Li, Z. Hydrologic risk analysis in the Yangtze River basin through coupling Gaussian mixtures into copulas. *Adv. Water Resour.* **2016**, *88*, 170–185. [CrossRef]
19. Vrac, M.; Naveau, P. Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resour. Res.* **2007**, *43*, W07402. [CrossRef]
20. Furrer, E.M.; Katz, R.W. Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resour. Res.* **2008**, *44*, W12439. [CrossRef]
21. Hundecha, Y.; Pahlow, M.; Schumann, A. Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes. *Water Resour. Res.* **2009**, *45*, W12412. [CrossRef]
22. Stendinger, J.R.; Vogel, R.M.; Foufoula-Georgiou, E. Frequency Analysis of Extreme Events. In *Handbook of Hydrology*, 1st ed.; Maidment, D.A., Ed.; McGraw-Hill: New York, NY, USA, 1993; Chapter 18, ISBN 0070397325.
23. Alam, M.A.; Emura, K.; Farnham, C.; Yuan, J. Best-Fit Probability Distributions and Return Periods for Maximum Monthly Rainfall in Bangladesh. *Climate* **2018**, *6*, 9. [CrossRef]
24. Tan, K.; Chu, M. Estimation of Portfolio Return and Value at Risk using a class of Gaussian Mixture Distributions. *Int. J. Bus. Financ. Res.* **2012**, *6*, 97–107.
25. Hass, M. Value-at-Risk via mixture distributions reconsidered. *Appl. Math. Comput.* **2009**, *215*, 2103–2119. [CrossRef]
26. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38.
27. McLachlan, G.J.; Krishnan, T. *The EM Algorithm and Extensions*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2008; pp. 1–40, ISBN 978-0-471-20170-0.
28. Gumbel, E.J. The return period of flood flows. *Ann. Math. Stat.* **1941**, *12*, 163–190. [CrossRef]
29. Thomas, H.A. Frequency of minor floods. *J. Boston Soc. Civ. Eng.* **1949**, *35*, 425–442.
30. Benjamin, J.R.; Cornell, C.A. *Probability, Statistics, and Decision for Civil Engineers*, 1st ed.; McGraw-Hill: New York, NY, USA, 1970; pp. 232–235, ISBN 978-0-486-78072-6.
31. Eslamian, S.S.; Feizi, H. Maximum Monthly Rainfall Analysis Using L-Moments for an Arid Region in Isfahan Province, Iran. *J. Appl. Meteorol. Climatol.* **2007**, *46*, 494–503. [CrossRef]
32. Ahsan, M.N.; Chowdhary, M.A.M.; Quadir, D.A. Variability and rends of summer monsoon rainfall over Bangladesh. *J. Hydrol. Meteorol.* **2010**, *7*, 1–17. [CrossRef]

33. Kapur, J.N. *Maximum-Entropy Models in Science and Engineering*, 1st ed.; Wiley Eastern Limited: New Delhi, India, 1989; pp. 44–47, ISBN 81-224-0216-X.
34. Christopher, M.B. *Pattern Recognition and Machine Learning*, 1st ed.; Springer: New York, NY, USA, 2016; pp. 430–439, ISBN 978-0387-31073-2.
35. Everitt, B.S.; Hand, D.J. *Finite Mixture Distributions*, 1st ed.; Chapman and Hall: London, UK, 1981; pp. 25–57, ISBN 0-412-22420-8.
36. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Proceeding in the Second International Symposium on Information Theory*; Petrov, B.N., Caski, F., Eds.; Akademiai Kiado: Budapest, Hungary, 1973; pp. 267–281.
37. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
38. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
39. Graedel, T.E.; Kleiner, B. Exploratory analysis of atmospheric data. In *Probability, Statistics and Decision Making in the Atmospheric Sciences*, 1st ed.; Murphy, A.H., Katz, R.W., Eds.; Westview: Boulder, CO, USA, 1985; pp. 1–43, ISBN 0865311528.
40. Yen, B.C. Risks in hydrologic design of engineering projects. *J. Hydraul. Div. Am. Soc. Civ. Eng.* **1970**, *96*, 959–965.
41. Sarker, A.A.; Rashid, A.K.M.M. Landslide and Flashflood in Bangladesh. In *Disaster Risk Reduction Approaches in Bangladesh*; Shaw, R., Mallick, F., Islam, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 165–189.