Vol. 7 Issue 2 Meteorology Hydrology and Water Management Research and Operational Applications June 2019

# Homogeneous regionalization via L-moments for Mumbai City, India

Amit Sharad Parchure, Shirish Kumar Gedam
*Indian Institute of Technology (IIT), Centre of Studies in Resources Engineering, Bombay 400076, India, e-mail: amit231279@gmail.com*

**Abstract.** This study identified homogeneous rainfall regions using a combination of cluster analysis and the L-moments approach. The L-moments of heavy rainfall events of various durations (0.25, 1, 6, 12, 24, 48, 72, 96, and 120 h) were analysed using seasonal (June-September) rainfall measurements at 47 meteorological stations over the period 2006-2016. In the primary phase of this study, the homogeneity of Mumbai as a single region was examined by statistical testing (based on L-moment ratios and variations of the L-moments). The K-means clustering approach was applied to the site characteristics to identify candidate regions. Based on the most appropriate distribution, these regions were subsequently tested using at-site statistics to form the final homogeneous regions. For durations above 1h, the regionalisation procedure delineated six clusters of similarly behaved rain gauges, where each cluster represented one separate class of variables for the rain gauges. However, for durations below 1h, the regionalisation procedure was not efficient in the sense of identifying homogeneous regions for rainfall. Furthermore, the final clusters confirmed that the spatial variation of rainfall was related to the complex topography, which comprised flatlands (below or at mean sea level), urban areas with high rise buildings, and mountainous and hilly areas.

**Keywords:** regional analysis, L-moments, tests for homogeneity, K-means clustering, principal components analysis

**Submitted** 12 September 2018, **revised** 2 April 2019, **accepted** 4 June 2019

## 1. Introduction

Rainfall flooding is one of the most dangerous natural hazards as it affects the economy, environment, and population. Some recent studies indicate that heavily urbanized megacities in low-lying coastal areas are hotspots for flooding (Hallegatte 2010). In Mumbai, which is one of the megacities along the coast of India, severe floods occur almost every year. Furthermore, the extreme rainfall event that occurred on July 26$^{th}$, 2005 demonstrated the high spatio-temporal variability of rainfall events in different areas over a 24-h period (Lokanadham et al. 2012). Therefore, it is vital to assess the regionalization of hydroclimatic variables such as flooding, evapotranspiration, and rainfall in order to optimize efficiency in design and reduce uncertainties.

Regionalization is generally used when rain gauge data are not available at a target site or to improve at-site (single) estimates, especially for short data records (Malekinezhad, Zare-Garizi 2014; Sun et al. 2015; Halbert et al. 2016; Requena et al. 2016). This approach involves "trading time for space" by pooling observations for stations with similar behavior. Various rainfall regionalization techniques have been developed and applied by researchers worldwide, for example, in Pakistan (Khan et al. 2017), Slovakia (Gaál et al. 2009), the Brazilian Amazon (Santos et al. 2015), Jeju Island, Korea (Kar et al. 2017), and mid-Norway (Hailegeorgis, Alfredsen 2017). In India, rainfall regionalization techniques that have been developed and applied include principal components analysis (Nair et al. 2013), correlation analysis (Sinha et al. 2013), cluster analysis (Ahuja, Dhanya 2012; Bharath, Srinivas 2015), neural networks (Saha et al. 2017), and shared nearest neighbor (Kakade, Kulkarni 2017). However, most of these applications have been on a national rather than regional scale. In addition, most of these cluster analysis studies have been conducted using a top-down approach, which explains the top-down control of the large-scale climatic attributes (e.g., mean annual precipitation, temperature, wind velocity, wind direction, and specific humidity) over regional hydrological processes and patterns (Hessburg et al. 2005). Based on these attributes, homogeneous regions are identified, which are expected to be reflected in the records of the hydrometeorological variables of interest. In this context, the L-moments approach is a promising technique, and is a well-known and widely used procedure for regionalization (Hosking, Wallis 1997; Ngongondo et al. 2011; Rahman et al. 2013). This method is relatively insensitive to outliers, and the parameter estimates are more reliable than conventional moment estimates, especially for small samples. Furthermore, the estimators of the L-moments are virtually unbiased (Smithers, Schulze 2001). In this study, the L-moments algorithm with site characteristics (objective) and site statistical (subjective, process-based) pooling techniques were used to cluster

rain gauges within the region into groups, thus augmenting the comparability of the rain gauges.

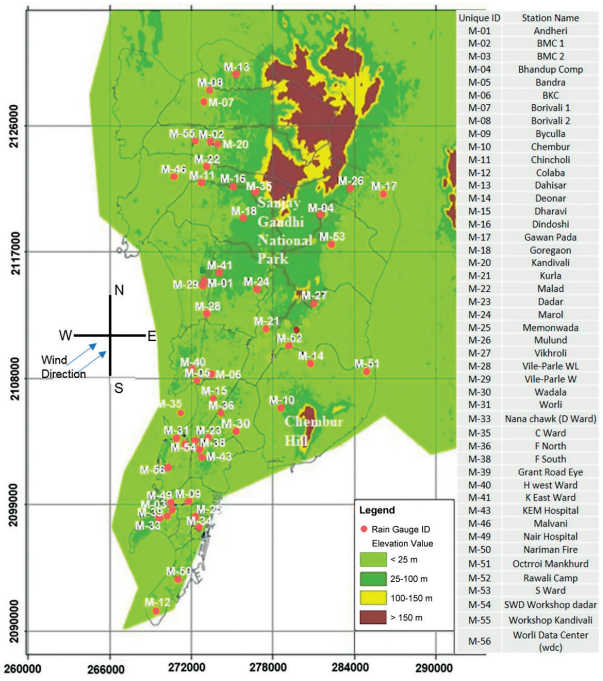
The research presented in this paper was motivated by the fact that most of the studies in this region have used data from two Indian Metrological Department (IMD) stations (Colaba and Santacruz) to investigate the formation and prediction of extreme events and estimation of design rainfall amounts. Furthermore, the July 26$^{th}$, 2005 event drove the local authorities (the Municipal Corporation of Greater Mumbai (MCGM)) to install a dense network of rain gauges to measure continuous and consistent rainfall. As a result, 60 rain gauges were installed that provided more than five years of steady, coherent observations. These rain gauges were installed without prior studies on the selection of optimal sites. Moreover, Parchure and Gedam (2018) used a priori knowledge of rainfall events, and considering the advantage of Self-Organising Maps (SOMs) for (e.g.), abstraction of attributes, displayed the distribution of each component, and for effective visualisation, analysed the clustered rain gauges in groups (regions). However, this is a time-consuming method when compared with the L-moment approach. Hence, the primary intent of this study is to identify homogeneous rainfall regions at the local scale using the L-moments approach.

The specific objectives of this study are as follows:

- To perform a regionalization of heavy rainfall (of 0.25, 1, 6, 12, 24, 48, 72, 96, and 120 h duration) using the L-moments approach.
- To check whether the study area behaves as a single homogeneous region.
- To identify homogeneous rainfall regions by combining the results obtained from the site characteristics and site statistics pooling techniques.
- To evaluate the performance of the top-down approach, i.e. the L-moments approach

## 2. Study area and data description

Mumbai is situated on the western coast of India and extends between 18.00°-19.20°N and 72.00°-73.00°E. The region has a humid, tropical climate, with monsoons that move in from the southwestern Indian Ocean from June to September. Initially, Mumbai was composed of a group of islands that have now been reclaimed to meet the demand for land. This reclamation has resulted in the region having a complex topography that comprises flatlands (below or at mean sea level), urban areas, and mountainous and hilly areas (for instance, the Sanjay Gandhi National Park that is located in the northern part has an elevation of up to 450 m above mean sea level). Sub-hourly precipitation data (at 15 min intervals) were




| Unique ID | Station Name |
|---|---|
| M-01 | Andheri |
| M-02 | BMC 1 |
| M-03 | BMC 2 |
| M-04 | Bhandup Comp |
| M-05 | Bandra |
| M-06 | BKC |
| M-07 | Borivali 1 |
| M-08 | Borivali 2 |
| M-09 | Byculla |
| M-10 | Chembur |
| M-11 | Chincholi |
| M-12 | Colaba |
| M-13 | Dahisar |
| M-14 | Deonar |
| M-15 | Dharavi |
| M-16 | Dindoshi |
| M-17 | Gawan Pada |
| M-18 | Goregaon |
| M-20 | Kandivali |
| M-21 | Kurla |
| M-22 | Malad |
| M-23 | Dadar |
| M-24 | Marol |
| M-25 | Memonwada |
| M-26 | Mulund |
| M-27 | Vikhroli |
| M-28 | Vile-Parle WL |
| M-29 | Vile-Parle W |
| M-30 | Wadala |
| M-31 | Worli |
| M-33 | Nana chawk (D Ward) |
| M-35 | C Ward |
| M-36 | F North |
| M-38 | F South |
| M-39 | Grant Road Eye |
| M-40 | H west Ward |
| M-41 | K East Ward |
| M-43 | KEM Hospital |
| M-46 | Malvani |
| M-49 | Nair Hospital |
| M-50 | Nariman Fire |
| M-51 | Octrroi Mankhurd |
| M-52 | Rawali Camp |
| M-53 | S Ward |
| M-54 | SWD Workshop dadar |
| M-55 | Workshop Kandivali |
| M-56 | Worli Data Center (wdc) |

Fig. 1. Names, unique codes, and locations of rain gauge stations throughout Mumbai (on SRTM elevation map)

acquired from the MCGM. These data were collected from 60 rain gauges for the southwest monsoon period (June-September) from 2006 to 2016. However, after accounting for missing data, newer installations, as well as the reliability, consistency, and operational period of each rain gauge, this study considered data from 47 of these gauges. Figure 1 shows the names, unique codes, and locations of these gauges on a Shuttle Radar Topography Mission (SRTM) elevation map.

## 3. Methodology

This section describes the methodology used to achieve the objectives of this study.

### 3.1. Screening of data

Continuous rainfall data from the rain gauges were analyzed, and a validation method was used to ensure the reliability of the data and identify suspect or incorrect values. The suspect or incorrect data were not modified but instead were flagged appropriately (e.g., as 'suspect' or 'missing'). A range test (Estévez et al. 2015) and double mass curve were applied subsequently. Only stations with more than three years of data were screened. This procedure resulted in the inclusion of 47 stations. A series of maximum rainfall amounts for durations of 0.25, 1, 3, 6, and 12 h and 1-5 d was acquired using a movable time window method.

Table 1. Threshold values used to identify heavy storms from sub-hourly data

| Duration [h] | 0.25 | 1 | 12 | 24 | 48 | 96 |
|---|---|---|---|---|---|---|
| Threshold [mm] | 10 | 25 | 52 | 64.5 | 90 | 118 |

Furthermore, because the threshold value affects the number of data points that are extracted (Pham, Lee 2015), the threshold values proposed by IMD for identifying heavy storms were used (Table 1). The discordancy test was used to identify and remove outliers; this test was also used to identify the appropriate datasets for regionalization. If a vector:

$$u_i = [t^{(i)}, t_3^{(i)}, t_4^{(i)}]^T \quad (1)$$

controls the L-moment ratio for the site $i$, then the discordancy measure may be characterized as:

$$D_i = \frac{1}{3}(u_i - \overline{u})^T S^{-1}(u_i - \overline{u}) \quad (2)$$

where $u_i$ is the vector of L-CV, L-Skewness, and L-Kurtosis, $S$ is the covariance matrix of $u_i$, and $\overline{u}$ is the mean vector of $u_i$.

### 3.2. Homogeneous rainfall region

The algorithm by Hosking and Wallis (1997) has been used to identify homogeneous rainfall regions in various countries, including Korea (Kar et al. 2017), Brazil (Carvalho et al. 2016), and Jakarta (Liu et al. 2015). The first step involves formation of candidate regions using cluster analysis of the site characteristics and testing the homogeneity of these proposed regions using at-site statistics (Castellarin et al. 2008; Malekinezhad, Zare-Garizi 2014). Site characteristics and site statistics are defined as follows (Gaál et al. 2009):

- Site characteristics are either quantities that are determined from the long-term climate of the site, (e.g., the mean yearly precipitation) or quantities that are known even before rainfall measurements are obtained (e.g., location, elevation, and other site physiographic properties).
- Site statistics are the measurements or any results of statistical processing of the rainfall data observed at the site.

#### 3.2.1. Site characteristics

The site characteristics (or variables) were prepared for each rain gauge station. The first three variables (latitude, longitude, and elevation) are geographical characteristics. The next 10 variables represent the long-term precipitation regime; these include the mean annual precipitation, mean monthly precipitation, maximum monthly precipitation, and mean annual number of wet days (i.e., with daily precipitation of 4 mm or more). The remaining seven variables describe the distance from the coast along different wind directions (0, 15, 30, 45, 60, 75, and 90° from the west, as shown in Figure 1). To reduce the number of variables, a principal component analysis (PCA) technique was applied with minimal loss of information. The PCA method identifies the most critical relationship structures among several variables. As a result, a few linear combinations of the original variables are used to describe the significant part of the overall variance. These optimized variables were used as inputs to the K-means cluster algorithm (MacQueen 1967), which uncovered the inherent structures in the data. The cluster algorithm starts by computing the centroid of each cluster and then calculates the distances between the current data vector and each of these centroids. The current vector is assigned to the cluster with the closest centroid. Because this is a dynamic method, vectors can change clusters after being assigned. This process is repeated until all the vectors have been assigned to a cluster, and the members of every cluster are closer to the centroid of their assigned cluster than to the mean of the other clusters (Genolini et al. 2016; Dullo et al. 2017). There is no specific technique for optimizing the number of clusters; therefore, the following coefficients were used:

- Connectivity (Handl et al. 2005) is the degree to which the observations are kept in the same group as their closest neighbors in the data space. The connectivity value ranges from zero to infinity and should be minimized.
- Dunn Index (Dunn 1973) consolidates the divergence among clusters and their measurements to gauge the most substantial number of clusters. A higher Dunn Index implies better grouping.
- Silhouette width (Rousseeuw 1987) measures the closeness of a data vector to the assigned cluster rather than a different one. The average dissimilarity between points demonstrates the structure of the data and consequently its possible groups (Rao, Srinivas 2008). Silhouette width values close to 1.0 indicate unusual groupings.

Rao and Srinivas (2008) noted that the Dunn Index is very sensitive to outliers; hence, in this study, the silhouette width was used first, followed by the Dunn Index, and finally the connectivity index. The optimal selection also maximized the quality of the clusters obtained.

#### 3.2.2. Site statistics

The L-moments method works as a linear function of probability weighted moments that are defined by the general form:

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k \binom{r-1}{k} E\{X_{r-k:r}\} \quad (3)$$

where $\lambda_r$ is the straight capacity of the $r^{th}$ L-moment of a distribution $X$, and $r$ is a non-negative integer with values 1-3. Using equation (3), the first four L-moments are composed as follows:

$$\lambda_1 = EX \quad (4)$$

$$\lambda_2 = \tfrac{1}{2}E(X_{2:2} - X_{1:2}) \quad (5)$$

$$\lambda_3 = \tfrac{1}{3}E(X_{3:3} - 2X_{2:3} + X_{1:3}) \quad (6)$$

$$\lambda_4 = \tfrac{1}{4}E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) \quad (7)$$

The L-moment ratios are defined as:

$$\tau_2 = \frac{\lambda_2}{\lambda_1},\ \tau_3 = \frac{\lambda_3}{\lambda_2},\ \tau_4 = \frac{\lambda_4}{\lambda_2} \quad (8)$$

where: $\tau_2$ is the measure of covariance (scale); $\tau_3$ is the measure of skewness (with range 0-1); $\tau_4$ is the measure of kurtosis (peakedness). These ratio estimators and their respective graphical charts are especially significant for identifying the distributional properties of skewed data.

3.2.3. Heterogeneity test

A homogeneous region is a set of sites whose probability distributions are approximately identical after rescaling their respective site variables. To project a homogeneous region, a statistical comparison of the site distributions of L-moment samples was performed. Hosking and Wallis (1993) recommended a statistical test of the heterogeneity ($H$) of a proposed homogeneous region. Accordingly, the heterogeneity was computed as:

$$H = \frac{V_{obs} - \mu_v}{\sigma_v} \quad (9)$$

Here, $\mu_v$ and $\sigma_v$ are the mean and standard deviation of the simulated data, respectively. $V_{obs}$ was obtained from the regional data, based on three V Statistics (V1, V2, V3) defined as follows:

$$V = \left\{\frac{\sum_{i=1}^{N} n_i(\tau_2^i - \tau_2^R)^2}{\sum_{i=1}^{N} n_i}\right\}^{\frac{1}{2}} \quad (10)$$

The region is sensibly homogeneous if $H < 1$, potentially homogeneous if $1 \leq H < 2$, and definitely heterogeneous if $H \geq 2$ (Hosking, Wallis 1993).

### 3.3. Goodness of fit of the regionalisation algorithm

To evaluate the performance of the L-moment approach, we compared it with an SOM regionalisation method (Parchure, Gedam 2018) using two regionalisation efficiency measures, namely, regionalisation efficiency (*RE*) and allocation efficiency (*AE*) (Núñez et al. 2016):

$$RE = \frac{HR}{TR} \times 100$$

$$AE = \frac{SAHR}{TS} \times 100$$

where: *HR* – number of homogeneous ($H1 < 2.0$) subregions ($n$); *TR* – total number of subregions (homogeneous + heterogeneous) ($n$); *SAHR* – stations allocated in homogeneous subregions ($n$); *TS* – total number of stations available in the regionalisation process ($n$).

Once the subregions have been delineated, the homogeneity analysis is performed on these subregions using Hosking and Wallis's (1997) $H1$ heterogeneity measure. The $H1$ and regionalisation efficiency measures were analysed using the rainfall depths of various durations.

Most of the statistical analyses and graphical illustrations of the results of this study were developed with the statistical program R-3.2.0, Orange Canvas 3.7.1 software, and MS Excel version 2007.

## 4. Results and discussion

This section presents the results of this study.

### 4.1. Mumbai as a single region

Mumbai was inspected as a single homogeneous region. The homogeneity test results, shown in Table 2, indicate that Mumbai could be regarded as a single homogeneous region for precipitation amounts of 1-4 d durations but not for those of less than 1 d duration. This distinction could be due to the occurrence of intense one-day duration storms. As such, these heavy storms are very unlikely to last for one or more consecutive days. This may be the explanation for the homogeneity of the other multiday precipitation amounts. Table 2 also shows another fascinating part of the heterogeneity investigation. The definitely heterogeneous ($H > 2.0$) behaviour occurs once again for the 5-day duration precipitation, perhaps due to changes in climatic parameters.

Table 2. Summary of homogeneity tests for the Mumbai region as a whole

| Rainfall time period [h] | *H* Test | | |
|---|---|---|---|
| | *H*1 | *H*2 | *H*3 |
| 0.25 | 12.7 | 12.88 | 9.77 |
| 1 | 12.03 | 13.35 | 11.71 |
| 6 | 5.33 | 0.24 | −2.58 |
| 12 | 3.39 | −0.82 | −2.64 |
| 24 | 1.47 | −1.51 | −2.93 |
| 48 | −0.47 | −2.32 | −2.52 |
| 72 | −1.43 | −2.99 | −3.33 |
| 96 | −1.5 | −2.78 | −2.77 |
| 120 | 12.88 | 14.77 | 12.31 |

### 4.2. Homogeneous rainfall regions

4.2.1. Site characteristics

The PCA method was used to optimize the input variables; this resulted in nine components that captured about 95% of the variance in the total input variables (Fig. 2).

The input variables for the cluster analysis determined by the PCA results are as follows:

- Longitude and Latitude: these variables describe the location of the gauges.
- Five wind direction angles: these angles are in the range 195°-255°, indicative of the southwesterly winds from the Arabian Sea that blow over the study area.
- Two precipitation variables: these are the mean annual precipitation and mean July precipitation (July is the wettest month of the year).

The optimum number of clusters was derived using the K-means cluster technique, the nine principal components, and the three internal indices described in section 3.2.1. Table 3 shows the values obtained

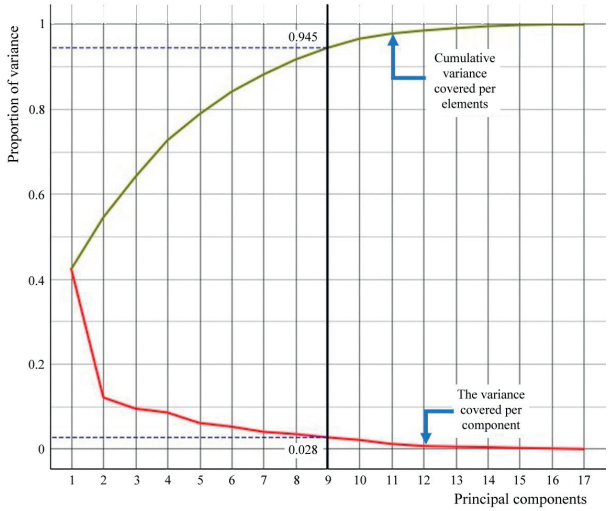



Fig. 2. Nine principal components captured about 95% of the variance in the total variables

Table 3. Results obtained from 2 to 8 cluster stages with K-means clustering method

| Internal measures | Connectivity | Dunn | Silhouette |
|---|---|---|---|
| 2 | 21.72 | 0.17 | 0.24 |
| 3 | 23.44 | 0.2 | 0.25 |
| 4 | 23.75 | 0.22 | 0.26 |
| 5 | 40.88 | 0.2 | 0.17 |
| 6 | 41.15 | 0.12 | 0.22 |
| 7 | 41.48 | 0.14 | 0.23 |
| 8 | 45.67 | 0.15 | 0.22 |

from the cluster stages 2-8 with the K-means cluster technique. Four candidate precipitation regions were identified. The homogeneity of each region was inspected for all of the selected durations. Figure 4 (left) shows the location and composition of each region (cluster). A compact overview of the most relevant variables of the regions is shown in Table 5. Another intriguing insight from Table 4 is that the dominant part of the region did not pass the homogeneity test. Figure 4 also shows that a few sites are scattered in the geographical space. These results were the bases for our conclusion that a combination of the site characteristics and site statistics must be used for homogeneous regional analysis. The specific outcomes of this analysis are as follows:

- Region SC#1 (region #1 for site characteristics) consists of 16 stations located in the area near the Arabian Sea, mainly in the southwestern part of Mumbai. Table 4 indicates that this region is heterogeneous for precipitation durations up to 6 h, which might be due to the effect of air masses from the territory of the Arabian Sea. Also, the detailed analysis of the data highlighted the extraordinary amount of rainfall measured at Worli (57.4-89.14 mm) and Wadala (55.12 mm) for 15 min events.
- Region SC#2 is the largest among the regions studied based on site characteristics. This region, which comprises 18 sites covering the southwestern part and extending further inland, represents low land, urban areas with high-rise buildings, and parts of the windward side of the Sanjay Gandhi National Park. The rain gauge cluster for this region consists of a few sites having minimum and maximum mean annual precipitation and mean July precipitation (Table 5). The region was identified as homogeneous according to the *H* test except for 15-min precipitation amounts.
- Region SC#3 is located in the eastern part of Mumbai and consists of five sites covering the leeward side of the Sanjay Gandhi National Park. The sites are fairly scattered at various altitudes; however, the mean

elevation is the highest among the four regions (Table 5). The cluster analysis isolated these sites principally because they have the highest station altitudes, which resulted in the heterogeneity of the region for rainfall durations up to 3 d (Table 4). The detailed analysis of the data also indicates significant variation in rainfall amount, which may be due to the funnelling action by the Chembur and Sanjay Gandhi National Park hills.

- Region SC#4 consists of 8 sites located predominantly in the western part of the city between the Arabian Sea and windward side of the Sanjay Gandhi National Park. This region was identified as homogeneous according to the H test (Table 4).

The L-moments graphs shown in Figure 3 indicate that three distinct groups of sites comprise the region SC#2 (top graph). These sites are located in two different geographical regions. The first group is made up of 4 sites from the southeastern parts (the lowlands) of Region SS#5: Kurla, Rawali camp, Deonar, and Octroi Mankhurd. The second group consists of 4 stations located near the foot of the Sanjay Gandhi National Park hills on the windward side (Region SS#6). The third group includes 10 sites from the lowland and urban parts of Region SS#2. The L-moments diagram (Fig. 3) of region SC#3 suggested moving the Vikhroli station to Region SS#5.

Therefore, based on the L-moments diagram, a further subdivision of Regions SC#2 and SC#3 was performed using site statistics. This analysis excluded the unusually wet sites (Worli and Wadala) from Region SC#1. The results indicate that the region becomes homogeneous for the one-hour precipitation amount, while the *H* value decreases from 3.87 to 3.18 for the 15-min rainfall amount.

4.2.2. Site statistics

The final precipitation regionalization was created using the site statistics and the results of the site characteristics analysis as inputs. The primary intent was to reflect the characteristics of the individual geomorphological units and consider an objective measure of similarity within the regions. Six regions were obtained from combining the best features of the two pooling strategies. The composition and geographical locations of these regions are shown in Figure 4 (right), and a synopsis of their physical and climatological characteristics is given in Table 5. In addition, Table 4 shows the heterogeneity measures for the selected rainfall durations. The specific outcomes of this analysis are as follows:

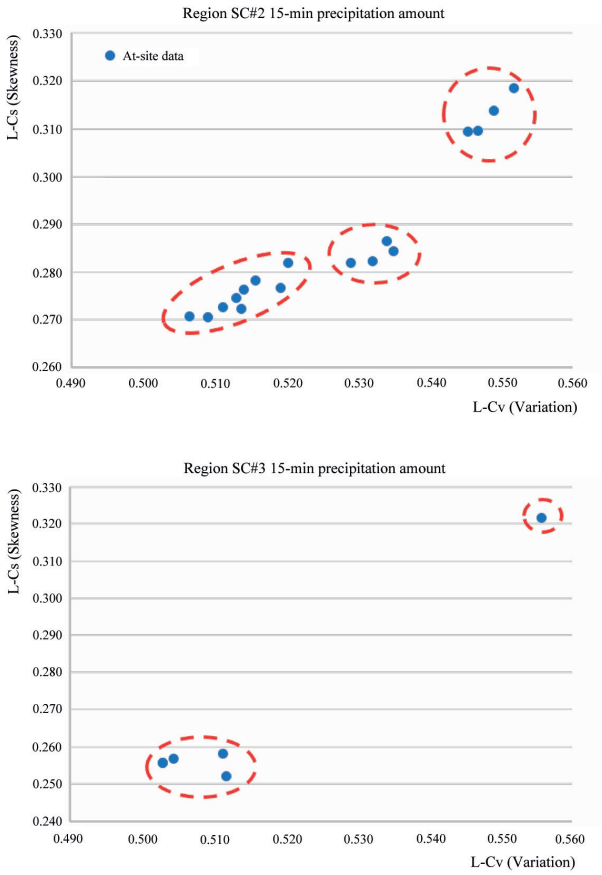



Fig. 3. L-moments graph for Regions SC#2 and SC#3 for 15 min precipitation duration

- Region SS#1 consists of 14 stations in the area near the Arabian Sea, mainly covering the southwestern area of Mumbai. This region is identical to Region SC#1, except for the two stations (Worli and Wadala) that recorded extraordinary rainfall events. The *H* test results for rainfall durations of 1 h and above suggested that this region is homogeneous.
- Region SS#2 comprises 10 sites, the majority of which are also in Region SC#2; the stations that are located near the mountain area and leeward side of Chembur Hill are excluded from Region SC#2. This region was identified as homogeneous except for the 15-min precipitation.
- Region SS#3 comprises four sites and stretches through the leeward side of the Sanjay Gandhi National Park hills to the east of Mumbai. This region was identified as homogeneous for most of the durations probably due to the shadowing effect of the hills, which reduces the rainfall.
- Region SS#4 is similar to Region SC#4 because of its high level of homogeneity.
- Region SS#5 consists of five sites located between the Sanjay Gandhi National Park and Chembur hills. These stations are grouped together due to the funnelling action of the hills. The *H* test results indicated

Table 4. Summary of homogeneity ($H$) test values of sites identified based on site statistics and site characteristics (in parentheses) for each region and rainfall duration

| Rainfall duration [h] | $H$ Test | Region SS#1 (SC#1) | Region SS#2 (SC#2) | Region SS#3 (SC #3) | Region SS#4 (SC#4) | Region SS#5 | Region SS#6 |
|---|---|---|---|---|---|---|---|
| 0.25 | $H1$ | 3.65 (3.87) | 2.64 (5.59) | 5.1 (6.27) | −0.72 (−0.72) | 2.08 | 5.65 |
| 1 | $H1$ | 0.94 (3.88) | −0.01 (0.8) | 1.42 (1.19) | −0.26 (−0.26) | 1.22 | 1.51 |
| 6 | $H1$ | 0.25 (0.15) | −1.19 (−1.71) | 0.79 (1.24) | −1.16 (−1.16) | −0.53 | −0.85 |
| 12 | $H1$ | −1.06 (−0.89) | −1.54 (−1.98) | 1.29 (1.81) | −0.35 (−0.35) | −1.26 | −0.68 |
| 24 | $H1$ | −1.95 (−1.76) | −1.45 (−2.28) | 0.9 (1.86) | 0.03 (0.03) | −1.6 | −0.8 |
| 48 | $H1$ | −3.12 (−2.97) | −1.52 (−2.04) | 1.1 (1.42) | −0.63 (−0.63) | −1.17 | −0.57 |
| 72 | $H1$ | −2.58 (−2.64) | −0.98 (−2) | 0.26 (0.84) | −0.66 (−0.66) | −0.85 | −1 |
| 96 | $H1$ | −2.8 (−2.62) | −1.63 (−2.04) | 0.93 (0.93) | −0.33 (−0.33) | −0.66 | −0.5 |
| 120 | $H1$ | −1.03 (−1.12) | −0.63 (−0.63) | −0.69 (1.25) | 0.48 (0.48) | −0.13 | 0.37 |

Table 5. Summary of relevant characteristics of sites identified based on site statistics and site characteristics (in parentheses) for each region

| Relevant characteristics | Region SS#1 (SC#1) | Region SS#2 (SC#2) | Region SS#3 (SC#3) | Region SS#4 (SC#4) | Region SS#5 | Region SS#6 |
|---|---|---|---|---|---|---|
| Number of stations | 14* (16) | 10 (18) | 4 (5) | 8 (8) | 5 | 4 |
| Range of $H$ [m] | 11.66-1.28 (11.66-31.28) | 12.31-37.2 (8.84-50.09) | 10.71-85.68 (10.71-85.68) | 13.86-24.99 (13.86-24.99) | 8.84-29.61 | 28.26-50.09 |
| Average of $H$ [m] | 21.77 (21.88) | 22.52 (23.96) | 40.71 (39.12) | 20.63 (21.6) | 16.93 | 37.3 |
| Range of Mean Annual Precipitation [mm] | 1791.68-2228.02 (1791.68-2429.21) | 1601.8-2393.33 (1601.8-2598.32) | 2137.4-2594.28 (1940.39-2594.28) | 1793.35-2616.77 (1793.35-2616.77) | 2218.32-2598.32 | 1905.05-2227.82 |
| Median of Mean Annual Precipitation [mm] | 1997.49 (2039.57) | 2181.02 (2219.2) | 2435.06 (2099.13) | 2132.66 (2227.76) | 2429.21 | 2102.46 |
| Range of Mean July Precipitation [mm] | 651.7-930.89 (651.7-947.86) | 637.37-954.51 (637.37-1139.48) | 824.11-1139.48 (675.73-1090.6) | 675.73-1090.6 (765.48-896.75) | 746.51-1034.15 | 725.77-922.36 |
| Median of Mean July Precipitation [mm] | 778.83 (797.7) | 845.58 (866.42) | 949.15 (853.88) | 803.16 (804.32) | 917.98 | 853.73 |

that this region is homogeneous for rainfall durations of 1 h and above except for the 15-min rainfall duration.

- Region SS#6 includes four sites on the windward side of Sanjay Gandhi National Park and covers the area west of Mumbai near the foot of the hills.

These results confirmed the effect of the complex topography of the study area on the spatial variation of rainfall. The topography includes the flatland near the Arabian Sea, the urban areas with high-rise buildings, and the mountains and hilly areas (particularly the Sanjay Gandhi National Park). In addition, the intense precipitation recorded over Region SS#2, which is located in the urban pocket, supports the findings of Paul et al. (2018) that associates extreme precipitation with urbanization.

4.2.3. Heterogeneity test

The six regions that were identified by the site statistics were also identified as potentially homogeneous regions according to the *H* test for rainfall durations of 1 h and above (Table 4). However, the spatial variation in the 15-min rainfall data resulted in the heterogeneity of the regions. The rapid reduction in spatial variation as the rainfall duration increases explains why the regions were identified as homogeneous for rainfall durations of 1 h and above. Because the *H* test is a somewhat subjective process, unusual sites may be removed and the sites regrouped to improve the homogeneity of the identified regions. However, further removal and regrouping of the sites was not performed for the following reasons:

- To avoid loss of valuable information on precipitation in the analysis of dissimilar sites.
- To avoid the unnecessary burden of evaluating extraordinary events, especially where removal of such sites produces minimal change in the *H* Value.

### 4.3. Goodness of fit of the regionalisation algorithm

Figure 5a shows the box-whisker plot of the *H*1 heterogeneity values of the two regionalisation methods (L-moment and SOM) for the rainfall with 0.25 h duration. The SOM method showed the lowest H1 dispersion; all of its values were below the critical value (*Hc* = 2). These results highlighted that the direct application of cluster analysis by itself does not guarantee an automatic delimitation of homogeneous regions, which supports the findings of Hosking and Wallis (1997), Ilorme and Griffis (2013) and Wazneh et al. (2015).

Similarly, Figure 5b shows the efficiency values (*RE* and *AE*). SOM showed higher *RE* (*AE*) efficiency (100%) as compared to the L-moment approach (20%). This method was based on the predefined region sizes and a function of MAP, which depicts consistency in the strong relationships between precipitation variability, L-moment-ratios, and MAP (Wallis et al. 2007; Núñez et al. 2011).

These results support the findings of Clarke (2010) and Núñez et al. (2016), highlighting the limitations of the L-moment approach: (a) Meteorological networks cannot fully represent the spatial continuum of the large-scale variables and attributes; (b) Cluster analysis has been used in general with the expectation of identifying homogeneous regions, but in physical terms, it remains to be established why such regions should be considered homogeneous. Hence, a subjective judgement in the regionalisation process is warranted, owing to a possibility of dissociation between the expression of Mother Nature (to paraphrase J.R. Wallis) and how the monitoring station networks and the associated statistical analysis procedures capture this expression (Hosking, Wallis 1997; Wazneh et al. 2015).

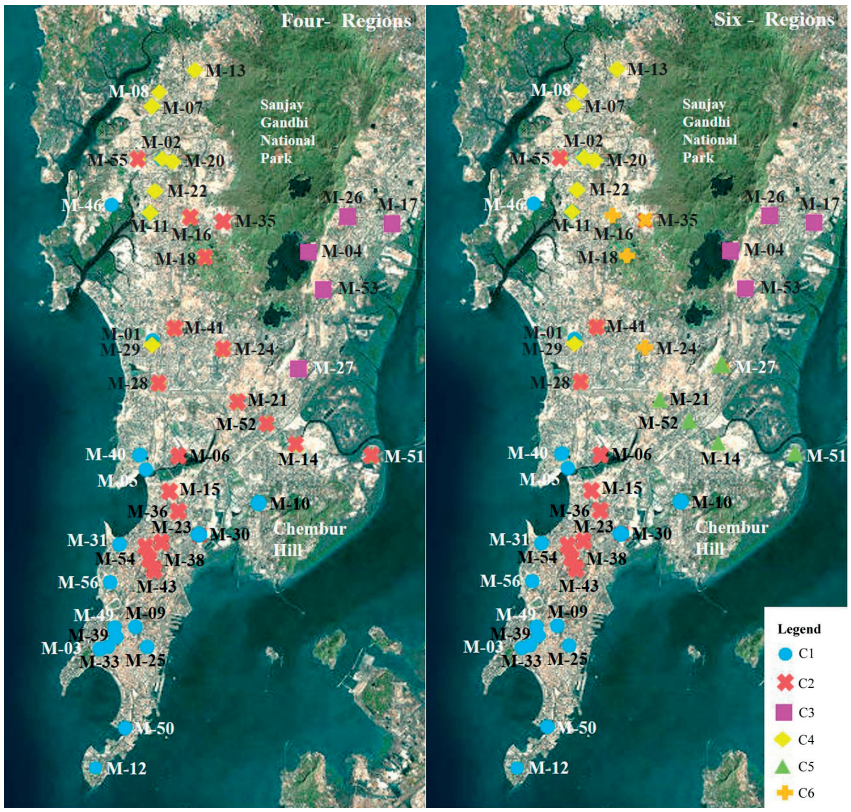



Fig. 4. Composition and location of regions using site characteristics (left) and refinement using site statistics (right)

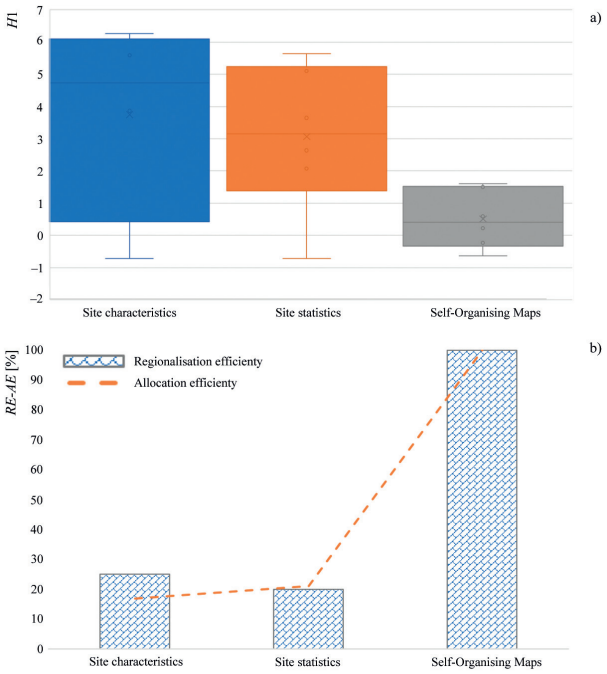



Fig. 5. Representations of efficiency evaluation: (a) $H1$ boxplots for various regionalisation methods; (b) regionalisation and allocation efficiency bar plots for various regionalisation methods

## 5. Conclusions

The main objective of this research was to identify the homogeneous rainfall regions of Mumbai based on a combined analysis of site characteristics and site statistics. The study also focused on the climatological conditions of the city using rainfall amounts of various durations (0.25, 1, 6, 12, 24, 48, 72, 96, and 120 h). The main results and conclusions are as follows:

1. Mumbai can be considered as a single homogenous region for precipitation events of 1-4 d durations but not for those of less than one day duration. This could be due to the occurrence of intense storms of one day or shorter duration.
2. To obtain the homogeneous regions, it is advisable to evaluate the variance in the homogeneous sub-regions obtained via L-moment regarding changes in the temporal scale of the response variable.
3. The combination of two techniques, namely cluster analysis (objective) and site statistical pooling (subjective) for identification of homogeneous region support the findings of Clarke (2010) and Núñez et al. (2016) and highlighted the limitations of these approaches.
4. The regions based on the site statistics approach were identified as potentially homogeneous according to the $H$ Test for rainfall durations of 1 h and above. However, spatial variation in the 15-min rainfall data resulted in the heterogeneity of the regions. The rapid reduction in spatial variation as the rainfall duration increases explains why the regions were identified as homogeneous for rainfall durations of 1 h and above. Because the $H$ Test is a somewhat subjective process, unusual sites may be removed, and the sites regrouped to improve the homogeneity of the identified regions. However, further removal and regrouping of the sites was not performed. These results support the SOM approach adopted by Parchure and Gedam (2018) to cluster rain gauges in groups (regions).
5. The six clusters of rainfall gauges obtained confirm that the spatial variation of rainfall is a result of the complex topography of Mumbai, which includes the flatland near the Arabian Sea, the urban areas with high-rise buildings, and the mountainous and hilly areas (particularly the Sanjay Gandhi National Park located in the northern part).

The limitations of the study are the following:

1. This study only identified homogeneous rainfall regions. A potential topic of future research would be the computation of appropriate probability distribution functions and design rainfall quantities along with their return periods.
2. Further research is needed to test the hypothesis of clusters using other climatological data (such as temperature, wind speed, and wind direction) that are measured by MCGM at the 47 stations and radar data obtained by the IMD.
3. The discussions of this study have alluded to possible influences from the complex orography and coastal proximity of Mumbai. The significance of these influences needs to be confirmed through future studies of multiple cases across different cities.

**Acknowledgments**

The authors thank the MCGM for providing rain gauge data and the Indian Metrological Department, Mumbai, for their comments and suggestions, which significantly contributed to improving the clarity of the paper.

Bibliography

Ahuja S., Dhanya C.T., 2012, Regionalization of rainfall using RCDA cluster ensemble algorithm in India, Journal of Software Engineering and Applications, 5 (8), 568-573, DOI: 10.4236/jsea.2012.58065

Bharath R., Srinivas V.V., 2015, Regionalization of extreme rainfall in India, International Journal of Climatology, 35 (6), 1142-1156, DOI: 10.1002/joc.4044

Carvalho J.R.P. de, Nakai A.M., Monteiro J.E.B.A., 2016, Spatio-temporal modeling of data imputation for daily rainfall series

in homogeneous zones, Revista Brasileira de Meteorologia, 31 (2), 196-201, DOI: 10.1590/0102-778631220150025

Castellarin A., Burn D.H., Brath A., 2008, Homogeneity testing: how homogeneous do heterogeneous cross-correlated regions seem?, Journal of Hydrology, 360 (1-4), 67-76, DOI: 10.1016/j.jhydrol.2008.07.014

Clarke R., 2010, On the (mis)use of statistical methods in hydro-climatological research, Hydrological Sciences Journal, 55 (2), 139-144, DOI: 10.1080/02626661003616819

Dullo T.T., Kalyanapu A.J., Teegavarapu S.V., 2017, Evaluation of changing characteristics of temporal rainfall distribution within 24-hour duration storms and their influences on peak discharges: case study of Asheville, North Carolina, Journal of Hydrologic Engineering, 22 (11), DOI: 10.1061/(ASCE) HE.1943-5584.0001575

Dunn J.C., 1973, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 3 (3), 32-57, DOI: 10.1080/01969 727308546046

Estévez J., Gavilán P., García-Marín A.P., Zardi D., 2015, Detection of spurious precipitation signals from automatic weather stations in irrigated areas, International Journal of Climatology, 35 (7), 1556-1568, DOI: 10.1002/joc.4076

Gaál L., Kyselý J., Szolgay J., 2008, Region-of-influence approach to a frequency analysis of heavy precipitation in Slovakia, Hydrology and Earth System Sciences, 12 (3), 825-839, DOI: 10.5194/hess-12-825-2008

Gaál L., Szolgay J., Lapin M., Faško P., 2009, Hybrid approach to the delineation of homogeneous regions for regional precipitation frequency analysis, Journal of Hydrology and Hydromechanics, 57 (4), 226-249, DOI: 10.2478/v10098-009-0021-1

Genolini C., Ecochard R., Benghezal M., Driss T., Andrieu S., Subtil F., 2016, kmlShape: An efficient method to cluster Longitudinal data (time-series) according to their shapes, PLOS One, 11 (6), DOI: 10.1371/journal.pone.0150738

Hailegeorgis T.T., Alfredsen K., 2017, Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for mid-Norway, Journal of Hydrology: Regional Studies, 9, 104-126, DOI: 10.1016/j.ejrh.2016.11.004

Halbert K., Nguyen C.C., Payrastre O., Gaume E., 2016, Reducing uncertainty in flood frequency analyses: a comparison of local and regional approaches involving information on extreme historical floods, Journal of Hydrology, 541 (Part A), 90-98, DOI: 10.1016/j.jhydrol.2016.01.017

Hallegatte S., 2010, Flood risks, climate change impacts and adaptation benefits in Mumbai. An initial assessment of socioeconomic consequences of present and climate change induced flood risks and of possible adaptation options, OECD Environment Working, Papers, 27, DOI: 10.1787/5km4hv6wb434-en

Handl J., Knowles J., Kell D.B., 2005, Computational cluster validation in post-genomic data analysis, Bioinformatics, 21 (15), 3201-3012, DOI: 10.1093/bioinformatics/bti517

Hessburg P.F., Kuhlman E.E., Swetnam T.W., 2005, Examining the recent climate through the lens of ecology: inferences from temporal pattern analysis, Ecological Applications, 15 (2), 440-457

Hosking J.R.M., Wallis J.R., 1993, Some statistic useful in region frequency analysis, Water Resources Research, 29 (2), 271-281, DOI: 10.1029/92WR01980

Hosking J.R.M., Wallis J.R., 1997, Regional frequency analysis: an approach based on L-moments, Cambridge University Press, Cambridge, UK, 224 pp.

Ilorme F., Griffis V.W., 2013, A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungauged sites for design flood estimation, Journal of Hydrology, 492, 151-162, DOI: 10.1016/j.jhydrol.2013.03.045

Kakade S.B., Kulkarni A., 2017, Seasonal prediction of summer monsoon rainfall over cluster regions of India, Journal of Earth System Science, 126 (34), DOI: 10.1007/s12040-017-0811-5

Kar K.K., Yang S.-K., Lee J.-H., Khadim F.K., 2017, Regional frequency analysis for consecutive hour rainfall using L-moments approach in Jeju Island, Korea, Geoenvironmental Disasters, 4 (18), DOI: 10.1186/s40677-017-0082-0

Khan S.A., Hussain I., Hussain T., Faisal M., Muhammad Y.S., Shoukry A.M., 2017, Regional frequency analysis of extremes precipitation using L-moments and partial L-Moments, Advances in Meteorology, DOI: 10.1155/2017/6954902

Liu J., Doan C.D., Liong S.-Y., Sanders R., Dao A.T., Fewtrell T., 2015, Regional frequency analysis of extreme rainfall events in Jakarta, Natural Hazards, 75 (2), 1075-1104, DOI: 10.1007/s11069-014-1363-5

Lokanadham B., Gupta K., Nikam V., 2012, Characterization of spatial and temporal distribution of monsoon rainfall over Mumbai, ISH Journal of Hydraulic Engineering, 15 (2), 69-80, DOI: 10.1080/09715010.2009.10514941

MacQueen J.B., 1967, Some methods for classification and analysis of multivariate observations, [in:] Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume 1: Statistics, L.M. Le Cam, J. Neyman (eds.), University of California Press, Berkeley, 281-297

Malekinezhad H., Zare-Garizi A., 2014, Regional frequency analysis of daily rainfall extremes using L-moments approach, Atmósfera, 27 (4), 411-427, DOI: 10.1016/S0187-6236(14)70039-6

Nair A., Mohanty U.C., Acharya N., 2013, Monthly prediction of rainfall over India and its homogeneous zones during monsoon season: a supervised principal component regression approach on general circulation model products, Theoretical and Applied Climatology, 111 (1-2), 327-339, DOI: 10.1007/s00704-012-0660-8

Ngongondo C.S., Xu C.Y., Tallaksen L.M., Alemaw B., Chirwa T., 2011, Regional frequency analysis of rainfall extremes in Southern Malawi using the index rainfall and L-moments approaches, Stochastic Environmental Research and Risk Assessment, 25 (7), 939-955, DOI: 10.1007/s00477-011-0480-x

Núñez J.H., Hallack- Alegría M., Cadena M., 2016, Resolving regional frequency analysis of precipitation at large and complex scales using a bottom-up approach: The Latin America and the Caribbean Drought Atlas, Journal of Hydrology, 538, 515-538, DOI: 10.1016/j.jhydrol.2016.04.025

Núñez J.H., Verbist K., Wallis J.R., Schaefer M.G., Morales L., Cornelis W.M., 2011, Regional frequency analysis for mapping drought events in north-central Chile, Journal of Hydrology, 405 (3-4), 352-366, DOI: 10.1016/j.jhydrol.2011.05.035

Parchure A.S., Gedam S.K., 2018, Precipitation regionalization using Self-Organizing Maps for Mumbai City, India, Journal of Water Resource and Protection, 10 (9), 939-956, DOI: 10.4236/jwarp.2018.109055

Paul S., Ghosh S., Mathew M., Devanand A., Karmakar S., Niyogi D., 2018, Increased spatial variability and intensification of extreme monsoon rainfall due to urbanization, Scientific Reports, 8 (3918), DOI: 10.1038/s41598-018-22322-9

Pham V.H., Lee B.R., 2015, An image segmentation approach for fruit defect detection using k-means clustering and graph-based algorithm, Vietnam Journal of Computer Science, 2 (1), 25-33, DOI: 10.1007/s40595-014-0028-3

Rahman M., Sarkar S., Najafi M.R., Rai R.K., 2013, Regional extreme rainfall mapping for Bangladesh using L-moment technique, Journal of Hydrologic Engineering, 18 (5), DOI: 10.1061/(ASCE)HE.1943-5584.0000663

Rao A.R., Srinivas V.V., 2008, Regionalization of watersheds. An approach based on cluster analysis, Water Science and Technology Library Series, 58, Springer Netherlands, 245 pp.

Requena A.I., Chebana F., Mediero L., 2016, A complete procedure for multivariate index-flood model application, Journal of Hydrology, 535, 559-580, DOI: 10.1016/j.jhydrol.2016.02.004

Rousseeuw P.J., 1987, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics, 20, 53-65, DOI: 10.1016/0377-0427(87)90125-7

Saha M., Mitra P., Nanjundiah R.S., 2017, Deep learning for predicting the monsoon over the homogeneous region of India, Journal of Earth System Science, 126 (54), DOI: 10.1007/s12040-017-0838-7

Santos E.B., Lucio P.S., Silva C.M.S., 2015, Precipitation regionalization of the Brazilian Amazon, Atmospheric Science Letters, 16 (3), 185-192, DOI: 10.1002/asl2.535

Sen S., Vittal H., Singh T., Singh J., Karmakar S., 2013, At-site design rainfall estimation with a diagnostic check for nonstationary: an application to Mumbai rainfall datasets, [in:] Proceedings of HYDRO 2013 INTERNATIONAL, 4-6 December 2013, Madras, India, 14 pp.

Sherly M.A., Karmakar S., Chan T., Rau C., 2015, Design rainfall framework using multivariate parametric-nonparametric Approach, Journal of Hydrologic Engineering, 21 (1), DOI: 10.1061/(ASCE)HE.1943-5584.0001256

Singh J., Sekharan S., Karmarkar S., Ghosh S., Zope P.E., Eldho T.I., 2017, Spatio-temporal analysis of sub-hourly rainfall over Mumbai, India: is statistical forecasting futile?, Journal of Earth System Science, 126 (38), DOI: 10.1007/s12040-017-0817-z

Sinha P., Mohanty U.C., Kar S.C., Dash S.K., Robertson A.W., Tippett M.K., 2013, Seasonal prediction of the Indian summer monsoon rainfall using canonical correlation analysis of the NCMRWF global model products, International Journal of Climatology, 33 (7), 1601-1614, DOI: 10.1002/joc.3536

Smithers J.C., Schulze R.E., 2001, A methodology for the estimation of short duration design storms in South Africa using a regional approach based on L-moments, Journal of Hydrology, 241 (1-2), 42-52, DOI: 10.1016/S0022-1694(00)00374-7

Sun X., Lall U., Merz B., Nguyen V.D., 2015, Hierarchical Bayesian clustering for nonstationary flood frequency analysis: application to trends of annual maximum flow in Germany, Water Resources Research, 51 (8), 6586-6601, DOI: 10.1002/2015WR017117

Wallis J.R., Schaefer M.G., Barker B.L., Taylor G.H., 2007, Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington States, Hydrology and Earth System Sciences, 11 (1), 415-442, DOI: 10.5194/hess-11-415-2007

Wazneh H., Chebana F., Ouarda T.B.M.J., 2015, Delineation of homogeneous region for regional frequency analysis using statistical depth function, Journal of Hydrology, 521, 232-244, DOI: 10.1016/j.hydrol.2014.11.068