Regional Taiwan rainfall frequency analysis using principal component analysis, self-organizing maps and *L*-moments

Lu-Hsien Chen and Yu-Ting Hong

ABSTRACT

The objective of this paper is to propose an approach, which consists of principal component analysis (PCA), self-organizing maps (SOM) and the *L*-moment method, for improving estimation of desired rainfall quantiles of ungauged sites. Firstly, PCA is applied to obtain the principal components. Then SOM is applied to group the rain gauges into specific clusters and the number of clusters can be objectively decided by visual inspection. Moreover, the *L*-moment based discordancy and heterogeneity are used to test whether clusters may be acceptable as being homogeneous. After the gauges are grouped into specific clusters, the homogeneous regions are then delineated. Finally, goodness-of-fit measure is used to select the regional probability distributions and the design rainfall quantiles with various return periods for each region can be estimated. The proposed approach is applied to analyze and quantify regional rainfalls in Taiwan. The proposed approach is a robust and efficient way for regional rainfall frequency analysis. Moreover, one can easily assign an ungauged site to a previously defined cluster according to a map of homogeneous regions. Therefore, the proposed approach is expected to be useful for providing the design rainfall quantiles with various return periods at ungauged sites.

Key words | *L*-moments, principal component analysis, regional rainfall frequency analysis, self-organizing map

Lu-Hsien Chen (corresponding author)
Yu-Ting Hong
Department of Leisure Management,
Taiwan Shoufu University,
Madou,
Tainan 72153,
Taiwan
E-mail: *lhchen@tsu.edu.tw*

INTRODUCTION

Rainfall frequency analysis, which involves the estimation of distributional parameters and the extrapolation of cumulative distribution functions to estimate extreme rainfall values, is very important in the areas of hydrology. Accurate estimation of rainfall frequency is needed in many hydraulic designs such as dams, culverts and urban drainage systems. For locations without observed data, regional rainfall frequency analysis is often needed to estimate *T*-year event magnitude. It attempts to respond to the need for rainfall estimation in ungauged sites and for improving the at-site estimate by using the available rainfall data within a region.

The aforesaid rainfall quantiles can then be estimated using an appropriate statistical model and an appropriate parameter estimation technique, provided that the data do not exhibit any persistence. Different approaches are presented which use conventional moments to extract order statistics such as mean, standard deviation, skewness and kurtosis. Due to problems arising from data quality, such as short record and outliers, conventional moments are problematic. Hosking & Wallis (1997) developed *L*-moments which are linear combinations of order statistics. The main advantage of *L*-moments over conventional moments is that they suffer less from the effects of sampling variability. They are more robust to outliers and virtually unbiased for small samples (Hosking & Wallis 1997). More recently, hydrologic researchers have focused on the *L*-moment approach and it is increasingly used in regional frequency analysis (Fowler & Kilsby 2003; Rao & Srinivas 2006a, b; Parida & Moalafhi 2008; Meshgi & Khalili 2009).

For a regional analysis to be successful, the identification of homogeneous regions first using the historical data is essential. The basic idea behind regional rainfall frequency

doi: 10.2166/nh.2012.032

analysis is to make use of similarities in the characteristics of rainfalls at different sites in a region. Consequently, regional homogeneity is an important requirement and a critical issue in such an analysis. Cluster analysis is traditionally employed to perform identification of homogeneous regions for regional rainfall frequency analysis. The clustering methods consist of hierarchical methods, like Ward's method, and the non-hierarchical methods, such as the *K*-means method (Mingoti & Lima 2006). However, each clustering method carries its own shortcomings. Especially, determination of an optimal number of clusters is a difficult task. Moreover, different choices of clustering methods often lead to different clustering results even though the same data sets are analyzed (Nathan & McMahon 1990) and so conventional cluster analysis may not be the best technique for regionalization in frequency analysis. These problems must be solved and self-organizing maps (SOM) can present an alternative to solve the problems.

The SOM can project high-dimensional input space on a low-dimensional topology so as to allow one to compute the number of clusters directly by sight. Having a capability to preserve the topological structure of data is the main advantage of the SOM algorithm (ASCE 2000a). This capability helps one to discover the relationships among complex data and to group data into clusters. Mangiameli *et al.* (1996) and Michaelides *et al.* (2001) showed the better performance of SOM in hydrologic clustering than that of conventional clustering techniques. Detailed reviews of SOM along with assessments of their application in water resources and hydrology can be found in ASCE (2000b) and Kalteh *et al.* (2008). In Taiwan, SOM has also been applied in hydrology, including flood forecasting (Chang *et al.* 2007; Yang & Chen 2009), groundwater modeling (Lin & Chen 2005), precipitation data (Hsu & Li 2010), typhoon-rainfall forecasting (Lin & Wu 2009) and design hyetographs (Lin & Wu 2007; Lin *et al.* 2010). Recently, SOM has been applied in delineation of homogeneous regions for regional flood frequency analysis (Hall & Minns 1999; Hall *et al.* 2002; Zhang & Hall 2004). For regional rainfall frequency analysis, Lin & Chen (2006) used the SOM, *K*-means method and Ward's method to identify the rainfall-homogeneous regions based on site characteristics, which include the site's geographic location, indicators of rainfall amount, and indicators of the distribution of the amounts through a year. They show that the SOM can identify the homogeneous regions more accurately as compared to the other two clustering methods.

In the past analyses, there is no theoretical principle for determining the optimum size of the output layer, and hence the output layer is kept large to ensure that the maximum number of clusters is formed from the training data. About the number of iterations, as a general rule, it must be at least 500 times the number of output neurons in the network (Haykin 1994). If all of the independent variables are used directly as input variables to the SOM, the method will need more time to obtain the feature map. In order to save computation time, transformed data resulting from principle component analysis (PCA) were used as input variables to the SOM in this paper. PCA is a linear transformation technique that provides a smaller set of uncorrelated variables (called components) from a set of correlated variables while maintaining most of the information in the original data set. PCA is often used as a preprocessing step to clustering (Everitt 1993), and it is in an attempt to reduce the number of variables. This factor is important because it helps to reduce future data collection costs. Usually, most of the variation in a large group of variables can be captured with only a few principal components.

The purpose of this paper is to carry out the regional rainfall frequency analysis for annual maximum daily rainfall in Taiwan using PCA, SOM and *L*-moment method. The paper is organized as follows. First, the theories of PCA, SOM and *L*-moment method are presented, respectively. Then its application on rainfall is addressed and finally as a case study, regional rainfall frequency by combining PCA, SOM and *L*-moments in Taiwan is mentioned. The PCA and SOM are applied to identify the homogeneous regions, and the *L*-moments are used for parameter estimation, homogeneity testing and selection of the regional distribution. Finally, the desired rainfall quantile estimates with different return periods for each homogeneous region are estimated and the characteristics of design rainfalls in each cluster are also discussed.

METHODOLOGY

Principal component analysis

Principal component analysis (PCA) is a well known linear optimization method that maximizes the explained variance

of a dataset by an orthogonal set of eigenvectors, and it has numerous applications in various science and engineering problems. PCA mathematically transforms a dataset into a reduced set of uncorrelated (i.e. orthogonal) variables which represent as much as possible of the information contained within the original data. The principal components (PCs), which are obtained through eigenanalysis of the correlation or covariance matrix, are selected in order of the amount of variance. The first PC is that component which has the greatest possible variance, the second PC has the second greatest variance, and so on. Eigenvalues describe the amount of variance explained by each PC, and thus decrease with each successive PC extracted. A detail of the PCA method can be found in Jolliffe (2002).

Self-organizing maps

The SOM is a learning algorithm that was originally proposed by Kohonen (1990). The SOM consists of one input layer and one output layer (Kohonen layer). The input layer of neurons is fully connected to the output layer. An attractive ability of the SOM is to map high-dimensional input space into low-dimensional space. The topological structure of the SOM can be one or two dimensional. Higher dimensions are acceptable but not common.

The SOM is trained using an unsupervised competitive learning algorithm which is a process of self organization. After SOM training is complete, a feature map is then obtained by labeling all winning neurons in the output space with the identities of corresponding input patterns. The feature map is two dimensional and composed of grids that represent neurons in the Kohonen layer. The winning neuron shows the topological location of the input pattern. The neighborhood of the winning neuron shows the statistical distribution of the input pattern. Once the clusters are formed in the feature map, the data records from each cluster can be sampled and the density map can then be constructed. Initially, the number of members in each grid of the feature map is counted. The number of members represents the frequency that the grid has been 'imaged' by specific input patterns. Then, every grid is labeled with an integer which is the number of members in that grid. If there are no members in a grid, the label is left blank. Hence, the feature map can show the topological relationships among the corresponding input patterns, and the density map can yield an objective number of clusters for input patterns. These are the advantages of SOM over the conventional clustering methods. For more details regarding the SOM learning processes please refer to Kohonen (2001).

Method of *L*-moments

The approach based on the theory of *L*-moments was first proposed by Wallis (1989), and then developed by Hosking & Wallis (1997). *L*-moments are summary statistics for probability distributions and data samples. They are analogous to ordinary moments, because of providing measures of location, dispersion, skewness, kurtosis, and other aspects of the shape of probability distributions or data samples, but are computed from linear combinations of the ordered data values. *L*-moments may be applied in four steps of the regional frequency analysis including screening of the data, identification of homogeneous regions, choice of a frequency distribution and estimation of the frequency distribution (Hosking & Wallis 1997). The main advantages of *L*-moments over conventional moments are that they are able to characterize a wider range of distributions, are less subject to bias in estimation and more robust to the presence of outliers in the data. Basically, *L*-moments are linear functions of probability weighted moments (PWMs):

$$\beta_k = \mathrm{E}\{X[F(x)]^r\} \tag{1}$$

where $F(x)$ is the cumulative distribution function of x. The first four *L*-moments expressed as linear combinations of PWM are:

$$\lambda_1 = \beta_0 \tag{2}$$

$$\lambda_2 = 2\beta_1 - \beta_0 \tag{3}$$

$$\lambda_3 = 6\beta_2 - 6\beta_1 + 6\beta_0 \tag{4}$$

$$\lambda_4 = 20\beta_3 - 30\beta_2 - 12\beta_1 - \beta_0 \tag{5}$$

The *L*-mean, λ_1, is a measure of central tendency and the *L*-standard deviation, λ_2, is a measure dispersion. Their ratio,

λ_2/λ_1, is termed the *L*-coefficient of variation, τ_2. The ratio λ_3/λ_2, is referred to as τ_3 or *L*-skewness, while the ratio λ_4/λ_2, is referred to as τ_4 or *L*-kurtosis. A detailed description of the *L*-moments can be found in Hosking & Wallis (1997).

Screening of data

The aim of this stage is to form groups of gauges that satisfy the homogeneity condition, those gauges with frequency distributions that are identical apart from a gauge–specific scale factors. This is usually carried out by dividing the sites into disjoint groups. Hosking & Wallis (1997) present a discordancy measure. In this approach, if any point does not appear to belong to the cluster of such points on the *L*-moment diagram, that is, is far from the center of the cluster, the site related to that point should be removed from the region due to a nonhomogeneity condition. Discordancy measure D_i of a site can be calculated by:

$$D_i = \frac{1}{3}(u_i - \bar{u})^{\mathrm{T}} A^{-1} (u_i - \bar{u}) \tag{6}$$

$$A = \frac{1}{N-1}\sum_{i=1}^{N}(u_i - \bar{u})(u_i - \bar{u})^{\mathrm{T}} \tag{7}$$

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i \tag{8}$$

where N is the total number of sites and u_i is defined as a vector containing the *L*-moment ratios for site I, and $\bar{u}$ and A are the group averages and sample covariance matrix of u, respectively. Generally, sites with *D*-statistics greater than 3 are considered to be discordant from the rest of the region.

Heterogeneity test for regions

For testing the regional homogeneity, a test statistic H, termed as heterogeneity measure was proposed by Hosking & Wallis (1997). Through a Monte-Carlo simulation exercise, it compares the inter-site variations in sample *L*-moments for the group of sites with what would be expected of a homogeneous region. A number of 500 data regions are generated based on the regional weighted average statistics using Kappa distribution. Heterogeneity measure H is computed as below:

$$H = \frac{(V - \mu_V)}{\sigma_V} \tag{9}$$

where V is weighted (standard deviation) *L*-coefficient of variation, τ_2, μ_V and σ_V are the mean and standard deviation of 500 values of V. The value of H-statistic indicate that the region under consideration is acceptably homogeneous when $H < 1$, possibly heterogeneous when $1 \le H < 2$, and definitely heterogeneous when $H \ge 2$ (Hosking & Wallis 1997).

Choosing the best-fit frequency distribution

In regional frequency analysis, a single frequency distribution is fitted to the data from several sites in a homogeneous region. Hosking & Wallis (1997) proposed an appropriate method for goodness-of-fit criterion based on *L*-kurtosis. The goodness-of-fit criterion, *Z*-statistic, is defined as:

$$Z^{\mathrm{DIST}} = \frac{(\tau_4^{\mathrm{DIST}} - t_4^{(R)} + \beta_4)}{\sigma_4} \tag{10}$$

where τ_4^{DIST} is average of τ_4 computed from simulation of a fitted distribution, σ^4 is standard deviation of τ_4 obtained from simulated data, β_4 and δ_4 are the bias and standard deviation of τ_4, respectively. The goodness-of-fit measure, Z, judges how well the simulated *L*-skewness and *L*-kurtosis values obtained from the observed data. A probability distribution with the smallest value of $|Z|$ is considered the best choice among possible distributions. At the significance level of $\alpha = 0.10$, the critical value of Z is 1.64, i.e., if a probability distribution whose $|Z| \le 1.64$, then it is assessed to be an acceptable distribution for representing sample data at $\alpha = 0.10$.

Estimation of rainfall quantiles

Using the regional parameters for the identified distribution, standardized quantiles for the region with specified return periods can be estimated, and they are then multiplied by the gauge specific annual average rainfall to obtain the desired rainfall quantiles for the rain gauge in question.

THE STUDY AREA AND DATA

In this paper, actual rainfall data of Taiwan are used. As shown in Figure 1, Taiwan is located at the southeast of Asia in the Western Pacific (between Japan and the Philippines) with a total area of 36,000 km^2. The shape of Taiwan is long and narrow, and the middle of Taiwan is the Central Mountain Range. The mountainous area with elevation higher than 1,000 m occupies 32% of the island, hills and plateaus of 100 m to 1,000 m cover 31% of the island, and the rest of the island is plain with elevation less than 100 m. The contour map of mean annual rainfall in Taiwan is shown in Figure 2. The mean annual rainfall in Taiwan is around 2,500 mm, it reaches 3,000 to 5,000 mm in the mountainous regions. However, up to 80 percent of that comes between May and October during the summer typhoon season in particular. The maximum one hour rainfall reaches 300 mm, the maximum one-day rainfall reaches l,748 mm which is 93.4% of the world record (1,870 mm). In comparison with the records in the world, the one-hour to 3-day maximum rainfalls in Taiwan are approximately 85 to 93% of the world records.

Figure 1 shows the study area and the locations of 127 rain gauges. Elevations of the rain gauges range from 3 to 2,540 m above sea level. The rainfall data are collected from computer archives of the Water Resources Agency. These rain gauges have 20 or more years of rainfall record and all gauges are currently operational. There are no significant and spatially uniform trends in extreme rainfall events over the period under study, so the basic assumption of the regional analysis, which relies on stationary processes, is not violated.

RESULTS AND DISCUSSIONS

Principal component analysis

In this section, a PCA is performed on the series of annual maximum daily rainfall values from 127 gauges in Taiwan. Rainfall records at 1 h intervals for the 127 gauges are available for 20 years from 1988 to 2007. Hence, there are 20 input values for each rainfall gauge used in the PCA. The PCs are ordered in such a way that the variance explained by the first PC is the greatest, the variance explained by the second one is smaller, and so on, until that of the last one is the smallest. The PCs represent the feature pattern

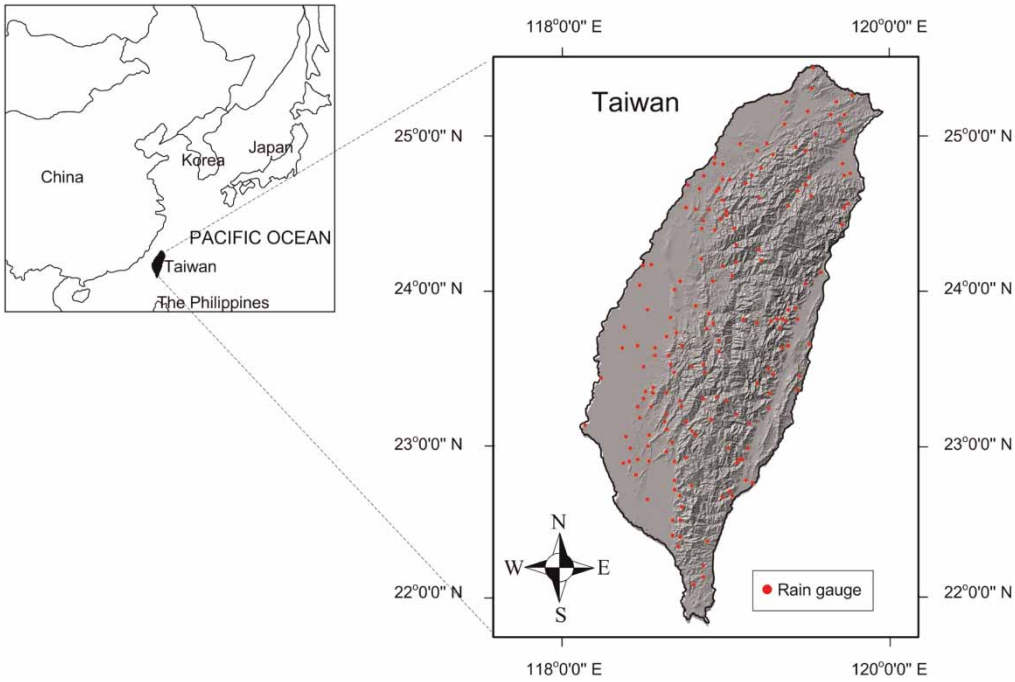

Figure 1 | The study area and the locations of rainfall gauges.

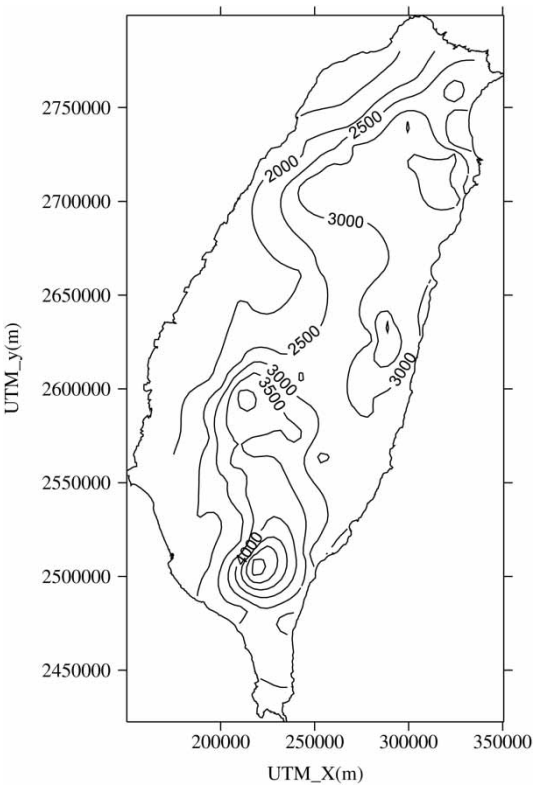

Figure 2 | The contour map of mean annual rainfall in Taiwan (mm).

of annual maximum daily rainfalls. Hence, designable feature pattern of annual maximum daily rainfalls can be determined when rain gauges are grouped into clusters. The results of the PCA are presented in Table 1. As shown in Table 1, one can find that the first nine PCs explain over 80% of the information.

SOM-based clustering

SOM is used to group the rain gauges into specific clusters based on the transformed data resulting from PCA and the geographic characters of the gauges, including gauge latitude (m), gauge longitude (m) and elevation (m). Because the scales of the data sets are very different and the clustering methods are very sensitive to such scale differences, the variables must be normalized to avoid different weights of data. The normalization is performed using the following equation:

$$E = \frac{G - G_{\min}}{G_{\max} - G_{\min}} \tag{11}$$

Table 1 | Results of principal components analysis on 127 annual maximum daily rainfall data

No. of principal component	Eigen value	Variance explained (%)	Total variance explained (%)
1	0.21	19.61	19.61
2	0.14	13.06	32.67
3	0.13	12.00	44.67
4	0.10	8.83	53.50
5	0.09	8.50	62.00
6	0.07	6.41	68.41
7	0.06	5.45	73.86
8	0.06	5.21	79.07
9	0.04	3.56	82.63
10	0.03	2.89	85.52
11	0.03	2.70	88.22
12	0.02	2.28	90.50
13	0.02	1.87	92.37
14	0.02	1.78	94.15
15	0.02	1.49	95.64
16	0.01	1.15	96.79
17	0.01	0.99	97.78
18	0.01	0.95	98.73
19	0.01	0.65	99.38
20	0.01	0.62	100.00

where E is the normalized value, and $G_{\min}$ and $G_{\max}$ are the minimum and maximum values in the data set, respectively. Equation (11) shows that these site characteristics are rescaled so that their values lie between 0 and 1. This normalization is done separately for each variable.

When SOM is applied to perform cluster analysis, a SOM of a small dimension is the first choice. If the SOM-based clustering result is reasonable and satisfactory, the cluster analysis is accepted. Otherwise, a SOM of a larger dimension is chosen to analyze input patterns and this situation continued until a satisfactory result is obtained. After a total of 98,000 iterations (500 times the number of neurons), the SOM is constructed. Once the feature map is available, the density map can then be constructed. Figure 3 presents the two-dimensional density map obtained on a network of 14×14 cells. The numbers in the lattices represent the number of clusters for input patterns. In other words, the density map can be divided into several regions by the

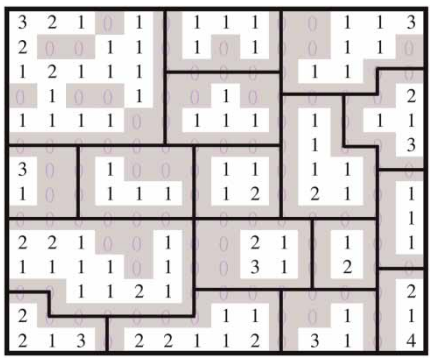

Figure 3 | The density map derived from the SOM of dimensions 14×14.

zero-numbered lattices. As shown in Figure 3, the map is divided into 17 regions. That is, the 127 rain gauges in Taiwan can be grouped into 17 clusters. The locations of the rain gauges in each cluster are shown in Figure 4.

Homogeneity test for regions

The results of discordancy measures and heterogeneity tests for the 17 clusters obtained by the SOM are summarized in Table 2. Discordancy measures for the 17 clusters are found between 0.07 and 2.42. The data range shows no discordancy for these clusters. In addition, Table 2 shows that the values

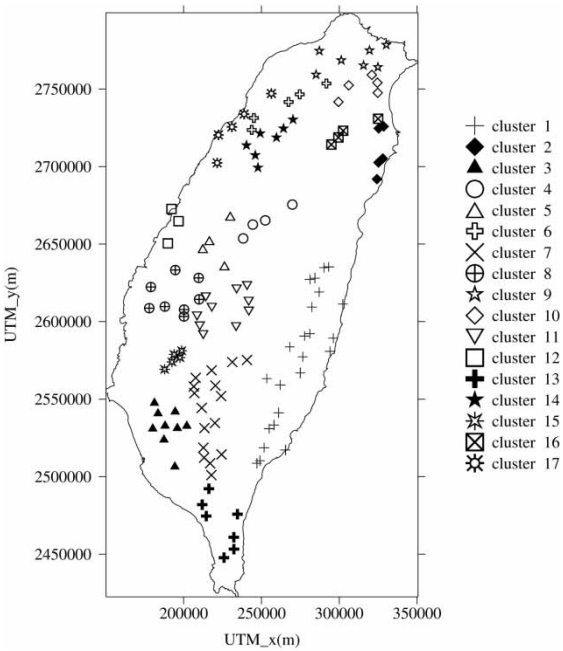

Figure 4 | Location of the sites in 17 clusters obtained by the SOM based on the principal components.

Table 2 | Results of discordancy measures and heterogeneity tests for the 17 clusters obtained by the SOM based on the principal components

No. of cluster	Number of gauges	Discordancy measure D	Heterogeneity measure H		
			H_1	H_2	H_3
1	23	0.22–2.28	−1.14	−0.06	−0.82
2	5	0.42–1.32	0.86	0.12	−0.26
3	9	0.04–1.51	−0.14	−0.19	0.42
4	4	1.00	−0.23	−0.41	−0.81
5	4	1.00	−1.26	−0.72	−0.33
6	5	1.00	−0.84	−0.40	−0.69
7	16	0.72–1.32	−0.72	−0.69	−0.80
8	8	0.07–1.73	0.34	−1.00	−1.09
9	7	0.42–1.90	−1.16	−1.22	−0.96
10	5	0.07–2.42	−1.20	−2.49	−2.04
11	10	0.10–1.69	−0.01	0.08	0.76
12	3	1.00	0.69	−0.70	−1.39
13	7	1.00	−0.54	−1.35	−1.34
14	7	0.46–1.63	0.76	−0.97	−1.50
15	5	0.27–2.11	−1.14	−2.78	−3.11
16	4	0.35–1.24	−1.00	−0.46	−0.64
17	5	0.48–1.61	0.41	−1.97	−2.47

of different heterogeneity measures, H_1, H_2 and H_3 for the 17 clusters are all found under 1. Hence, for the SOM method all regions are 'acceptably homogeneous'. These results demonstrate that the 17 regions are sufficiently homogeneous.

Delineation of homogeneous regions

When the rain gauges are grouped into specific clusters using the SOM, the next step is to delineate the homogenous region for each cluster. Because the rainfall characteristics at ungauged sites are unknown, a logical assumption is that the correlation of rainfall characteristics between two locations increases with decreasing distance. In this paper, Taiwan is divided equally into grid squares of 500 m × 500 m. The grid latitudes and longitudes are prior calculated and the grids are then formed. Finally, the cluster to which each grid belongs can be determined. The homogeneous regions for the 17 clusters obtained by PCA and the SOM in Taiwan are shown in Figure 5. A gauged site or ungauged site can be assigned to a proper homogeneous region according to the map of homogeneous regions (Figure 5), and regional rainfall frequency analysis will be performed.

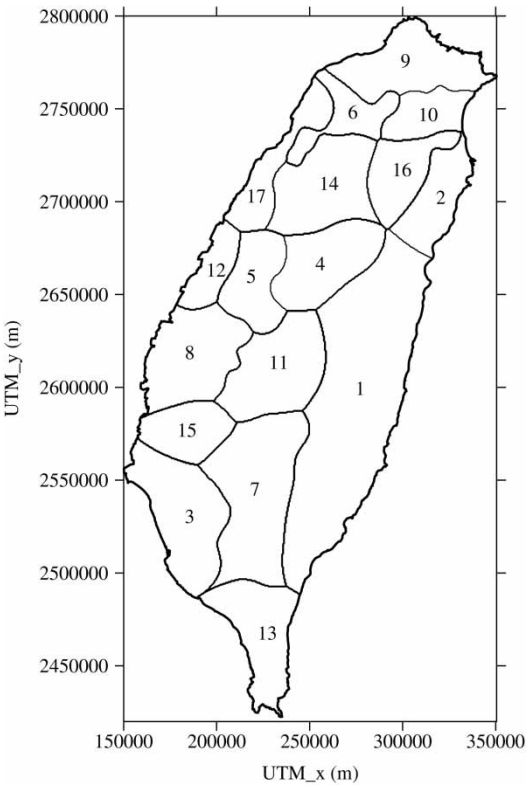

Figure 5 | The 17 homogeneous regions obtained by PCA and the SOM in Taiwan.

Choosing the best-fit frequency distribution

After confirming the homogeneity of the 17 regions, an appropriate distribution needs to be selected for the regional rainfall frequency analysis. In other words, all gauges in a homogeneous region should have the same population *L*-moments. The identification of an appropriate regional distribution for the 17 regions is based on the *Z*-statistic. The selection is carried out by comparing the moments of the candidate distributions with the average moments statistics derived from the regional data. The best fit to the observed data indicates the most appropriate distribution. A number of five three-parameter distributions, i.e. Generalized logistic (GLO), Generalized extreme value (GEV), three-parameter lognormal distribution (LN3), Pearson type III distribution (PE3) and Generalized Pareto distribution (GPA) are fitted to the regions. *Z*-statistics computed using Equation (11) with the total data, reveal that five distributions are the possible candidates which can describe the observed data well. The *Z*-statistics concerning with some statistical distributions for the delineated homogeneous regions are shown in Table 3. According to the analysis of results of goodness-of-fit test (*Z*-statistic), one can find that some regions have only one acceptable distribution, but in some regions the all the candidates are acceptable. The distribution, which produces the least *Z*-statistics from among the other plausible distributions, is chosen as the underlying distribution, because it indicates that this distribution fitted the data better than others. For example, the values of *Z*-statistic related to GEV, LN3 and PE3 distributions for Region 1 are less than the critical *Z*-statistic value (1.64). Therefore, these distributions may be used in regional frequency analysis for the regions, and the GEV distribution which has the least *Z*-statistic (0.72) is considered as the best-fit distribution. Chosen distribution type and function for each homogeneous region is presented in Table 4. It can be seen that the GPA distribution is acceptable most often, in nine of the homogeneous regions.

Once Taiwan has been divided into 17 homogeneous regions, the data for each region is analyzed and a best-fit distribution is chosen and fitted for each region. In Taiwan, there are

Table 3 | Results of goodness-of-fit measures for each region

	|Z|				
No. of region	**GLO**	**GEV**	**LN3**	**PE3**	**GPA**
1	3.90	0.72[a]	1.23[a]	1.09[a]	5.38
2	4.84	2.79	3.35	3.34	0.86[a]
3	4.21	2.41	2.45	2.15	1.29[a]
4	3.21	2.77	2.36	1.66	1.53[a]
5	3.05	2.22	1.96	1.45[a]	0.28[a]
6	4.58	3.11	3.16	2.93	0.09[a]
7	4.08	1.95	1.95	1.55[a]	2.48
8	5.55	4.04	3.85	3.34	0.72[a]
9	2.78	1.38[a]	1.32[a]	0.99[a]	1.60[a]
10	5.85	4.32	4.39	4.18	1.19[a]
11	4.38	3.52	3.41	3.10	1.62[a]
12	1.68	0.94[a]	0.78[a]	0.43[a]	0.74[a]
13	1.25[a]	2.05	2.28	2.74	3.90
14	3.36	2.81	2.64	2.31	1.52[a]
15	0.82[a]	0.43[a]	0.04[a]	0.86[a]	0.73[a]
16	1.46[a]	0.68[a]	0.47[a]	0.06[a]	1.13[a]
17	0.62[a]	0.06[a]	0.30[a]	0.92[a]	1.42[a]

[a]The distribution may be accepted as a regional distribution.

Table 4 | Chosen distribution type and function for each homogeneous region

No. of region	Probability distribution type	Cumulative distribution function
1	GEV	$F(x) = e^{(0.4268x-1.353)^{5.814}}$
2	GPA	$F(x) = 1-(1.1572-0.6634x)^{0.9756}$
3	GPA	$F(x) = 1-(1.2563-0.6133x)^{1.8018}$
4	GPA	$F(x) = 1-(1.024-0.0625x)^{25}$
5	GPA	$F(x) = 1-(1.1621-0.3777x)^{3.6364}$
6	GPA	$F(x) = 1-(1.2857-0.6493x)^{1.7483}$
7	PE3	$F(x) = G(5.565x+0.6229,\ 6.188)$
8	GPA	$F(x) = 1-(1.1862-0.4786x)^{2.4213}$
9	PE3	$F(x) = G(5.568x-0.0057,\ 5.0612)$
10	GPA	$F(x) = 1-(1.1266-0.4983x)^{1.692}$
11	GPA	$F(x) = 1-(1.0655-0.3543x)^{2.4631}$
12	PE3	$F(x) = G(2.9548x-0.1677,\ 2.7871)$
13	GLO	$F(x) = [1+(1.0142x+0.0660)^{-4.6512}]^{-1}$
14	GPA	$F(x) = 1-(1.0374-0.258x)^{3.5336}$
15	LN3	$F(x) = \phi(-1.3569\ln x-0.2888)$
16	PE3	$F(x) = G(2.8571x-0.2971,\ 2.56)$
17	GEV	$F(x) = e^{-(0.5146x+0.6064)^{-6.2893}}$

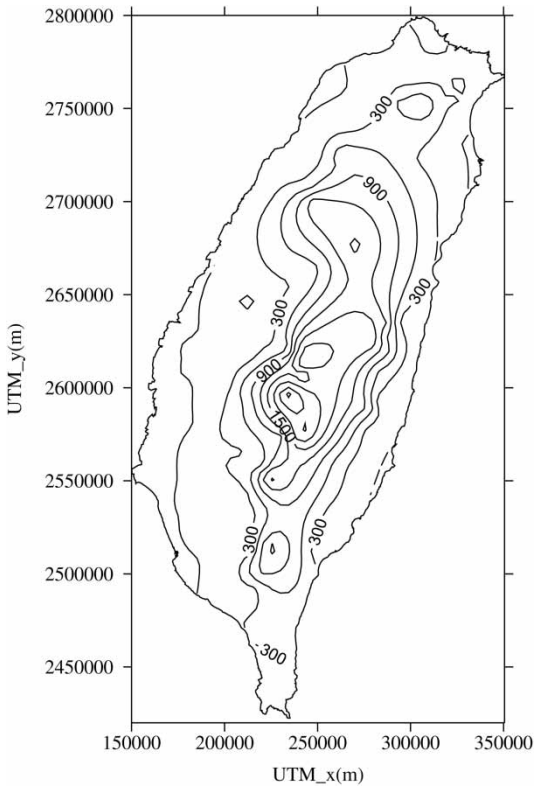

Figure 6 | The contour map of Taiwan (m).

different distributions (GPA, GEV, PE3, GLO, and LN3) with different parameters for the 17 regions. However, according to Figure 5 and Table 4, one can see that the contiguous regions have different probability distribution types. For example, the regions 1, 7 and 13 are contiguous in the south-east of Taiwan (see Figure 5), but there are different probability distribution types such as GEV, PE3 and GLO for the three regions (see Table 4). The reason is that the distribution of water resources in Taiwan is uneven both temporally and spatially. On average, approximately 80% of annual rainfall occurs during the wet period, from May to October, with typhoons and south-western convective storms accounting for most of this rainfall. The period from November to April is the dry season, and accounts for only around 20% of mean annual rainfall. Furthermore, the island of Taiwan is sectioned into two distinguished parts by the Central Mountain Range (see Figures 2 and 6). Foothills from the Central Mountain Range lead to tablelands and coast plains in the western (Region 7) and south Taiwan (Region 13). The eastern Taiwan (Region 1) is characterized by long and narrow gorges. Therefore, in regions 1, 7 and 13, the three different distributions are used, although the three regions are contiguous in the south-east of Taiwan.

Regional rainfall quantile estimation

The next step in regional rainfall frequency is to estimate rainfall quantiles in the homogeneous region. The desired rainfall quantile estimates with different return periods for each homogeneous region are laid out in Table 5. As indicated in Table 5, the 5-year rainfall depths is between 224.0 mm (Region 17) and 540.2 mm (Region 7); the 10-year rainfall depths is between 277.4 mm (Region 17) and 630.2 mm (Region 7); the 20-year rainfall depths is between 315.1 mm (Region 12) and 707.4 mm (Region 7); the 50-year rainfall depths is between 355.0 mm (Region 12) and 796.7 mm (Region 7); the 100-year rainfall depths is between 381.8 mm (Region 12) and 857.3 mm (Region 7). According to the above results, one can find that the maximum T-year rainfall depths of one-day duration in Taiwan occur in Region 7, which belongs to the mountain areas of southwest Taiwan. On the other hand, the minimum T-year rainfall depths of one-day duration in Taiwan occur in Region 12 and Region 17, which are located on the southwest coast

Table 5 | Quantile estimates with different return periods for each homogeneous region

	Rainfall quantiles (mm)				
Region	$T = 10$ yr	$T = 20$ yr	$T = 50$ yr	$T = 100$ yr	$T = 200$ yr
1	408.4	483.8	551.3	632.5	688.9
2	447.3	528.0	599.3	683.8	741.7
3	324.4	368.9	406.4	449.1	477.6
4	331.6	394.8	449.9	514.9	559.4
5	266.8	309.5	346.3	389.2	418.3
6	267.4	303.7	334.4	369.5	393.0
7	540.2	630.2	707.4	796.7	857.3
8	277.4	322.4	361.1	406.1	436.6
9	310.6	363.1	408.1	460.3	495.6
10	427.9	499.5	560.0	629.0	675.1
11	474.0	566.8	646.1	737.8	799.8
12	239.5	280.4	315.1	355.0	381.8
13	422.2	507.7	592.3	705.6	793.3
14	435.6	520.8	593.8	678.4	735.8
15	372.8	467.9	565.4	700.1	806.4
16	452.0	535.7	607.3	690.1	746.2
17	224.0	277.4	335.3	421.3	495.1

of Taiwan. The contour map of Taiwan is shown in Figure 6. According to Figure 6 and Table 5, one can find that the T-year rainfall depths increase with the increasing elevation of the gauges. Especially near the Central Mountain Range, the T-year rainfall depths become greater. In other words, altitude has a great influence on regional rainfall frequency analysis in Taiwan. For future hydrologic designs, one can determine which ungauged site belongs to which region in Figure 5 according to the location of the site, and then the desired rainfall quantiles with various return periods for ungauged site are obtained according to Table 5.

According to Figure 6 and Table 5, it is observed that the regions 15, 8 and 3 are all on the west coast of Taiwan and there are no elevation differences, but the 100 year return period daily precipitations for regions 15, 8 and 3 are 700, 406 and 449 mm, respectively. The 100-year rainfall depth for Region 15 is more than 200 mm greater than other two similar regions (Region 8 and Region 3) because there is no rainfall gauge on the west coast of Region 15 (see Figure 4). That is, there are only five gauges near the Central Mountain Range in Region 15. According to Figure 6, one can find that the five gauges are located at higher elevations. Therefore, the 100-year rainfall depth for Region 15 is greater than that for Regions 8 and 3. It is suggested that, in order to overcome these problems, the rainfall gauges shall be installed on the west coast of Region 15.

SUMMARY AND CONCLUSIONS

The estimation of annual maximum rainfall in a region where no data is available is very important for engineering hydrologic design. The main purpose of the study aims to investigate the regional rainfall frequency for ungauged sites. The PCA, SOM and L-moments are applied for regionalization of annual maximum rainfall in Taiwan. This study is based on 20 or more years of annual rainfall data at 127 rain gauges across Taiwan. First, PCA is applied to obtain the PCs. It is found that the first nine principal components explain over 80% of the information. Based on the transformed data resulting from PCA and the geographic characters of the gauges, the SOM is used to group the rain gauges into specific clusters. A two-dimensional density map indicates that the 127 rain gauges can be grouped into 17 clusters. That is, the 17 homogeneous regions for regional frequency analysis can be delineated. One can determine which ungauged site belongs to which region according to the location of the site, and then the desired rainfall quantiles with various return periods for ungauged site can be estimated. Moreover, the L-moment based discordancy, and heterogeneity are used to test whether regions may be acceptable as being homogeneous and goodness-of-fit measures is used to select the regional probability distributions of rainfalls. The results show that the 17 regions are sufficiently homogeneous, and the best regional probability distributions for 17 regions are determined. Finally, the design rainfall quantiles with various return periods for each region are estimated. In the proposed approach, the design rainfall quantiles at the point of interest can be easily determined once the homogeneous regions are available. Moreover, the results show that the desired rainfall quantiles with various return periods increase with the increasing elevation of the gauges in Taiwan. In our proposed approach, a gauged site or ungauged site can be assigned to a proper region

according to the map of homogeneous regions (Figure 5), and then the regional rainfall frequency analysis will be performed easily. Hence, the proposed approach is recommended as an alternative to regional rainfall frequency analysis because it can not only objectively determine the suitable number of homogeneous regions, but also produce reasonable desired rainfall quantiles.

REFERENCES

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000a Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering-ASCE* **5**, 115–123.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000b Artificial neural networks in hydrology. II: Hydrologic applications. *Journal of Hydrologic Engineering-ASCE* **5**, 124–137.

Chang, F. J., Chang, L. C. & Wang, Y. S. 2007 Enforced self-organizing map neural networks for river flood forecasting. *Hydrological Processes* **21**, 741–749.

Everitt, B. S. 1993 *Cluster Analysis*. John Wiley & Sons, New York.

Fowler, H. J. & Kilsby, C. G. 2003 A regional frequency analysis of United Kingdom extreme rainfall from 1961 to 2000. *International Journal of Climatology* **6**, 2103–2135.

Hall, M. J. & Minns, A. W. 1999 The classification of hydrologically homogeneous regions. *Hydrological Sciences Journal* **44**, 693–704.

Hall, M. J., Minns, A. W. & Ashrafuzzaman, A. K. M. 2002 The application of data mining techniques for the regionalization of hydrological variables. *Hydrology and Earth System Sciences* **6**, 685–694.

Haykin, S. 1994 *Neural Networks: A Comprehensive Foundation*. IEEE Press, New York.

Hosking, J. R. M. & Wallis, J. R. 1997 *Regional Frequency Analysis: An Approach Based on L-moments*. Cambridge University Press, New York.

Hsu, K. S. & Li, S. T. 2010 Clustering spatial–temporal precipitation data using wavelet transform and self-organizing map neural network. *Advances in Water Resources* **33**, 190–200.

Jolliffe, I. T. 2002 *Principal Component Analysis*, 2nd edition. Springer, New York.

Kalteh, A. M., Hjorth, P. & Berndtsson, R. 2008 Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. *Environmental Modelling and Software* **23**, 835–845.

Kohonen, T. 1990 The self-organizing map. *Proceedings of the Institute of Electrical and Electronics Engineers* **78**, 1464–1480.

Kohonen, T. 2001 *Self-Organizing Maps*. Springer-Verlag, Berlin.

Lin, G. F. & Chen, L. H. 2005 Time series forecasting by combining the radial basis function network and the self-organizing map. *Hydrological Processes* **19**, 1925–1937.

Lin, G. F. & Chen, L. H. 2006 Identification of homogeneous regions for regional frequency analysis using the self-organizing map. *Journal of Hydrology* **324**, 1–9.

Lin, G. F. & Wu, M. C. 2007 A SOM-based approach to estimating design hyetographs of ungauged sites. *Journal of Hydrology* **339**, 216–226.

Lin, G. F. & Wu, M. C. 2009 A hybrid neural network model for typhoon-rainfall forecasting. *Journal of Hydrology* **370**, 450–458.

Lin, G. F., Wu, M. C., Chen, G. R. & Liu, S.J. 2010 Construction of design hyetographs for locations without observed data. *Hydrological Processes* **24**, 481–491.

Mangiameli, P., Chen, S. K. & West, D. 1996 A comparison of SOM neural network and hierarchical clustering methods. *European Journal of Operational Research* **93**, 402–417.

Meshgi, A. & Khalili, D. 2009 Comprehensive evaluation of regional flood frequency analysis by *L*- and *LH*-moments. I. A re-visit to regional homogeneity. *Stochastic Environmental Research and Risk Assessment* **23**, 119–135.

Michaelides, S. C., Pattichis, C. S. & Kleovoulou, G. 2001 Classification of rainfall variability by using artificial neural networks. *International Journal of Climatology* **21**, 1401–1414.

Mingoti, S. A. & Lima, J. O. 2006 Comparing SOM neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research* **174**, 1742–1759.

Nathan, R. J. & McMahon, T. A. 1990 Identification of homogeneous regions for the purposes of regionalisation. *Journal of Hydrology* **121**, 217–238.

Parida, B. P. & Moalafhi, D. B. 2008 Regional rainfall frequency analysis for Botswana using *L*-moments and radial basis function network. *Physics and Chemistry of the Earth* **33**, 614–620.

Rao, A. R. & Srinivas, V. V. 2006a Regionalization of watersheds by Hybrid cluster analysis. *Journal of Hydrology* **318**, 37–56.

Rao, A .R. & Srinivas, V.V. 2006b Regionalization of watersheds by Fuzzy cluster analysis. *Journal of Hydrology* **318**, 57–79.

Wallis, J. R. 1989 *Regional frequency studies using L-moments*. IBM Research Report RC14597, New York, USA.

Yang, C. C. & Chen, C. S. 2009 Application of integrated back-propagation network and self organizing map for flood forecasting. *Hydrological Processes* **23**, 1313–1323.

Zhang, J. Y. & Hall, M. J. 2004 Regional flood frequency analysis for the Gan-Ming River basin in China. *Journal of Hydrology* **296**, 98–117.

First received 12 April 2010; accepted in revised form 25 February 2011