International Journal of Integrated Engineering, Vol. 9 No. 2 (2017) p. 18-26

Identification of Homogeneous Areas for Drought Frequency Analysis

Siti Nazahiyah Rahmat[1,*], **Niranjali Jayasuriya**[2], **Muhammed Bhuiyan**[2]

[1]Faculty of Civil and Environmental Engineering, Universiti Tun Hussein Onn Malaysia (UTHM), Batu Pahat, Johor, Malaysia
[2]School of Engineering, RMIT University, GPO Box 2476, Melbourne, Victoria 3001, Australia

Received 31 January 2017; accepted 9 April 2017; available online 10 April 2017

Abstract: Owing to high spatial and temporal rainfall variability, rationale water management decision-making is complex. Hence, it is essential to identify homogeneous areas to assist water management. This paper focusses on separating the study area into homogeneous groups to predict the risk of occurrence of a drought event. The severity-duration-frequency (SDF) curves were developed to determine the relationship between the probability of a drought occurring with a certain severity and frequency at the selected stations in Victoria, Australia. Two techniques namely cluster analysis and modified Andrews curve were used in grouping study area that have similar climate characteristics with respect to risk of occurrence of drought (i.e. rainfall variability). Based on the results, mean seasonal precipitations (i.e. summer and spring) were found to be the most important parameters in clustering droughts. The study area was divided into six clusters and they adequately covered the study area. A mean drought frequency curve was developed for each homogeneous group to determine the probability of vulnerability to a drought event with a certain severity. The advantage of separating stations into homogenous groups based on similar drought characteristics is that it eliminates the necessity to carry out a detailed drought characteristic analysis for any location of interest.

Keywords: Cluster analysis, drought, modified Andrews curve, regionalisation, severity-duration-frequency (SDF) curves

1. Introduction

The multivariate classification methods have been widely used for the grouping of water quality [1] and the regionalisation of streams [2-3]. However, for drought characterisation [4], their application has been very limited to date.

Common methods used for regionalising catchment parameters are classified into three which are regression between individual calibrated parameters and catchment characteristics, catchment spatial proximity and catchment similarity of physical properties [5]. Clustering approach [6] and the spatial method, including Kriging interpolation [7] are the examples of catchment spatial proximity methods. [8] concluded that the methods based on spatial proximity alone performed significantly better than any of the regression methods based on catchment attributes. [9] concluded that spatial proximity and a combination of physical similarity methods performed best. Overall, there is no universal method exists at present that performs best for all conditions and this remains a subject of investigation. It is worthwhile to continue the research and test methods in different regions.

Drought Severity-Duration-Frequency (SDF) curves was developed by [10] in Victoria, Australia. The development of the frequency curves is based on the precipitation values which were computed based on the Standardised Precipitation Index (SPI) developed by [11]. The formula to calculate SPI can be used to calculate the precipitation threshold as well as the precipitation deficit. In the current study, the precipitation deficit estimate is based on the SPI thresholds. Based on the SDF curves that were developed, some stations gave similar values hence showed similar curves (see Fig. 1). Furthermore, these curves could not be applied at new stations and another curves had to be established.

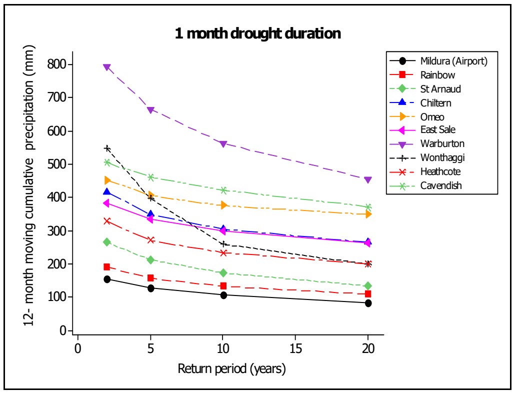


Fig. 1 The 12-month cumulative rainfall (mm) plot for the 10 stations [11]

While much work has been done on rainfall and runoff regionalisation, there is still no comprehensive study on

**Corresponding author:nazahiya@uthm.edu.my*

penerbit.uthm.edu.my/ojs/index.php/ijie

grouping methods for droughts. Therefore, this paper will focus on separating the study area into homogeneous groups that have similar drought signatures (spatial and temporal rainfall variation) using cluster analysis and modified Andrews curve [12]. An additional 60 stations were selected to comprehensively cover the State of Victoria and to attain a more accurate outcome. The SDF curves will be derived for each identified clusters. The advantage of separating stations into homogenous groups is that it eliminates the necessity to carry out a detailed drought characterisation for any location of interest. The characteristics of this station will determine its best match with the existing cluster groups. In the present study, the clustering approach was applied and modified Andrews curves were used for visualizing similarity in catchment characteristics within the groups. The analyses were carried out using precipitation as the basic parameter, making the analysis compatible with the SPI as its computation only requires rainfall data to identify drought.

This paper is organised as follows: Section 2 presents a description of the study area, significant information on the selected rainfall stations and methods of analysis. Section 3 provides the application of cluster and modified Andrews curve methods to cluster regions based on the SPI. This section also describes the development of the drought SDF curves. Finally, Section 4 presents the main conclusions drawn.

2. Methodology

2.1 Study Area and Data

The focus of this study is the State of Victoria, Australia. Victoria is located in south-east Australia. The 70 rainfall stations selected for this study is shown in Fig. 1. Initially, nine variables (i.e. mean monthly precipitation, mean annual precipitation, mean monthly and mean annual precipitation coefficient of variations and mean seasonal precipitation (summer, autumn, winter and spring) and elevation) were chosen to classify the selected stations using regionalisation methods. These variables (with the exception of elevation) were selected to make the analysis compatible with the SPI as its computation only requires rainfall data to identify drought. For the computation of seasonal precipitation, the seasons were described as summer (from December to February), autumn (from March to May), winter (from June to August) and spring (from September to November).

2.2 Methods of Analysis

Initially, cluster analysis was carried out to separate the study area into homogenous areas followed by modified Andrews curves to refine the stations in each cluster. However, before applying cluster analysis, step-wise regression needs to be carried out to identify the most important variables with respect to risk of occurrence of drought. From the cluster analysis, it is possible to obtain the optimum number of homogenous groups. Modified Andrews curves were then developed for each station and compared. Therefore, a quick visual observation with regard to the appropriateness of the station belonging to a specific group can be carried out. These two methods are discussed below.

2.2.1 Cluster Analysis

Cluster analysis methods provide means for classifying a given population into groups, based on similarity or closeness measures. This principle can be mathematically quantified by means of the objective principle of the distance [13]. The Euclidean distance commonly gives the similarity between two samples and a distance can be represented by the difference between analytical values from the samples [14]. The agglomerative hierarchical methods have been widely used for clustering. In these methods, the groups are formed by merging the objects into bigger and bigger clusters. In the present study, the hierarchic agglomerative cluster applied was the complete linkage whereas the distance elaboration was performed by adopting the squared Euclidean distance. For example, the squared Euclidean distance (D^2) between location 1 and location 2 is calculated from normalised values as follows:

$$D^2 = (Z_{MSu1} - Z_{MSu2})^2 + (Z_{MSp1} - Z_{MSp2})^2 + \cdots \qquad (1)$$

where Z_{MSu1} and Z_{MSu2} are the normalised values of mean summer precipitation at locations 1 and 2. Similarly, Z_{MSp1} and Z_{MSp2} are similar values of mean spring precipitation. The results of a cluster analysis are normally illustrated using a dendogram diagram that enables a clear visualisation of the similarity between the studied objects. Selection of the number of clusters was based on several trials to ensure a high percentage in the similarity levels.

2.2.2 Modified Andrews Curve

Andrews curve or plot [15] provides a graphical comparison or visual observations of homogeneous groups. One of the disadvantages is that they are not able to preserve order. For example, the shape of the curves will be completely different if we change the order of variables [12]. In the Andrews functions, variables are used as the coefficients of the trigonometric function. Hence, the statistical variation of the data is intermixed with the periodic variation of the sine and cosine waves and making the plots harder to interpret. Therefore, [12] proposed a new function as an alternative to an Andrews plot. Each variable is assigned as a coefficient to a sine term as well as a cosine term. This approach is used in the present study, and is discussed in detail below.

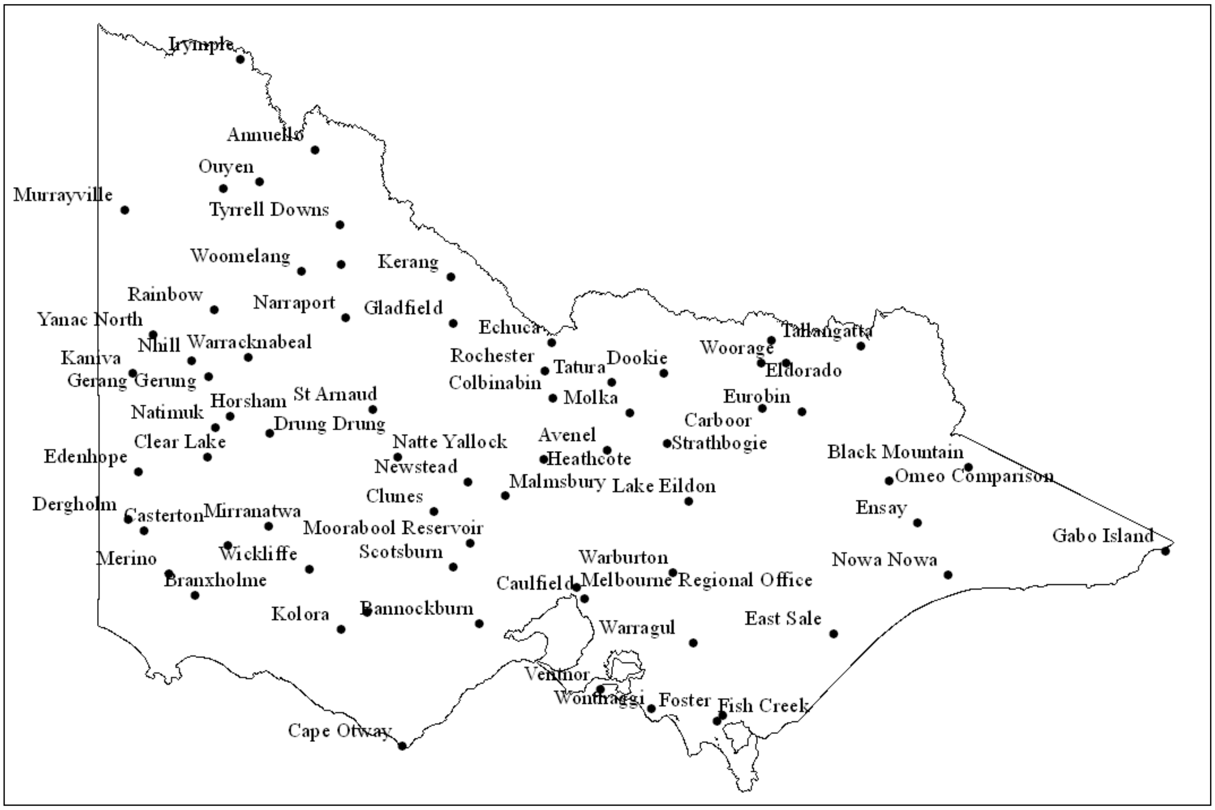


Fig. 2 Meteorological stations selected for the study

A point in a moving three-dimensional image is presented via the function:

$$g_y(t) = \frac{1}{\sqrt{2}}\{y_1 + y_2\,(\sin(t) + \cos(t)) \quad (2)$$
$$+y_3\ (\sin(t) - \cos(t)) - y_4\,(\sin(2t) + \cos(2t))$$
$$+y_5\,(\sin(2t) - \cos(2t)) + \ldots\}$$

The variables $y_1, y_2, \ldots$represent each of the variables used to characterise the catchment and the function $g_y(t)$ is the full range of t values (-π to +π). In a modified Andrews plot, a series of points ($g_y(t)$ values) between -π to π for a catchment is drawn. Groups of similar catchments will perform as a band of closely spaced curves, otherwise it might be presumed to fit in to a different group [15].

3. Data Analysis

The first step in classifying groups is to undertake stepwise linear regression analysis. This analysis was carried out using SPSS Statistics 17 between SPI and nine selected variables (as listed in Section 2.1) to determine the important variables. Equation 3 illustrates the step-wise regression equation obtained from the independent and dependent variables. Of all nine variables, mean summer precipitation (M_{Su}) and mean spring precipitation (M_{Sp}) were identified as the important variables related to SPI.

$$SPI = -3.208 + 0.03M_{Su} - 0.014M_{Sp} \quad (3)$$

where M_{Su} is the mean summer precipitation and M_{Sp} is the mean spring precipitation. These two variables were then weighted by determining the magnitude of their β-coefficients which are the coefficients of the stepwise regression model based on standardised catchment characteristics [2, 16]. In the current study, M_{Su} appeared as the most important variable and followed by M_{Sp} with weights of 1.263 and -0.866, respectively.

A cluster analysis was carried out before applying an Andrews curve for catchment grouping. Minitab Statistical Software Version 16 was used to perform the cluster analysis. As mentioned in the previous section, complete linkage and squared Euclidean distance were used to determine the number of homogeneous groups for 70 stations selected with similar drought characteristics. From this analysis, six cluster groupings with different drought patterns were formed with Cluster 1 (25 stations), Cluster 2 (24 stations), Cluster 3 (11 stations), Cluster 4 (4 stations), Cluster 5 (5 stations) and Cluster 6 (1 station), respectively.

In this study, the modified Andrews curve [12] for each station was obtained by applying the variables of each station in Equation 2.

$$g_y(t) = \frac{1}{\sqrt{2}}\{M_{Su} + M_{Sp}\,(\sin(t) + \cos(t))\} \quad (4)$$

where M_{Su} is the mean summer precipitation, and M_{Sp} is the mean spring precipitation. Appendix 1 depicts modified Andrews curves for all stations. Stations fitting in to a particular group appear as a band of closely spaced curves. Otherwise, the identified band is then removed from the plot and assumed to belong to another group. These plots clearly demonstrate the heterogeneous nature of catchments when all the 70 stations were considered. It is important to refine the cluster groups to ensure that all the curves in a cluster fall into a narrow band. In refining the cluster groups, outliers were removed from each cluster. An outlier is defined as a curve which is located at a far distance from the rest of the curves in a cluster. The curves of the outliers were compared with curves in other clusters to obtain a better fit, or a separate group of curves was combined to form a new cluster.

Mean modified Andrews curve of each cluster group was calculated by taking the mean of all the dependent catchments in the corresponding cluster. The mean group curve for each group is given in Fig. 4 and the equations are as follows:

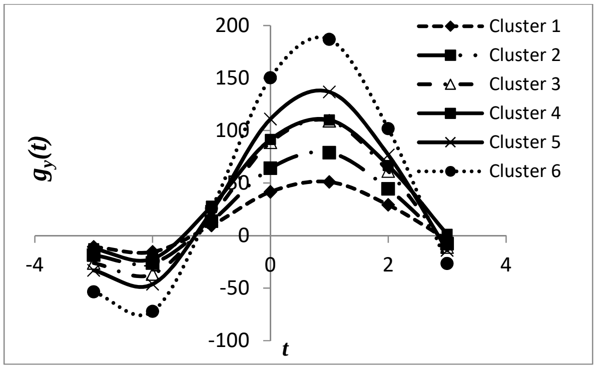


Fig. 4 Mean modified Andrews curves for different clusters

For Cluster 1:

$$g_y(t) = \frac{1}{\sqrt{2}}\{24 \; + 35\,(\sin(t) + \cos(t))\} \tag{5}$$

For Cluster 2:

$$g_y(t) = \frac{1}{\sqrt{2}}\{36 \; + 55\,(\sin(t) + \cos(t))\} \tag{6}$$

For Cluster 3:

$$g_y(t) = \frac{1}{\sqrt{2}}\{49 \; + 76\,(\sin(t) + \cos(t))\} \tag{7}$$

For Cluster 4:

$$g_y(t) = \frac{1}{\sqrt{2}}\{60 \; + 69\,(\sin(t) + \cos(t))\} \tag{8}$$

For Cluster 5:

$$g_y(t) = \frac{1}{\sqrt{2}}\{61 \; + 96\,(\sin(t) + \cos(t))\} \tag{9}$$

For Cluster 6:

$$g_y(t) = \frac{1}{\sqrt{2}}\{77 \; + 135\,(\sin(t) + \cos(t))\} \tag{10}$$

The mean group curve for each group could be used to identify the cluster of an unknown station. Though the allocation of membership can be determined by a visual comparison, an objective measure must be derived. In the current study, the sum of the squares of the differences was used:

$$SS = \Sigma\left(S_i - U_j\right)^2 \tag{11}$$

where (S_i) is the *ith* group signature and (U_j) is the unknown of the *jth* station under consideration. The station is assigned to the cluster with the lowest value of the differences.

The entire classification is mapped in Fig. 5 below. While developing the mean group curve, two stations from Clusters 1 and 2 and one station from Clusters 3 and 4 were kept as independent stations. There is no guideline to determine the number of independent stations. It is based on subjective matter depending on the quality of dataset. For Cluster 1, Gladfield and Kerang stations were kept as independent stations and for Cluster 2, Kolora and Dergholm were chosen.

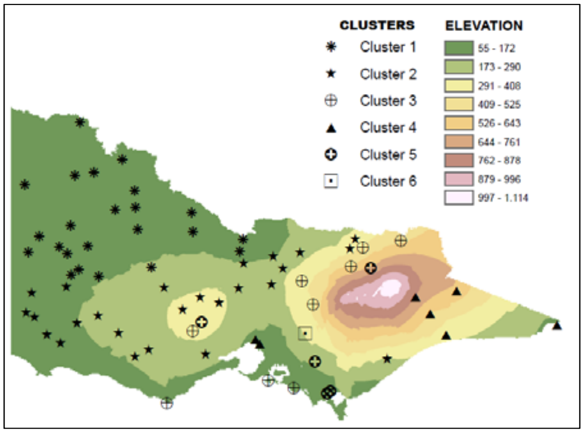


Fig. 5 Classification of stations based on regionalisation methods and elevations

3.1 Development of Severity-Duration-Frequency SDF Curves

Otto *et al.* [10] developed the drought severity-duration-frequency (SDF) curves to determine the relationship between the probability of a drought occurring with a certain severity and frequency at 10 locations. The development of the SDF curves were computed based on SPI (in a 12-month time scale) drought class boundaries. Instead of using SPI values, the moving cumulative precipitation thresholds were used to compute the severity of droughts. The Partial Duration Series (PDS) was used to analyse the time series. Minimum 12-month moving cumulative precipitation values and the drought durations (i.e. 1, 2, 3, 4 and 5 months) corresponding to return periods of 2, 5, 10 and 20 years, respectively, were developed. Using the same methodology, the SDF curves were then developed for the remaining 60 stations.

The mean SDF (*SDF_{mean}*) for each group were calculated. For example, in the case of Cluster 5, the SDF value of the mean group curve was calculated by taking the average of SDFs at Eurobin, Foster, Fish Creek, Warragul and Moorabool. The Log Pearson Type III (LPIII) distribution was used to develop *SDF_{mean}* group curve for each cluster. The signature or the mean SDF (*SDF_{mean}*) for each group was then computed to assist determination of the possibility of drought with a certain severity occurring in a catchment with a homogeneous signature in the future (see Appendix 2).

3.1.1 The accuracy of the developed mean SDF (SDF_{Mean}) of the cluster

To check the accuracy of the developed SDF_{Mean} for each cluster, the percentage error between the values from the SDF of individual stations were compared with the corresponding values obtained from SDF_{Mean} for a particular homogenous cluster. The percentage error was calculated using Equation 12:

$$\% \, Error = \left(\frac{SDF_{station} - SDF_{Mean}}{SDF_{station}}\right)100 \qquad (12)$$

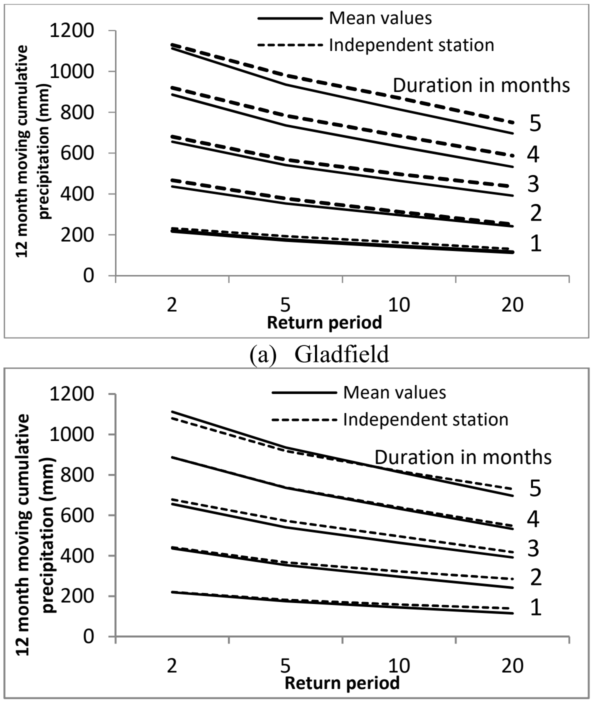
where SDF_{Mean} is the recurrence interval SDF value from the mean SDF of the cluster and $SDF_{station}$ is the recurrence interval SDF value from the dependent and independent stations

In Australia, Chiew et al. [17] carried out a study to calibrate and verify the rainfall-runoff models using streamflow data over 183 catchments. The verification results show that the errors in the mean annual runoff are less than 20 percent in more than half the catchments and therefore taken 20% as the limit. Boughton and Chiew [18] in their study to develop linear regression equations for the estimation of average annual runoff on ungauged catchments estimated that two-thirds of the estimates of average annual runoff were within ±25% of the actual value. Hence, they have taken the percentage error band as 25%. In the current study, most of the values at each station (87%) showed the errors were less than 20%. Therefore, the threshold value of ±20% error was selected as acceptable for this analysis.

The percentage error values obtained from the dependent and independent stations for all clusters and durations are summarised in Table 1. For instance, Cluster 1 has the highest number of stations assigned. 23 stations were taken to develop the SDF_{Mean} and two catchments were kept as independent catchments for verification. Fig. 6 illustrates the SDF values of independent stations (Gladfield and Kerang) and the SDF_{Mean} for Cluster 1. The solid lines in the figures indicate the SDF_{Mean} curves and the dotted lines show the curves for the independent stations. The percentage error values obtained between the recurrence interval from $SDF_{Independent\ station}$ and the recurrence interval from SDF_{Mean} are tabulated in Table 1.

Table 1 Number of times the percentage error (%) values were less and more than 20% for all clusters and durations

	Dependent stations			Independent stations		
Cluster	No. of stn.	Number of times		No. of stn.	Number of times	
		< ±20%	> ±20%		< ±20%	> ±20%
1	21	375	45	2	40	3
2	23	422	38	2	20	1
3	9	129	11	-	-	-
4	7	131	9	-	-	-
5	5	90	10	-	-	-



(a) Gladfield

(b) Kerang

Fig. 6 SDF curves of independent stations and SDF_{Mean} for Cluster 1

Based on the results in Table 1, most of the errors obtained for dependent and independent stations are within ±20%. For the independent stations, only 4 points out of 64 points (6%) show the error percentages greater than ±20%. It can be concluded that the SDF_{Mean} curves were successfully developed and can be used for long-term planning purposes, such as irrigation supply allocations.

3.2 The use of modified Andrews curves for identifying the cluster and the specific set of SDF curves

Homogeneous clusters and derived modified Andrews curves with their unique signatures provide an approach to the determination of drought characteristics. Hence, it is unnecessary to carry out a detailed drought characteristic analysis or develop new curves to determine the SDF relationship for any location of interest. To estimate the required drought characteristics of a certain location, a sufficient amount of rainfall data should be available from a neighbouring climatic station. In the present study, 70 stations across Victoria were selected and comprehensively cover the state of Victoria.

For a new station the group membership of which is unknown, a summary of the procedure to identify the homogeneous group and SDF curves is as follows:

(1) Determine the variables of mean seasonal precipitation (i.e. summer and spring).

(2) Plot the modified Andrews curve for the station using the equation below:

$$g_y(t) = \frac{1}{\sqrt{2}}\{M_{Su} + M_{Sp}\,(\sin(t) + \cos(t))\}$$

where M_{Su} is the mean summer precipitation and M_{Sp} is the mean spring precipitation.

(3) Match with the derived mean cluster modified Andrews curves.

(4) Select the SDF curve of the particular cluster.

4. Conclusions and Recommendations

The catchments used in the analysis from Victoria, Australia were separated into homogenous groups or clusters subject to similar climatic characteristics related to SPI. It enables the group membership to be identified for a new station. In the current study, stepwise linear regression analysis was carried out to select and weight the most appropriate variables. Mean seasonal precipitations (i.e. summer and spring) were found to be the most important parameters in clustering droughts. This finding is important especially when climatic data is very limited at a new station. However, it has to be tested in other regions for validations.

The following conclusions are based on separating the catchments into homogenous groups:

- Cluster analysis and modified Andrews curves were selected as the appropriate techniques to identify homogenous groups with respect to risk of drought occurrence based on rainfall characteristics. Both techniques showed good agreement in classifying the selected stations with six clusters were formed.
- In the current study, six clusters were identified as having similar drought characteristics. Also, mean SDF curves was developed for each cluster for various return periods.

Further analysis is carried out to develop ways of clustering at un-measured locations or those with limited data (i.e. regression analysis and spatial interpolation). The measurable characteristics of these stations will determine their best match with the existing cluster groups.

References

[1] Andrews, D. F. Plots of High-Dimensional Data. *Biometrics*, 28(1), (1972), 125-136.

[2] Khattree, R. and Naik, D. N. Andrews plots for multivariate data: some new suggestions and applications. *J. of Statistical Planning and Inference*, 100(2), (2002), 411-425.

[3] Kizza, M., Guerrero, J.-L., Rodhe, A., Xu, C. and Ntale, H. K. Modelling catchment inflows into Lake Victoria: Regionalisation of the parameters of a conceptual water balance model. *Hydrology Research*, 44(5), (2012), 789-808.

[4] Laaha, G. and Bloschl, G. A comparison of low flow regionalisation methods—catchment grouping. *J. of Hydrology*, 323(1–4), (2006), 193-214.

[5] Lyra, G. B., Oliveira-Júnior, J. F. and Zeri, M. Cluster analysis applied to the spatial and temporal variability of monthly rainfall in Alagoas state, Northeast of Brazil. *Int. J. of Climatology*, (2014).

[6] McKee, T. B., Doesken, N. J. and Kleist, J. The relationship of drought frequency and duration to time scales. In: Proc. 8th Conf. on Applied Climatol, 17-22 January, *Americ Meteorol Soc. Mass,* (1993), 179-184.

[7] Merz, R. and Bloschl, G. Regionalisation of catchment model parameters. *J. of Hydrology*, 287(1–4), (2004), 95-123.

[8] Nathan, R. J. and McMahon, T. A. Identification of homogeneous regions for the purposes of regionalisation. *J. of Hydrology*, 121(1–4), (1990), 217-238.

[9] Nazahiyah, R., Niranjali, J. and Bhuiyan, M. Development of Drought Severity-Duration-Frequency (SDF) Curves in Victoria, Australia. *Australian Journal of Water Resources*, (2015).

[10] Otto, M.. Multivariate Methods: In: R. Kellner, J. M. Mermet, M. Otto and H. M. Widmer, Eds., Analytical Chemistry. Wiley-VCH,Weinheim, (1998).

[11] Parajka, J., Merz, R. and Blöschl, G. A comparison of regionalisation methods for catchment model parameters. *Hydrol. Earth Syst. Sci.*, 9(3), (2005), 157-171.

[12] Ragno, G., Luca, M. D. and Ioele, G. An application of cluster analysis and multivariate classification methods to spring water monitoring data. *Microchemical Journal*, 87(2), (2007), 119-127.

[13] Raziei, T., Bordi, I. and Pereira, L. S. A precipitation-based regionalization for Western Iran and regional drought variability. *Hydrol. Earth Syst. Sci.*, 12(6), (2008), 1309-1321.

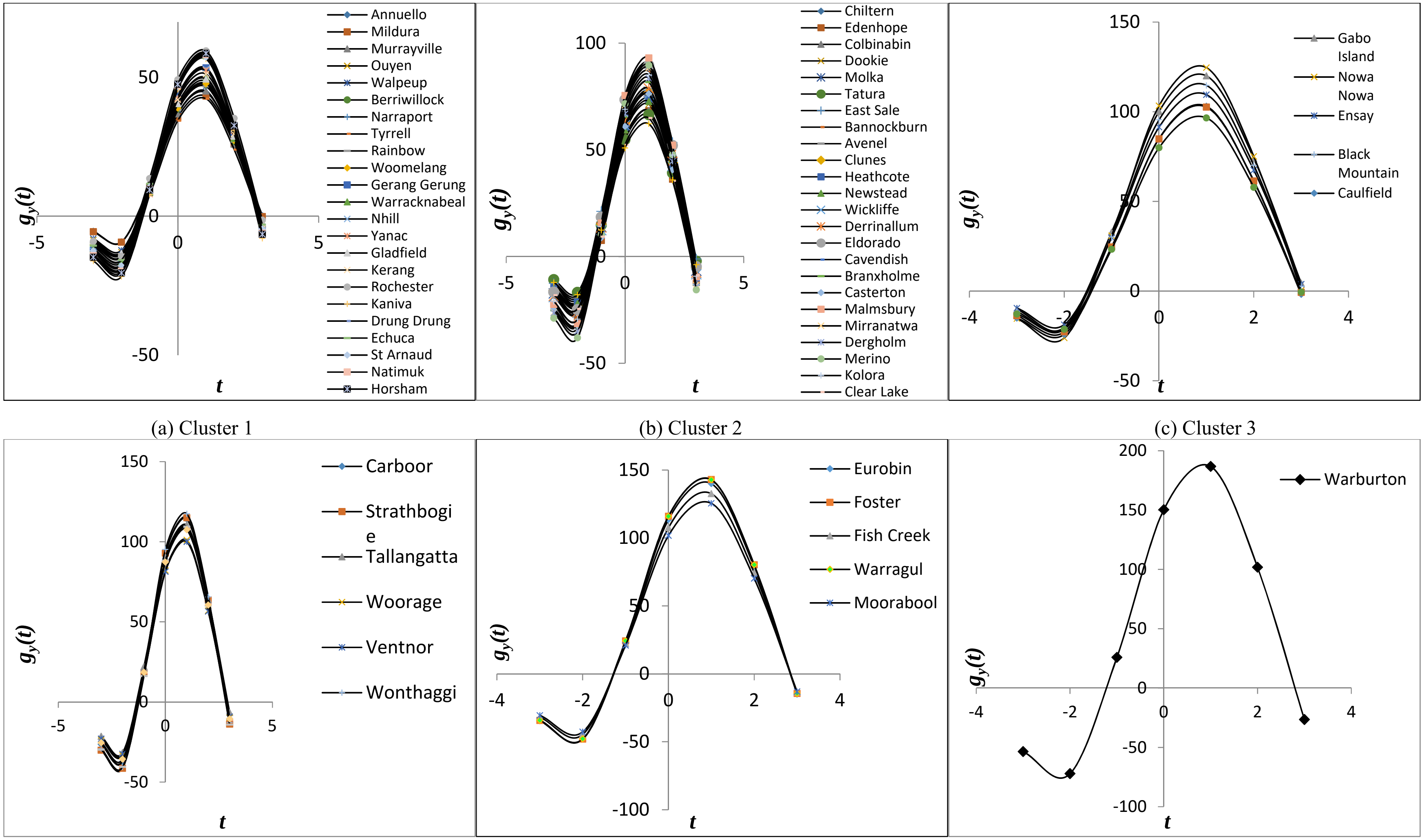
[14] Vezza, P., Comoglio, C., Rosso, M. and Viglione, A. Low Flows Regionalization in North-Western Italy. *Water Resources Management*, 24(14), (2010), 4049-4074.

[15] Yusof, F., Hui-Mean, F., Suhaila, J., Yusop, Z. and Ching-Yee, K. Rainfall characterisation by application of standardised precipitation index (SPI) in Peninsular Malaysia. *Theoretical and Applied Climatology*, 115(3-4), (2014), 503-516.

[16] Zhao, J., Fu, G., Lei, K. and Li, Y. Multivariate analysis of surface water quality in the Three Gorges area of China and implications for water management. *J. of Environmental Sciences*, 23(9), (2011), 1460-1471

[17] Chiew, F., Vaze, J., Viney, N., Jordan, P., Perraud, J.-M., Zhang, L., Teng, J., Young, W., Penaarancibia, J., Morden, R., Freebairn, A., Austin, J., Hill, P., Wiesenfeld, C. and Murphy, R. Rainfall-runoff modelling across the Murray-Darling Basin. A report to the Australian Government from the CSIRO *Murray-Darling Basin Sustainable Yields Project*. CSIRO, Australia: (2008), 66.

[18] Boughton, W. and Chiew, F. (2008). Estimating runoff inungauged catchments from rainfall, PET and the AWBM model Griffith University, Brisbane, Australia.



(a) Cluster 1

(b) Cluster 2

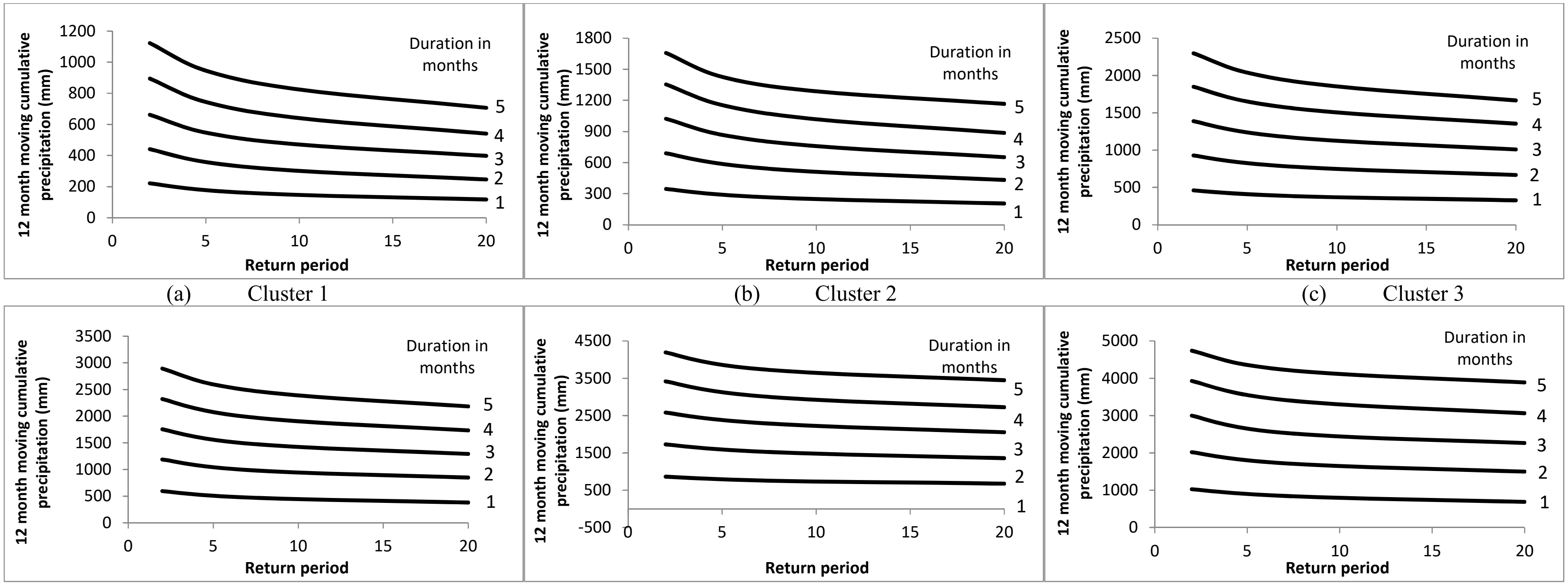
(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

(f) Cluster 6

Appendix 1 Modified Andrews curves for all clusters after refinement



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

(f) Cluster 6

Appendix 2 The 12-month moving cumulative precipitation (mm) plot for each cluster