Hydrologic Regionalization of River Basins of Kerala State

Fathima Hesna M S, Reeba Thomas
Civil Engineering Department,
Government Engineering College Trichur
Thrissur, Kerala
mshesna@gmail.com,reebakurien@yahoo.com

Abstract— **Hydrologic regionalization can be defined as the process of systematically arranging streams or catchments into groups that are most similar with respect to the characteristics of their flow parameter. It can be employed for importing data from gauged sites to ungauged regions where gauging stations may be absent due to topographical or physiographical reasons. A number of techniques have been developed for hydrologic regionalization including artificial neural networks, clustering algorithms, fuzzy and neuro-fuzzy analysis. This study presents an approach that combines two regionalization techniques in order to obtain homogeneous regions and compute streamflow at desired ungauged sites. First, an agglomerative hierarchical clustering algorithm is used to determine the number of homogeneous regions and to define the homogeneous regions. Second, the flow duration curve (FDC) method is used to compute streamflow discharge at desired ungauged sites in homogeneous regions. The combination of this approach is then applied to 44 river basins of Kerala state having 82 river gauging stations. From these sub-basins homogeneous regions are identified and streamflow discharge can be calculated at ungauged sites. It is found that these 82 river gauging stations of Kerala can be grouped into six homogeneous regions. FDCs, constructed for each homogeneous region, fit to monthly discharge data with correlation coefficients between 97 and 100% for all the clusters. Statistical tests are conducted to affirm the same. The Specific Water Yield map of the entire Kerala state is obtained by classifying the entire area into six clusters on the basis of specific discharge in these regions.**

Keywords—regionalization; hierarchical clustering; watersheds; Kerala.

I. INTRODUCTION

The state of Kerala is blessed with two important monsoon seasons: namely South- West and North-East monsoon and have 44 rivers driving this water to the sea. The water resource is finite and this widely present resource has a higher degree of mismatch in both temporal and spatial aspects. These rivers are entirely monsoon-fed and many of them shrink into rivulets or dry up completely during summer. Hence it is clear that the availability of water and its management have a higher significance in the state. Watershed Management is a framework to integrate this natural resource, its management addresses the issues of degradation of natural resources, soil erosion, landslides, floods, frequent droughts and desertification, low agricultural productivity, poor water quantity and quality and so on. The watershed management and water harvesting schemes for their accomplishment require the hydrologic characteristics of the entire state. The strategy of hydrologic classification is to ascribe watersheds to groupings or classes, so as to maximize the similarity between the members of each group and minimize the similarity between groups. Classifications process has frequently been applied by hydrologists seeking to extend insights gained from well-gauged regions to ungauged or sparsely gauged regions or rivers. i.e. this classification ultimately leads to regionalization which can be employed for importing data from gauged sites to ungauged sites.

Sivakumar et al. [1] reviewed several attempts on hydrological classification. Hall and Minns [2] did their classification of data sets using variety of modern informatic tools, such as artificial neural networks and fuzzy set. Olden et al. [3] gives a ecological and methodological approaches towards the hydrological classification. Ley et al. [4] in their study define hydrological classification by high similarities of their indices and analysis is done using Self- Organizing Maps (SOM). Patil and Stieglitz [5] suggested that for catchments under variable flow conditions, flow duration curve method (FDC) can be used effectively for the Hydrologic similarity analysis and the method is less sensitive to the exceptional years of flood or drought in the record. Isik and Singh [6] put forward an approach that combines agglomerative hierarchical clustering algorithm, k-means partitioning method Flow duration curve (FDC) method to compute streamflow discharge at desired ungauged sites in homogeneous regions thus resulting regionalization. Hansen and DeLattre[7] specified that Ward's algorithm gives a better result in hierarchical clustering as other linkages may not be well separated and are ambiguous.

II. STUDY AREA AND DATA DESCRIPTION

For this study, the entire state of Kerala, which occupies the southernmost part of India is selected. Kerala having an area of 38863sq.km, lies between a latitude of 8^o 18'N and 12^o 48'N and a longitude of 74^o52'E and 77^o22'E. Kerala's climate is mainly wet and maritime tropical, heavily influenced by the seasonal heavy rains brought mainly by the North-East and South-West monsoon. Kerala is blessed with 44 rivers including major and minor ones. All of these rivers arise from the Western Ghats and 41 flows towards the west to join the Arabian sea while three rivers mainly, Kabani, Bhavani and Pambar flows toward east to join the Bay of Bengal.

For the study, the main parameter required is the specific discharge of all the gauging stations based on which the classification is done. Stream gauging stations in Kerala are maintained mainly by the Hydrology Department and Central Water Commission. The daily discharge data are collected for a period of 30-40 years depending on the availability of data. The daily discharge is summed up to obtain the monthly discharge and averaged over the period of availability of data to obtain the monthly average discharge value for each stream gauging station selected as the outlet point. Though there are more than 125 gauging stations in Kerala, the discharge values for only 82 stations could only be obtained. The DEM is obtained from the Kerala Forest Research Institute (KFRI), Peechi, Kerala.

III. METHODOLOGY

By the principle of classification, it is possible to identify the hydrologically similar regions of an area. For some of these areas, parameters which are unknown are obtained from the similar basins for which data is available which is otherwise called as regionalization. And for the estimation of these parameters at unguaged sites in an area, Singh et al. [8] developed a regression equation of power form for gauged watersheds as in "(1)".

$$Q = kW_1^{a1}W_2^{a2} \ldots W_n^{an} \quad (1)$$

where Q -streamflow discharge; $W_1, W_2, \ldots, W_n$ - watershed and climatic characteristics; k, $a_1, a_2, \ldots, a_n$ - empirical coefficients. But in almost all practical situations the drainage areas place the foremost role. Also the other parameters are not easily available and hence "(1)" is modified as "(2)" considering the drainage area alone.

$$Q = kA^a \quad (2)$$

Murdock and Gulliver [9] used the "(3)" for the estimation of streamflow at an ungauged site which can be extrapolated from streamflow at a gauged site located within the same watershed.

$$Q_u = (A/A_u)Q \quad (3)$$

where Q_u-estimated streamflow at an ungauged site; Q-streamflow at a gauged site; A_u-watershed area at the ungauged site; A-watershed area at the gauged site.

To obtain the specific discharge at each of the gauging stations, the drainage area corresponding to the station are to be obtained. The drainage area are delineated using the 'Hydrology Tool' of ARCGIS 10.1. On obtaining the daily discharge data over a period of 30- 40 years for each of the 82 gauging station, the monthly average for the stations are found using a code in MATLAB 7.8. Thus the average monthly discharge value for each station(Q) and its corresponding drainage area(A) are derived and the specific discharge(q) for each station for each month are found using "(4)".

$$q = Q/A \quad (4)$$

The data matrix is obtained by arranging the monthly specific discharge values in 12 rows and gauging stations in 82 columns and thus a data matrix of 82 x 12 is prepared. Standardization of the features prior to analysis ensures that undue weights are not attributed to the features with highest absolute numbers. Hence in the study, the standardization is done by converting the data in to standard normal distribution using "(5)".

$$Z_{i,j} = (X_{i,j} - X_{mean})/\sigma \quad (5)$$

Thus it is possible to convert a group of variables having any mean and standard deviation to random variables that have a standard normal distribution with mean zero and standard deviation 1.

In order to obtain homogeneous regions, the agglomerative hierarchical clustering algorithm is used. The flow duration curve method can be used to compute streamflow at desired ungauged sites.

IV. CLUSTER ANALYSIS

Hierarchical clustering method is a procedure for transforming a proximity matrix into a sequence of nested partitions and is classified into two: agglomerative and divisive algorithm. An agglomerative algorithm for hierarchical clustering starts with disjoint clustering, which places each of the N objects in an individual cluster. The divisive algorithm does the reverse. As per the literatures Ward's algorithm for hierarchical clustering performs better.

The Ward's algorithm is as follows: Assuming that there are N elements to cluster, begin with N clusters consisting exactly of one entity per cluster. Search the similarity matrix for the most similar pair of clusters. Reduce the number of clusters by one through merging the most similar pairs of clusters. Perform these steps until all clusters are merged. At

each stage the objective is to find those two clusters whose merger gives the minimum increase in the total 'within group error' sum of squares. 'Within group error' sum of squares is given as in "(6)" and "(7)".

$$W = \sum_{K=1}^{P} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(X_{kij} - \acute{X}kj\right)^2 \quad (6)$$

$$\acute{X}_{kJ} = \frac{1}{N} \sum_{i=1}^{N} X_{kij} \quad (7)$$

where W- total 'within group error' sum of squares; P - the number of clusters; N - the number of gauging stations in each cluster; M - the number of months; X_{kij}- j^{th} month at the i^{th} gauging station in the k^{th} cluster; $\bar{X}_{kj}$- average value of the i^{th} elements at the j^{th} variable in the k^{th} cluster.

The Ward's Algorithm is executed using MATLAB 7.8. In the program Euclidean distance between the variables which is used as a similarity measure is found out. Based on which they are clustered and represented as the cluster diagrams called dendrograms. The Euclidean distance can be defined as in "(8)".

$$d_{a,b} = \sqrt{\sum_j \left(X_{aj} - X_{bj}\right)^2} \quad (8)$$

where $d_{a,b}$=Euclidean distance between X_{aj} and X_{bj} over the available data points.

The dendrogram does not provide cluster assignments by itself and hence the number of clusters to be formed must be interpreted by the user. In addition to the visual assessment using the dendrogram, a statistical factor called as Root Mean Square Standard Deviation (RMSSTD) is used. RMSSTD is the mean square root of the variance of each variable of the new cluster formed by merging two clusters and summed over all variables and is a constant value for each number of clusters to be formed. RMSSTD is represented mathematically as given in "(8)".

$$RMSSTD = \frac{1}{M} \sqrt{\sum_{K=1}^{P} \sum_{i=1}^{N} \sum_{j=1}^{M} \left(X_{kij} - \acute{X}kj\right)^2} \quad (8)$$

where the notations are as same as in "(6)". In order to yield the best result we look for a significant increase of the RMSSTD value and the procedure should stop before the last step of increase. i.e the optimum number of cluster is the one corresponding to a jerk in the plot of RMSSTD vs Number of Clusters.

V. FLOW DURATION CURVE

The flow-duration curve is a cumulative frequency curve that shows the percent of time specific discharges are equaled or exceeded during a given period. The percentage exceedance is found using the formula as in "(9)".

$$p_i = 1 - P(Qi <= q) \text{ or } p_i = P(Qi > q) \quad (9)$$

where p_i- i/(i+1)=exceedance probability; Q_i - equivalent to plotting the ordered observations q_i.

The FDC is drawn using the 'curve fitting tool' of MATLAB 7.8 using the average of monthly specific discharge data and corresponding percentage of exceedance as inputs. The curves are fitted with a confidence interval of 95%. To represent each of the clusters a representative equation is derived. This is done using regression analysis.

VI. ANALYSIS AND RESULTS

For the regionalization procedure, the stream gauging stations selected are shown in as in "Fig. 1" and the corresponding watersheds are shown in "Fig.2". From the daily streamflow data, average monthly data for each of the 82 stations are calculated and the streamflow matrix is converted into specific streamflow matrix using "(4)". The corresponding area for calculation is obtained using ARCGIS 10.1. Thus a specific streamflow data matrix of size 82 x12 is generated and is standardized using "(5)". The mean and standard deviation for the data matrix and standardized data matrix is as in "Table 1".

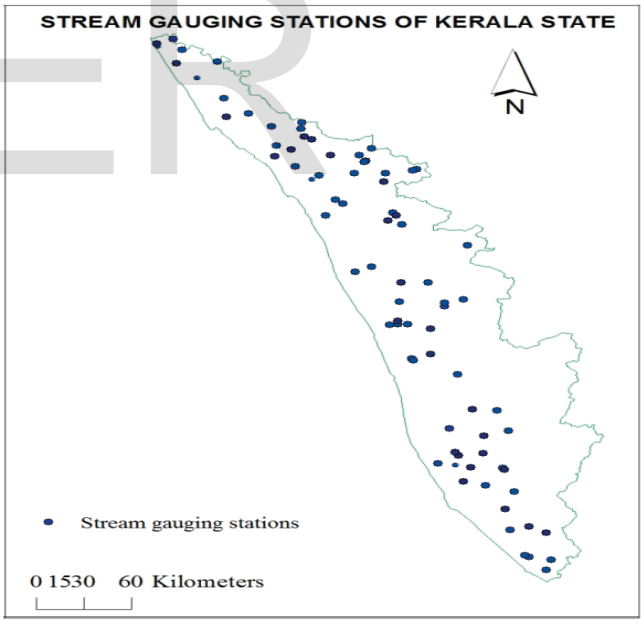

Fig.1 Location map of Stream gauging stations of Kerala.

TABLE 1. Mean and standard deviation of data matrix and standardized data matrix

	Data Matrix	Standardized Data Matrix
Mean	0.1080	0
Standard Deviation	0.2825	1

Ward's method for hierarchical clustering is executed using Matlab using the standardized data matrix as the input. The output from the Ward's algorithm is in the form of a dendrogram, which give the different clusters formed by the gauging stations based on the specific discharge value. The clusters are visually interpreted by cutting the dendrogram horizontally so that the distance between each of the cluster is significant. From the visual interpretation of the dendrogram the optimum numbers of stations are obtained as six which gives the minimum 'within the group' sum of squares error. The 'within the group' sum of squares error is given as in "Fig. 4" and the optimum number of clusters is identified as six. The clustered map is as in "Fig.3" and is tabulated as in "Table 2".

TABLE 2. OUTPUT OF WARD'S ALGORITHM

Cluster Number	Number of Member Stations
1	15
2	32
3	2
4	4
5	27
6	2

To verify the optimum number of clusters, a statistical test, 'Root Mean Square Standard Deviation' as mentioned by Isik and Singh (2008) is conducted. The output is given in "Table 3". The same is plotted and shown in "Fig.5".

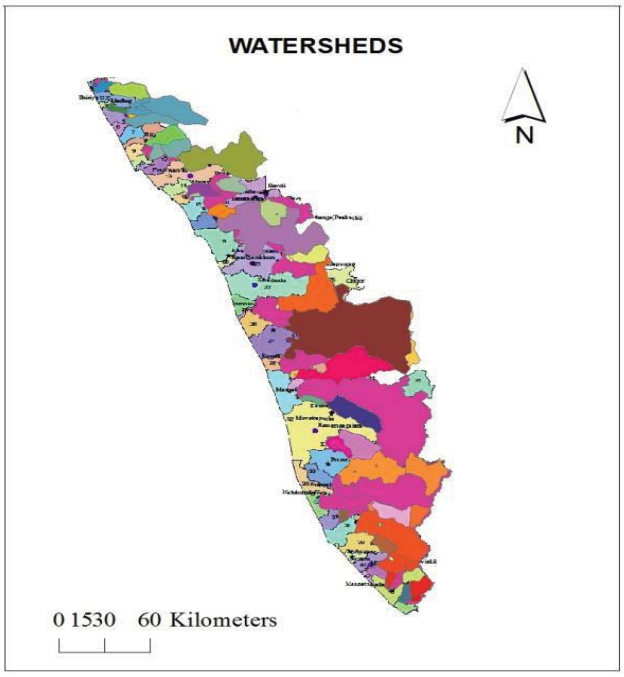

Fig.2 Delineated watersheds of Kerala

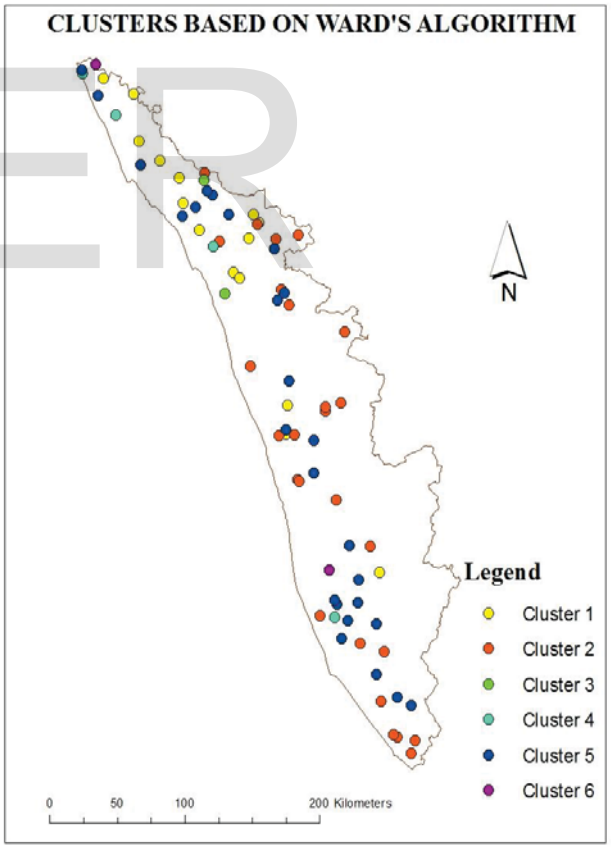

Fig.3 Output of Ward's algorithm

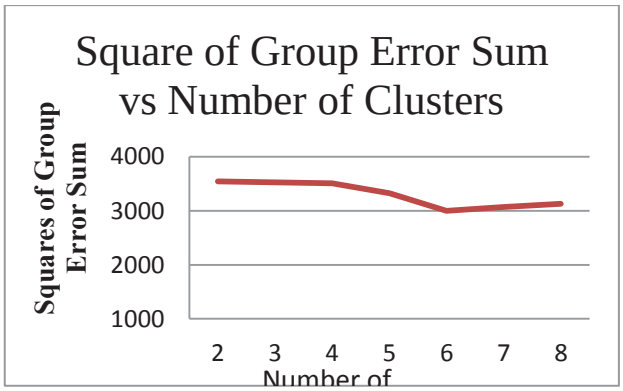

Fig.4 Plot showing within the group sum of square errors vs number of clusters

TABLE.3 RMSSTD values for each clusters

Number of Cluster	RMSSTD
2	4.962
3	4.9511
4	4.9379
5	4.8043
6	4.5633
7	4.6162
8	4.6614

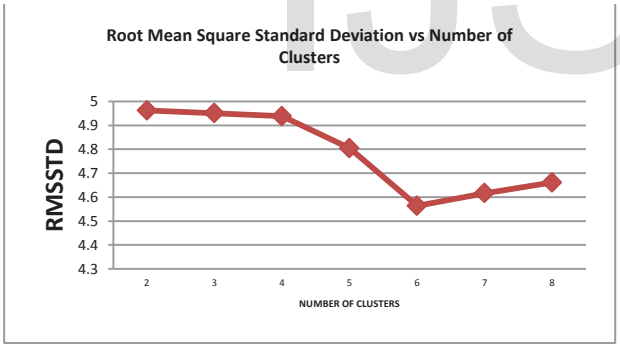

Fig. 5 Plot of RMSSTD vs Cluster number

The flow duration curve for each cluster is prepared for representing the flow duration curve of its individual members and is as given in "Fig.6"-"Fig.11". The best fit for the FDC is obtained from the 'Curve Fitting Toolbox'. The Regression equation, Coefficient of Regression(R^2), Sum of Square Error(SSE) and the Root Mean Square Standard Deviation (RMSE) for each clusters are obtained.

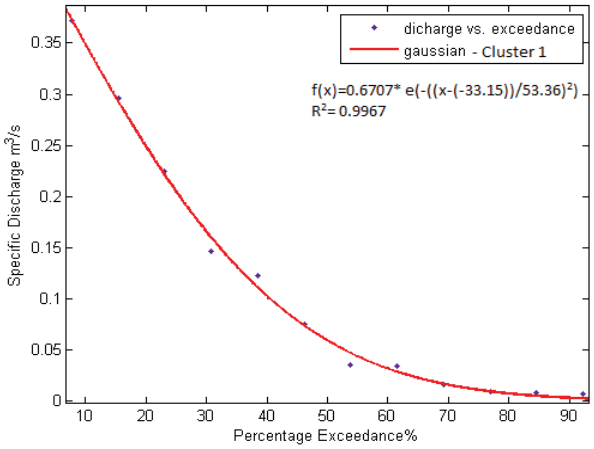

Fig. 6.FDC for cluster 1

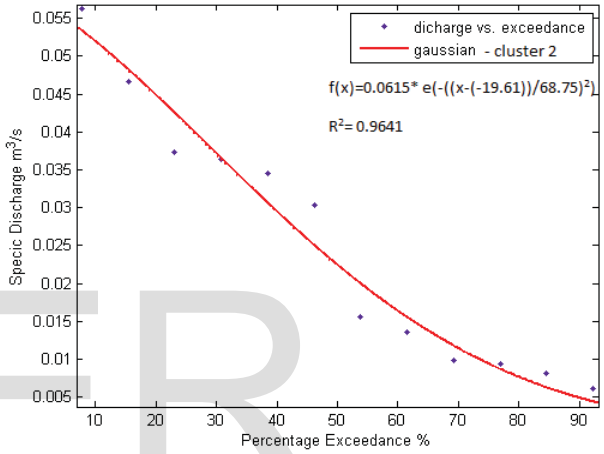

Fig. 7.FDC for cluster 2

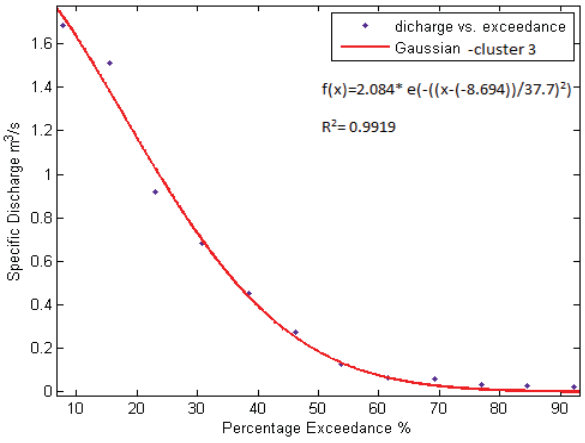

Fig. 8.FDC for cluster 3

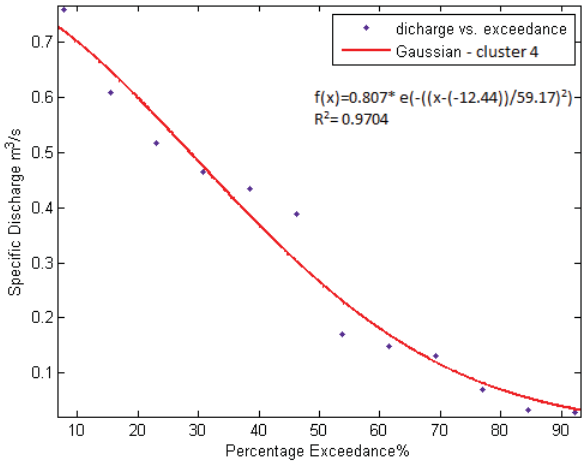

Fig. 9.FDC for cluster 4

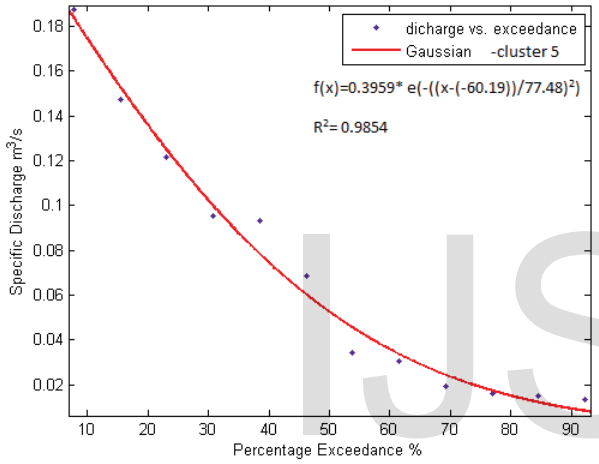

Fig. 10.FDC for cluster 5

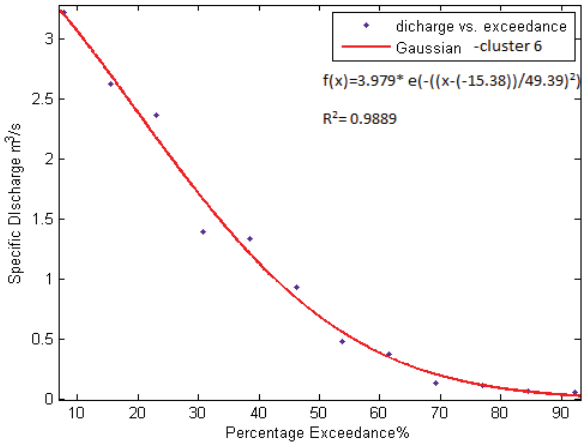

Fig. 11.FDC for cluster 6

From the analysis of each regression equation, it can be seen that all of the regression equations provide a best fit for the flow duration curves obtained for each cluster. A higher degree Gaussian equation was selected to obtain a lower RMS error and a higher value for regression coefficient. The coefficient of regression for each of the cluster range between 98% to 99.7% and thus providing a best fit. The sum of square errors(SSE) and the Root Mean Square Error(RMSE) are also in the permissible limit as they range between 10^{-5} and 10^{-1}.

On identification of the homogeneous regions all over Kerala state, a specific water yield map is prepared as in"Fig.12". The clustering using the Ward's algorithm provided a partial map of the state due to the absence of stream gauging stations towards the mouth of the rivers. Thus to obtain the regionalized map of the entire state, the clusters are extrapolated towards the coastal area on the assumption that the mouth of the rivers too belongs to the same cluster as that of the immediate upstream cluster.

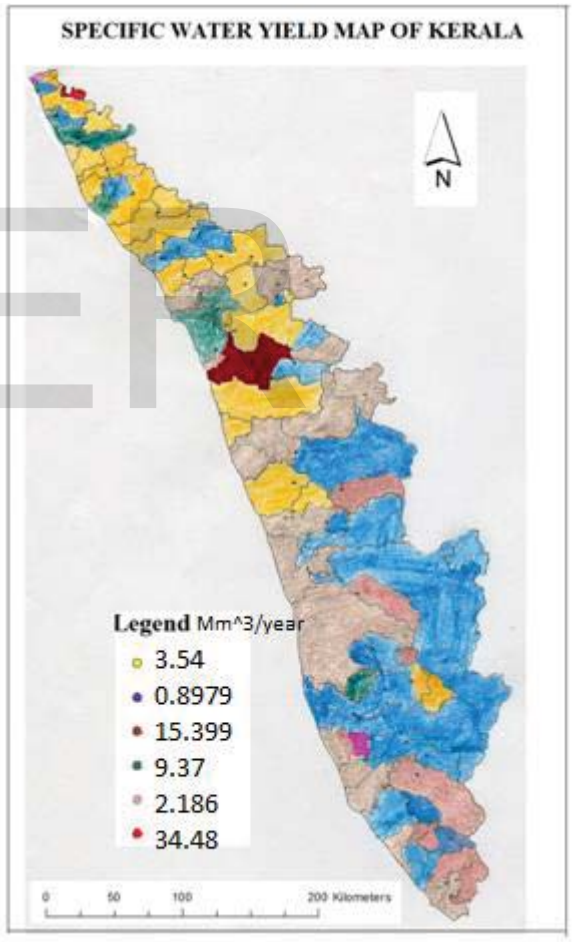

Fig. 12. Specific Water Yield Map of Kerala

VII. CONCLUSION

After conducting the study on hydrologic regionalization of river basins of Kerala state, several conclusions are drawn. The developed approach consists of two techniques in order to group homogeneous regions. First, an agglomerative hierarchical clustering is used to determine the number of homogeneous regions and later to cluster the gauging stations. Then the flow duration curve (FDC) is generated for each cluster as a representative one for the group members.

Using the Ward's algorithm the entire state is grouped in to six homogeneous regions based on the monthly specific discharge. These clusters are checked using several statistical tests and is proved accurate. For each of these clusters, the flow duration curves and regression equations with a confidence of 95% are developed. Further, the specific water yield map of the state is prepared cluster wise. Once thespecific water yield map is prepared it is possible to estimate the specific yield of any desired point of unknown discharge

References

[1] Sivakumar B. M., Vijay P. Singh, Ronny Berndtsson, Shakera K. Khan.: 'Catchment Classification Framework in Hydrology: Challenges and Directions', Journal of Hydrologic Engineering, April 2013

[2] Hall M. J. and Minns A. W. (1999): 'The classification of hydrologically homogeneous regions.' Hydrologic Science Journal., Vol. 44, pp.693–704.

[3] Olden J. D., Kennard M. J., and Pusey B. J. (2012): 'A framework for hydrologic classification with a review of methodologies and applications in ecohydrology.' Ecohydrology, Vol. 5, pp. 503–518.

[4] Ley R., Casper M. C., Hellebrand H., and Merz R.(2011): 'Catchment classification by runoff behaviour with self-organizing maps (SOM).' Hydrologic Earth System and Science, Vol. 15, pp. 2947–2962.

[5] Patil S. and Stieglitz M. (2011). 'Hydrologic similarity among catchments under variable flow conditions.' Hydrologic Earth System and Science, Vol. 15, pp. 989–997

[6] Isik S. and Singh V. P. (2008). 'Hydrologic regionalization of watersheds in Turkey.' Journal Hydrologic Engineering, ASCE, 13(9), 824–834.

[7] Hansen P and DeLattre M (1978): 'Complete Link Cluster Analysis by Graph Colouring', Journal of American Statistical Association, Vol. 73, pp. 397-403

[8] Singh R. D., Mishra S. K., and Chowdhary H.(2001):'Regional flow duration models for large number of ungauged Himalayan catchmentsfor planning microhydro projects.' Journal Hydrology Enineering,Vol. 6, pp. 310–316.

[9] Murdock R. U. and Gulliver J. S. (1993):"Prediction of river discharges at ungauged sites with analysis of uncertainty." Journal Water ResourcesPlanning Management, 119, 473–487