Estimation of Design Flow in Ungauged Basins by Regionalization

Yu, P.-S., H.-P. Tsai, S.-T. Chen and Y.-C. Wang

Department of Hydraulic and Ocean Engineering, National Cheng Kung University, Taiwan
E-mail: yups@ncku.edu.tw

Keywords: *Design flow, regionalization, ungauged basins, cluster analysis*

EXTENDED ABSTRACT

Estimating design flow in ungauged basins is a task frequently encountered in the design and planning of hydraulic and water resources engineering. Regionalization is a way to deal with this issue. In this study, a regional formula for peak flows was established using gauged flows and basin topographic characteristics in order to estimate the design flows in ungauged areas within the homogeneous region. At first, principal component and cluster analyses were used to classify the small sub-basins, delineated by this work, in Western Taiwan into three homogeneous regions. Due to the limitation of data availability and the region characteristics, only two homogeneous regions, Region 1 and Region 2, were used to build regional design flow formulae, but Region 3 is excluded from this study. The regional formula with respect to the area crossing Regions 1 and 2, named Crossing Region, and the formula regarding all regions, named Whole Region, were established to assess the performance of the proposed regionalization method. Crossing Region indicates that the catchment of a station in Region 1 encompasses both areas of Regions 1 and 2. Therefore, Region 1 means that a station in Region 1 encloses a catchment area only in Region 1. Whole Region is a combination of Regions 1 and 2.

Regional regression functions of design flow with respect to different return periods were developed, with topographic inputs of basin area, longest stream length, mean elevation and mean slope. Moreover, the coefficient of the regression function is found to have high correlation with return periods. Therefore, this study integrated these regression functions with different return periods into a single regional design flow formula for easy implementation by engineers. Statistical tests demonstrate that the design flows are significantly related to the topographic variables at 5% significance level, pertaining to formulae of Region 1 and Crossing Region. But the significant test on the formula of Region 2 shows insignificance, presumably due to insufficiency of data from gauged stations in Region 2. The mean relative errors of simulations pertaining to Region 1 and Crossing Region are lower than the regional formula with respect to Whole Region. The results of cross validation show that the regional formulae in Region 1 and Crossing Region have good applicability, and prove that the delineation of homogeneous regions can enhance the performance of regional formulae to estimate design flow.

1. INTRODUCTION

Estimation of design flow is necessary in many tasks associated with water resources management and engineering. As planning areas in water resources projects are often ungauged, regional analysis is used to solve this problem. The hydrologic regionalization technique is to infer required data in ungauged catchments from neighbour catchments where hydrologic data have been collected (e.g. Nathan and McMahon, 1990; Bullock and Andrews, 1997; Hall and Minns, 1999). A common assumption in the regional analysis is that catchments within a homogeneous region so that they may behave in a similar hydrologic fashion. To identify objectively homogeneous hydrologic regions, cluster and principal component analyses are commonly used (Mosley, 1981; Waylen and Woo, 1984; Gottschalk, 1985; Burn 1989). Then, multiple regression is applied to establish the relationship between design flows at ungauged sites and the catchment characteristics (Sanborn and Bledsoe, 2006).

This paper delineates homogeneous regions by cluster and principal component analyses so that all sub-basins in a cluster have similar hydrologic characteristics. The regional formulae for estimating design flows at ungauged site were then developed for each homogeneous region. This method was then applied at ungauged sites to estimate the design flow, and the cross validation is used to assess the performance of the proposed method.

2. STUDY AREA AND DATA SET

The western part of Taiwan, which is the most populous and economically developed, is selected as the study area. Gauged stations which have annual peak discharge records of at least 20 years are collected from Water Resources Agency and Taiwan Power Company. These stations with proper data for following research are distributed over 14 watersheds, with a total area of nearly 14,000 km2. Figure 1 shows the study area of 14 watersheds.

3. 3. REGIONALIZATION METHOD

3.1. 3.1. Defining sub-basins and clustering variables

Cluster analysis is used in this work to delineate hydrologically homogeneous regions. The delineation of homogeneous regions is based on a number of small sub-basins generated from digital elevation model (DEM) data referred to a properly set threshold value in this study. Geographic information system (GIS) software, ARC/INFO and ARCVIEW 3.2, were employed to generate sub-basins, and the topographic characteristics of each sub-basin, such as area, form factor, mean elevation and mean slope, can be derived using GIS tools. As a result, there are 965 sub-basins (Figure 2) in the study area to be grouped by cluster analysis.

Clustering variables should be suitably chosen before the cluster analysis is applied to group the sub-basins. The purpose is to define homogeneous regions solely based on topographic characteristics such that each region has similar peak flow response. Therefore the clustering variables were chosen as form factor, mean elevation and mean slope of sub-basins.

Figure 1. Study area

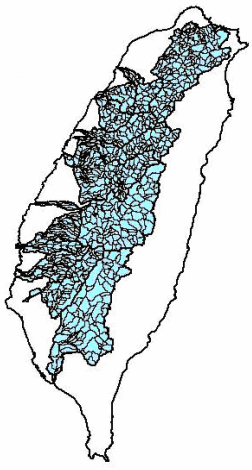

Figure 2. Location of sub-basins

3.2. Principle component analysis and cluster analysis

As those selected clustering variables may have significant correlation with each other, the principal component analysis is applied to derive the principal components, which are independent of each other, as the input variables of cluster analysis. Table 1 shows that the first two components comprise nearly 95% of the total variance of the original clustering variables. Therefore the first two components can be used as the input variables of cluster analysis. Table 2 shows the matrix of loading factor of two principal components, and indicates that the first principal component describes the geographical gradient and the second principal component explains the form of a sub-basin.

Cluster analysis was applied using those two principal components as clustering variables and the dendrogram is presented in Figure 3. Three homogeneous regions were selected so that a better grouping of sub-basins can be derived. Figure 4 shows that the location of homogeneous regions corresponds to upstream, midstream and downstream area, named Region 1, Region 2 and Region 3, respectively.

This study found that Region 3 is not suitable for the following analysis due to two reasons. First, Region 3 is in the downstream flat area, so the pre-process skill for filling depressions by increasing the elevation of lower grids in the DEM may alter the drainage network in flat areas. Second, Region 3 is much urbanized so that using only topographic characteristics to calculate peak flow is not appropriate in this region. Therefore only two homogeneous regions, Region 1 and Region 2, were used to build regional design flow formulae.

Table 1. Percentage of variance and cumulative variance

Principle component	Percentage of variance (%)	Percentage of cumulative variance (%)
1	62.2	62.2
2	32.4	94.6
3	5.4	100.0

Table 2. Loading factor of varimax

Original variables		Principal component Y_1	Y_2
F	Form factor	-0.058	0.998
E	Mean elevation	0.959	-0.034
S	Mean slope	0.956	-0.078

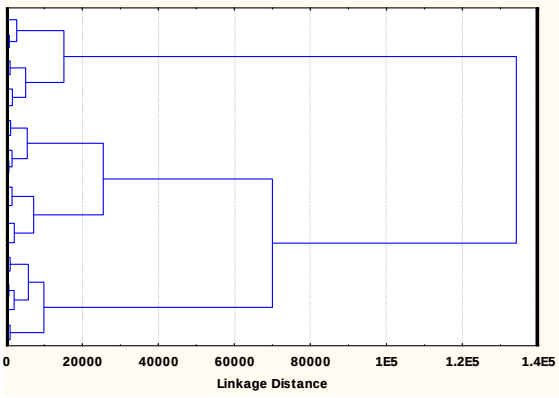

Figure 3. Dendrogram of cluster analysis

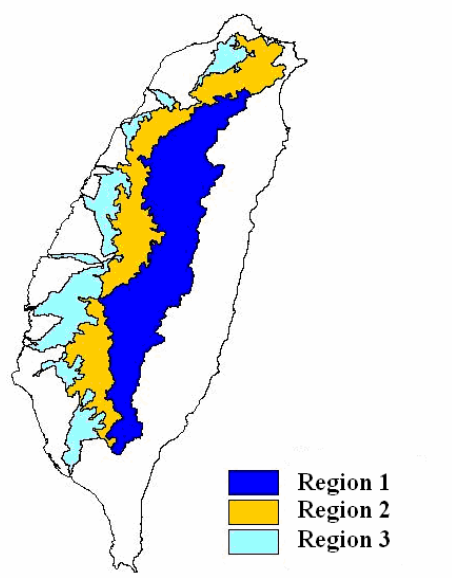

Figure 4. Homogeneous regions

The number of stations in Regions 1 is 16. The stations in Region 2 can be further divided into two cases. One is that the catchment of a station encloses an area only in Region 2, and the other is that the catchment of a station overlaps both Regions 1 and 2. Stations of the former case are in the same homogeneous area and the case is named Region 2, while stations of the latter case are not in the same homogeneous area and the case is termed Crossing Region. The numbers of stations in Region 2 and Crossing Region are 9 and 19 respectively.

4. REGIONAL DESIGN FLOW FORMULA

4.1. Frequency analysis

Peak discharge data of all 44 stations were fitted to probability distributions of Generalized Extreme Value, Extreme Value Type I, Pearson Type III Distribution, Log-Pearson Type III and Three-

Parameter Log-normal distributions. The Chi-squared test was used to examine the fitness of distributions, and the best fitted distribution was determined by the standard error (SE) criterion. Then the peak flows for different return periods can be estimated by the probability distribution function.

4.2. Regional formula

Regional peak flow formulae for different return periods (2, 5, 10, 20, 25, 50, 100 and 200 years) and for three regions (Region 1, Region 2 and Crossing Region) were built individually. The predictors are selected as basin area, main stream length, mean elevation and mean slope, considering the convenience of data acquirement and the consistence of variables used in cluster analysis. The form of the nonlinear regional formulae is in Eq. (1), which is determined by regression analysis.

$$Q_p = 10^{c_0} A^{c_1} L^{c_2} E^{c_3} S^{c_4} \quad (1)$$

where Q_p (m^3/s) is the peak flow; A (km^2) is the basin area; L (m) is the main stream length; E (m) is the mean elevation, and S is the mean slope.

The coefficients from multiple regression analysis are listed in Table 3. Coefficient of determination of Region 1 is greater than 0.9, while those of Region 2 and Crossing Region are respectively about 0.4 and 0.7. Significance test of regression functions demonstrated that the design flow and topographic variables of Region 1 and Crossing Region are statistically significant with a significance level of 5%, while regional regression functions in Region 2 are not statistically significant. The poor result in Region 2 is presumably due to insufficiency of data from gauged stations.

The established regional formulae were used to calculate the peak flow. The mean relative error was used to assess model performance.

$$RE(\%) = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|\hat{Q}_i - Q_i\right|}{Q_i} \times 100\% \quad (2)$$

where RE is the mean relative error in percentage, $\hat{Q}_i$ (m^3/s) is the estimated flow using the regional formulae; Q_i (m^3/s) is the design flow calculated from probability distribution function for the ith station, and n is the number of stations.

The mean relative errors in Region 1, Region 2 and Crossing Region are respectively 20%, 35% and 28%. Figure 5 shows the scatter diagram of the simulation results for return period of 50 years. The results for other return periods demonstrate similar results as 50-yr return period.

4.3. Cross validation of regional formula

The performance of the regional formulae is further assessed by the use of a cross validation method. where each station is in turn reserved as the validation station, while all the other stations are used to build the regional regression formula. Results of cross validation for return period of 50 years is shown in Figure 6. The mean relative errors of all return periods in Region 1, Region 2 and Crossing Region are respectively 37%, 123% and 46%. Figure 6 shows that some stations in Region 2 have large error so that the error of cross validation in Region 2 is not satisfactory.

4.4. Integration of the regional design formulae

From Table 3, it can be identified that the coefficients have good correlation with return period. Therefore these formulae of different return periods can be further integrated into a single regional flow formula, resorting to making the coefficient as a function of return period. The form of the integrated formula is expressed as

$$Q_p(T) = 10^{c_0(T)} A^{c_1(T)} L^{c_2(T)} E^{c_3(T)} S^{c_4(T)} \quad (3)$$

where T is the return period.

Results of the regression function of the coefficient as a function of return period are listed in Table 4. The values of coefficient of determination indicate the results of regression are very good. The mean relative errors of the single regional flow formula in Region 1, Region 2 and Crossing Region are respectively 22%, 35% and 30%, which are only slightly poorer than those estimated by original formulae.

4.5. Comparison with the regional formulae

In order to examine the performance of regionalization, this study also established a design flow formula regarding all stations, without clustering the stations. This regional formula with respect to the whole area is named Whole Region. Using this formula of Whole Region to estimate the peak flows in Region 1, Region 2 and Crossing Region results in the mean relative errors as 40%, 75% and 36%, respectively, which is much poorer

Table 3. Coefficients of regression function for different return periods

Region	coefficient	Return period							
		2	5	10	20	25	50	100	200
Region 1	C_0	12.45	12.54	12.83	13.20	13.30	13.72	14.17	14.62
	C_1	0.69	0.75	0.77	0.78	0.78	0.79	0.79	0.80
	C_2	0.33	0.23	0.18	0.14	0.13	0.10	0.08	0.05
	C_3	-2.15	-2.31	-2.38	-2.45	-2.46	-2.52	-2.58	-2.63
	C_4	-3.84	-3.17	-3.02	-2.97	-2.96	-2.99	-3.05	-3.12
	R^2	0.90	0.92	0.92	0.93	0.93	0.93	0.92	0.91
Region 2	C_0	0.81	1.12	1.35	1.56	1.62	1.79	1.95	2.06
	C_1	0.54	0.48	0.44	0.41	0.39	0.36	0.31	0.27
	C_2	-0.06	-0.08	-0.10	-0.10	-0.10	-0.10	-0.09	-0.07
	C_3	0.54	0.38	0.25	0.11	0.07	-0.06	-0.18	-0.30
	C_4	-0.29	0.14	0.41	0.66	0.73	0.95	1.17	1.38
	R^2	0.32	0.37	0.40	0.44	0.45	0.48	0.52	0.55
Crossing Region	C_0	-0.18	0.30	0.71	1.08	1.20	1.55	1.89	2.20
	C_1	0.96	0.91	0.88	0.84	0.83	0.80	0.76	0.73
	C_2	0.24	0.12	0.02	-0.07	-0.09	-0.18	-0.26	-0.34
	C_3	-2.65	-1.92	-1.43	-0.97	-0.83	-0.39	0.04	0.46
	C_4	5.33	4.12	3.25	2.42	2.17	1.38	0.61	-0.15
	R^2	0.73	0.70	0.70	0.71	0.71	0.72	0.73	0.74

Table 4. Regression results of coefficients as a function of return period

Region	coefficient	Regression function	R^2
Region 1	C_0	$0.066(\ln(T))^2 + 0.096 \ln(T) + 12.294$	0.99
	C_1	$-0.007(\ln(T))^2 + 0.064 \ln(T) + 0.656$	0.98
	C_2	$-0.009(\ln(T))^2 - 0.111 \ln(T) + 0.393$	0.99
	C_3	$-0.011(\ln(T))^2 - 0.165 \ln(T) - 2.051$	0.99
	C_4	$-0.098(\ln(T))^2 + 0.708 \ln(T) - 4.191$	0.93
Region 2	C_0	$0.023(\ln(T))^2 + 0.413\ln(T) + 0.527$	0.99
	C_1	$-0.058 \ln(T) + 0.577$	0.99
	C_2	$0.007(\ln(T))^2 - 0.043 \ln(T) - 0.034$	0.99
	C_3	$-0.185 \ln(T) + 0.671$	0.99
	C_4	$-0.021 (\ln(T))^2 + 0.482 \ln(T) - 0.6$	0.99
Crossing Region	C_0	$0.522\ln(T)-0.515$	0.99
	C_1	$-0.050\ln(T)+0.990$	0.99
	C_2	$-0.125\ln(T)+0.314$	0.99
	C_3	$0.670\ln(T)-3.022$	0.99
	C_4	$-1.182\ln(T)+6.030$	0.99

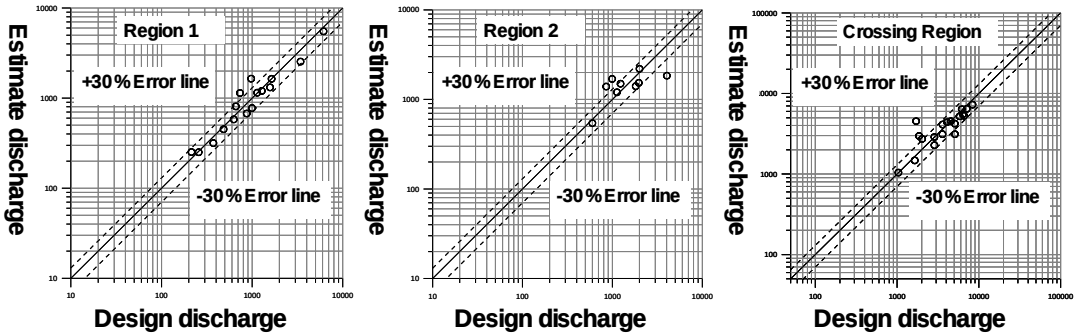

Figure 5. Simulation by regional formulae

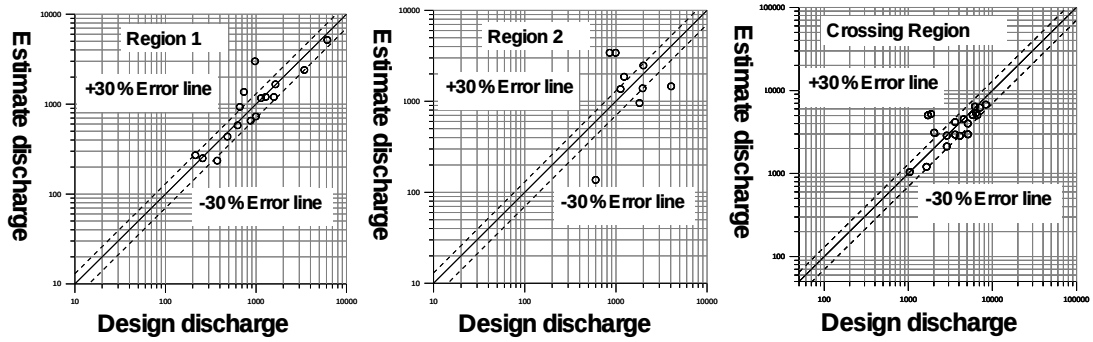

Figure 6. Results of cross validation

than those of 20%, 35% and 28% calculated by regional formulae. This shows that regionalization enhances the performance of design flow formula.

5. CONCLUSIONS

This paper established the regional formulae in a homogeneous region to offer design flow of ungauged areas. Many sub-basins were delineated and then principal component analysis and cluster analysis were applied to group these sub-basins into three homogeneous regions. Subsequently, regional formulae were constructed in Region 1, Region 2 and Crossing Region. This study found that the coefficients of regression function have high correlation with return periods, so that regional formulae of different return periods were integrated into a single regional formula. This kind of formula is practical and convenient for engineers to use.

Simulation and cross validation results demonstrate that regional formulae in Region 1 and Crossing Region show good capability to estimate design flow, but the regional formula in Region 2 show poor results presumably due to insufficient data in this region. Finally, a formula was built using all stations in the whole region. Comparison of the performance of these formulae demonstrates that regionalization does enhance the performance of regional formula.

6. REFERENCES

Bullock, A., Andrews, A.J. (1997) Southern African FRIEND – International collaboration in the hydrological science of flow regimes. *Sustainability of Water Resources under Increasing Uncertainty*, **240**, 133–143.

Burn, D.H. (1989) Cluster analysis as applied to regional flood frequency. *Journal of Water Resources Planning and Management*, **115(5)**, 567–582.

Gottschalk, L. (1985) Hydrological regionalization of Sweden. *Hydrological Sciences Journal*, **30**, 65–84.

Hall, M.J., Minns, A.W. (1999) The classification of hydrologically homogeneous regions. *Hydrological Sciences Journal*, **44(5)**, 693–704.

Mosley, M.P. (1981) Delimitation of New Zealand Hydrologic Regions. *Journal of Hydrology*, **49(1-2)**, 173–192.

Nathan, R.J., McMahon, T.A. (1990) Identification of homogeneous regions for the purpose of regionalization. *Journal of Hydrology*, **121**, 217-238.

Sanborn, S.C., Bledsoe. B.D. (2006) Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon. *Journal of Hydrology*, **325**, 65-84.

Waylen, P.R., Woo, M.K. (1984) Regionalization and prediction of floods in the Fraser River catchment. *Water Resources Bulletin*, **20**, 941-949.