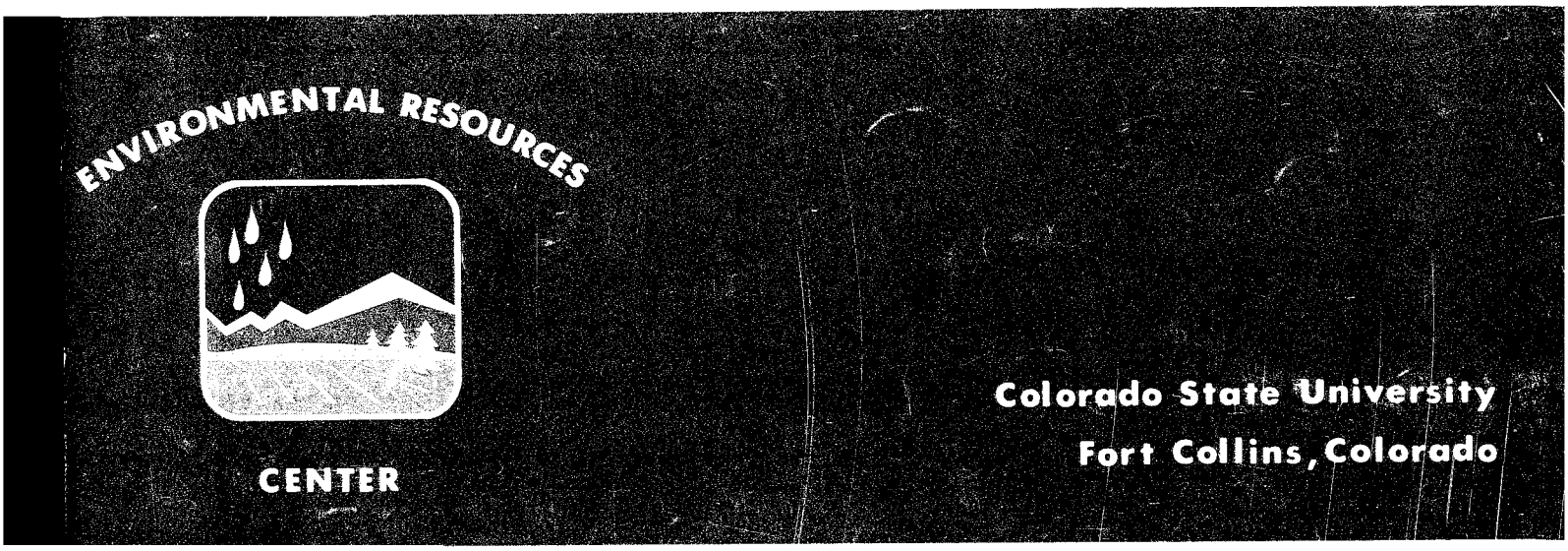
AN APPLICATION OF MULTI-VARIATE ANALYSIS IN HYDROLOGY

by
V. Yevjevich,
M. Dyhr-Nielsen and E. F. Schulz

August 1972



Completion Report Series
No. 35

AN APPLICATION OF MULTIVARIATE ANALYSIS IN HYDROLOGY

Completion Report

OWRR Project No. A-009-COLO
Period July 1, 1969 to June 30, 1972

by

V. Yevjevich, M. Dyhr-Nielsen and E. F. Schulz
Department of Civil Engineering
Colorado State University

Submitted to

Office of Water Resources Research
U. S. Department of Interior
Washington, D. C. 20240

August 31, 1972

The work upon which this report is based was supported by funds provided by the U. S. Department of Interior, Office of Water Resources Research Act of 1964; and pursuant to Grant Agreement Nos. 14-31-0001-3006, 14-31-0001-3206, 14-31-0001-3506.

Colorado Water Resources Research Institute
Colorado State University
Fort Collins, Colorado 80521

Norman A. Evans, Director

CER72-73VY-MD-EFS3

ABSTRACT

Multivariate analysis was used to make a selection of some of the more meaningful physical parameters dealing with the response of a small watershed to flood producing rainfall. Factor Analysis, Principal Component Analysis and a Correlation Coefficient Matrix was utilized. The list of 24 parameters was reduced to a list of 8 parameters. This reduction results in a very material economy in the encoding of relevant geomorphological data in flood analysis.

AN APPLICATION OF MULTIVARIATE ANALYSIS IN HYDROLOGY

Introduction

The watershed physiography leave an unmistakable influence on the size and timing of the flood response.

It was proposed to apply the principles of multivariate analysis to the task of selection or ordering of the various physical parameters being assembled and used in the CSU small watershed data file. The purpose of collecting high quality rainfall-runoff events was to obtain research data for use in research work in small watershed response to flood producing rainfall. At the present time 24 different measurements are made from a topographic or other maps of the watershed. These result in the computation of 40 different physiographic parameters. Many of these are redundant and the cost to quantify the hydrologic data file could be reduced if only the most meaningful parameters are selected for retention in the future.

Many of the physiographic parameters have been proposed by researchers in geology and geomorphology. As a better understanding of the basic hydrological laws evolved, some of the parameters proposed in the earlier research work were supplanted by newer more efficient or more efficient parameters. Thus, some of the parameters currently being evaluated in the small watershed flood program are remnants of an earlier obsolete concepts.

Examination of any one of the general schematic diagrams depicting the Hydrologic Cycle illustrates the complexity and interrelated nature of the elements of the hydrologic system. Figure 1 is a pictorial representation of the hydrologic cycle from Wilson (1969).

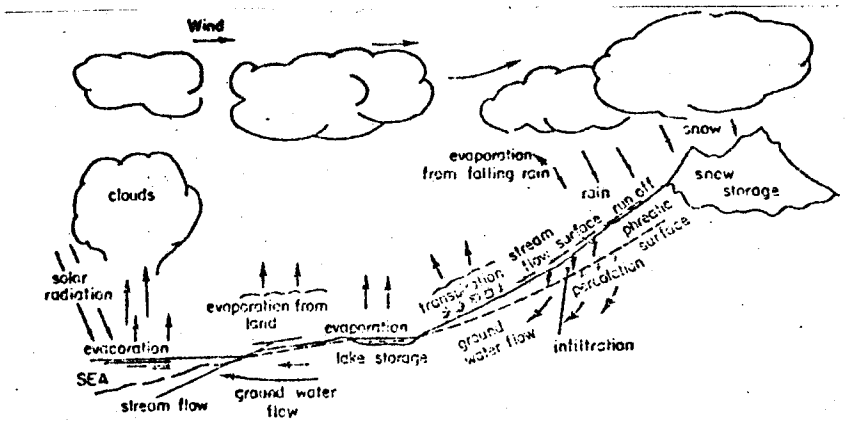


Figure 1. Hydrologic Cycle.

Moore and Claborn in a paper in Yevjevich (1971) show an organizational diagram of the hydrologic cycle which was prepared to outline the computer program for the University of Texas Watershed Model which is shown as Figure 2.

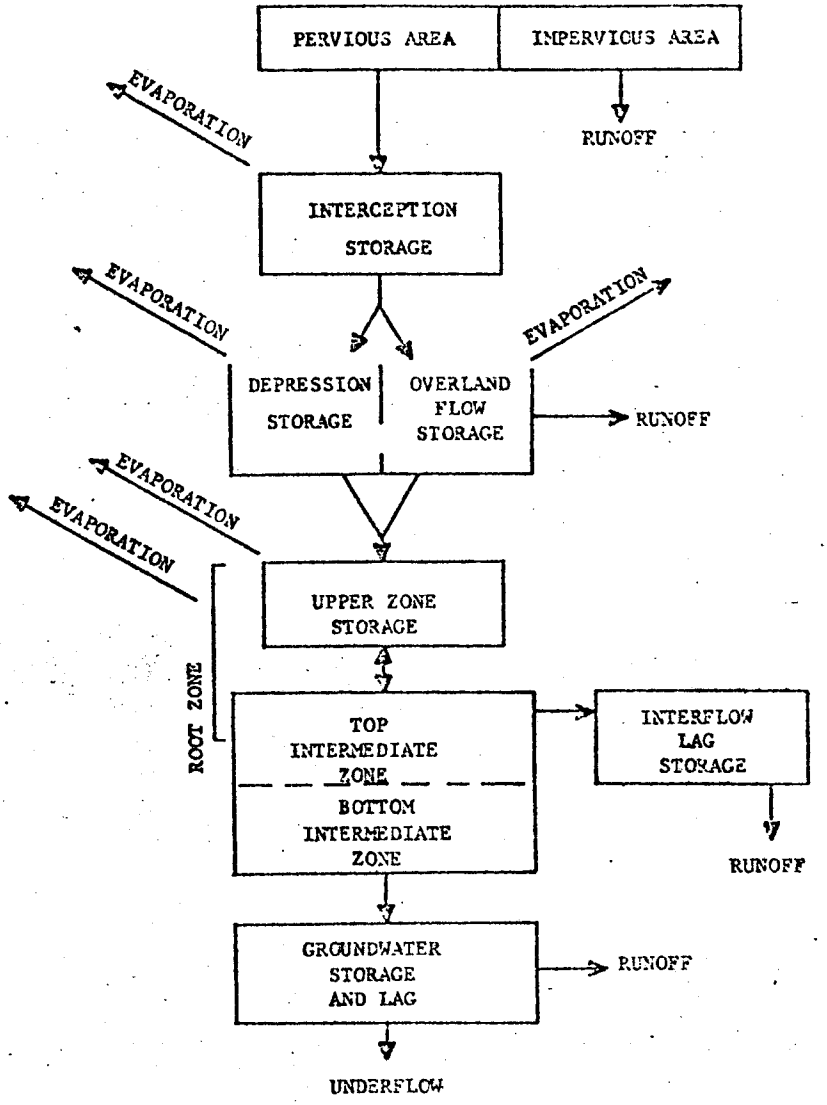


Figure 2. Schematic Diagram of University of Texas Watershed Model.

The interrelationship among the variables shown in Figure 2 shows that the relationship between watershed characteristics and the hydrologic response is complicated and almost impossible to attack the problem of evaluating the parameters in a strictly physical or deterministic framework. It has been popular to attempt to use the technique of multiple regression to study these relationships. In this investigation, it has been decided to use the technique of multivariate analysis. Dyhr-Nielsen (1971) applied the principles of multivariate analysis to the study of selection of the most meaningful physical parameters. The complexity of the hydrologic system makes it impossible to reduce the problem analysis to a completely deterministic form. However, Dyhr-Nielsen attempted to use known physical relationships about the watershed runoff process to guide the multivariate analysis. A second objective was to try to reduce the intercorrelation of the parameters or at least select those which were only weakly correlated.

Multivariate Methods Used

Various types of multivariate methods are well adapted to the problem of the interdependence of variables and the analysis of data obtained from interdependent data. These methods of multivariate analysis were applied to this problem:

1) Factor Analysis
2) Principal Component Analysis,
3) Correlation Coefficient Matrix

A linear additive model was assumed to represent the system. Some cases on non-linear response can be accommodated by employing a transformation and applying the linear theory to the transformed

variable. This technique will produce a linear transformed function of a power function. Many of the variables in hydrology seem to follow power functions.

Correlation Coefficient Matrix - While the correlation coefficient is a statistical parameter, it is often used to find the coefficients relating two deterministic variables. This is possible because the correlation coefficient is a measure of the linear dependence among two populations of variables. The correlation coefficient is defined as the dimensionless product moment or the ratio of the covariance of the two variables to the square root of the product of the two variances:

$$\rho(x,y) = \frac{\text{cov}\ (x,y)}{\sqrt{\text{var}(x)\ .\ \text{var}(y)}}$$

The variables x and y are linearly uncorrelated if $\rho = 0$. When $\rho = 1$ or $\rho = -1$, the variables x and y are perfectly correlated through a linear relationship and the variables are presumed to be deterministically related. A given value of x determines exactly the value of y. If the correlation coefficient, ρ, has values between 0 and 1, the correlation coefficient is a measure of the linear dependency because ρ^2 is the part of the total variance of a variable which can be explained through a linear relation to the other variable.

In general, ρ is not known exactly, but is estimated from the sample in which case the equation is rewritten:

$$r = \frac{s_{xy}}{s_x \cdot s_y} \quad , \text{ where}$$

$$s_{xy} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \quad ,$$

s_x and s_y are the sample values of the standard deviation.

It is of interest to determine when $\rho = 0$ and when $\rho = 1$ or -1. Since the sampling error presents an exact computation of the population correlation coefficient, all sample values within the tolerance limits around $\rho = 0$ are considered not significantly different from zero. Numerous tests for $\rho = 0$ have been developed. In this investigation the statistic

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$ has been employed.

Testing the linear dependence of two variables requires knowledge about the error introduced in measuring the dependent variable. Any measuring error is superimposed over the true relationships and obscures the ability to discern the true relationship. An exact evaluation of this error is not possible, but if the effect of measurement and sampling error is estimated some general areas of determination can be defined. For example, if the measurement errors produce uncertainty of the true value of the variable to the extent, the ratio of the unexplained variance to the total variance is 20%, sample correlation coefficients greater than 0.90 can be considered not significantly different from 1.

In this investigation a major objective was to reduce the number of variables being stored and investigated by a selection of the most statistically influential and physically logical variables. Any variable that can be functionally related to another variable is

considered superfluous and should be eliminated from the data set and replaced by its function. In the case of the *Correlation Coefficient Matrix*, the strategy will be to seek out those variables having correlation coefficients greater than 0.90.

In a situation where two variables are only highly correlated with each other, the selection is somewhat subjective. Other criteria for selection could be based on economy of data acquisition, measurement accuracy and reliability, physical relevance or hydrologic principles. In this type of situation, the correlation coefficient itself contains no information about which variable should be eliminated.

A different case exists when a set of more than two parameters are highly correlated. The parameter, which has the largest sum of squared correlation coefficients, is the one that explains the maximum amount of variance of the other parameters through functional relationships and should be retained.

In this investigation, several highly correlated parameters exist. The later criterion was used to select the superfluous variable.

Principal Components - The principal component technique has been developed to provide a simpler description of the variation of the variables. The description is framed in terms of linear combinations of the observed variables. The variables are mutually independent and obtained under the condition that the first component explains the greatest possible amount of the variance and covariance in the correlation matrix. The second component explains the maximum possible amount of the remaining variance and so on. The variance is concentrated upon the first component. This results in the reduction of the number of variables necessary to explain the variation of the complete set of

observations with a relatively modest loss of explained variance. The method was first developed by Hotelling (1933) and has been thoroughly discussed by Kendall (1957), Harman (1960) and Morrison (1967).

The first principal component is found by forming a linear combination of the observations:

$$\underline{Y_1} = \underline{a_1'}\ \underline{x}$$

with a variance

$$SY_1^2 = \underline{a_1'}\ \underline{\underline{R}}\ a_1$$

where

$\underline{\underline{R}}$ is the correlation coefficient matrix.

The variance, SY_1^2, is optimized under the constraint that the vectors are normalized so that

$$a_1 \cdot a_1' = 1 \quad .$$

By introducing the constraint as a Lagrange multiplier, λ_1, and differentiating with respect to $\underline{a_1}$;

$$\frac{\partial}{\partial \underline{a_i}} \{SY_1^2 + \lambda_1 (1 - a_1' \cdot a_1)\} =$$

$$2 \{\underline{\underline{R}} - \lambda_1 \underline{\underline{I}}\}\ \underline{a_1} \quad .$$

The optimum is achieved when the derivative is zero. Then the first principal component is the solution to the vector equation:

$$\underline{R} - \lambda_1 \underline{I} \quad \underline{a_i} = 0 \quad .$$

The solution to this equation is the eigenvalue where λ_1 is the eigenvalue and $\underline{a_1}$ is the corresponding eigenvector. To determine which of the eigenvalues should be selected, premultiply the equation above by a_i. Since $a_i\ a_1 = 1$, it follows that;

$$\lambda_1 = \underline{a_1'}\ \underline{R}\ a_1 = SY_1^2 \quad .$$

To maximize the variance, SY_1^2, the value of λ_1 must be selected as the largest eigenvalue of $\underline{R}$ and its corresponding eigenvector is the first principal component. At the same time the explained variance of the component is found to be equal to λ_1.

The second principal component is found by maximizing the variance of:

$$\underline{Y_2} = \underline{a_2'}\ \underline{x} \quad ,$$

subject to the constraints that:

$$a_2'\ a_2 = 1 \quad \text{and}$$

$$a_1'\ a_2 = 0 \quad .$$

This turns out to be the eigenvector corresponding to the second greatest eigenvalue of the correlation matrix where the eigenvalue

equals the explained variance. The remaining principal components are found in their turn from the other characteristic vectors.

A major problem in the use of principal component analysis is to determine which component to retain. Two criteria can be used for selecting the principal component:

1) Sum of the explained variance of all the retained components or,
2) Relative amount of the total variance the retained component explains.

The first criterion also has an additional benefit in reduced rank regression studies, where a trade off is made between the reduction in the number of variables and the corresponding decrease in explained variance.

The importance of a single component is of interest in any search for new significant parameters. The elimination of a variable should occur when the explained variance is less than a given (or assumed) critical value. In the case of standardized variables, the criterion is set at unity corresponding to the variance of one of the observed variables. Another criterion would be to eliminate all other variables after a significant decrease (say 50%) in the explained variance of the component occurs. This will give a group of the most important components, but it can only be applied when a significant change occurs when adding an additional variable. If the eigenvalues are decreasing without jumps, it will not provide any assistance in the selection of signficant variables.

The sample distributions for the eigenvalues have been developed for principal components drawn from normal multivariate variables

(Bartlett, 1950). Asymptotic expressions for the tolerance limits around the population eigenvalues have been found and tests of the equality of a subset of eigenvalues are present. These particular tests are not of much value in the present study.

The advantages of the principal components technique is that it develops a new set of mutually independent parameters that can be determined on the basis of the observed parameters. The number of principal components can be made smaller than the number of original parameters with a limited loss of accuracy. It has a sound mathematical background as it is developed as an optimization with constraints.

In order to apply the method of *Principal Components*, it is necessary that measurements of all of the variables be used. The method, therefore, is applicable only to measurements already available. Furthermore, it is difficult to attach any interpretation to the components. When the components are used in a regression analysis, the equations are transformed to the terms of the original variables. Snyder, (1962), has said that the regression equations based on principal components gives more meaningful results. However, this conclusion is entirely empirical.

Factor Analysis - In a factor analysis, the original variable is replaced by a new variable called the factor. It is assumed that the observations are linear functions of the common factors and that each variable is represented by a function of a number of unobservable common factor variates. The common factors generate the covariances among the observations, while the specific term contributes only to the variance of their particular responses.

A model employing factors can be presented as:

$$\underline{x} = \underline{\underline{A}} \cdot \underline{y} + \underline{\varepsilon} \quad ,$$

where

$\underline{x}$ is the variable,

$\underline{y}$ is the common factor,

$\underline{\varepsilon}$ is the specific factor variate,

$\underline{\underline{A}}$ is a coefficient matrix.

The coefficient matrix $\underline{\underline{A}}$ determines the linear relationship between the variable, $\underline{x}$, and the common factor $\underline{y}$. The coefficients in the coefficient matrix, $\underline{\underline{A}}$ are called factor loadings. It can be shown that the loadings are the covariance between a factor and the particular variable. Hence high loadings are an indication of high correlation between a factor and a variable. Suppose that the common factors, $\underline{y}$, are normally distributed, standardized, independent variables and that $\underline{\varepsilon}$ is equal to zero. Under these conditions, the covariance matrix $\underline{\underline{\Sigma}}$ of the observations can be generated by the loading matrix $\underline{\underline{A}}$ through the relation:

$$\underline{\underline{\Sigma}} = \underline{\underline{A}} \cdot \underline{\underline{A}}' \quad ,$$

where $\underline{\underline{A}}'$ is the transpose of $\underline{\underline{A}}$. This is a fundamental property of the loading matrix. The solution of this equation is not unique because if the loading matrix, $\underline{\underline{A}}$ is multiplied with an orthogonal matrix $\underline{\underline{T}}$:

$$\underline{\underline{A}}\ \underline{\underline{T}}\ (\underline{\underline{A}}\ \underline{\underline{T}})' = \underline{\underline{A}}\ \underline{\underline{T}}\ \underline{\underline{T}}'\ \underline{\underline{A}}' = \underline{\underline{A}} \cdot \underline{\underline{A}}'$$

and

$$\underline{\Sigma} = \underline{A} \cdot \underline{A} = \underline{A}\,\underline{T}\,(\underline{A}\,\underline{T})' \quad .$$

By choosing different orthogonal transformations, an infinite number of loading matrices are obtained all having the same covariance matrix.

The sum of the variances of the squared square loadings within each column of the factor matrix was proposed by Kaiser (1958) as a method of developing an evaluation criterion. By maximizing this criterion through orthogonal rotations of an initial factor matrix, a simple structured matrix can be found. Kaiser's criterion is stated mathematically:

$$v = \frac{1}{p} \sum_{j=1}^{m} [p \sum_{i=1}^{p} a_{ij}^{4} - (\sum_{i=1}^{p} a_{ij}^{2})^{2}] \quad ,$$

where a_{ij} = the loadings

p = a weighting factor.

Kaiser, 1958, called the criterion, v, the varimax criterion which is optimized during the selection procedure.

The initial values of the factor matrix used in the vector rotation are obtained from the coefficient matrix for the principal components. The coefficients, $\underline{a}_j$, in the jth component are scaled with the square root of the corresponding eigenvalue (which is the explained variance), $\sqrt{\lambda_j}$, to form a new vector, $\underline{a}_j \sqrt{\lambda_j}$, and a new matrix, $\underline{L}$. From this it follows that:

$$\underline{\Sigma} = \underline{L}\,\underline{L}' \quad .$$

This technique of employing the principal component provides a "factorization" of the correlation matrix, $\underline{\Sigma}$. The use of the

principal components provides a useful and convenient start for the varimax rotation. Otherwise, the Factor Analysis and Principal Component techniques are different.

The use of Factor Analysis in hydrology has been subject to considerable discussion and according to Yevjevich (1972), the use of the procedure has been the subject of some criticism. Several studies (Rice, 1967; Eiselstein, 1967; Lewis, 1968) have interpreted the loading matrix as coefficients on the observed variables and the factors as linear combinations of these. This concept appears to be more empirical than statistically rigorous. Wallis (1965a,b) has been one of the leading advocates of the application of factor analysis in hydrology. Matalas and Reiher (1967) subjected the application of factor analysis to hydrological problems to a critical review. In 1968, Wallis changed the name of the procedure to "Anti Factor Analysis".

In essence, the procedure is a stepwise rejection technique. A varimax rotated factor matrix is computed and for each factor only the variables that correspond to the two highest loadings greater than 0.90 are retained. On the basis of the remaining variables, a new varimax is computed. The low loaded variables are again removed. This continues until all variables are connected with high loadings.

The technique evidently functions because high loadings express a close correlation between factor and variable and therefore the variable can be used as a descriptor of the factor. This conclusion is based only on empirical results.

One of the objectives of this investigation is the study of various geomorphic parameters on the basin flood response to rainfall. The existing geomorphological parameters are highly interrelated. Hopefully

the antifactor technique of Wallis will aid selecting the most significant variables for retention.

Watershed Parameters

The data used in this investigation were collected in the Small Watershed Data Assembly Program at Colorado State University. The geomorphological parameters were obtained from a series of measurements from 7 1/2 minute quadrangle sheets of the U. S. Geological Survey or from similar scale detailed topographic maps of the Agricultural Research Service or U. S. Forest Service. The logic for the selection of the parameters computed from the topographic measurements and the procedure followed in obtaining the data were described in a report by Laurenson, Schulz and Yevjevich (1963) and latter revised by Yevjevich and Holland (1967).

The data used in this investigation were obtained from 188 small watersheds located over the entire United States and therefore represent a sample drawn from a very large range in climatic and geological conditions. A brief listing of the geomorphological parameters follows. The reader is referred to Yevjevich and Holland, (1967) for a more detailed description.

Area and Length Parameters

1. Watershed Area, A (square miles),
2. Watershed Perimeter, P (miles),
3. Main Stream Length, L (miles,
4. Total Length of Extended Streams, L_s (miles),
5. Channel Length to Center of Area, L_c (miles),

Stream Slope Parameters

6. Total Fall in Main Stream, H (feet),
7. Stream Slope, $S_1 = H/L$ (feet/mile),
8. Stream Slope, S_2 (feet/mile),
9. Stream Slope, S_3 (feet/mile),
10. Stream Slope, S_4 (feet/mile),

Overland Flow Length

11. Overland Slope, $R_1 = \frac{c\Sigma Lcon}{A}$ (feet/mile),
12. Overland Slope, R_2 (feet/mile),
13. Overland Slope, R_3 (feet/mile),
14. Overland Slope, R_4 (feet/mile),
15. Overland Slope, R_5 (feet/mile),
16. Relief Ratio, R_6 (feet/feet),

Basin Shape Parameters

17. Longest Dimension of Watershed, L_L (miles),
18. Average Width of Watershed, $W = A/L_L$ (miles),
19. Form Factor, $F = A/L_L^2$,
20. Compactness Coefficient, $C = .28\ P\sqrt{A}$,

Stream Network Shape Parameters

21. Average Travel Distance, L_t (miles)
22. Dimensionless Mean Travel Distance,

$$L_m = L_t/\sqrt{A}$$

23. Standard Deviation of Travel Distance,

$$S_d = S_t/\sqrt{A}$$

24. Drainage Density, $D_d = L_s/A$ (mile/sq. mi).

The mean values of these 24 watershed parameters for the 188 watersheds together with their standard deviation are given in the next table:

Table I

Mean Values and Standard Deviation
Watershed Parameters

Parameter	Mean	Standard Deviation	Parameter	Mean	Standard Deviation
A	6.07	9.02	H	12.27	22.14
P	9.65	8.00	s_1	312	376
L	3.98	3.37	s_2	236	318
L_s	13.87	18.24	s_3	274	685
L_c	2.02	1.77	s_4	266	361
R_1	1086	1009	D_d	4.32	4.69
R_2	1039	997	W	1.02	.819
R_3	1070	986	F	.334	.231
R_4	924	850	C	1.37	.310
R_5	833	812	L_t	2.20	1.86
R_6	.063	.079	L_m	1.09	.325
			s_d	.48	.18
			s_t	.98	.83

Results of Multivariate Analysis

The analysis of the interdependence between the watershed parameters was based on two mathematical models - a) Simple Linear Additive Model and b) Multiplicative Model based on a logarithmic transformation of the linear variables.

Correlation Coefficients - A correlation coefficient matrix, r_{ij}, has been computed both for the linear and for the log-transformed

parameters. Those correlation coefficients which were found to be signficantly different from zero are shown in Table II and III.

The test of significance was based on the statistic:

$$t = \frac{r^2\sqrt{N-2}}{\sqrt{1-r^2}}$$

where t follows the Student distribution with $N-2$ degrees of freedom. At a 5% level $t_{0.05} = \pm 1.96$,

and $r_{0.05} = \pm .142$

Comparison of the correlation coefficients for similar positions in the matrix in Tables II and III shows that in general the multiplicative model yields higher correlation than the linear model. To gain a better insight into the parameters, they were divided into groups defined as: *Length Parameters, Stream Slope Parameters and Overland Flow Slope Parameters.*

Length Parameters - The watershed area, A, and the length parameters, L, L_s L_c, P, W, were highly correlated. The explained variance, r^2, between the area, A, and the length, L_c, is lowest among all of the length parameters and L has the highest explained variance. The length of the main stream, L, is the length variable retained.

A strong inverse correlation was expected between the main stream length, L, and the stream slope parameters. This was not found to be the case. Evidently this expected relationship was obscured by the wide difference in geologic conditions present in the sample used herein. If a sample is obtained from a more homogeneous physiographic region, it is expected that the L and stream slopes would be more highly correlated.

Stream Slopes - For the linear model, only the stream slopes, S_1, S_2 and S_4 were found highly correlated to each other. In contrast, the log transformed stream slope parameters, all have correlation coefficients greater than 0.96 which means any slope parameter would explain more than 92% of variance of any other slope parameter. A criterion based upon the maximum sum of the coefficient of determination, $\Sigma\ r_1^2$ is used as a basis of selection

Slope Parameter	$\Sigma\ r^2$	
S_1	3.832	
S_2	3.872	Maximum value
S_3	3.815	
S_4	3.867	

The stream slope parameter, S_2 has the maximum value of the selection criterion, $\Sigma\ r^2$; however, the parameter S_4 has only a slightly smaller value. Because of ease in determination it was decided to select the parameter S_4. All the other slope parameters could be estimated from regression equations from S_4.

Overland Slopes - The overland slope parameters - R_1, R_2, R_3, R_4, R_5, R_6 - form a group of highly correlated parameters. The relief ratio, R_6, exhibits quite different correlation coefficients from the others and therefore will be retained. The selection criterion for the other parameters are,

Overland Slope Parameter	Σr^2 Linear Form	Σr^2 Log Transformed
R_1	4.739*	4.792*
R_2	4.669	4.749
R_3	4.735	4.786
R_4	4.580	4.727
R_5	4.594	4.675

* Selected for retention

Watershed Shape Parameters - The watershed shape parameters - F, C, L_m and s_d - are linearly independent or very weakly correlated with the stream and overland slope parameters and a selection can not be made.

Principal Components - The principal components of the correlation matrices shown in Tables II and III have been determined for both the original and the log transformed parameters. Only loadings greater than 0.100 are shown in the results.

The data in Table IV are the loadings of the components for each of the 24 watershed parameters. Also shown are the explained variances for each of the components. The first 12 components explain 98% of the variance. Dyhr-Nielson (1971) limited the interpretation to the first 4 components having contributed variance of a single component greater than 1.0. The first three principal components are associated with stream slope, overland slope, stream length and watershed slope characteristics.

Because there were several parameters representing each group of physically based parameters, a reduced set of parameters was selected on the basis of the parameter loadings within the components in Table IV. Table V gives the principal components of the reduced set of

Table II

Correlation Coefficients Between Geomorphic Parameters
Only coefficients significantly different from 0 are shown

	A	L	L_s	L_c	L_t	s_t	s_d	P	H	S_1	S_2	S_3	S_4	D_d	W	F	C	L_m	R_1	R_2	R_3	R_4	R_5	R_6
A	1.000	.873	.821	.792	.668	.831	.254	.935	.115	-.169	-.171		-.168	-.279	.887		.507	.388						
L		1.000	.786	.956	.986	.974	.465	.951	.171	-.163	-.167		-.165	-.322	.730	-.324	.454	.585						
L_s			1.000	.727	.794	.784	.362	.803		-.170	-.177		-.174	-.150	.761	-.151	.483	.473						
L_c				1.000	.969	.940	.484	.866	.168					-.303	.654	-.330	.407	.632	.148		.148			
L_t					1.000	.971	.445	.947	.175	-.151	-.156		-.157	-.323	.748	-.299	.443	.597						
s_t						1.000	.552	.921	.158	-.162	-.159		-.166	-.316	.702	-.333	.459	.605						
s_d							1.000	.334								-.495	.568	.809						
P								1.000		-.199	-.203	-.152	-.201	-.351	.853	-.209	.525	.469						
H									1.000	.662	.670	.574	.686			-.163		.143	.680	.683	.682	.627	.566	.272
S_1										1.000	.964	.672	.979		-.177				.836	.830	.837	.828	.802	.739
S_2											1.000	.644	.968		-.182				.791	.786	.792	.772	.746	.687
S_3												1.000	.713	.173					.512	.524	.513	.508	.472	.299
S_4													1.000		-.186				.793	.790	.795	.789	.757	.692
D_d														1.000	-.392									
W															1.000		.438	.237						
F																1.000	-.264	-.489	-.190	-.190	-.196	-.163	-.164	
C																	1.000	.600						
L_m																		1.000	.163	.153	.163			
R_1																			1.000	.983	.997	.941	.943	.678
R_2																				1.000	.982	.931	.934	.668
R_3																					1.000	.942	.944	.680
R_4																						1.000	.970	.688
R_5																							1.000	.714
R_6																								1.000

Table III

Correlation Coefficients Between Log Transf. Geomorphic Parameters
Only coefficients significantly different from 0 are shown.

	A	L	L_s	L_c	L_t	s_t	s_d	P	H	S_1	S_2	S_3	S_4	D_d	W	F	C	L_m	R_1	R_2	R_3	R_4	R_5	R_6
A	1.000	.959	.877	.915	.968	.944	.330	.977	.387	-.236	-.280	-.294	-.270	-.619	.908	-.161	.322	.414	.145					
L		1.000	.870	.963	.989	.985	.500	.962	.435	-.211	-.249	-.262	-.244	-.533	.789	-.382	.381	.574	.179		.173			
L_s			1.000	.815	.872	.851	.362	.864	.361	-.223	-.269	-.271	-.263	-.259	.758	-.238	.341	.453	.152		.150			
L_c				1.000	.961	.956	.518	.918	.443	-.170	-.208	-.228	-.203	-.535	.748	-.387	.346	.605	.199	.160	.196	.153		
L_t					1.000	.976	.468	.965	.434	-.205	-.246	-.261	-.241	-.549	.818	-.331	.362	.575	.185	.144	.182	.146		
s_t						1.000	.559	.949	.428	-.206	-.244	-.260	-.243	-.531	.774	-.386	.392	.580	.183		.175	.145		
s_d							1.000	.381	.267						.186	-.503	.499	.788	.150		.144			
P								1.000	.379	-.251	-.295	-.302	-.287	-.579	.863	-.229	.402	.463						.159
H									1.000	.763	.723	.713	.728	-.266	.288	-.260		.345	.886	.874	.862	.843	.812	.704
S_1										1.000	.976	.961	.978		-.230				.853	.872	.853	.835	.831	.696
S_2											1.000	.975	.986		-.269				.795	.821	.796	.774	.774	.674
S_3												1.000	.970		-.292				.790	.815	.791	.772	.768	.646
S_4													1.000		-.256		-.144		.802	.825	.802	.784	.781	.673
D_d														1.000	-.644									
W															1.000	.210	.212	.194						-.188
F																1.000	-.296	-.605	.178	-.163	-.188	-.159	-.152	-.153
C																	1.000	.550						
L_m																		1.000	.212	.188	.209	.167	.147	
R_1																			1.000	.984	.997	.962	.951	.804
R_2																				1.000	.983	.955	.950	.816
R_3																					1.000	.961	.950	.805
R_4																						1.000	.983	.790
R_5																							1.000	.791
R_6																								1.000

Table IV

Principal Components of Watershed Parameters

(Only first components are given. Only coefficients greater than 0.10 are shown)

Component	1	2	3	4	5	6	7	8	9	10	11	12
A		.311	-.199	.149			.114	-.198				
L		.377			-.178					-.166		
L_s		.294		.170	.129	.201				.771	.191	-.172
L_c		.325			-.232		.114	.179		-.263	.117	
L_t		.338			-.166					-.160		
s_t		.335			-.159			.127			-.271	-.127
s_d		.189	.507		.134	-.228		.289		.214	-.574	-.162
P		.332	-.148					-.124		-.196		
H	-.248			.335	-.321	-.175	-.290		.531		.150	-.406
S_1	-.322						.238		.135			.172
S_2	-.310					-.137	.263		.280		-.193	.281
S_3	-.220			.475	-.141	-.320	.149		-.713		.103	-.160
S_4	-.314			.135		-.148	.273		.149		-.109	.170
D_d		-.113	.293	.536		.712		.184		-.163		.109
W		.273	-.335		.246			-.175		.151		.262
F		-.116	-.454	.170	.411	-.182		.700		-.178		
C		.206	.261	.209	.625	-.135	-.221	-.340		-.310		-.221
L_{im}		.229	.432		.102	-.196	.111	.256			.593	.399
R_1	-.325						-.230					
R_2	-.324						-.237					
R_3	-.326						-.226					
R_4	-.319			-.114		.116	-.199		-.139		-.121	
R_5	-.314			-.173	.116	.151	-.200		-.203			
R_6	-.252			-.330	.202	.231	.572				.199	-.522
Variance	8.61	8.26	1.96	1.15	.82	.72	.54	.49	.35	.27	.20	.16
Var. %	35.9	34.4	8.2	4.8	3.4	3.0	2.3	2.1	1.5	1.1	0.9	0.7
Cum. Var. %	35.9	70.3	78.5	83.3	86.7	89.7	91.9	94.0	95.5	96.6	97.4	98.1

Table V

Principal Components for a Reduced Set of Parameters

Component No.	1	2	3	4	5	6	7	8	9	10
A	.406		-.278			-.160		.101		-.347
L	.423		-.111		-.117	-.223		-.178		-.424
L_s	.396		-.169	.142	.144	-.184	-.102	-.222	.160	.768
s_d	.285	.120	.456			.279		-.328	.106	
H		.407	-.129	.295	-.467			-.177	-.678	
S_3		.371	-.104	.420	-.401	.235	-.159	.171	.613	
S_4	-.112	.486	-.118		.311		.323		.133	
D_d	-.133		.296	.634	.455	-.385	-.228	-.132		-.175
W	.358		-.403					.142		
F	-.164	-.167	-.426	.182	.281	.591	-.106	-.496		-.129
C	.324		.195	.194	.262	.465		.600	-.274	
L_m	.331	.153	.372			.192		-.305		-.194
R_1		.489	-.133	-.180	.334		.356			
R_6		.373		-.422			-.804			
Variance:	4.64	3.09	1.86	1.04	.80	.66	.52	.43	.35	.19
Var. % :	33.1	22.1	13.3	7.4	5.7	4.7	3.7	3.1	2.5	1.4
Cum. Var. %:	33.1	55.3	68.6	76.1	81.7	86.5	90.2	93.3	95.8	97.2

The 4 last components are not shown
Only coefficients greater than 0.10 are shown.

parameters. As in Table IV, only coefficients greater than 0.10 are shown in Table V. The first 10 components accounted for 97% of the explained variance.

The principal components for the log transformed variables are shown in Table VI. The transformed variables had about 10% higher explained variance in the first three components than in Table IV. The first component is associated with slope of both stream and overland flow. The second component is associated with stream length, slope and watershed shape, while the third component is primarily associated with drainage density and with watershed shape. The loadings within the first three components were used to aid in the selection of parameters for the reduced set.

Table VII shows the principal components and the loadings for each parameter for the selected reduced set of parameters. Interpretation of the results is not conclusive. Intuitively, it might be said that the physical model for the relationships is multiplicative rather than additive. The log-transformed parameters explained more variance in the first three components.

Factor Analysis - The factor analysis was carried out in two parts - Factor Analysis and Anti-Factor Analysis. A varimax rotation of a set of factors found on the basis of the principal component solution was performed. The loadings for the original parameters are shown in Table VIII. Only the first eight factors have correlation coefficients with the parameters exceeding 0.70 ($r^2 = 0.49$).

Most of the slope parameters are highly correlated with the first common factor. It is obvious that the first common factor is most

Table VI

Principal Components of Log Transformed Variables

Component	1	2	3	4	5	6	7	8	9	10	11	12
A		-.316	.169								-.111	.107
L		-.322		.110						-.110		
L_s		-.288		.166	.394	.148	-.283		.100	.676	.100	-.354
L_c		-.311		.121	-.120					-.516	.202	-.647
L_t		-.322			-.192					-.127		.106
s_t		-.320						.209				
s_d		-.171	-.433	-.213		.538	.281	.479	.181	.118		
P		-.319	.107			-.103				-.134		.294
H	.300						-.239	.160	-.295			.145
S_1	.287	.136					-.156					
S_2	.274	.149			-.142		-.284	.109				
S_3	.271	.152			-.102		-.298		-.185		-.145	
S_4	.275	.148			-.139		-.270				-.109	-.100
D_d		.174	-.336	.259	.669	.293	-.222			-.397		.158
W		-.271	.337	-.186		.213					-.167	.336
F		.103	.448	-.476	.195	.490		-.169		-.166		-.166
C		-.148	-.316	-.720	.263	-.446	-.215		.139			
L_m		-.191	-.440	-.115	-.139	.232		-.706	-.342		-.121	
R_1	.300				.134		.213				.389	.122
R_2	.313				.118		.186				.330	.106
R_3	.313				.136		.210				.409	
R_4	.305				.175		.329				-.411	-.139
R_5	.301				.195		.366				-.482	-.206
R_6	.276				-.136		-.149	-.345	.811			.162
Variance	9.65	8.97	2.19	.78	.73	.45	.39	.20	.18	.08	.08	.05
Var. %	40.2	37.4	9.14	3.3	3.0	2.0	1.6	.8	.8	.3	.3	.2
Cum. Var. %	40.2	77.6	86.7	90.0	93.0	95.0	96.7	97.5	98.3	98.6	99.0	99.1

The last 12 components are not shown
Only coefficients greater than 0.10 are shown.

Table VII

Principal Components for a Reduced Set of Log-Transformed Parameters

Component No.	1	2	3	4	5	6	7	8	9	10
A	-.405	.135	-.203	.112						.222
L	-.426	.094		.190						.297
L_s	-.377	.112		.304	-.433			-.164	.123	
S_d	-.274		.401	-.196	.166	-.615	-.516	-.186		
H	-.255	-.411	-.128				-.163	.375	.346	
S_4		-.517	-.115	-.104				.126	.502	
D_d	.228		.372	.293	-.725	-.178				.389
W	-.335	.167	-.364	-.101	-.130	-.177				.388
F	.117	.107	-.465	-.480	-.311	-.466	.181			
C	-.228		.306	-.687	-.313	.511		-.115		
L_m	-.308		.402		.124	-.255	.741	.314		
R_1	-.155	-.470	-.109		-.147		-.136	.253	-.774	-.112
R_6		-.503					.301	-.771		.130
Variance :	5.02	3.42	2.11	.73	.65	.45	.19	.17	.10	.05
:	38.69	65.02	81.32	86.93	91.96	95.49	96.96	98.26	99.09	99.55
Var. % :	38.69	26.33	16.30	5.61	5.03	3.53	1.47	1.30	.83	.46

The 3 last components are not shown
Only coefficients greater than 0.10 are shown.

Table VIII

Varimax Rotated Factors of 24 Watershed Parameters

Component	1	2	3	4	5	6	7	8	9	10	11	12
A		.934										
L		.960										
L_s		.824								-.532		
L_c		.916										
L_t		.965										
s_t		.932										
s_d			.910									
P		.967										
H							.715					
S_1	.884								-.141			
S_2	.840											.429
S_3				.836								
S_4	.842			.278			.155					
D_d						-.955						
W		.831			.195	.161						
F								-.936				
C					.852							
L_m			.732					.201			.437	
R_1	.980											
R_2	.971											
R_3	.986											
R_4	.966											
R_5	.966											
R_6									-.664			

The last 12 factors are not shown
Only loadings greater than 0.10 are shown.

highly correlated with the overland slope. The watershed area and the stream length parameters are highly correlated with the second common factor.

The loadings for the log-transformed parameters are shown in Table IX. Comparing the loadings in Tables VIII and IX demonstrates the more favorable values for the log-transformed parameters. Not only are the correlation coefficients higher, but the values are more logically grouped by common factors.

The results of the factor analysis is that:

1) There are four watershed properties which are associated with the underlying factors a) *Slope* b) *Length of Streams* c) *Length of Overland Flow (Drainage Density)* and d) *Watershed Shape*,

2) These four properties appear to be independent,

3) These factors cannot be observed directly.

Anti-Factor Analysis - In Anti-Factor Analysis only those parameters having the highest loadings are retained. The list of 24 parameters shown in Table IX was reduced to 14 parameters. The varimax rotated factors for the reduced set of parameters is shown in Table X. Factor 1 had four parameters. The two parameters having the highest loadings were retained, thereby eliminating both stream length parameters - L, L_s. Table XI shows the retained 12 parameters and their loadings.

Anti-Factor Analysis was also completed on the log-transformed parameters. A reduced set of parameters were selected from the parameters and loadings shown in Table IX. The reduced list is shown in Table XII. As before the two parameters having the highest loadings

Table IX

Varimax Rotated Factors of 24 Log-Transformed Parameters

Component	1	2	3	4	5	6	7	8	9	10	11	12
A		.984										
L		.960										
L_s		.915			-.241					-.303		
L_c		.923										-.226
L_t		.968										
s_t		.942										
s_d						.898						
P		.968										
H	.866											
S_1	.954											
S_2	.921											
S_3	.912											
S_4	.925										-.179	
D_d					-.842							
W		.884										
F			-.936									
C				.929								
L_m			.342	.266		.574		.547				
R_1	.962											
R_2	.969											
R_3	.962											
R_4	.946						.261					
R_5	.940						.308					
R_6	.881							-.441				

The last 12 factors are not shown
Only loadings greater than 0.10 are shown.

were selected for Common Factor 1 and 2. This reduced the list of parameters retained to 8. The loadings for the reduced log-transformed parameters are shown in Table XIII.

Table X

Varimax Rotated Factors of Reduced Set of Parameters

Factor:	1	2	3	4	5	6	7	8	9	
A	.959									A
L	.842									L
L_s	.857									L_s
s_d			.932							s_d
H									.893	H
S_3					-.913					S_3
S_4		.920								S_4
D_d				.971						D_d
W	.913									W
F						-.942				F
C								.841		C
L_m			.840							L_m
R_1		.943								R_1
R_6							-.915			R_6

It is possible to form Principal Components of the 8 retained log-transformed parameters. The principal components of the 8 retained parameters is shown in Table XIV. Only 5 of the components would explain 93.1% of the variance. No interpretation of these components was attempted.

Table XI

Varimax Rotated Factors of 12 Parameters

Factor	1	2	3	4	5	6	7	8	9	
A				.941						A
s_d	.933									s_d
H								-.908		H
S_3					.931					S_3
S_4			.926							S_4
D_d						-.973				D_d
W				.936						W
F							-.943			F
C									.844	C
L_m	.864									L_m
R_1			.938							R_1
R_6		-.953								R_6

Only loadings greater than 0.50 are shown.

Table XII

Varimax Rotated Factors of Reduced Set of Log Transf. Variables

Factor	1	2	3	4	5	6	7	
A	.935							A
L	.897							L
L_s	.945							
s_d						.930		s_d
H		.884						H
S_4		.938						S_4
D_d					-.917			D_d
W	.881							W
F			-.940					F
C				.937				C
L_m						.636	.550	L_m
R_1		.944						R_1
R_6		.908						R_6

Only loadings greater than 0.50 are shown.

Table XIII

Varimax Rotated Factors of 8 Log Transformed Variables

Factor	1	2	3	4	5	6	
A	.855						A
L_s	.970						L_s
s_d						.914	s_d
S_4		-.940					S_4
D_d					-.972		D_d
F			-.956				F
C				.951			C
R_1		-.954					R_1

Only loadings greater than 0.50 are shown.

Table XIV

Principal Components of 8 Log Transformed Variables

Component	1	2	3	4	5	6	7	8
A	-.524	-.503	-.336	-.157				.756
L_s			-.145	-.392	-.440	.759	-.268	-.545
s_d	-.336	-.145	.389	.174	.255			
S_4	-.157	-.392	-.157			.122	-.686	.129
D_d		-.440	.568	-.331	-.610	.487		.322
F			-.480	.431	-.472	-.406		
C		-.268	.313	.698	-.333			
R_1	.756	-.545	-.192		-.166		.672	
Cum Var. %:	35.3	57.5	76.8	85.4	93.1	98.4	99.6	100.0

Only coefficients greater than 0.10 are shown.

Physical Foundations

The response of a watershed to flood producing rainfall is controlled by physical laws - of potential energy, of kinetic energy, of frictional resistance, of surface storage, of infiltration, of evaporation, of channel hydraulics. A complete analytical treatment based on physical laws is a hopelessly complex problem.

In view of the complexity of the physical problem, a complete analytical treatment seems improbable. Dyhr-Nielsen (1971) has classified three different approaches in the analysis of flood response of a natural watershed. The earliest analytical approach is sometimes called the "*black-box*" technique including the unit hydrograph concept of Sherman (1932), Snyder (1938), and Dooge (1959). These are conceptualizations and although they have some qualitative meaning in the physical world. These concepts are not basically derived from basic laws of physics. Their relationships to catchment characteristics are developed by statistical tools.

The second analytical approach is called the "*grey-box*" technique. In the grey box technique elements of the hydrologic cycle are derived from fundamental physical models, but many of the required input variables or input parameters are not usually measured or measureable. To make practical use of the valuable insight provided by the analysis, it is necessary to make use of "effective" parameters. Thus, there is a sounder evaluation of the purely empirical methods typical in the "*black-box*" technique. An example of this type of analysis is the method of computing the runoff hydrograph using the kinematic wave model described by Schaake (1971) where the runoff hydrograph is computed from a basic physical model, but infiltration from the storm rainfall is account for with rather arbitrary estimates.

The "*white box*" approach is based on rather complete mathematical or physical model or representation of the natural watershed. Because of the complex nature of the hydrologic cycle, elements or parts are considered in the development of the physical models. An example of a "*white box*" technique is the application of the kinematic wave theory to the computation of the surface runoff hydrograph beginning with a physical model based on the equations of flow within the sheet of surface detention.

Comparison between the parameters which we might expect from a kinematic wave application to the watershed with those final selected by the different types multivariate of multivariate analysis is given in Table XV.

Conclusions

Although the parameters being evaluated in the CSU Small Watershed Data File have evolved from geomorphology, they did not represent direct measurements of parameters derived from basic physical laws.

There is a correlation between some of the parameters and basic physical variables. Since the variables are interrelated some of the methods of multivariate analysis proved useful in the selection of one variable or parameter from a group of highly correlated parameters representing the same physical watershed property.

Only when the response functions are developed on the basis of the laws of physics can the relationships between watershed measurements and the watershed flood response be found analytically. The response functions have not been derived except for very simple homogeneous watersheds.

Table XV

Watershed Selected from 24 Parameters

Analytical Model Parameters	Correlation Coef.		Principal Component		Factor Analysis		Anti Factor Analysis	
	Natural	Log. Form	Natural	Log. Form	Natural	Log. Form	Natural	Log. Form
Area and Length Parameters						A	A	A
L_s	L	L	L	L	L	L		L_s
L_c	L_c				L_s			
Slope Parameters							S_3	S_4
					S_1	S_2	S_4	R_1
$S_{surf.}$	S_4		S_4	S_4	S_3	R_2	R_1	
				R_5	R_1	R_5	R_6	
$S_{chan.}$	R_1	R_1	R_1	R_6	R_6	R_6	H	
Watershed Shape and Stream Network Parameters								
$n_{surf.}$	F	F	D_d	D_d	F	F	C	F
	C	C			C	C	W	C
							D_d	D_d
$n_{chan.}$	L_m	L_m			D_d	D_d		
						L_m	L_m	
	S_d	S_d			S_d	S_d	S_d	S_d

The correlation coefficient matrix provided a means of grouping parameters into groups. The correlation coefficients provided a worthwhile starting point from which to continue the analysis.

The technique of principal components provided a means of a reduction of 24 principal components to 12 principal components while only accepting a 2% loss in explained variance. The three most important components could be recognized as 1) a combined stream and overland slope parameter, 2) an area and stream length parameter and 3) a watershed shape parameter.

A factor analysis of the variables was based on a varimax rotation generated loadings on common factors. This technique likewise grouped the measured parameters together in grouping which could be identified with physical watershed properties.

The varimax technique called antifactor analysis provides a stepwise screening to reduce the original group of variables (or parameters) to a minimum. For the power function model (multiplicative model), this procedure reduced the number of variables from 24 to 8.

The remaining watershed parameters are:

1. A, watershed area, square miles,
2. L_s, total length of extended streams, miles,
3. F, form factor, A/L^2,
4. C, compactness coefficient, $.28P/\sqrt{A}$,
5. D_d, drainage density, L_s/A, miles per square mile,
6. S_d, dimensionless standard deviation of travel distance, $S_t/\sqrt{A}$,
7. S_4, stream slope, feet per mile,
8. R_1, overland slope, $c\Sigma L_{cont.}/A$, feet per miles.

The investigation has provided a basis for reducing the cost of obtaining and encoding relevant flood and geomorphological data for small watershed floods.

References Cited

Bartlett, M. S., (1950), "Tests of Significance in Factor Analysis", British Journ. Psych. (Stat. Sec.), n. 3, pp. 77-85, 1950.

Dyhr-Nielsen, M., (1971), "Analysis of Interrelationships Between Geomorphic Parameters of Small River Basins", Unpublished Class Report, Hydrology Program, Colorado State University, Fort Collins, 60 p.

Eiselstein, L. M., (1967), "A Principal Component Analysis of Surface Runoff Data from a New Zealand Alpine Watershed", Paper No. 61, Proc. Int. Hydrol. Symp., Fort Collins, Colorado.

Harman, H. H., (1960), "Modern Factor Analysis", Univ. of Chicago Press, Chicago.

Hotelling, H., (1933), "Analysis of a Complex of Statistical Variables into Principal Components", Journ. Educ. Psych., v. 24, 1933.

Kaiser, H. F., (1958), "The Varimax Criterion for Analytic Rotation in Factor Analysis", Psychrometrica, v. 23, no. 3, pp. 187-200.

Kendall, M. G., (1961), A course in Multivariate Analysis, Haffner, New York.

Laurenson, E. M., E. F. Schulz and V. Yevjevich, (1963), "Research Data Assembly for Small Watershed Floods", Engr. Res. Cen., Colo. State Univ., Fort Collins.

Lewis, G., (1968), "Selected Multivariate Statistical Methods Applied to Runoff Data from Montana Watersheds", Unpublished M.Sc. Thesis, Montana State Univ., Bozeman.

Matalas, N. C. and B. J. Reiher (1967), "Some Comments on the Use of Factor Analysis", Wat. Resourc. Res., v. 3, no. 1, pp. 213-224.

Morrison, D. F., (1967), Multivariate Statistical Methods, McGraw Hill Book Co., New York.

Rice, R. M., (1967), "Multivariate Methods Useful in Hydrology", Paper No. 60, Proc. Int. Hydrol. Symp., Fort Collins, Colo.

Schaake, John C. (1971), "Deterministic Urban Runoff Model", in Treatise on Urban Water Systems, Colo. State Univ., Ft. Collins, pp. 357-383.

Snyder, W. M., (1962), "Some Possibilities for Multivariate Analysis in Hydrologic Studies", Journ. Geophys. Res., v. 67, no. 2, pp. 721-729.

Wallis, J. R., (1965a), "A Factor Analysis of Soil Erosion and Stream Sedimentation in Northern California", Unpublished Ph.D. Thesis, Univ. of California, Berkeley.

Wallis, J. R., (1965b), "Multivariate Statistical Methods in Hydrology - A Comparison Using Data of Known Functional Relationship", Wat. Resour. Res., v. 1, no. 4, pp. 447-461.

Wallis, J. R., (1968), "Factor Analysis in Hydrology - An Agnostic View", Wat. Resourc. Res. v. 4, no. 3, pp. 521-527.

Wilson, E. M., (1969), Engineering Hydrology, Macmillan, London, 182 p.

Yevjevich, V. and M. E. Holland, (1967), "Research Data Assembly for Small Watershed Floods, Part II", Colo. State Univ., Fort Collins.

Yevjevich, V., editor (1971), Systems Approach to Hydrology, Proc. 1st Bilateral U. S.-Japan Seminar in Hydrology, Water Resources Publications, Fort Collins, Colorado, 464 p.

Yevjevich, V., (1962), Probability and Statistics in Hydrology, Water Resources Publications, Fort Collins, Colorado, 302 p.