Fisheries and Environment Canada
Pêches et Environnement Canada

Principal Components Analysis on Glacier-Climatological Data for Sentinel Glacier, British Columbia

S. Fogarasi and O. Mokievsky-Zubok

SCIENTIFIC SERIES NO. 95
(Résumé en français)

INLAND WATERS DIRECTORATE,
WATER RESOURCES BRANCH,
OTTAWA, CANADA, 1978.

Fisheries and Environment Canada
Pêches et Environnement Canada

Principal Components Analysis on Glacier-Climatological Data for Sentinel Glacier, British Columbia

S. Fogarasi and O. Mokievsky-Zubok

SCIENTIFIC SERIES NO. 95
(Résumé en français)

INLAND WATERS DIRECTORATE,
WATER RESOURCES BRANCH,
OTTAWA, CANADA, 1978.

Cat. No. En36-502/95

ISBN 0-662-10202-9

Contents

Tables

Illustrations

Abstract

Multivariate analysis is being carried out on glacier-climatological data in two stages. This paper describes the first stage, which is an application of principal components technique to the available data. The multicollinear climatological variables are transformed into orthonormal principal components. The characteristics and the interpretation of these principal components are described. Data transformations and the ensuing results are explained. The transformed new variables, the component vectors, are interpreted in terms of weather types. Recommendations are given for the second stage of the analysis, which is an application of multiple regression on principal components aimed at the estimation of runoff.

Résumé

L'application de l'analyse à plusieurs variables aux données de climatologie glaciaire s'effectue en deux étapes. Le présent rapport traite de la première étape qui consiste à appliquer la technique des principales composantes aux données accessibles, et selon laquelle les variables climatologiques multicollinéaires sont transposées en composantes principales orthonormales. Le rapport décrit les caractéristiques et l'interprétation des principales composantes, explique les transformations des données et des résultats obtenus et fournit une interprétation des nouvelles variables, les vecteurs composants, en fonction des types de conditions météorologiques. En dernier lieu, des recommandations sont formulées sur la deuxième étape de l'analyse en vue d'appliquer la régression multiple aux composantes principales afin d'évaluer le taux de ruissellement.

List of Symbols

$[C]$	component correlation matrix of 5 x 5
$\vec{C}_i = \sqrt{\lambda_i}\,\vec{y}_j$	i-th component vector
$\vec{C}_j$	j-th component vector
C_{ik}	k-th element of the i-th component vector
C_{jk}	k-th element of the j-th component vector
$[D]$	data matrix
$[D^T]$	transpose of the data matrix
I	unit matrix
$\vec{j}$	unit column vector
$\vec{j}^T$	unit row vector
ℓ_i	loadings of the components, i=1, ..., 5
n	size of the sample
P	precipitation
$\overline{P}$	freezing
r_{ij}	correlation coefficient of the i-th and j-th variables
$[R]$	correlation matrix
R^5	fifth dimension of the vector space V
RO	runoff
$\overline{RO}$	complement of runoff or reduced runoff
$[R_{v-c}]$	variance-covariance matrix
S	sunny
$\overline{S}$	cloudy
S_{x_i}	standard deviation of the i-th variable
S_{x_j}	standard deviation of the j-th variable
T	warm
$\overline{T}$	complement of T or cool
V	vector space
W	windy
$\overline{W}$	calm
x_i	i-th variable, i = 1, 2, ..., 5
$\overline{x}_i$	mean of the i-th variable
x_{ik}	k-th element (observation) of the i-th variable
x_j	j-th variable
$\overline{x}_j$	mean of the j-th variable
x_{jk}	k-th element of the j-th variable
$\mathrm{cov}(x_i, x_j)$	covariance of the i-th and j-th variables
$\vec{y}_j$	j-th eigenvector, j = 1, 2, ..., 5
y_{ij}	i-th element of the j-th eigenvector
Z_i	standard Z score of the i-th value
$\overline{Z}_{x_i}$	average of the standardized Z score of the variable x_i
$\bigcap_{i=1}^{5}$	intersection of a set of five items
λ_i	i-th eigenvalue
$\lvert\ \rvert$	determinant of a matrix

Principal Components Analysis on Glacier-Climatological Data for Sentinel Glacier, British Columbia

S. Fogarasi and O. Mokievsky-Zubok

INTRODUCTION

Glacier runoff, an important contributor to water balance, depends heavily on the interaction of various climatological variables influencing glacier melt. Linear regression techniques frequently used in runoff estimations do not usually reveal the intricate association between runoff and some of the weather-related variables. Regression techniques are unsuccessful as long as the variables are strongly dependent, therefore, the principal components technique has been applied in this study. This analysis essentially transforms the sets of dependent climatic variables into a new set of independent statistics, even though the original variables may be highly interrelated.

How the new independent variables, called components, are created out of dependent variables is sufficiently described by Anton (1973) and Kendall (1965) in a chapter on linear transformation. For readers with elementary statistical and linear algebraic background who are not completely familiar with principal-components analysis, the general principles of technique have been outlined.

By transforming a set of variates into principal components, the interrelationship of original variables may be explored more easily. The components, as independent variables, can then be used in a multiple regression model. The components are not equivalent to the original variables but are, to some degree, associated either with a single variable or with a group of variables. After the linear transformation, however, the total variance associated with the original data is preserved in the resultant independent new components. The principal components analysis assumes that all the variation in a given population is contained within the variables used; therefore this analysis is deterministic.

LOCATION AND PHYSICAL SETTING

The set of data used in the forthcoming analysis was collected on Sentinel Glacier, B.C., which has been intensively studied by Inland Waters Directorate, Department of the Environment. Sentinel Glacier (49°54'N and 122°59'W) is located 70 km north of Vancouver, in the Coast Mountain Range and within the boundaries of Garibaldi Provincial Park. The glacier occupies a basin facing north descending and ending within two kilometres of Garibaldi Lake (1478 m asl). The elevation range of the glacier is from 1550 m to 2100 m. The ice cover of the basin is about 2 km^2 but diminishing constantly. The area of the whole basin, which drains into the Sentinel meltwater stream, is 5 km^2.

INSTRUMENTATION AND OBSERVATION

Glaciological studies at Sentinel Glacier began in 1966 when the glacier was included in the International Hydrological Decade (IHD) program (Mokievsky-Zubok, 1973). The observation station was located 500 m in front of the glacier tongue at 1540 m asl. The station was equipped with short-term climatological instruments: mercury thermometer, thermohygrograph, totalizing anemometer, sunshine recorder, solarimeter and rain gauge. Daily climatological data used in this analysis were recorded at Sentinel Glacier during June, July, August and September, 1970-1974. Days with incomplete data have been omitted from the analysis. A total of 547 complete daily sets of data were available. Each set of complete data was made up of twelve variables as shown in Table 1.

Table 1. Daily Climatological Variables on Sentinel Glacier Used in the Analysis.

Identification No.	Variables	Dimension
1	Runoff	$10^3 m^3$
2	Precipitation	mm
3	Mean temperature	C°
4	Maximum temperature	C°
5	Minimum temperature	C°
6	Calculated mean temperature	C°
7	Melting degree day temperature	C°
8	Relative humidity	%
9	Daily total sunshine	h
10	Global radiation	$10^6 Jm^{-2}$
11	Mean daily cloudiness	1/10
12	Daily total wind run	km

Variables listed as numbers 4, 5, 8, 9, 10 and 12 were obtained from the data recorded directly by the instruments. Runoff, variable No. 1, was calculated from continuous observations of water level and the rating curve. The latter was based on regular discharge measurements conducted either by the personnel at the station or by the Water Survey of Canada (Station No. 08GA056). The gauging site is located 800 m downstream from the glacier tongue. Precipitation, variable No. 2, was measured with Pluvius or MSC (Meteorological Services of Canada) type of rain gauges or calculated in proportion 1/10 from snowfalls, which are given in millimetres of water equivalent values. Temperature data were obtained from four daily readings of a mercury thermometer at 0800, 1200, 1600 and 2000 Pacific Standard Time (PST), and from the weekly graph of a Leopold-Stevenson thermohygrograph. Mean temperature, variable No. 3, was calculated from a graph of hourly readings for each day. The thermohygrograph readings were adjusted to the four control readings and averaged. Calculated mean temperature, variable No. 6, was obtained by averaging the maximum and minimum daily temperatures. Melting degree day temperature, variable No. 7, was calculated by averaging the positive hourly temperatures. Cloudiness, variable No. 11, was estimated in tenths 4 times a day. The mean of the four daily observations was used in this analysis as a single daily value.

PREPARATION OF DATA

As mentioned earlier, the input data consisted of 547 sets of observations and each set contained 12 values. That is, the input matrix [D] had an order of 547 x 12. Initially, equal weight or importance was given to each variable. Therefore, Z scores were formed for each entry of the raw data because the Z scores are nondimensional. The formula

$$Z_i = \frac{x_i - \bar{x}_i}{S_{x_i}}, \qquad i = 1,2, \ldots, n = 547 \quad (1)$$

transforms the variables to zero mean ($\bar{Z}_{x_i} = 0$) and unit standard deviation ($S_{x_i} = 1$). Each Z score tells us how many standard deviations an entry is above or below the mean of the variable to which it belongs.

Out of this huge standardized data matrix a square and symmetric array had to be obtained to make it suitable for the principal components program. Therefore, a variance-covariance matrix, $[R_{v-c}]$, was formed by using a numerical method developed by Krumbein and Graybill (1965). This method is basically a matrix multiplication whereby the variance-covariance matrix is given as

$$[R_{v-c}] = \frac{1}{n-1}\left[[D^T]\,[D] - \frac{1}{n}\left([D^T]\,\vec{j}\right)\left(\vec{j}^T\,[D]\right)\right] \quad (2)$$

$$\text{where } \vec{j} = \underset{n\times 1}{\begin{bmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{bmatrix}} \quad \text{and } \vec{j}^T = \underset{1\times n}{[1\ 1\ .\ .\ .\ 1]}$$

The order of $[R_{v-c}]$ is in fact determined by the first term on the right because the second term on the right must be of the same order as the first term. Consequently, the difference,

$$\frac{1}{546}\left[\underbrace{[D^T]\ [D]}_{12\times 547\ \ 547\times 12} - \frac{1}{547}\underbrace{\left([D^T]\ \vec{j}\right)\left(\vec{j}^T\ [D^T]\right)}_{12\times 547\ \ 547\times 1\ \ \ 1\times 547\ \ 547\times 12}\right] \quad (2a)$$

results in a $[R_{v-c}]$ of the order of 12 x 12.

In this case, investigation is directed to the relationship between the twelve variables. In other words, when the columns of the variables are examined against each other, an R-mode analysis is done.

The present investigation is also done in R-mode. When the investigation is aimed at the relationship between the rows of the data matrix, [D], a Q-mode analysis is performed. If our analysis were in Q-mode, the order of the variance-covariance matrix would have been 547 x 547, generated by

$$\underbrace{\underset{547\times 12}{[D]}\ \ \underset{12\times 547}{[D^T]}}$$

The essential difference between the modes is in the sequence of multiplication with $[D^T]$. R-mode analysis shows the interrelation of variables for a point within a time-scale and the Q-mode analysis is a useful tool to show these relationships in an areal setting.

Even for an R-mode analysis the original data matrix, [D], of the order 547 x 12 is just too large for the analysis. Under the rules of the R-mode analysis, a variance-covariance matrix of 12 x 12 was obtained by using a CDC 7400 high-speed computer. To reduce computer storage space, the calculation was performed on partial matrices of the order of 12 x 12. The variance-covariance matrix thus obtained is shown in Table 2. Each entry in this

Table 2. Variance-Covariance Matrix, $[R_{v-c}]$, of the Twelve Variables Based on the Data Collected on Sentinel Glacier for June-September, 1970-1974.

Variables	Runoff	Preciptn.	Mean temp.	Max. temp.	Min. temp.	Calctd. mean temp.	Melting deg. days	RH %	Sunshine h	Global radiation	Mean cloudiness	Wind run
Runoff	4088.0											
Preciptn.	124.7	67.8										
Mean temp.	103.5	- 9.1	17.3									
Max. temp.	106.5	-13.2	21.5	29.1								
Min. temp.	113.9	- 4.6	14.0	16.5	14.3							
Calctd. mean temp.	110.1	- 8.8	17.7	22.8	15.4	19.1						
Melting deg. days	100.1	- 9.1	17.0	21.0	13.7	17.3	16.7					
RH	- 59.0	38.3	-41.8	-53.6	-29.8	-41.7	-41.5	276.2				
Sunshine	49.1	-15.0	12.0	16.9	6.9	11.9	12.0	-46.2	19.9			
Global radtn.	132.4	-22.6	16.2	22.4	8.7	15.6	15.9	-47.3	25.4	74.8		
Mean cloud.	- 22.7	11.2	- 9.8	-13.8	- 5.9	- 9.9	- 9.8	42.1	-14.6	-16.2	14.1	
Wind run	1657.5	183.8	102.3	89.8	134.6	111.7	98.1	-42.4	-49.5	-15.1	71.3	19179.8

symmetric matrix can be interpreted as a result of a bivariate manipulation. The off-diagonal entries are the covariances of two respective variables. Also, each entry is extracted from the original 547 x 12 sample matrix and each entry is equivalent as if it were calculated with the well-known formula:

$$\mathrm{cov}(x_i, x_j) = \frac{\sum_{i=1}^{n=547} \left[(x_i - \bar{x}_i)(x_j - \bar{x}_j)\right]}{547 - 1}, \; j = 1, 2, \ldots, 12. \quad (3)$$

Because the magnitude of the nominator is influenced by the size of the sample, the covariation formula is divided by 546. Along the diagonal i = j, therefore, the expression

$$\mathrm{cov}(x_i, x_i) = \frac{\sum_{k=1}^{547} (x_{ik} - \bar{x}_i)^2}{546} \quad (4)$$

becomes the variance formula

$$S^2_{x_i} = \frac{\sum_{k=1}^{547} (x_{ik} - \bar{x}_i)^2}{546} \quad (5)$$

Thus, variances are found along the diagonal and covariances off the diagonal.

The variance-covariance matrix can be checked statistically for the independence of the variables. The independence of any two variates can be determined by the value of their covariance. When the $\mathrm{cov}(x_i, x_j) = 0$, the x_i and x_j variables are independent. Table 2 shows that not one pair of variables can be considered independent. Yeates (1974) demonstrated that with the use of a standardizing factor, $1/(S_{x_i} S_{x_j})$, which is the reciprocal value of the product formed by the two relevant standard deviations of the two variables, the variance-covariance matrix can be converted into a correlation coefficient matrix

$$r_{ij} = \frac{\mathrm{cov}(x_i, x_j)}{S_{x_i} S_{x_j}}. \quad (6)$$

The variance-covariance matrix thus converted into 12 x 12 correlation matrix is shown in Table 3. Shaded areas enclose entries with high correlations. In other words, the great resemblance between mean, maximum and minimum, calculated mean and melting degree day temperatures in fact indicates that we have here similar variables but with different names. These variables, except melting degree day, were therefore omitted from further consideration. Because only one relative humidity value was given for each day, it was considered unrepresentative and also omitted. Finally, a remnant 5 x 5 correlation matrix was used as an input into the principal components program. This final input of 5 x 5 symmetric matrix is shown in Table 4. Values above the lines are variances-covariances; below these lines are the correlation values. The assumption is that with the omission of the four temperature-related variables and the relative humidity a grouping of identical variables on the same components is eliminated.

Table 3. Correlation Matrix, [R], of the Twelve Variables Based on Data Collected on Sentinel Glacier for June-September, 1970-1974.

Variables	Runoff	Preciptn.	Mean temp.	Max. temp.	Min. temp.	Calctd. mean temp.	Melting deg. days	RH %	Sunshine h	Global radiation	Mean cloudiness	Wind run
Runoff	1											
Preciptn.	0.238	1										
Mean temp.	0.386	-0.264	1									
Max. temp.	0.309	-0.298	0.948	1								
Min. temp.	0.469	-0.148	0.877	0.804	1							
Calctd. mean temp.	0.391	-0.243	0.958	0.959	0.921	1						
Melting deg. days	0.382	-0.271	0.987	0.949	0.879	0.959	1					
RH	-0.056	0.281	-0.599	-0.598	-0.472	-0.571	0.610	1				
Sunshine	0.170	-0.407	0.635	0.695	0.403	0.601	0.650	-0.618	1			
Global radtn.	0.241	-0.320	0.449	0.482	0.266	0.412	0.451	-0.331	0.656	1		
Mean cloud.	-0.093	0.359	-0.614	-0.673	-0.409	-0.592	-0.629	0.667	-0.853	-0.495	1	
Wind run	0.008	0.162	0.176	0.120	0.256	0.183	0.173	-0.018	-0.079	-0.013	0.135	1

Table 4. Variance-Covariance and Correlation Matrix of the Five Variables Based on Data from Sentinel Glacier for June, July, August and September, 1970-1974.

	Runoff, $10^3 m^3$	Precipitation, mm	Melting degree day, C°	Global radiation $10^6 Jm^{-2}$	Wind run, km
Runoff	4088.1 / 1				
Precipitation	124.7 / 0.238	61.8 / 1			
Melting degree day	100.1 / 0.382	-9.1 / 0.264	16.7 / 1		
Global radiation	132.4 / 0.241	-22.6 / 0.3201	15.9 / 0.451	74.8 / 1	
Wind run	1657.5 / 0.0087	183.8 / 0.162	98.1 / 0.173	-15.1 / 0.013	19179.8 / 1

It was expected that as the original dependent data matrix was converted into its own independent component matrix, by the principal components analysis, each column vector of the component matrix could be identified with one variable. Statistically and algebraically, this 5 x 5 correlation matrix is equivalent to the 547 x 5 data matrix. This correlation matrix now appears as if it were taken from the same population because each variable in it has the same mean and the same variance as a result of standardization.

Principal components analysis can be performed both on variance-covariance matrix and on correlation matrix, but the results would be different. It is, however, generally accepted that correlation matrix fits an R-mode analysis and variance-covariance matrix, a Q-mode analysis.

For further analysis, it is mathematically convenient to have a standardized symmetric correlation matrix. Because our final 5 x 5 input matrix preserved the characteristics of the original data matrix of (547 x 5), therefore, all the algebraic and statistical inferences made on this 5 x 5 matrix are also valid for the original data matrix.

PRINCIPAL COMPONENT ANALYSIS ON THE CORRELATION MATRIX

Algebraically, the 5 x 5 correlation matrix can be described and characterized with a single scalar value, called determinant, and with a fifth degree polynomial. The roots of this polynomial equation are called the latent roots, characteristic values, or eigenvalues (λ_i, and in this case $1 \leqslant i \leqslant 5$). The characteristic equation in the well-known matrix notation

$$|R - \lambda I| = 0 \tag{7}$$

can also be written as the characteristic equation,

$$\begin{vmatrix} r_{11} - \lambda_i & r_{12} & \cdots\cdots & r_{1n} \\ r_{21} & r_{22} - \lambda_i & \cdots\cdots & r_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ r_{n1} & r_{n2} & \cdots\cdots & r_{nn} - \lambda_i \end{vmatrix} = 0 \tag{8}$$

and be solved for the eigenvalues.

The eigenvalues obtained from the fifth degree polynomial and the probability of their upper and lower bounds at 95% certainty are shown in Table 5.

Table 5. Eigenvalues, Their Associated Percentages of Variation in the Original Raw Data and the Probability of the Upper and Lower Bounds at $\alpha = 0.05$ Level. These Values were Calculated from the 5 x 5 Correlation Matrix.

Order of eigenvalue	Eigenvalues	% variation in the data accounted for by λ_i	95% certainty that λ_i ranges between the values indicated
λ_1	1.822	36.4400	$P(1.63 \leqslant \lambda_1 \leqslant 2.07) = 0.95$
λ_2	1.405	28.0995	$P(1.26 \leqslant \lambda_2 \leqslant 1.59) = 0.95$
λ_3	0.8497	16.9933	$P(0.76 \leqslant \lambda_3 \leqslant 0.96) = 0.95$
λ_4	0.5335	10.6693	$P(0.48 \leqslant \lambda_4 \leqslant 0.61) = 0.95$
λ_5	0.3899	7.7976	$P(0.35 \leqslant \lambda_5 \leqslant 0.44) = 0.95$

Using symmetric square matrices so that their traces, the sums of the diagonal values, are equal to the sums of the eigenvalues, is a recognized advantage.

$$\text{Trace} = \sum_{i=1}^{5} \lambda_i = \sum_{i=1}^{5} r_{ii} = 5, \qquad (9)$$

where ii is along the diagonal.

Since the total variance of the original data is retained throughout this algebraic manipulation, the percentage of variance associated with any i-th eigenvalue explains a certain percentage of the variation in the original data matrix, as

$$\% = \frac{\lambda_i}{5} \times 100 \qquad (10)$$

These percentages are listed in Table 5, where the values are arranged in a descending order.

If there were two variables and hence a two-vector space, then the first eigenvalue could be interpreted as the length of the longitudinal or principal axis, and the second eigenvalue as the shorter secondary axis of an elongated ellipse that comprises the probability contours of the two variates. For three variables the three eigenvalues would indicate an ellipsoid; for five variables no geometric illustration would be possible.

By substituting the eigenvalues into characteristic matrix and expanding the matrix row-wise with the signed cofactors, the appropriate eigenvectors can be obtained. Only one eigenvector, $(\vec{y}_j)$, is associated with each eigenvalue. These eigenvectors (Table 6) are orthogonal (Anton, 1973); that is, they are independent from each other. They span the $V = R^5$ space and they form a basis for this five-dimensional space. The total variance rule is preserved because the deviations of the individual data vectors from the eigenvectors are normally distributed (King, 1969). The eigenvectors derived from the standardized data are normalized; that is, the eigenvalues are of unit length. Their norm is

$$\left\| \vec{y}_j \right\| = \sqrt{\sum_{i=1}^{5} \vec{y}_{ij}^{\,2}} = 1 \qquad (11)$$

The five eigenvectors in the sequence of their significance, as associated with the eigenvalues, are listed in Table 6. Eigenvectors can also be considered the resultants of the standardized data vectors.

Table 6. Eigenvalues and Their Associated Eigenvectors in an Increasing Order from Right to Left.

$\lambda_5 = 0.3899$	$\lambda_4 = 0.5335$	$\lambda_3 = 0.8497$	$\lambda_2 = 1.4050$	$\lambda_1 = 1.8220$
$\vec{y}_5$	$\vec{y}_4$	$\vec{y}_3$	$\vec{y}_2$	$\vec{y}_1$
-0.5705	0.1946	-0.4532	-0.5032	0.4169
0.5580	-0.2626	-0.3229	-0.6597	-0.2752
0.5812	0.5072	0.0827	-0.0147	0.6308
0.0828	-0.7728	-0.1538	0.2166	0.5703
-0.1358	-0.1967	0.8095	-0.5101	0.1652

By forming the scalar product of the square root of each eigenvalue with its eigenvector, another orthogonal set of vectors called component vectors is obtained, so that

$$\left[\sqrt{\lambda_5}\ \vec{y}_5 \quad \sqrt{\lambda_4}\ \vec{y}_4 \ \ldots \ \sqrt{\lambda_1}\ \vec{y}_1\right] = [C] \qquad (12)$$

where the first eigenvector is weighted with the square root of the first eigenvalue, $\vec{y}_2$ with $\sqrt{\lambda_2}$, and so on. These newly formed five sets of vectors are also called component correlation matrix, where the first component, $\sqrt{\lambda_1}\vec{y}_1$, is the principal component followed by the second, third, fourth and fifth components. These component vectors $\vec{C}_i$ are also orthogonal and independent from each other; that is, the inner or dot product of any two component vectors results in zero:

$$\left\langle \vec{C}_i, \vec{C}_j \right\rangle = \sum_{k=1}^{5} C_{ik} C_{jk} = 0 \qquad (13)$$

if $j = i$

i and $j = 1, 2, \ldots, 5$

The component correlation matrix [C] formed by the component vectors are also orthogonal, square and nonsingular. The entries or elements of the component vectors are called loadings or factor loadings. Each loading indicates the proportion of variation in the variables associated with the variation in the component. Loadings should be interpreted as correlation coefficients, or geometrically each loading is regarded as the orthogonal projection of the variables on the normalized eigenvectors. The five component vectors are listed in the same increasing order as the eigenvalues. Counting the vectors from right to left, the last column is the principal component and the first column is the fifth component (Table 7). The loadings of each variable are listed row-wise.

Table 7. Component Correlation Matrix of the Order of 5 x 5 Where $\vec{C}_1$ is the Principal Component Followed by the Second, Third, Fourth and Fifth Components.

Components					
$\vec{C}_5$	$\vec{C}_4$	$\vec{C}_3$	$\vec{C}_2$	$\vec{C}_1$	Variables
-0.3562	0.1421	-0.4178	-0.6014	0.5627	Runoff
0.3484	-0.1918	-0.3039	-0.7820	-0.3715	Precipitation
0.3629	0.3704	0.0763	-0.0174	0.8514	Melting degree day
0.0517	-0.5645	-0.1418	0.2567	0.7699	Global radiation
-0.0848	-0.1437	0.7462	-0.6046	0.2230	Wind run

INTERPRETATION OF THE COMPONENTS LOADINGS

The component loadings, or factor loadings, essentially indicate the proportion of variation in the variables that is associated with the variation in the component. This interpretation is done with respect to the loadings, the principle of interpretation being that variables having high correlations or loadings on a component will help to identify that component. The question of how many components should be considered is determined by the following rule of thumb. The components associated with the eigenvalues $\leqslant 1$, or each component that accounts for at least 5% of the total variance, should be considered. Loadings can be interpreted geometrically as well. For example, the coefficient of 0.8514 in the principal component (Table 7) is in fact an orthogonal projection of the variable "melting degree day" on the normalized eigenvector $\vec{y}_1$.

As we can see from Table 7, the principal component is heavily loaded on melting degree day (0.8514), global radiation (0.7699), and runoff (0.5627). The second component is identifiable with lack of precipitation (-0.7820), the third component with wind (0.7462), and so on. As all the variables used in the analysis are weather related, therefore, each component is associated with the simultaneous ensemble of variables or their components. Hence, each component is considered to be the intersection of the loadings, associated with the events. In other words, each component can be written in the form

$$\vec{C}_i = \bigcap_{i=1}^{5} \ell_i \qquad (14)$$

where ℓ_i is loading.

During the interpretation of the components, the following conditions (events) were associated with the loadings: runoff (RO), reduced runoff ($\overline{RO}$), warm (T), cool ($\overline{T}$), sunny (S), cloudy ($\overline{S}$), windy (W), calm ($\overline{W}$), precipitation (P), and finally, no precipitation or freezing ($\overline{P}$). The five components verbally described are as follows:

i) Warm, sunny day, high runoff, little breeze, and no precipitation, i.e. in set notation: $\vec{C}_1 = (T, S, RO, \overline{W}, \overline{P})$.
ii) Cool, cloudy, and calm with freezing and much reduced runoff, i.e. $\vec{C}_2 = (\overline{T}, \overline{S}, \overline{W}, \overline{P}, \overline{RO})$.
iii) Very windy, just above freezing, cloudy, increased evaporation and reduced runoff, i.e. $\vec{C}_3 = (W, T, \overline{S}, \overline{P}, \overline{RO})$.
iv) Fog or cloud over the glacier, above freezing, slight increase in runoff, no precipitation and calm, i.e. $\vec{C}_4 = (\overline{S}, T, RO, \overline{P}, \overline{W})$.
v) Mild, calm day with cloud cover precipitation and reduced runoff, i.e. $\vec{C}_5 = (T, \overline{W}, \overline{S}, P, \overline{RO})$.

Summarizing briefly, the principal component is the nice weather component; second component is the freezing type weather; third component is identified as windy weather; fourth component is foggy weather; and the fifth component is precipitation.

Owing to the normalized and orthogonal eigenvectors, there are two solutions for the eigenvectors and their respective component vectors: a positive and a negative one. Therefore, the sign of each component can be reversed and the number of weather conditions associated with each component can be doubled, except when one or more components are disregarded for reasons given later. The sign of each $\vec{C}_i$ vector in Table 7 can be reversed and each loading value will still be valid. For brevity, the opposite interpretation of the components is omitted.

The first couple of components accumulate certain variation in each variable and the degree of this variation is obtained by Yeates' method (Yeates, 1974). Accordingly, the squares of loadings are summed for each variable per each component and these sums are expressed as percentages. This measure, called the adequacy of solution for the

loadings, which can be written for each variable, is:

$$\text{Adequacy of solution} = \sum_{j=1}^{n} (\ell_{ij})^2 \times 100 \qquad (15)$$

$$i=1, 2, \ldots, 5$$

where i indicates the variables and j the component vectors (Table 8).

Table 8. Adequacy of Solution of the Loadings for Each Variable on Each Component.

Variables	Efficiency of solution for the first 3 components %	first 4 components %	all 5 components %
Runoff	70.878	72.897	99.999
Precipitation	84.189	87.867	100.0
Melting degree day	73.100	86.819	99.989
Global radiation	67.874	99.740	100.0
Wind run	97.210	99.274	99.994

NOTE: The percentages in some cases did not add up to 100% because of the roundoff error by the computer.

On the basis of eigenvalues and the associated percentages as shown in Table 5, the cutoff point for considering components can be either component 3 or 4. Yeates (1974) recommends a third possibility for locating the cutoff point, which is to choose the point along the curve of the cumulative proportion of total variation where the curvature changes sharply (Fig. 1). In this case, however, the cutoff point cannot be determined because of the uniform change of curvature along the cumulative frequency curve.

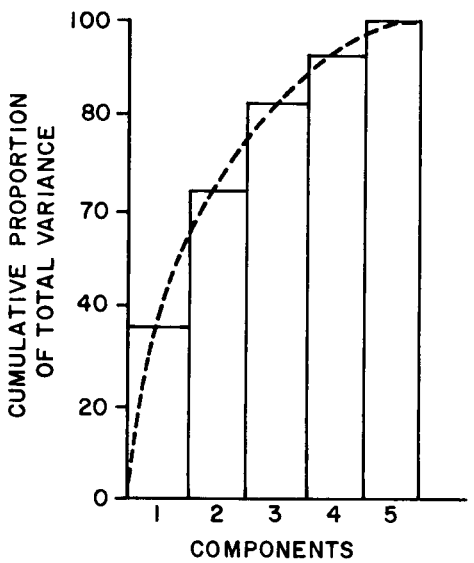

Figure 1. Cutoff point for components. It cannot be determined because there is no sharp change of curvature along the cumulative frequency curve.

CONCLUSION

The daily climatological observations may provide sufficient support for mass budget studies, but for energy-budget analyses these twelve variables seem to be irrelevant. It is believed, however, that with the exception of precipitation the other three variables can explain the amount of snowmelt. As soon as precipitation comes into the picture, either in liquid or solid phase, the snowmelt rates are drastically changed. In this exercise only the water-equivalent precipitation was used and its phases were disregarded. High albedo conditions following snowfall can completely cancel the influence of high global radiation.

Normality of the variables is not an absolute necessity for principal components analysis, but when inferences are to be made and the inferences require normally distributed variables, then this assumption cannot be relaxed. In multiple regression studies the least squares theory does not require normality. To make inferences based on the estimates, however, normal theory is used.

Precipitation values subjected to chi-square test indicated a distribution that is far from normal (Table 9). (In this chi-square analysis of precipitation, zero values were omitted.) In fact, precipitation did not show normal distribution in either liquid or solid phase. It is suggested that precipitation values should be divided into snow and rain groups, and the frequency distribution of each should be transformed into normal distribution by a suitable function (Panofsky and Brier, 1968). However, omitting cloudiness and sunshine duration from the analysis also gives non-normal, j-shaped distribution (Essenwanger, 1976; Fogarasi and Strome, 1976). According to the chi-square test, cloudiness is the second less normally distributed variable. Consequently, it is not surprising that the observed frequencies of sunshine duration differ significantly from the normally distributed frequencies. Therefore, cloud data can be utilized only after transformation. Finally, it should be stressed that runoff itself is a complicated variable and is not quite normally distributed either (Table 9). The histogram of the runoff data shows bimodal distribution (Fig. 2). One mode is at 50 x $10^3 m^3$ and the other one at 125 x $10^3 m^3$. The two modes are frequently looked upon as indicators which are associated with two different processes: snowmelt and perhaps precipitation.

It is generally accepted that net radiation is a significant component in snowmelt processes. Unfortunately,

Table 9. Chi-Square Goodness-of-Fit Test for Normality of Seven Variables.

Variables	Calculated X^2	Tabulated X^2 at significance level 0.1%	0.01%	Distribution at level 0.1%	0.01%
Runoff	17.13	15.1	20.5	Not normal	Normal
Precipitation	99.18	15.1	20.5	Not normal	Not normal
Cloudiness	99.02	15.1	20.5	Not normal	Not normal
Melting degree day	6.36	15.1	20.5	Normal	Normal
Global radiation	16.69	15.1	20.5	Not normal	Normal
Wind run	17.13	15.1	20.5	Not normal	Normal
Sunshine duration	26.58	15.1	20.5	Not normal	Not normal

NOTE: The tabulated X^2 statistics are shown for $\alpha = 0.1\%$ and 0.01% significance levels with five degrees of freedom. The null hypothesis (H_0 = observed frequencies are normally distributed) is rejected with a 99% confidence when the calculated X^2 > tabulated X^2.

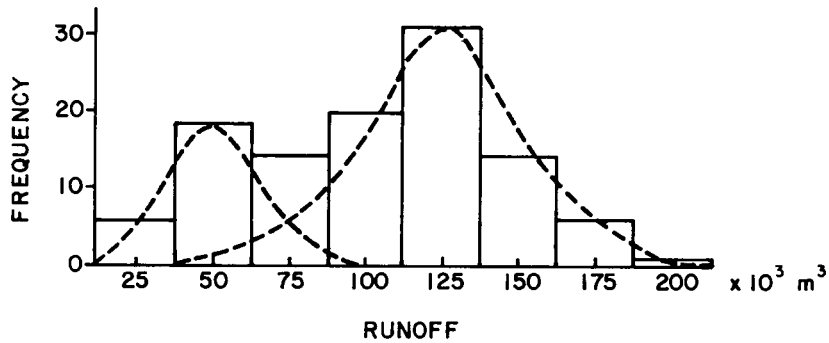

Figure 2. Frequency polygon of runoff data for Sentinel Glacier, B.C., June-September, 1970-1974.

this variable was not recorded on Sentinel Glacier. Net radiation can, however, be approximated from local observations, neighbouring weather stations, or from satellite pictures. With the use of net radiation, a good prediction is expected from a multiple regression analysis performed on principal components. For this type of analysis, it is suggested that expensive local field observations could be discontinued and research workers could rely completely on climatological and satellite information available from the Atmospheric Environment Service.

In light of the preceding discussion, to expect the use of these variables in a statistical model to explain the amount of runoff would, perhaps, be too naive. For the second stage of the analysis, it is therefore suggested that with normally distributed variables and with the omission of days with precipitation, the snowmelt can be estimated for selected days if the normalized runoff values are multiply regressed on the independent principal components (Massey, 1965).

ACKNOWLEDGMENTS

The authors would like to thank J. Glynn and E. Langham, Glaciology Division, Environment Canada; L. Derikx, Natural Sciences and Engineering Research Council of Canada; and G. Cavadias, Faculty of Management, McGill University, for their valuable comments.

REFERENCES

Anton, H., 1973. Elementary Linear Algebra. John Wiley & Sons Inc., New York, London, Toronto, 296 pp.

Essenwanger, O., 1976. Applied Statistics in Atmospheric Science. Part A. Frequencies and Curve Fitting. Elsevier Scientific Publishing Co., Amsterdam, Oxford, New York.

Fogarasi, S. and Strome, M., 1976. Cloud Climate at the Head of Inugsuin Fiord, N.W.T., Summer 1968. Musk-Ox, Vol. 18, pp. 76–85.

Kendall, M.G., 1965. A Course in Multivariate Analysis. Charles Griffin & Co. Ltd., London.

King, L.J., 1969. Statistical Analysis in Geography. Prentice-Hall Inc., Englewood Cliffs, N.J. 288 pp.

Krumbein, W.C. and Graybill, F.A., 1965. An Introduction to Statistical Models in Geology. McGraw-Hill Book Co., New York, Toronto, London. 475 pp.

Massey, W.F., 1965. Principal components regression in exploratory statistical research. Journal of the American Statistical Association, Vol. 60, No. 309, pp. 234-56.

Mokievsky-Zubok, O., 1973. Study of Sentinel Glacier, British Columbia, Canada, within the International Hydrological Decade Program. Procedures and Techniques. Department of the Environment, Inland Waters Directorate, Water Resources Branch, Ottawa, Ontario. Technical Bulletin No. 77.

Panofsky, H.A. and Brier, G.W., 1968. Some Applications of Statistics to Meteorology. Pennsylvania State University, University Park, Pa., College of Earth and Mineral Sciences. 224 pp.

Yeates, M., 1974. An Introduction to Quantitative Analysis in Human Geography. McGraw-Hill Book Co., New York, Toronto, London. 300 pp.