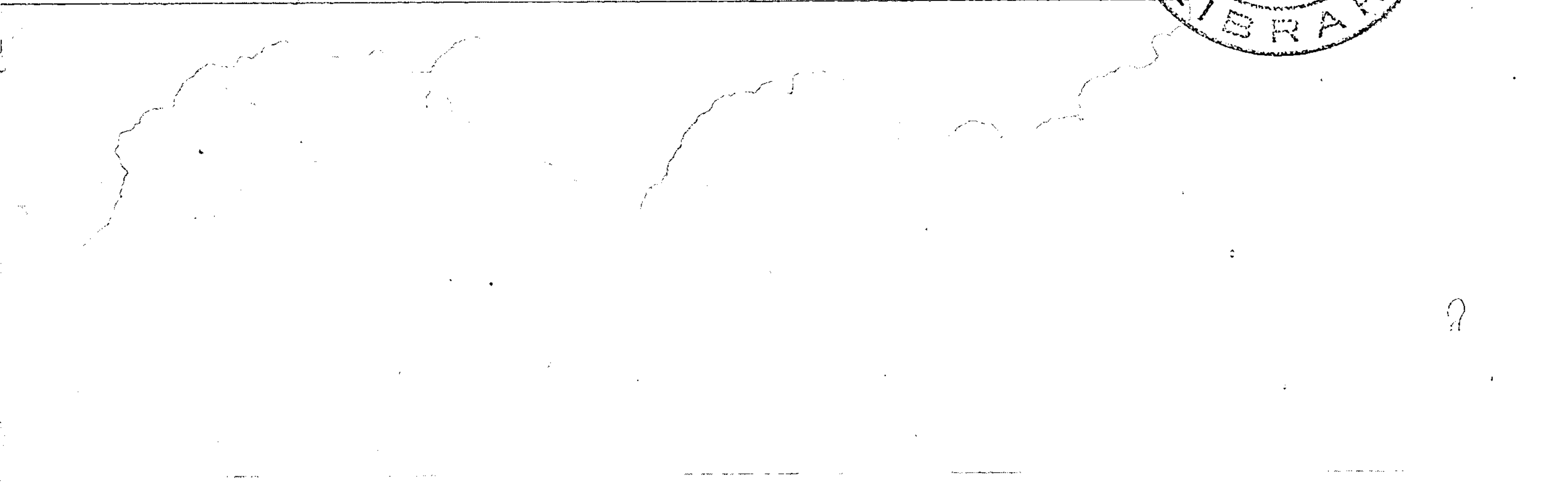
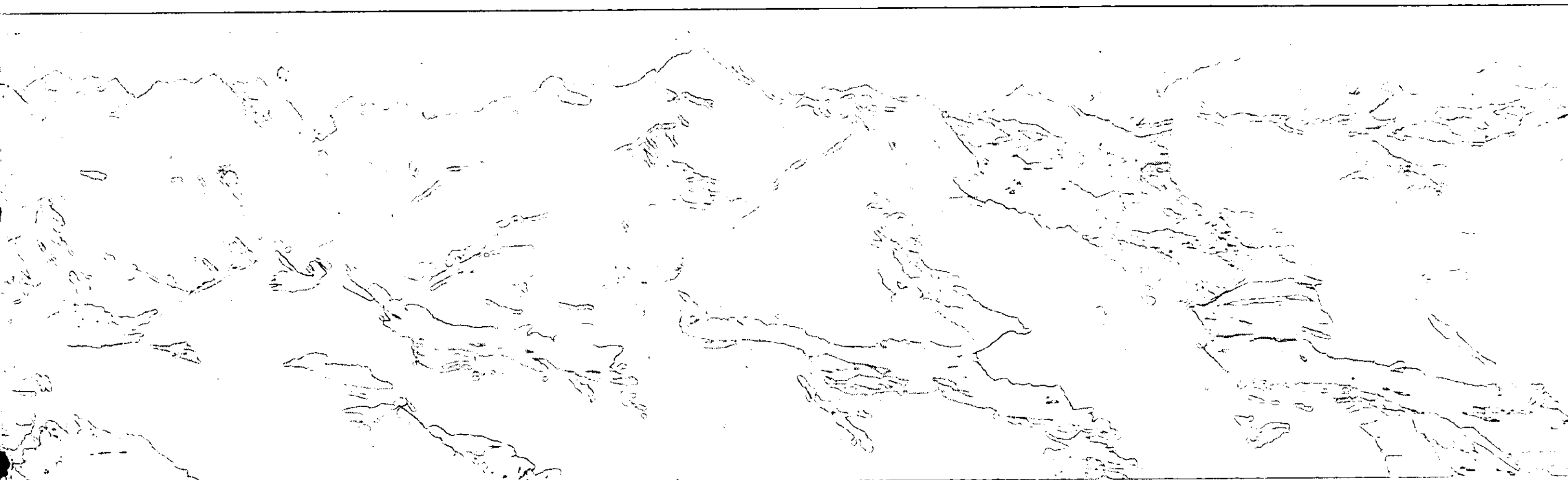
Environment Canada — Environnement Canada

Streamflow Regionalization in British Columbia, No. 2

Regression of Mean Annual Flows on Physiographic Parameters

R. M. Leith





REPORT SERIES NO. 46
(Résumé en français)

INLAND WATERS DIRECTORATE, PACIFIC REGION,
WATER RESOURCES BRANCH,
VANCOUVER, BRITISH COLUMBIA, 1976.

Environment Canada Environnement Canada

Streamflow Regionalization in British Columbia, No. 2

Regression of Mean Annual Flows on Physiographic Parameters

R. M. Leith

REPORT SERIES NO. 46
(Résumé en français)

INLAND WATERS DIRECTORATE, PACIFIC REGION,
WATER RESOURCES BRANCH,
VANCOUVER, BRITISH COLUMBIA, 1976.

Cat. No.: En 36-502/64

ISBN 0-662-00352-7

Contract No. 07KX.KL210-6-4526
THORN PRESS LIMITED

CONTENTS

FIGURES

TABLES

ABSTRACT

Mean annual flows have been regressed on physiographic parameters for 62 hydrometric stations in British Columbia. The equation for unit mean annual flow, UMAF,

$$UMAF = 10.4 + 0.0319\ NPOSI - 0.172 \times 10^{-2}\ ELEV - 0.131 \times 10^{-2}\ DSNW - 0.0567\ RAFOR + 0.909\ RASWP + 0.758 \times 10^{-4}\ SEW$$

explained 61% of the total variation and had a standard error of estimate of 1.71 cfs/sq.mi.

This equation was developed by backward elimination of variables and was tested by examination of residuals, by geographic plot, by runs test and by tests of normalcy.

In the process of building the equation, subsamples of the 62 stations were taken and regression equations developed. This allowed the identification of important parameters, other than those which appear in the above equation. For example, Barrier Height to North, BHN, appeared in five of the seven intermediate equations.

RÉSUMÉ

Les débits annuels moyens ont été régressés selon des paramètres physiographiques à 62 stations hydrométriques de la Colombie-Britannique. L'équation du débit unitaire moyen par année (UMAF),

$$UMAF = 10.4 + 0.0319\ NPOSI - 0.172 \times 10^{-2}\ ELEV - 0.131 \times 10^{-2}\ DSNW - 0.0567\ RAFOR + 0.909\ RASWP + 0.758 \times 10^{-4}\ SEW$$

a pu expliquer 61 % de la variation totale et a comporté une erreur normale d'estimation de 1.71 pcs/mi^2.

On a obtenu cette équation par l'élimination rétrogressive des variables et on l'a éprouvée par l'examen des restes, la mise sur tableaux géographiques, la mise à l'essai des séries et les essais de normalité.

Au cours de l'élaboration de l'équation, on a retenu des sous-échantillons des 62 stations et mis au point des équations de régression, ce qui a permis de déterminer des paramètres importants autres que ceux qui figurent dans l'équation ci-dessus. Par exemple, la hauteur de la barrière vers le nord, BHN, est apparue dans cinq des sept équations intermédiaires.

Streamflow Regionalization in British Columbia, No. 2
Regression of Mean Annual Flows on Physiographic Parameters

R. M. Leith

1. INTRODUCTION

This report describes the second in a series of regionalization studies in British Columbia. Regionalization refers to grouping data in such a manner that analysis benefits by increased accuracy. In this study mean annual flow is regressed against physiographic parameters for 62 basins throughout British Columbia. Mean annual flow is the mean for calendar year flows over the period of record for a particular station where only complete years of record are considered.

2. DISCUSSION

The stations selected for the study met four conditions:

i) recorded natural or near natural flow;

ii) had at least five complete years of record;

iii) recorded flows from basins with drainage areas between 60 and 900 square miles;

iv) recorded flows from basins for which average basin physiographic parameters were available.

With regard to requirement i), near natural flow refers to flow which suffers pondage or other mild artificial changes. An example is Station 08NH004, Goat River near Erickson, which has pondage from 1932 to 1952 for the Erickson Power Plant.

The periods of record are not of equal length nor are they continuous as shown in Appendix 1. Five years was considered a minimum necessary to define the mean annual flow. Mean annual flows include 1973 data where available.

Physiographic parameters were extracted from 1:250,000 topographic maps on a 10 km x 10 km grid and were averaged over each basin. The lower limit in condition iii) ensures that there will be several values to average, and the upper limit restricts the station to a regional sampling.

Regionalization by regression provides a tool for evaluating the existing network in terms of standard error of estimate. The procedure also identifies significant physiographic parameters and provides an estimate of network density required to achieve accuracy goals.

However, a regression equation should be examined analytically. An F-value will indicate whether an equation is a good predictor; the square of the multiple correlation coefficient shows the extent to which the fitted equation explains the variation in data; the standard error of estimate or the square root of the residual mean square will provide a measure of the accuracy of the prediction; and residuals contain detailed information about the relation of the regression model to the data. In this report residual is defined as observed value of unit mean annual flow minus the regression estimate.

If this model is correct, the residuals represent observed errors. In applying regression technique, assumptions are made that errors are independent, have zero mean, a constant variance and follow a normal distribution.

The regression equation should be tested for lack of fit, to see if the model is complete, but this requires replication of data (Reference 2, page 28).

Skew, kurtosis and probability of error of residuals provide indicators of normalcy (Reference 1). A geographical plot of residuals and a runs test examine spatial independence of residuals (Reference 2). Plots of residuals against other parameters such as calculated mean annual flow or drainage area examine the equation for bias. These tests of residuals provide indicators as to the applicability of the regression equation to ungauged basins and whether the study region should be subdivided.

3. PROCEDURE

After the stations had been selected and the mean annual flows computed, several random samples of stations were taken; four of 16 stations, two of 32 stations and one of 48 stations. Finally, a full 62-station sample was considered. For each sample a regression equation was developed for unit mean annual flow, that is, mean annual flow divided by the drainage area. This transformation was made to remove spurious correlation of flow with drainage area.

The equations were developed by consideration of correlation matrix of unit mean annual flow and physiographic parameters. Those physiographic parameters which correlated highly with unit mean annual flow were selected and examined for intercorrelations. If parameters were highly intercorrelated, the parameter with the highest F-value was selected as a possible regression variable. Several correlation runs were made to check on a parameter being highly correlated with a combination of other parameters. Scatter plots of the parameters were examined to see if any transformations were indicated.

Equations were developed using the backward elimination procedure of TRIP, a Triangular Regression Package of the University of British Columbia. Checks were made on eliminated parameters to see if they had become significant at a later step. The variables in the final equations were significant at a 0.05 level, in other words, there is a 95% chance of their being significant.

After an equation was developed, it was applied to the stations not used in its development. This is called the validation process. The standard error of validation was derived from the residuals of the validation process.

The final equation for 62 stations was tested by examining skew and kurtosis of the residuals, a geographical plot of residuals, plots of residuals against observed unit flows and drainage areas. The geographical distribution of residuals was further examined by runs tests.

Twenty-four stations which had not met the conditions for selection as development stations were used not as a test but rather as a flexibility indicator for the equation.

4. RESULTS AND OBSERVATIONS

The buildup of the 62-station equation is shown in Table 1.

TABLE 1. Regression Equations for Unit Mean Annual Flow (UMAF) in cubic feet per second per square mile

16-Station Samples

1st $R^2 = 0.819$ S.E.E. = 1.07 cfs/sq.mi.

$$UMAF = 4.74 + 0.624\ RALKE - 0.192 \times 10^{-4}\ SENW$$

2nd $R^2 = 0.336$ S.E.E = 2.47 cfs/sq.mi.

$$UMAF = 3.06 + 2.41\ RASWP$$

3rd $R^2 = 0.711$ S.E.E. = 1.50 cfs/sq.mi.

$$UMAF = 0.678 + 2.43\ RASWP + 0.986 \times 10^{-3}\ BHN$$

4th $R^2 = 0.839$ S.E.E. = 1.30 cfs/sq.mi.

$$UMAF = -1.25 + 0.043\ SLP\% - 0.464 \times 10^{-2}\ DSSW + 0.113\ BHN$$

TABLE 1. (continued)

32-Station Samples

1st $R^2 = 0.715$ S.E.E. = 1.36 cfs/sq.mi.

$$UMAF = 9.00 + 0.0315\ NPOSI - 0.135 \times 10^{-2}\ ELEV - 0.0927\ RAFOR + 0.706 \times 10^{-3}\ BHN$$

2nd $R^2 = 0.566$ S.E.E. = 1.82 cfs/sq.mi.

$$UMAF = 7.78 + 0.0446\ NPOSI - 0.151 \times 10^{-2}\ ELEV - 0.108\ RAFOR + 0.988 \times 10^{-3}\ BHN$$

48-Station Sample

$R^2 = 0.556$ S.E.E. = 1.72 cfs/sq.mi.

$$UMAF = 4.83 - 0.0388\ RAFOR + 0.963\ RASWP + 0.752 \times 10^{-3}\ BHN - 0.984 \times 10^{-5}\ SENW - 0.0298\ SSE$$

62-Station Sample

$R^2 = 0.611$ S.E.E. = 1.71 cfs/sq.mi.

$$UMAF = 10.4 + 0.0319\ NPOSI - 0.172 \times 10^{-2}\ ELEV - 0.131 \times 10^{-2}\ DSNW - 0.0567\ RAFOR + 0.909\ RASWP + 0.758 \times 10^{-4}\ SEW$$

A guide to the abbreviations and an explanation of physiographic parameters is provided in Appendix 2. Regression coefficients are given to three significant figures, but some input data is to two significant figures, for example, relative area of forests.

Figure 1 shows standard error of validation as a function of number of samples used in developing the equation. Quantitatively, extrapolation of this graph is dangerous in that numerical estimates would be unreliable, but qualitatively it does appear worthwhile in terms of a decrease in standard error to increase the number of stations used in the analysis.

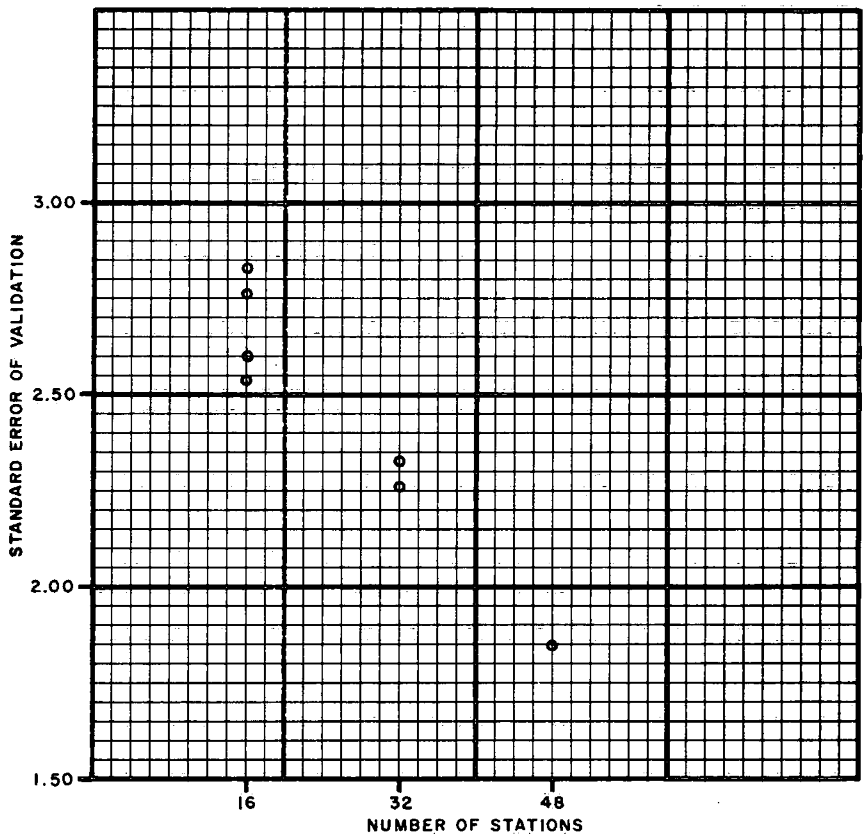


Figure 1. Standard error of validation as a function of number of stations used in developing the regional equation.

Results of the 62-station equation residuals are shown in Table 2 and Figures 2 and 3. The values of the coefficients of skew and kurtosis and the probability plot (Figure 2), are strong indications of the normalcy of the residuals. On the probability plot there are two outliers, Stations 08HA001 Chemainus River near Westholme and 08NF001 Kootenay River at Kootenay Crossing. The physiographic parameters for both stations seem reasonable when checked against those for neighbouring stations and both have reasonably long periods of record of the order of 20 years, so that mean annual flow should be well defined.

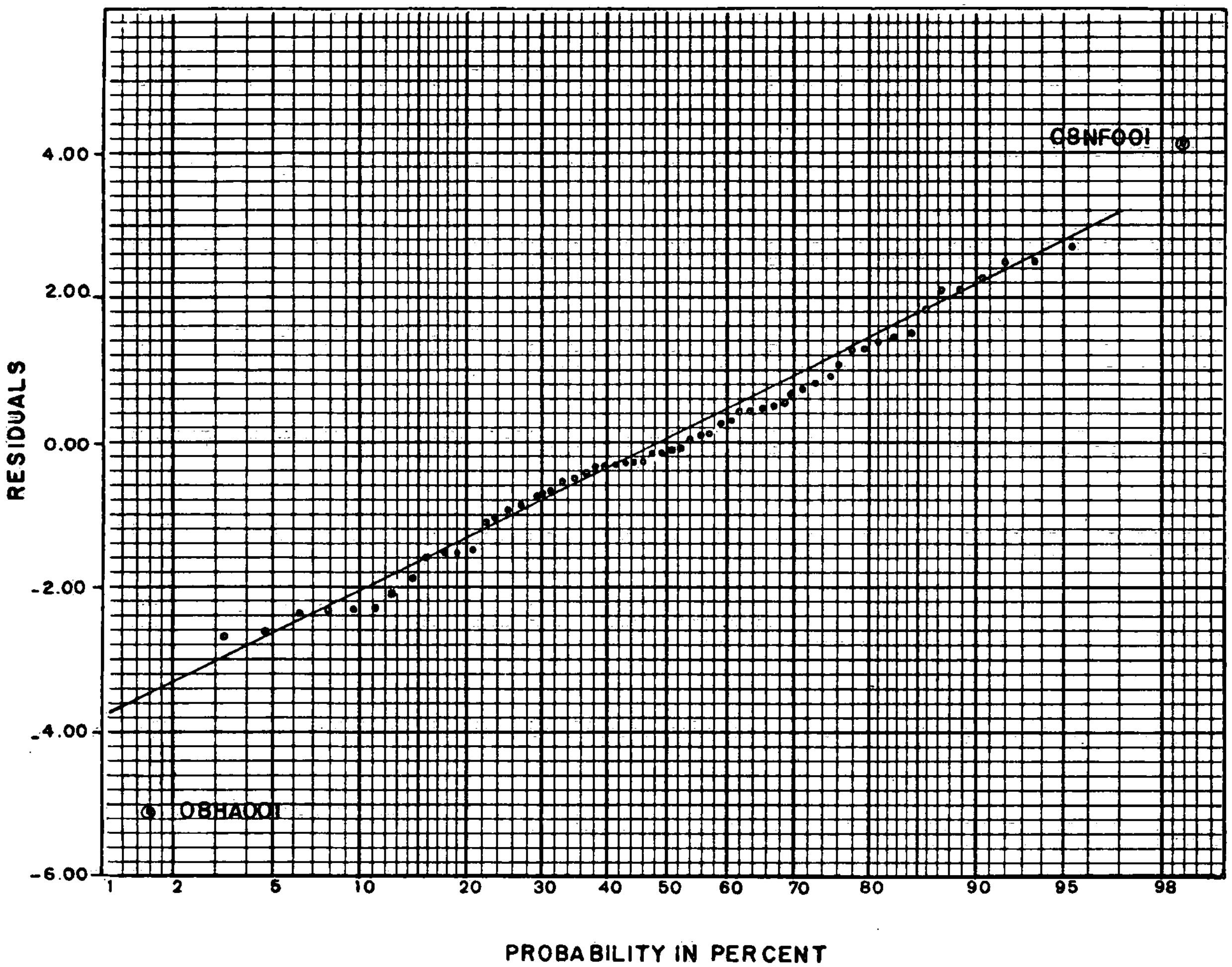


Figure 2. Probability plot of residuals for 62-station equation.

TABLE 2. Test Results for Regional Regression Equation

Coefficient of Skew	-0.0140	For a normal sample	0
Coefficient of Kurtosis	3.56	For a normal sample	3
Number of Stations		$n = 62$	
Number of Positive Residuals		$n_1 = 29$	
Number of Negative Residuals		$n_2 = 33$	
Number of Runs		$u = 27$	
Normal Deviate		$z = -1.08$	

Probability of obtaining a normal deviate of -1.08 or less is 0.14.

Note: Definitions of the terms used are provided in Appendix 3.

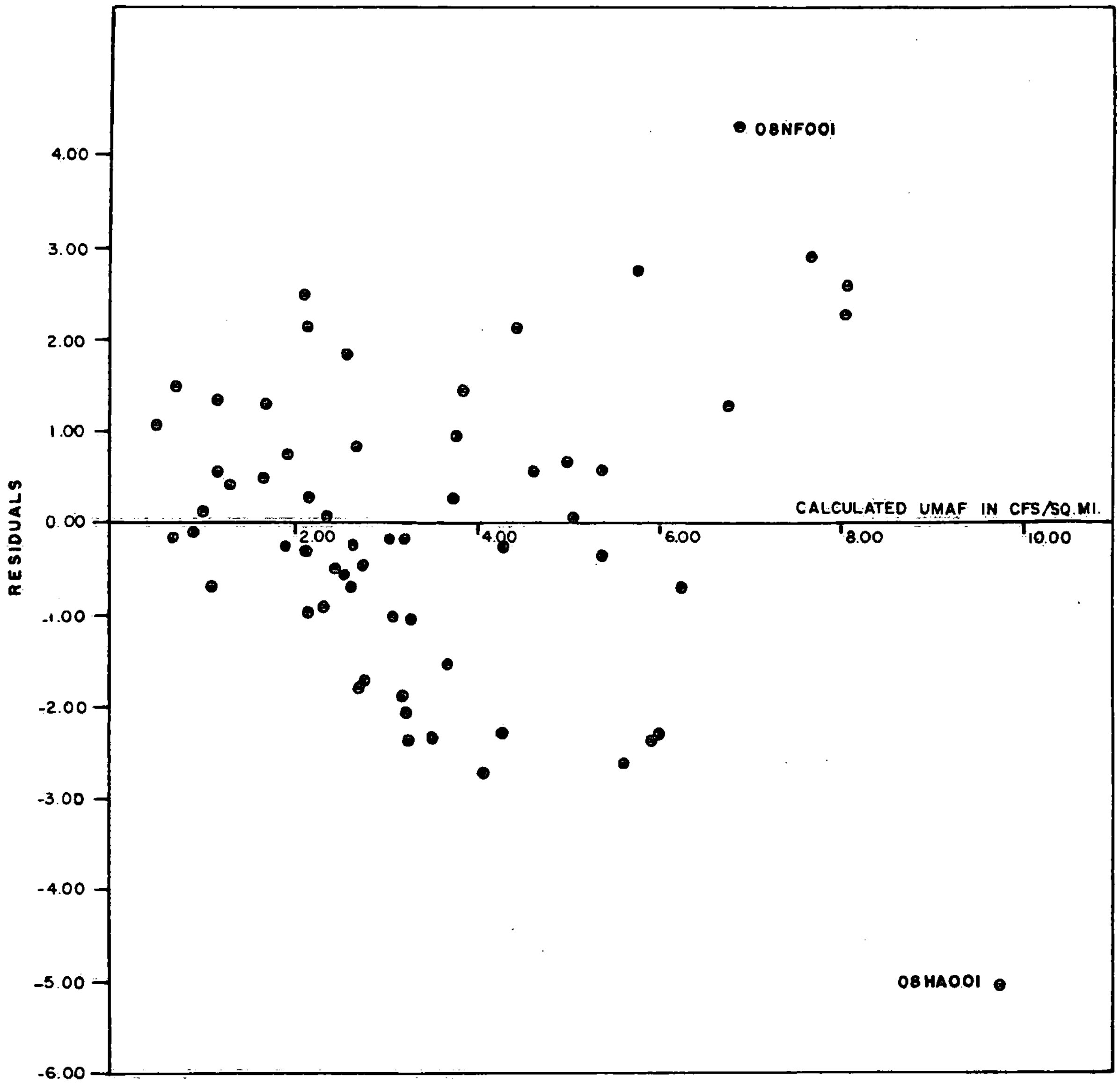


Figure 3. Residuals for the 62 stations plotted against calculated mean annual flows.

Figure 3 indicates possibility of bias, large residuals occur with large unit mean annual flows. If the two outliers, 08HA001 Chemainus River near Westholme and 08NF001 Kootenay River at Kootenay Crossing are removed, the assumption of uniform variance appears better.

Figure 4 provides a geographical plot of the residuals. The runs test was made by following the lines connecting the stations. Although there are many negative residuals in the southeast corner of the province, statistical tests indicated no significant concentration, so that subdivision of the province was not attempted.

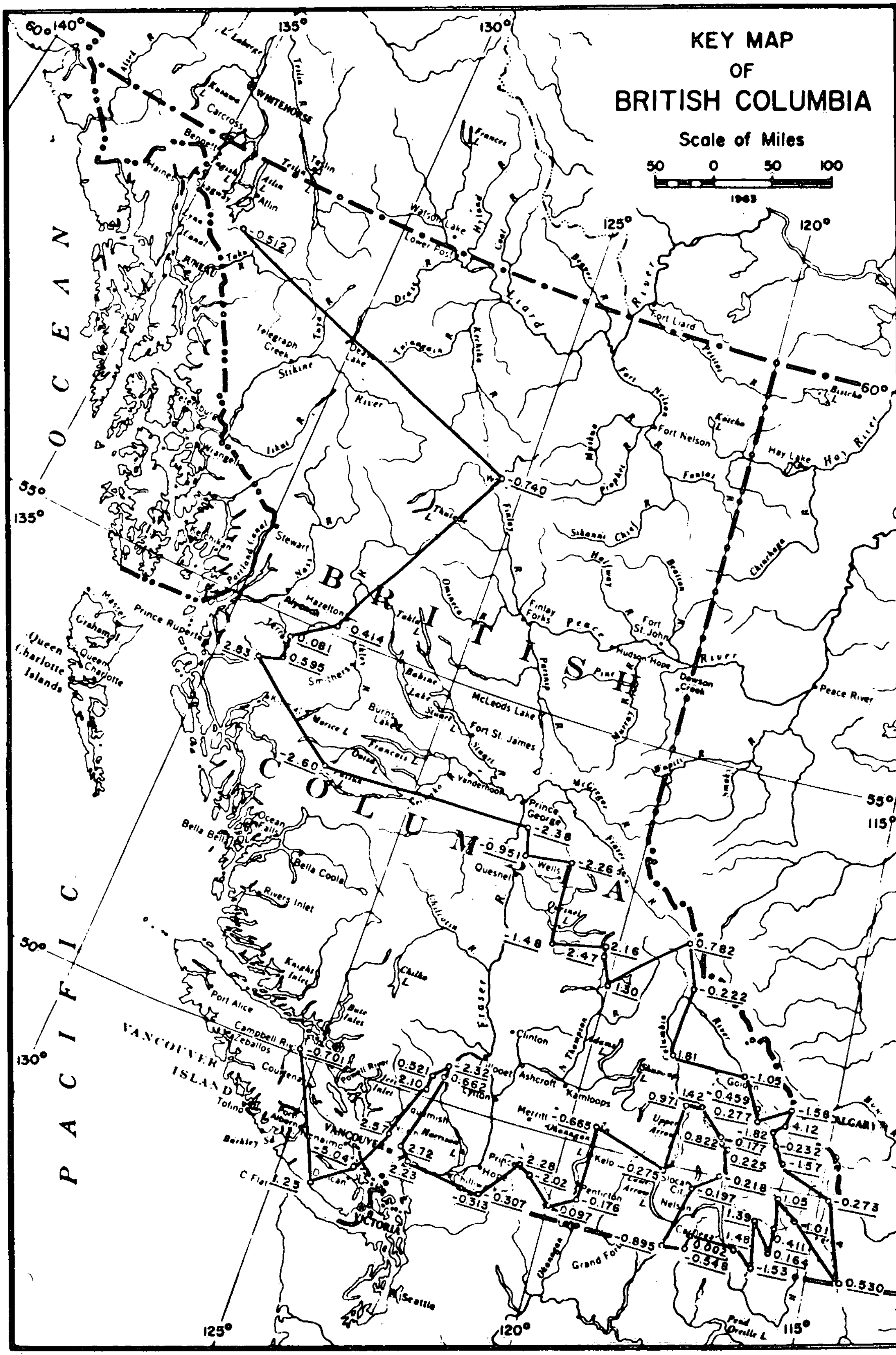


Figure 4. Geographical plot of residuals for 62-station equation.

TABLE 3. Flexibility Indicator

Station		Unit Mean Annual Flow Obs.	Unit Mean Annual Flow Cal.	Residual	Area Sq.Mi.	Years of Record	Flow Type*
08NH084	Arrow Creek	2.354	1.799	0.555	30	10	N
08NE073	Blueberry Creek	1.083	2.133	-1.05	56	6	N
08NC002	Canoe River	2.896	3.226	-0.330	1350	7	N
08GA031	Capilano River	10.500	5.185	5.31	76	25	R
08JC005	Chilako River	0.345	5.252	-4.91	1310	12	N
08LA007	Clearwater River	4.254	2.172	2.08	1140	16	N
08LA001	Clearwater River	2.015	1.966	0.049	4320	32	N
08NK012	Elk River	1.707	3.911	-2.20	1360	25	N
08KA004	Fraser River	2.436	3.221	-0.785	6950	20	N
08MG004	Green River	5.309	3.132	2.18	55	26	N
08NM053	Kelowna Creek	0.215	6.272	-6.06	85	6	R
08NL010	Keremeos Creek	0.346	0.948	-0.602	67	21	R
08KB003	McGregor River	4.489	3.436	1.05	1840	13	N
08LA008	Mahood River	0.646	2.604	-1.96	1820	11	N
08NM116	Mission Creek	0.627	4.409	-3.78	312	7	R
08NM016	Mission Creek	0.725	4.774	-4.05	240	4	R
08NL023	Otter Creek	0.489	4.804	-4.31	260	23	R
08KH001	Quesnel River	1.913	2.418	-0.505	2290	39	N
08KH006	Quesnel River	1.761	2.070	-0.309	4450	28	N
08MG006	Rutherford Creek	6.581	5.132	1.45	62	23	N
08GA013	Seymour River	10.193	4.871	5.32	59	18	R
08MH056	Slesse Creek	5.710	2.301	3.41	63	12	N
08NJ013	Slocan River	2.424	2.766	-0.342	1270	48	N
08EF003	Zymoetz River	4.058	3.963	0.095	1200	9	N

* R - Regulated
  N - Natural

Table 3 shows the results when the 62-station equation was applied to basins which did not meet the flow requirements in Section 2. The equation does not work well on regulated basins or on basins with drainage areas less than 60 square miles, but does work well on large basins. Basins slightly larger than those used in developing the equation sometimes agree well and sometimes deviate markedly, for example, Chilako River near Prince George and Zymoetz River near Terrace.

5. CONCLUSIONS AND RECOMMENDATIONS

The results indicate regionalization by regression on physiographic parameters does increase accuracy in that the standard error of estimate (1.71) for unit mean annual flow is less than the standard deviation (2.61) for the 62-station sample. The change in accuracy with network density cannot be properly assessed with the sample size used in this study as the number of stations in the validation samples is limited. It is obvious that the number of stations in the analysis should be greater than 62.

A larger sample would also allow examination of subdivision of the province. More stations would be available if a smaller grid were used, so the study should be repeated when the 2 km x 2 km British Columbia data file becomes operational.

The important physiographic parameters were identified not just by their occurrence in one equation but through the frequency of occurrence in the equation buildup. Now that the significant parameters have been identified, the sampling of those parameters by the network of stations can be examined.

Appendix 4 provides a guide to the use of the regional equation on an ungauged basin. The basin should have average physiographic parameters within limits provided.

Even then the equation may not provide good results, reference the two outliers of the residual study. Regression is a statistical technique and provides best answers in the centre of the range of values. This is why the large basins gave good results in the flexibility examination.

REFERENCES

1. Solomon, S.I. Chapter 12, Parameter Regionalization and Network Design. Proceedings of the Institute on Application of Stochastic Methods to Water Resource Problems. Colorado State University, Fort Collins, Colorado, 1975.

2. Draper, N.R., and Smith, H. Applied Regression Analysis. John Wiley and Sons Inc., 1966.

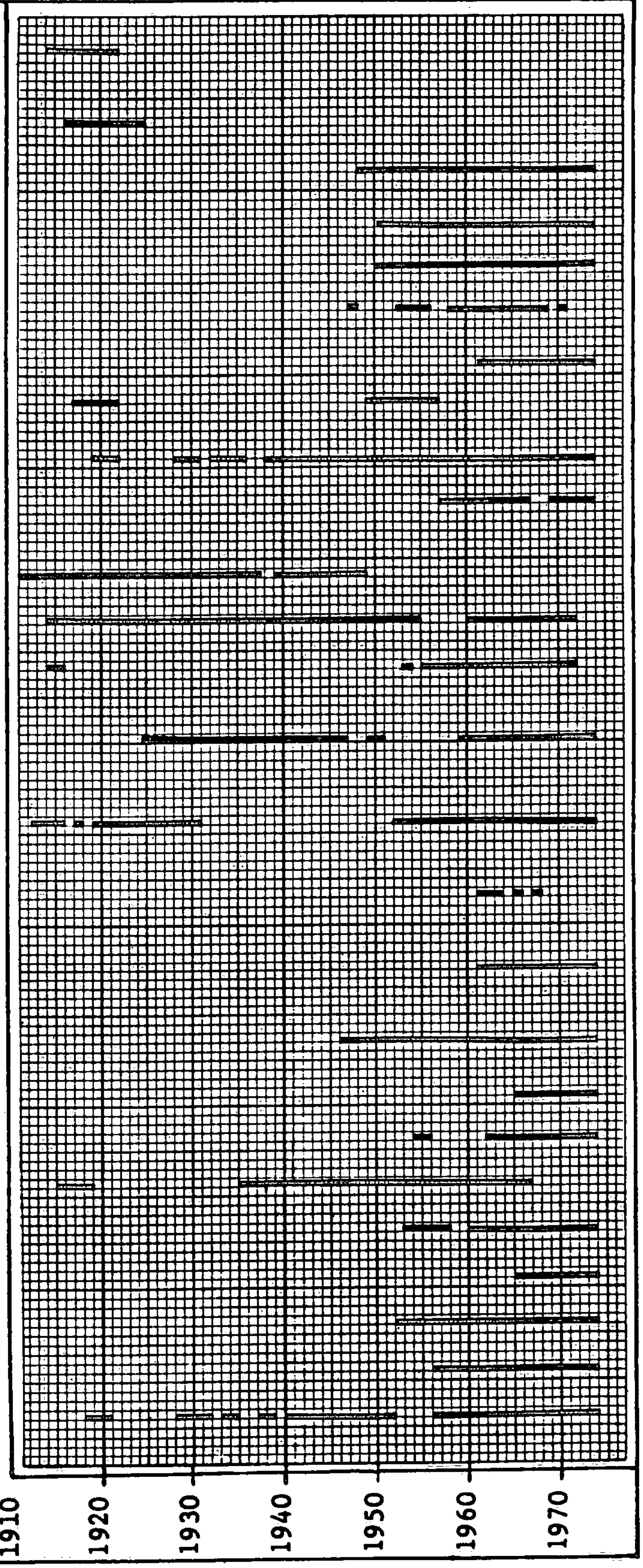
3. Leith, R.M. Streamflow Regionalization in British Columbia, No. 1. Regression of Mean Annual Floods on Physiographic Parameters. Report Series No. 40. Inland Waters Directorate, Pacific Region, Water Resources Branch, Vancouver, B.C., 1975.

4. Water Survey of Canada. Historical Streamflow Summary British Columbia to 1973. Inland Waters Directorate, Water Resources Branch, Ottawa, 1974.

5. Water Survey of Canada. Surface Water Data Reference Index Canada 1974. Inland Waters Directorate, Water Resources Branch, Ottawa, 1975.

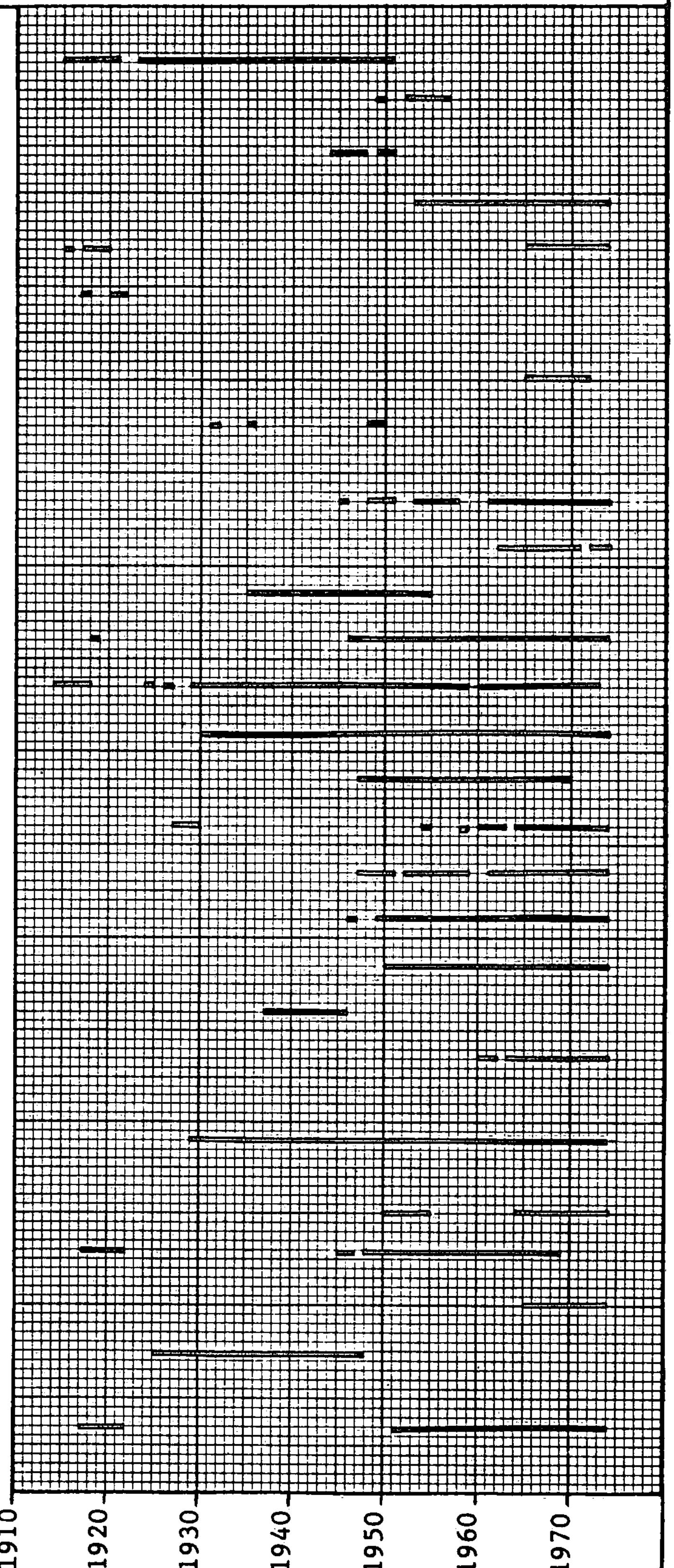
APPENDIX 1. Periods of Record for Stations Used in Developing Equations

Station		Period of Record
08ND001	Akolkolex River near Revelstoke	
08NM014	Alouette River at Outlet of Alouette Lake	
08NL004	Ashnola River near Keremeos	
08NE077	Barnes Creek near Needles	
08NE039	Big Sheep near Rossland	
08MG008	Birkenhead River at Mount Currie	
08KD001	Bowron River near Wells	
08NA001	Bugaboo Creek near Spillimacheen	
08NG002	Bull River near Wardner	
08KE015	Cale Creek near Red Rock	
08HD001	Campbell River at Outlet of Campbell Lake	
08GA010	Capilano River above Intake	
08HA001	Chemainus River near Westholme	
08MH016	Chilliwack River at Outlet of Chilliwack Lake	
08MH001	Chilliwack River at Vedder Crossing	
08LA009	Clearwater River at Inlet to Clearwater Lake	
08LA013	Clearwater River at Outlet of Hobson Lake	
08NA045	Columbia River near Fairmont Hot Springs	
08KE009	Cottonwood River near Cinema	
08ND009	Downie Creek near Revelstoke	
08NH001	Duncan River near Howser	
08NK016	Elk River near Natal	
08EG012	Exchamsiks River near Terrace	
08NP001	Flathead River at Flathead	
08KA007	Fraser River at Red Pass	
08NH004	Goat River near Erickson	



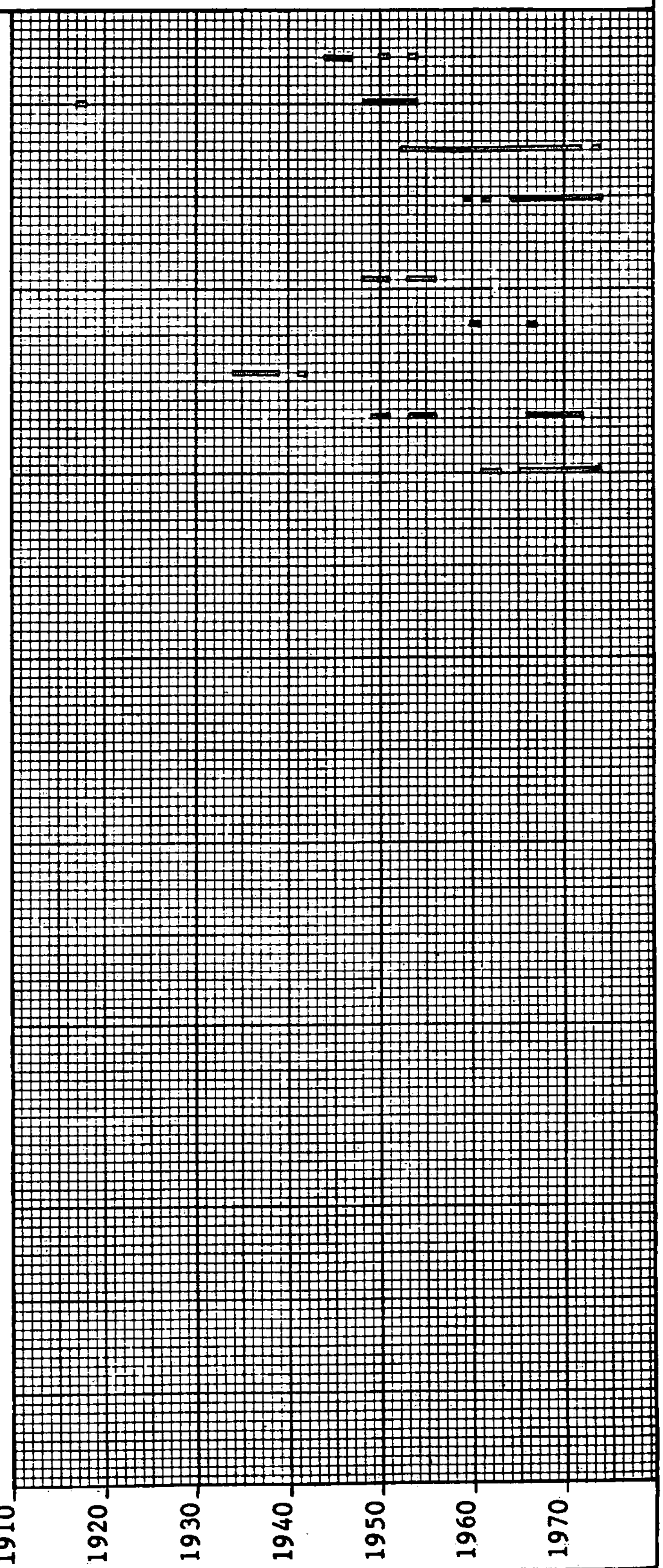
APPENDIX 1. (cont'd)

Station		Period of Record
08MG003	Green River near Pemberton	
08KH007	Horsefly River at Horsefly	
08NA005	Horsethief Creek near Wilmer	
08NE001	Incomappleux River near Beaton	
08NH005	Kaslo River below Kemp Creek	
08NA006	Kicking Horse River at Golden	
08EF004	Kitseguecla River near Skeena Crossing	
08EG006	Kitsumkalum River near Terrace	
08NF001	Kootenay River at Kootenay Crossing	
07EA002	Kwadacha River near Ware	
08NH066	Lardeau River at Gerrard	
08NH007	Lardeau River at Marblehead	
08MG005	Lillooet River near Pemberton	
08NH006	Moyie River at Eastport	
08NH034	Moyie River at Moyie	
08LA004	Murtle River above Dawson Falls	
08NG012	St. Mary River at Wycliffe	
08NG046	St. Mary River near Marysville	
08NE074	Salmo River near Salmo	
08NE044	Salmo River near Waneta	
08HA010	San Juan River near Port Renfrew	
08GA030	Seymour River near North Vancouver	
08NG051	Skookumchuck Creek near Skookumchuck	
08NJ014	Slocan River at Slocan City	
08BB002	Sloko River near Atlin	
08MG007	Soo River near Pemberton	
08NA011	Spillimacheen River near Spillimacheen	



APPENDIX 1. (cont'd)

Station		Period of Record
08NA012	Toby Creek near Athalmer	
08NL008	Tulameen River at Coalmont	
08NL024	Tulameen River at Princeton	
08NM015	Vaseux Creek above Dutton Creek	
08NF004	Vermilion River near Radium Hot Springs	
08NM046	Whiteman Creek near Vernon	
08JA003	Whitesail River near Ootsa Lake	
08NC001	Wood River near Donald	
08EG011	Zymagotitz River near Terrace	



APPENDIX 2. Physiographic Parameters

Parameters	Symbols	Units	Explanation
Drainage Area	AREA	Square Miles	Total drainage area for the basin
Grid Coordinates	I, J	Dimensionless	Coordinates for the centre of gravity of the basin
Elevation	ELEV	Feet	Average elevation of the basin
% Slope x 10	SLP%		Basin slope averaged over the squares included in the basin
Azimuth	SLPAZ	Degrees	Angle between the west-east direction and the horizontal projection of the line of steepest descent of the local slope plain
Distance to Sea			Distance from centre of gravity of basin to the sea in the north, the northwest, west, and southwest directions
North	DSN		
Northwest	DSNW		
West	DSW	Kilometres	
Southwest	DSSW		
Relative Area			Percentage of the area of the basin occupied by lakes, forests, swamp, glaciers and built-up areas. Note: ΣRA does not always equal 100.
Lake	RALKE		
Forest	RAFOR		
Swamp	RASWP	Dimensionless	
Glacier	RAGLC		
Urban	RAURB		
Barrier Height			Difference between average elevation of the basin and highest elevation encountered in the north, northwest, west and southwest directions until the ocean is reached
North	BHN		
Northwest	BHNW		
West	BHW	Feet	
Southwest	BHSW		

APPENDIX 2. (cont'd)

Parameters	Symbols	Units	Explanation
Shield Effect			Sum of elevation differential of all ascending stretches of terrain encountered when travelling from ocean shore at north, northwest, west, southwest directions to corresponding point
North	SEN		
Northwest	SENW		
West	SEW	Feet	
Southwest	SESW		
Signed Slope			Takes into account general configuration of the terrain
Northeast	SSNE		
East	SSE	Feet/Kilometre	
Southeast	SSSE		

Note:

Further information and references on these parameters may be found in Hydrometric Network Planning Study for Western and Northern Canada Report 5019-1-70 November 1970 by the Shawinigan Engineering Company Limited, Section 4.2.1. page 33.

APPENDIX 3. Definitions of Statistical Terms

coefficient of skew $\frac{N}{N-1} \frac{S3}{(S2)^{1.5}}$

coefficient of kurtosis $\frac{(N)^3}{(N-1)(N-2)(N-3)} \frac{S4}{(S2)^2}$

where N is the number of items in sample.
S2 is the sum of the square of the residuals.
S3 is the sum of the cube of the residuals.
S4 is the sum of the fourth power of the residuals.

correlation matrix

the matrix of correlation coefficients of the physiographic parameters and unit mean annual flow produced by TRIP.

runs and runs test

for the purpose of runs tests, a run is defined as the number of sign changes in residual along a path connecting stations. Referring to Figure 4, -0.512, -0.740, 0.414, 0.081, 0.595, 2.83, -2.60 has 3 runs.

for the runs test n = number of stations
n_1 = number of positive residuals
n_2 = number of negative residuals
u = number of runs

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1$$

$$\sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

$$z = (u - \mu + 0.5)/\sigma$$

standard error of estimate

the square root of the sum squared about regression divided by its degrees of freedom.

$$\left[\frac{\sum_{i=1}^{N} (\text{RESIDUAL})_i^2}{N - M - 1}\right]^{\frac{1}{2}}$$

N is the number of observations used in establishing equation.

M is the number of independent variables in the regression equation.

standard error of validation

same as standard error of estimate except N is the number of observations used in the validation sample.

F-value

$$F = \frac{\text{mean square due to regression}}{\text{mean square due to residual variation}}$$

$= t^2$ for single independent variable. This statistic tests significance of regression coefficient b_i.

APPENDIX 4. Statistics for the Significant Physiographic Parameters

	Mean	Standard Deviation	Maximum	Minimum
AREA	399	256	940	62
NPOSI	170	32	212	62
ELEV	4727	1410	7017	1136
DSNW	2349	1152	3450	226
DSW	615	266	990	64
RAFOR	73	20	99	11
RASWP	0.19	0.57	2	0
SEW	37432	21026	71300	2070
SESW	60488	43901	128200	1850
UMAF	3.46	2.60	10.99	0.37