APPLICABILITY OF CANONICAL CORRELATION IN HYDROLOGY

by

PADOONG TORRANIN

November 1972

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
Fort Collins, Colorado

58

APPLICABILITY OF CANONICAL CORRELATION IN HYDROLOGY

by

Padoong Torranin*

HYDROLOGY PAPERS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO 80521

November 1972 **No. 58**

*Post-Doctoral Research Associate, Department of Civil Engineering, Colorado State University, Fort Collins, Colorado.

TABLE OF CONTENTS

CHAPTER PAGE

ACKNOWLEDGEMENTS . . . iv

ABSTRACT . . . iv

PREFACE . . . iv

I INTRODUCTION . . . 1

1.1 Application of Multivariate Analysis in Hydrology . . . 1
1.2 Relevance of Canonical Correlation Analysis in Hydrologic Investigation . . . 2
1.3 Objective of the Study . . . 3
1.4 Selection of Sets of Dependent Variables for the Two Examples of Long-range Prediction . . . 3

II MATHEMATICAL TECHNIQUES USED IN THE ANALYSIS . . . 5

2.1 Autocorrelation Analysis . . . 5
2.2 Model of Sequentially Dependent Time Series . . . 5
2.3 Canonical Correlation Analysis . . . 5

III ASSEMBLY AND PROPERTIES OF DATA . . . 9

3.1 Data for the Analysis of Precipitation Forecast . . . 9
Precipitation . . . 9
Sea surface temperature of the Pacific Ocean . . . 11

3.2 Data for the Analysis of Snowmelt Runoff Forecast . . . 12
Snowmelt runoff . . . 12
Method of computation of indices of snowmelt runoff . . . 12
Fall and winter precipitation index . . . 15
Snow water equivalent index . . . 15

IV APPLICATION OF CANONICAL CORRELATION . . . 17

4.1 Results of Analyses of Historical Data . . . 17
Coastal precipitation forecast . . . 17
Snowmelt runoff forecast . . . 17

4.2 Examples of Forecast by Using Canonical Correlation Analysis . . . 18
Coastal precipitation forecast . . . 19
Snowmelt runoff forecast . . . 20

V CONCLUSIONS . . . 23

BIBLIOGRAPHY . . . 24

APPENDIX A - Canonical Correlation Analysis . . . 25

APPENDIX B - Precipitation Stations Selected . . . 29

APPENDIX C - List of Selected Symbols . . . 30

ACKNOWLEDGEMENTS

The material in this paper is a portion of Ph.D. dissertation submitted by the writer to Colorado State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The research leading to the dissertation financially sponsored by the U.S. National Science Foundation under grant GK-11564 (Large Continental Droughts). This financial support and the graduate research assistantship that made possible the writer's studies are gratefully acknowledged.

The writer appreciates the suggestions, comments and the encouragement to conduct this research given by his major professor and advisor, Dr. Vujica Yevjevich, Professor of Civil Engineering. Special thanks are expressed to Dr. Mohammed M. Siddiqui, Professor in the Department of Mathematic and Statistics; his suggestions and guidance concerning the statistical part of the research are highly appreciated.

ABSTRACT

The potential for application of canonical correlation analysis to hydrologic problems is demonstrated by two problems in long-range hydrologic prediction: (1) forecast of monthly precipitation of three large areas of the West Coast of the United States, and (2) forecast of seasonal snowmelt runoff for three gaging stations in the Flathead River Basin in Montana.

Canonical correlation analysis is found to be effective in investigating linear correlation between two or more three-dimensional hydrologic processes, in which the set of time series of each process are mutually correlated, in addition to a relatively high correlation between the processes themselves. The main advantages of using this technique concern the significance testing of the linear correlation between the processes, the reduced effort in the correlation analysis, and particularly for the prediction problem as it concerns the construction of a confidence region of the simultaneous predicted values. Though not demonstrated in the examples, canonical correlation analysis can also be used for selecting significant data observation stations for use in the correlation analysis.

A set of forecasts is made for each prediction problem by using the canonical correlation analysis of the historical data. Results of these forecasts indicate that the precipitation prediction is not reliable, while the runoff due to seasonal snowmelt can be well predicted.

PREFACE

In hydrology, most realistic relationships involve a large number of random variables, since a process in three or four dimensions must often be related to one or more processes in three or more dimensions. As a consequence, the multivariate distributions and analyses of sets of hydrologic random variables represent the best approach in deriving hydrologic relationships of a probabilistic type. There are several types of multivariate analyses that may be suitable for deriving these relations. Currently, the technique most used in hydrology is the multiple regression and correlation analysis, mainly for prediction purposes. Many cases of application of principal components analysis in treating multivariate hydrologic problems are also available in the literature. Multivariate factor analysis has been tried on several problems with a relatively limited success. When a set of mutually correlated variables must be related to another set of mutually dependent random variables, analysis by canonical correlation seems to represent the most suitable multivariate technique.

The Ph.D. dissertation by Padoong Torranin explores the feasibility of using canonical correlation analysis to establish relationships between two sets of random variables which are not only correlated among the sets, but also dependent within each set. This case occurs frequently in hydrology. Although the two examples selected for this study treat only problems of the prediction type, the potential application of canonical correlation in hydrology transcends the application for forecasting purposes. The results of the study show that a good potential exists for this technique to be applied in various areas of hydrology.

The study has been carried out under the research project "Large Continental Droughts," sponsored by the U. S. National Science Foundation, Grant No. GK-11564, at Colorado State University, Department of Civil Engineering, Graduate and Research Hydrology and Water Resources Program. One research aspect of this project is an inquiry into the predictability of large continental droughts. Because droughts are slowly evolving natural disasters, long range prediction in hydrology, say over several months or years, seems not to be feasible except in the case of snow and water already accumulated on the ground. Large continental droughts of long duration, given severity and large areal coverage fall into the category of deterministically unpredictable hydrologic phenomena, except in exceptional cases of already accumulated snow, underground and/or surface water in river basins. Application of canonical correlation analysis in this study represents an attempt not only to analyze the potential of this technique, but also to obtain information on long-range hydrologic prediction as it is related to droughts. There is a need to throw more light on whether large droughts are a predictable or an unpredictable phenomenon, in the classical sense of deterministic hydrologic predictions.

It is expected that this study will give an impetus to other trials and a fair chance for the further application of canonical correlation analysis in hydrology. This analytical method needs to be tested in various hydrologic problems for which the relationships of mutually dependent sets of random variables are required.

Vujica Yevjevich
Professor-in-Charge of
Hydrology and Water
Resources Program
Department of Civil
Engineering
Colorado State University
Fort Collins, Colorado

October 1972

CHAPTER I

INTRODUCTION

This chapter briefly explains different forms of multivariate analysis and their uses in hydrology. Potential of applications of canonical correlation analysis in hydrology are reviewed. Objective of this study is described along with general approach to accomplish it.

1.1 Application of Multivariate Analysis in Hydrology

Multivariate analysis as a statistical approach for the investigation of the relation within a set (or among several sets) of random variables is not a new development. In fact, one such method was originated in the early 1900s in the form of principal components analysis by Karl Pearson. However, an attempt at more effective application of the multivariate analysis to hydrology was made by W. M. Snyder in 1962 (Synder (1962)). He singled out some properties of multivariate analysis which may be advantageously used in hydrology. Besides the favorable statistical properties associated with various forms of multivariate analysis, one very useful property is that they allow an investigation of a hydrologic phenomenon simultaneously at many locations. Regional investigation of a hydrologic phenomenon, or finding relationship among hydrologic phenomena on a regional basis, can be made conveniently by the multivariate analysis approach.

Most of multivariate analysis may be considered counterparts of univariate statistical methods commonly used in hydrology. The mean and variance of a single random variable in univariate analysis are replaced by a vector of means and a matrix of covariances of the corresponding vector of random variables in multivariate analysis. Besides the well used multiple correlation analysis, three other multivariate analyses are applied in hydrology with varying degrees of frequency and/or success. These include principal components analysis, factor analysis, and canonical correlation analysis. Basically, each of these analyses involves a linear transformation of the original set or sets of random variables into new ones such that the transformed variables have certain required properties.

In the principal components analysis the transformed variables, called the principal components, are mutually linearly uncorrelated. Each of these variables has a maximized variance, arranged from highest to lowest. Compared to the number of original variables, fewer of the principal components explain a high percentage, say 90 to 95 percent, of the variances of original variables.

Instead of maximizing the variance of each set of components, the canonical analysis linearly transforms the two sets of random variables, where the variables in each set may be mutually correlated, into the two sets of transformed variables, called canonical variables, in such a way that pairwise linear correlations between certain pairs of the two sets of canonical variables are maximized. By those transformations, the canonical variables of each set become mutually uncorrelated, while each of them becomes uncorrelated with all the canonical variables of other set except for the one variable with which it has a maximized correlation.

The principal components and factor analysis are somewhat related because one may be considered as an approach to the problem in the opposite direction of the other. In order to avoid the problem of physical interpretations of the derived principal components, a factor analysis may be used. A small number of physical factors related to the set of random variables are proposed such that each random variable can be expressed as a function of these factors. If the factors are selected arbitrarily from the physical properties of a problem, the factor analysis is usually considered as a subjective approach. However, the principal components analysis has been used in assisting with the identification of factors in a method of factor analysis called Varimax, proposed by Kaiser (1958). This method modifies the derived principal components into factors in such a way that each factor is uncorrelated with the others, and is highly related to only a few of the original random variables. Each of these factors expressed only some particular attribute of the set of original random variables. Therefore, they perform the function which the proposed subjective factors were set out to do, that is, to physically represent some joint properties of the original set of random variables.

Since the introduction of the multivariate analysis to hydrology most of the applications involve the use of the principal components and factor analysis. The purpose of most of the applications was to use the analysis to arrive at a new set of random variables which has some required statistical properties suitable for further analysis. One such application would be to find a new set of mutually uncorrelated random variables to be used as a set of independent variables in a multiple correlation analysis (Snyder (1962), Anderson and Westl (1965), Eiselstein (1967), Diaz, Sewell, and Shelton (1968), Marsden and Davis (1968), Veitch and Shepherd (1971)). Another application is in economizing the analysis concerned with a large number of random variables that are mutually correlated. The principal components or factor analysis are used to derive a smaller number of transformed random variables which have a high percentage of the variation of the set of original random variables (Dawdy and Feth (1967), Nimmannit and Morel-Seytoux (1969)). Another interesting field of application of principal components is to make use of some pertinent statistical properties of the principal components analysis in generating series of a hydrologic process for such a purpose like investigating droughts on an areal basis.

Although canonical correlation analysis is potentially as useful as the other multivariate analysis, so far this type of analysis has been applied infrequently in hydrology. Its applications in other fields such as psychology, economics, and education are no less than the applications of other multivariate analysis. Some of the applications are given as examples in Kendall (1957) in the form of canonical correlation analysis between reading tests and arithematic tests of school children, between the prices of beef steers and hogs and meat consumption for the United States, between qualities of Canadian Hard Red Spring wheat and the flour made from it, etc. In hydrology, Rice (1967) proposed the use of canonical analysis in estimating parameters of storm hydrographs, Nimmannit and Morel-Seytoux (1969) used this analysis in a study of the effects of weather modification on runoff on a regional basis.

Canonical correlation analysis often results in high linear correlation between pairs of canonical variables which are linear transformations of the original variables. Therefore, a qualitative description of the two types of random variables can be reliably made. In hydrology, however, the numerical values of the original variables are required, and not the values of canonical variables. Since this information is not readily given by the canonical analysis, this may be one reason for its infrequent use in hydrology.

1.2 Relevance of Canonical Correlation Analysis in Hydrologic Investigation

Most of the processes involved in hydrologic investigations can be considered to be three-dimensional. They vary along x and y coordinates as well as along a time axis. For example, the sea surface temperature of the Pacific Ocean varies with latitude and longitude and it varies with time. The same is true for the monthly precipitation of the U.S. West Coast. When correlation analysis is made between a pair of the three-dimensional hydrologic processes, each process is usually divided into many time series at subareas, then the correlation analysis is applied between the two sets of time series of the processes.

A set of hydrologic variables observed at points in an area, or at nearby areas which are hydrologically similar, are usually related. Examples of such correlated sets are snow water equivalent observed at points in a river basin, runoffs from nearby basins, precipitation of adjacent areas, etc. Therefore, when a set of hydrologic variables affects one variable in another set, it is very likely that it also affects other variables in that set as well. Hence, the correlation analysis between two hydrologic processes usually becomes the correlation analysis between two sets of variables which are mutually correlated in each set as well as between the sets.

One approach to this problem of correlation analysis is to use the multiple correlation analysis between each individual variable in the set of dependent variables and all variables in the set of independent variables. This approach has two drawbacks: the number of analyses used is as many as the number of the dependent variables; and the sampling distribution of the correlation coefficient generally used for the significance testing of the coefficient cannot be used due to the mutual correlation of the set of independent variables.

Another approach which can be used effectively for this problem is canonical correlation analysis, especially when independent variables for each of the dependent variables are more or less the same. For example, snowmelt runoffs of watersheds which are hydrologically similar and close together may depend on the same set of indices representing inflow of water into the basins, wetness of the basins, etc. In this case, the correlation analysis between the two sets of variables can be made with only one application of the canonical correlation analysis. The test of significance of the correlation coefficient between the two sets of variables is not affected by the mutual correlation of each set of variables.

As concerns the economic aspect of the canonical correlation analysis in data observation of the hydrologic variables used in the analysis, the technique can be used to select only small number of independent variables which make significant contribution to the correlation between the two sets of variables. As described previously, the hydrologic variables of the set of independent variables usually used in the analysis are mutually correlated. If all the variables are used, some of them may be considered as redundant variables which cause unnecessary reduction in the degree of freedom of the correlation analysis. The contribution of each independent variable may be judged from the magnitude of the coefficient of the linear combination of that variable (an element of the matrix γ_i of Eq. 2.25) which is used for computing the canonical variables which are highly significantly correlated with the canonical variable of the set of dependent variables. If the magnitude of the coefficient is very small compared with those of the other independent variables, that variable may be omitted from the analysis. This usually reduces the number of the independent variables significantly, so expense of maintaining observation stations which make only small contributions to the analysis can be reduced, or reallocated to improve the quality of the data from the more significant stations.

Since a correlation matrix of the set of dependent variables is used in the canonical correlation analysis, the values of the variables computed from the set of the independent variables by using the canonical correlation analysis relate among themselves in such a manner as to preserve the characteristic of their correlations as observed in the historical data.

Because of the maximized correlation between the first pair of the computed canonical variables, the linear relationship between the pair is very reliable. It has been shown by Rice (1969) that the values of the set of dependent variables computed by using all possible pairs of the canonical variables (with a transformation technique which is described later) are mathematically the same as the results of a multiple correlation analysis for each of the dependent variables. Therefore, with a much reduced effort of analysis the canonical correlation analysis gives results that have the same accuracy as those of multiple correlation analysis.

One outstanding advantage of using canonical correlation analysis is in the construction of a confidence region for the computed dependent variables. When variables within a set of hydrologic variables are computed simultaneously, their variations around the computed values to be expected are also very useful information. In the case where these variables are mutually correlated, the joint confidence region of all the variables can be conveniently constructed by using canonical correlation analysis.

Therefore, the correlation analysis between two or more hydrologic processes can be effectively made by using canonical correlation analysis. In this procedure the set of dependent variables consists of the variables which are to be computed, while the set of independent variables consists of the variables which affect the variations of the variables in the former set (which may be considered as the causes of the dependent variables). Usually the set of dependent variables is from the same hydrologic process, while the set of independent variables may be formed from many processes selected in such a way that they affect the dependent variables to a high degree.

1.3 Objective of the Study

The main objective of this study is to demonstrate the potential of the application of the multivariate canonical correlation analysis to hydrologic problems. The field of long-range hydrologic prediction is used as an example of the application by applying the analysis to two prediction problems: the forecast of monthly precipitation of three coastal areas of the United States as shown in Figure 1, and the forecast of seasonal runoff from snowmelt measured at three river gaging stations of river basins in Montana as shown in Figure 2. As far as the accuracy of the long-range forecast is concerned, the selected examples may be considered as extreme cases. For most of the river basins, the forecast of snowmelt runoff can be made with sufficient accuracy as required for the purpose of water resources planning in the basins. On the other hand, reliability of the long-range precipitation forecast at present is still questionable, despite intensive study and research in this field.

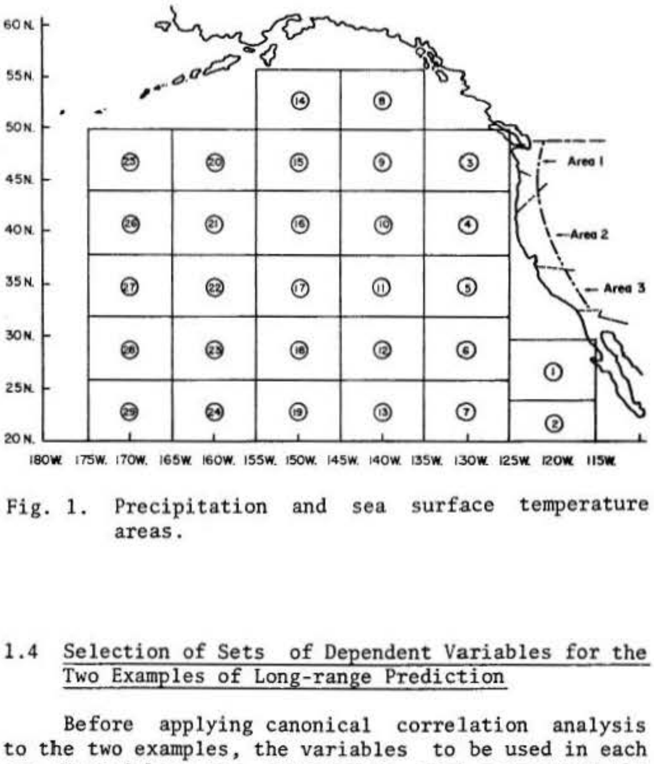


Fig. 1. Precipitation and sea surface temperature areas.

1.4 Selection of Sets of Dependent Variables for the Two Examples of Long-range Prediction

Before applying canonical correlation analysis to the two examples, the variables to be used in each set of variables are selected in such a way that the two sets of variables are significantly correlated. The selections are based on the physical background of each problem. The procedure for each example is as follows.

For the long-range precipitation forecast, a technique of lag cross correlation is used to investigate a linear correlation between the coastal precipitation and some ot its prior causative factors. These factors are the sea surface temperature of the nearby Pacific Ocean and other processes as explained in Torranin (1972). His investigation leads to the conclusion that the significant lag cross correlation exists only between the summer coastal precipitation and the sea surface temperature of some of the 29 areas of the nearby Pacific Ocean, shown in Figure 1. This cross correlation is relatively small, so that the numerical forecast of the coastal precipitation by the lag cross correlation with the sea surface temperature as the forecasting variables is of low reliability. However, this example of forecast of monthly coastal precipitation is used in this investigation only for the purpose of demonstrating the feasibility of technique of canonical correlation analysis for hydrologic predictions.

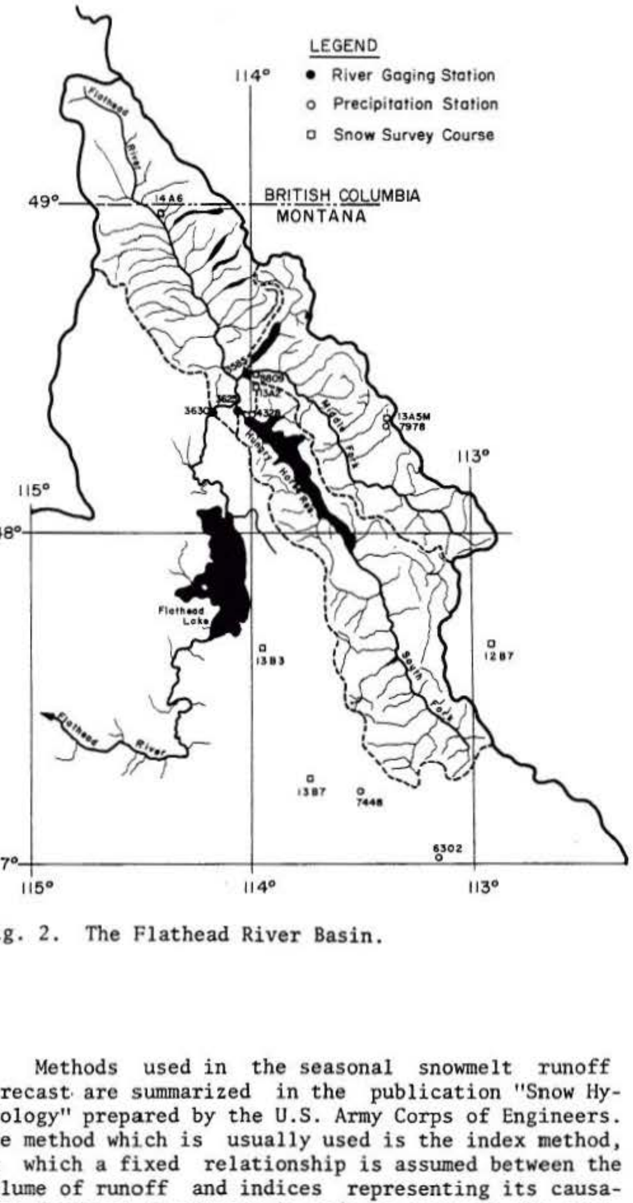


Fig. 2. The Flathead River Basin.

Methods used in the seasonal snowmelt runoff forecast are summarized in the publication "Snow Hydrology" prepared by the U.S. Army Corps of Engineers. One method which is usually used is the index method, in which a fixed relationship is assumed between the volume of runoff and indices representing its causative factors; no attempt is made to evaluate the quantitative contribution of each causative factor. The fixed relationship is obtained by the use of a statistical technique, mostly by the multiple correlation analysis, based on the historical records available.

Factors which affect the seasonal snowmelt runoff are broadly classified as supply and loss. The supply for a given season is comprised mainly of precipitation. The major loss is due to evaporation and evapotranspiration from the basin. Other losses which may be significant in a particular river basin are those due to deep percolation and retention as soil moisture.

The indices usually used in forecasts of seasonal snowmelt runoff are: the winter precipitation index and/or the snow water equivalent index, which represent the major supply to the basin; the evapotranspiration index, which represents the major loss, and the antecedent moisture index which represents the soil moisture condition of the basin. In a basin where the significant amount of precipitation occurs during the snowmelt period, an additional index of the spring-summer precipitation may be included. The forecast covers the period April through July. At the forecast date, the spring precipitation index and the evapotranspiration index are not known. If these two indices are used in the forecast, their values must be first estimated, usually by using either the means or some percentile values. In this study, only those indices that are available by the forecast date are used. It will be shown that the accuracy of results obtained by using this technique of forecast is still acceptable. The indices used in this study include; the fall precipitation index, the winter precipitation index, and the snow water equivalent index as of April 1.

Chapter II

MATHEMATICAL TECHNIQUES USED IN THE ANALYSIS

This chapter summarizes the mathematical techniques used in the study. The summary is intended to be concise and convenient as a rapid reference for the presentations given in this study. For more detailed information about the techniques used, the reader is referred to the appropriate references given in the bibliography at the end of this paper.

2.1 Autocorrelation Analysis

Autocorrelation is used in this study as the method for investigating dependence among the time series. The population autocorrelation coefficient of a continuous time series X_t is defined as

$$\rho_\tau = \text{Cov}\ (X_t, X_{t+\tau})/\text{Var}\ X_t \quad , \qquad 2.1$$

in which Cov $(X_t, X_{t+\tau})$ is the covariance between X_t and $X_{t+\tau}$, Var X_t is the variance of X_t , the subscripts t and $t + \tau$ indicate the times at which X is taken and ρ_k is the lag time. For discrete series X_i the value of ρ_τ is estimated from a sample of size N and the discrete lags $k = 1, 2, \ldots,$ by using the open series approach by

$$r_k = \frac{\text{Cov}\ (X_i, X_{i+k})}{(\text{Var}\ X_i \cdot \text{Var}\ X_{i+k})^{1/2}} \qquad 2.2$$

or by

$$r_k = \frac{\frac{1}{N-k} \sum_{i=1}^{N-k} x_i x_{i+k} - \frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k} x_i\right)\left(\sum_{i=1}^{N-k} x_{i+k}\right)}{\left[\frac{1}{N-k} \sum_{i=1}^{N-k} x_i^2 - \frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k} x_i\right)^2\right]^{1/2} \cdot \left[\frac{1}{N-k} \sum_{i=1}^{N-k} x_{i+k}^2 - \frac{1}{(N-k)^2}\left(\sum_{i=1}^{N-k} x_{i+k}\right)^2\right]^{1/2}} \qquad 2.3$$

For serially uncorrelated time series, the sampling distribution of r_k has an expected value Er_k and a variance Var r_k given as

$$Er_k = -1/(N-k+1) \ , \qquad 2.4$$

and

$$\text{Var}\ r_k = \frac{(N-k+1)^3 - 3(N-k+1)^2 + 4}{(N-k+1)^2 \ [(N-k+1)^2 - 1]} \qquad 2.5$$

For a value of N larger than 30, the sampling distribution of r_k may be approximated by a normal distribution. The 95 percent confidence limits of the serially uncorrelated time series can be computed by

$$T_{1,2}(k) = Er_k \pm 1.96\ (\text{Var}\ r_k)^{1/2} \qquad 2.6$$

Therefore, the dependence in sequence of any time series can be investigated by comparing the sample correlogram, given as the plot of r_k versus k, of the discrete series with the expected correlogram of serially uncorrelated time series. A time series may be considered to be serially uncorrelated if its sample correlogram lies within the confidence limits, and/or if only a small percentage of r_k values defined by the confidence limit probability lies outside these limits.

2.2 Model of Sequentially Dependent Time Series

The dependent model of time series usually used in hydrologic investigation, especially where the phenomenon under investigation has a storage or carryover effect, are approximately of the autoregressive or Markov linear type model. The first order linear autoregressive model, often as the first or rough approximation, is

$$X_i = \rho_1 X_{i-1} + \delta_i \quad , \qquad 2.7$$

in which δ_i is the sequentially independent stochastic component, and ρ_1 is the autoregressive coefficient estimated by the sample first serial correlation coefficient r_1.

The expected correlogram of the first-order Markov model is

$$r_k = r_1^{\ k} \ . \qquad 2.8$$

A method used in this study for testing the goodness of fit of the first order linear Markov model is by the "whitening" procedure. The stochastic component δ_i of the fitted model is computed from the available sample, and if the δ-series is not significantly different from a sequentially independent series, the model of Eq. 2.7 and 2.8 is accepted. The investigation of sequential independence may be also made by using the correlogram technique as described in the previous section.

2.3 Canonical Correlation Analysis

This analysis is usually used in the correlation analysis between two sets of random variables. It searches for a linear combination of each set of variables, such that the correlation between a linear combination, called the canonical variable, of the first set and the linear combination of the second set is maximized. Then a second pair of canonical variables, one from each set, is sought in such a way that the correlation between them is the maximum of all correlation between the linear combinations, uncorrelated with the first pair of canonical variables. The number of pairs of canonical variables is equal to the minimum of the number of original random variables of the two sets. Hopefully, but not necessarily, the first pair of canonical variables will have very high correlation (say 0.90). If this is the case, only the first pair of canonical variables need be used for the description of the correlation between the two original sets of random variables.

The analysis is very effective in investigating whether there is any linear correlation between the two sets of variables, because it maximizes the correlation between linear combinations of variables in each set. In using this analysis, generally, each set of variables as a whole, not each individual member of the set, is of interest to an investigator. However, the analysis becomes more meaningful if the canonical variables have some physical significance. As an example, if the coefficients of the linear combination of each set are all positive, it can be concluded that a weighted averages of the two sets of random variables are highly correlated. Details of this analysis can be found in statistical texts such as Anderson (1958) and Kendall (1957), and a summary of the canonical analysis as given in Appendix A.

Canonical analysis has three particular properties which are of interest with respect to application to forecasting problems. First, since the correlation between the first pair of canonical variables is the maximum, the maximum contribution of the set of independent variables used in the forecast can be estimated. Also, the linear regression equation derived for the canonical variables can be used to forecast the canonical variables of the dependent variables with greater reliability. Second, by using this analysis the forecast values have the same correlation among themselves as those of their historical record. Third, since pairs of the canonical variables usually are uncorrelated, the confidence region of the forecast canonical variables, as well as the forecast variables themselves, is easy to construct and is more reliable than using the other statistical multivariate techniques.

Let $X^{(1)}$ be a column vector of dependent variables with p_1 components, such as the precipitation at the three coastal areas in the first problem studied. Let $X^{(2)}$ be a column vector of independent variables with p_2 components, such as the series of causative factors of sea temperature in this problem. For the sake of convenience in description, let $p_1 \leq p_2$.

Steps used in the canonical analysis between $X^{(1)}$ and $X^{(2)}$ are summarized as follows:

(1) First the covariance matrix of the matrix X,

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} = \begin{bmatrix} x(1) \\ x(2) \\ \vdots \\ x(p_1) \\ x(p_1+1) \\ \vdots \\ x(p_1+p_2) \end{bmatrix} \qquad 2.9$$

is computed as

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} & \cdots \hat{\sigma}_{1p_1} & \hat{\sigma}_{1(p_1+1)} & \cdots \hat{\sigma}_{1(p_1+p_2)} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} & \cdots \hat{\sigma}_{2p_1} & \hat{\sigma}_{2(p_1+1)} & \cdots \hat{\sigma}_{2(p_1+p_2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\sigma}_{p_1 1} & \hat{\sigma}_{p_1 2} & \cdots \hat{\sigma}_{p_1 p_1} & \hat{\sigma}_{p_1(p_1+1)} & \cdots \hat{\sigma}_{p_1(p_1+p_2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{\sigma}_{(p_1+p_2)1} & \hat{\sigma}_{(p_1+p_2)2} & \cdots \hat{\sigma}_{(p_1+p_2)p_1} & \hat{\sigma}_{(p_1+p_2)(p_1+1)} & \cdots \hat{\sigma}_{(p_1+p_2)(p_1+p_2)} \end{bmatrix} \qquad 2.10$$

in which $\hat{\sigma}_{ii}$ is the variance of the i-th variable, x(i) of the matrix X of Eq. 2.9, given by

$$\hat{\sigma}_{ii} = \frac{1}{N} \sum_{\ell=1}^{N} (x_\ell(i) - \overline{x}(i))^2 , \qquad 2.11$$

with $x_\ell(i)$ the ℓth value of the series of N values of x(i),

$$\overline{x}(i) = \frac{1}{N} \sum_{\ell=1}^{N} x_\ell(i) , \qquad 2.12$$

and $\hat{\sigma}_{ij}$ the covariance between x(i) and x(j) of the matrix of Eq. 2.9.

$$\hat{\sigma}_{ij} = \frac{1}{N} \sum_{\ell=1}^{N} (x_\ell(i) - \overline{x}(i))(x_\ell(j) - \overline{x}(j)) \qquad 2.13$$

with

$$\hat{\sigma}_{ij} = \hat{\sigma}_{ji} . \qquad 2.14$$

(2) The partition of the covariance matrix $\hat{\Sigma}$ is made as follows:

$$\hat{\Sigma} = \begin{bmatrix} \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & \hat{\Sigma}_{22} \end{bmatrix} . \qquad 2.15$$

$$\Sigma_{11} = \begin{bmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{12} \cdots \hat{\sigma}_{1p_1} \\ \hat{\sigma}_{21} & \hat{\sigma}_{22} \cdots \hat{\sigma}_{2p_1} \\ \vdots & \vdots \\ \hat{\sigma}_{p_1 1} & \hat{\sigma}_{p_1 2} \cdots \hat{\sigma}_{p_1 p_2} \end{bmatrix} , \qquad 2.16$$

$$\hat{\Sigma}_{12} = \begin{bmatrix} \hat{\sigma}_{1(p_1+1)} & \hat{\sigma}_{1(p_1+1)} \cdots & \hat{\sigma}_{1(p_1+p_2)} \\ \hat{\sigma}_{2(p_1+1)} & \hat{\sigma}_{2(p_1+2)} & \hat{\sigma}_{2(p_1+p_2)} \\ \vdots & & \\ \hat{\sigma}_{p_1(p_1+1)} & \hat{\sigma}_{p_1(p_1+2)} & \hat{\sigma}_{p_1(p_1+p_2)} \end{bmatrix} . \qquad 2.17$$

$$\hat{\Sigma}_{21} = \hat{\Sigma}_{12}^{T} , \qquad 2.18$$

in which $\hat{\Sigma}_{12}^{T}$ is the transpose of $\hat{\Sigma}_{12}$, with

$$\hat{\Sigma}_{22} = \begin{bmatrix} \hat{\sigma}_{(p_1+1)(p_1+1)} & \hat{\sigma}_{(p_1+1)(p_1+2)} & \cdots \hat{\sigma}_{(p_1+1)(p_1+p_2)} \\ \hat{\sigma}_{(p_1+2)(p_1+1)} & \hat{\sigma}_{(p_1+2)(p_1+2)} & \cdots \hat{\sigma}_{(p_1+2)(p_1+p_2)} \\ \vdots & \vdots & \vdots \\ \hat{\sigma}_{(p_1+p_2)(p_1+1)} & \hat{\sigma}_{(p_1+p_2)(p_1+2)} & \cdots \hat{\sigma}_{(p_1+p_2)(p_1+p_2)} \end{bmatrix} . \qquad 2.19$$

(3) The canonical correlations are computed by solving the system of equations:

$$\begin{vmatrix} -\lambda\hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & -\lambda\hat{\Sigma}_{22} \end{vmatrix} = 0 . \qquad 2.20$$

This system of equations is solved for the first p_1 largest roots as

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_{p_1} ,$$

where λ_i is the ith canonical correlation coefficient, or the linear correlation coefficient between the ith pair of canonical variables.

(4) Let α_i and γ_i be the column vector of coefficients for the ith pair of canonical variables which corresponds to the canonical correlation coefficient λ_i. The column vectors α_i and γ_i are obtained by the solution of the following system of equations:

$$\begin{bmatrix} -\lambda_i \hat{\Sigma}_{11} & \hat{\Sigma}_{12} \\ \hat{\Sigma}_{21} & -\lambda_i \hat{\Sigma}_{22} \end{bmatrix} \cdot \begin{bmatrix} \alpha_i \\ \gamma_i \end{bmatrix} = 0 \quad , \qquad 2.21$$

subjected to the conditions,

$$\alpha_i^T \hat{\Sigma}_{11} \alpha_i = 1 \qquad 2.22$$

$$\gamma_i^T \hat{\Sigma}_{22} \gamma_i = 1 \ . \qquad 2.23$$

(5) The ith pair of canonical variables are computed by

$$U_i = \alpha_i^T X^{(1)} \qquad 2.24$$

and

$$V_i = \gamma_i^T X^{(2)} \quad , \qquad 2.25$$

in which U_i and V_i represent the ith canonical variable of the set of dependent and independent variables, respectively.

The derivations which lead to these steps are described in Appendix A.

If X is multivariate normally distributed, then U_i and V_i of Eqs. 2.24 and 2.25 are also normally distributed. Since the linear correlation between U_j and V_j, for $i = j$, is maximized, the values of V_j computed from the observed values of the group of independent variables, $X^{(2)}$, by using Eq.2.25 can be used for the forecast of U_i by the linear regression equation between U_i and V_j. The use of the linear regression equation becomes now more reliable because of the maximized correlation thus obtained.

Let $\hat{U}_i$ be a forecast value of U_i from the linear regression equation between U_i and V_i, and e_i^2 be the variance of a single forecast of $\hat{U}_i$ for the value of V_i used, i.e. the square of the error.

Therefore, for each observed value of

$$V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_{p_1} \end{bmatrix} \quad , \qquad 2.26$$

the forecast value $\hat{U}$,

$$\hat{U} = \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \vdots \\ \hat{U}_{p_1} \end{bmatrix} \quad , \qquad 2.27$$

is made with the variance of a single forecast E,

$$E = \begin{bmatrix} e_1^2 \\ e_2^2 \\ \vdots \\ e^2_{p_1} \end{bmatrix} \qquad 2.28$$

Equations 2.26, 2.27 and 2.28 are equivalent to the following statements:

$$U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_{p_1} \end{bmatrix} \qquad 2.29$$

is multivariate normally distributed with a mean matrix $\hat{U}$,

$$\hat{U} = \begin{bmatrix} U_1|V_1 \\ U_2|V_2 \\ \vdots \\ U_{p_1}|V_{p_1} \end{bmatrix} \quad , \qquad 2.30$$

and with a covariance matrix E,

$$E = \begin{bmatrix} e_1^2 & 0 & 0 \ \dots \ 0 \\ 0 & e_2^2 & 0 \ \dots \ 0 \\ \vdots & \vdots & \\ 0 & 0 & 0 \ \dots \ e^2_{p_1} \end{bmatrix} \qquad 2.31$$

in which the symbol $U_i|V_i$ means the value of U_i given V_i. Equation 2.31 is realistic because U_i and V_j are uncorrelated for $i \neq j$.

These properties of the canonical analysis make possible the construction of a joint confidence region for the forecast value of U, as well as for the dependent variables themselves.

From Eq. 2.24,

$$U = \alpha^T X^{(1)} \quad , \qquad 2.32$$

in which

$$\alpha = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_{p_1}] \quad . \qquad 2.33$$

Therefore,

$$X^{(1)} = (\alpha^T)^{-1} \ U \ , \qquad 2.34$$

in which $(\alpha^T)^{-1}$ is the inverse of the matrix α^T.

Equation 2.34 can be used to transform the forecast canonical variable, $\hat{U}$, to the original dependent variables. If

$$U \sim N[\hat{U}, E] \ , \qquad 2.35$$

the symbol ~ means "distributed as," and $N[\hat{U}, E]$ means "multivariate normal distribution with a mean vector $\hat{U}$ and a covariance matrix E." Then, the quadractic form Q(U),

$$Q(U) = (U - \hat{U})^T E^{-1} (U - \hat{U}) \qquad 2.36$$

is distributed as the chi-square distribution with p_1 degree of freedom. The proof of Eq. 2.36 is shown in Appendix A.

Equation 2.36 can be used to construct a confidence region for the forecast value $\hat{U}$, which is a spheriod in p_1 dimensional space.

Also, since $U \sim N[\hat{U}, E]$, it is shown in Anderson (1958, p. 19) that

$$X^{(1)} = (\alpha^T)^{-1} U \sim N[(\alpha^T)^{-1} U, (\alpha^T)^{-1} E\{(\alpha^T)^{-1}\}^T] \ ,$$

or

$$X^{(1)} \sim N[U^*, E^*] \ . \qquad 2.37$$

Therefore, the quadratic form $Q^*(X)$,

$$Q^*(X) = (X^{(1)} - U^*)^T E^{*-1} (X^{(1)} - U^*) \ , \qquad 2.38$$

is distributed as the chi-square distribution with p_1 degree of freedom.

Equation 2.38 can be used to construct a confidence region for the forecast value of the original dependent variables, $X^{(1)}$, which are transformed back from the forecasted canonical variables $\hat{U}$.

For the case that $X^{(1)}$ has a multivariate normal distribution, Anderson (1958) presented a joint probability distribution of the square of the p_1 canonical correlation coefficients when the population values are zero (Eq. A-31 in Appendix A). The marginal cumulative distribution of the square of the ith sample canonical correlation coefficient is derived from the joint probability distribution, as shown in Appendix A, for i = 1, 2 and 3. The marginal cumulative distributions for $p_1 = 3$, N = 63, p_2 = 13, 14, 15 and for $p_2 = 3$, and N = 30 are shown in Fig. 3, a to d, respectively. These curves can be used for testing the significance of the computed sample canonical correlation coefficients.

The computer routine BMDX75M of the Biomedical Computer Program is used in this study for the canonical correlation analysis; detailed explanations are given in the programs manual (Dixon, 1970).

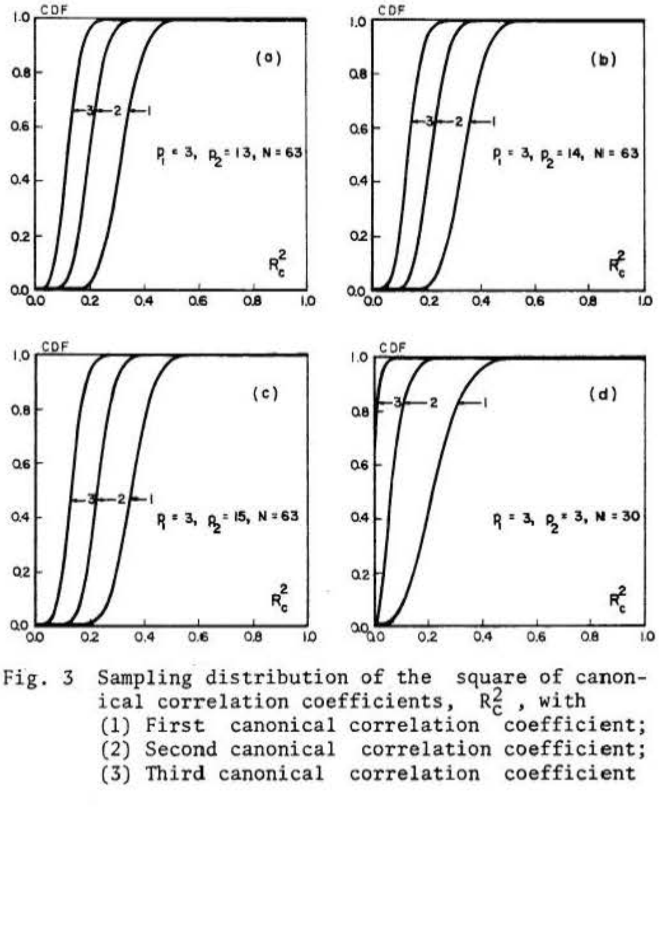


Fig. 3 Sampling distribution of the square of canonical correlation coefficients, R_c^2 , with
(1) First canonical correlation coefficient;
(2) Second canonical correlation coefficient;
(3) Third canonical correlation coefficient

CHAPTER III

ASSEMBLY AND PROPERTIES OF DATA

This chapter treats in detail the data used in this study, and particularly concerning their source, length of record, computation, representativeness, and some of their physical and statistical properties.

Monthly time series of variables used in this study are mostly of the periodic-stochastic type. The periodic component is the result of astronomic cycles. The stochastic component, the occurrence of which is governed by the laws of chance, results from many random processes in nature, especially the atmospheric random processes. The monthly series of these processes, therefore, are not stationary; their properties change from month to month. According to Roesner and Yevjevich (1966), the values of each of the 12 calendar months can be considered as those coming from different populations, each with its population mean, μ_τ, and its population standard deviation, σ_τ, and with τ varying from 1 to 12 representing January through December. Second-order stationary time series means that the mean and covariance of the series do not vary with time and approach their population values with a probability unity when time goes to infinity. The second-order stationary components of these monthly series can be computed from values of the original non-stationary time series by

$$\varepsilon_{p,\tau} = (X_{p,\tau} - m_\tau)/s_\tau \ , \qquad 3.1$$

in which $\tau = 1, 2,\ldots, 12$; $p = 1, 2,\ldots, N$, $X_{p,\tau}$ is the value of the original series for the month τ of the year p, N is the number of years of data, and m_τ and s_τ are the sample estimates of μ_τ and σ_τ, respectively. The values of m_τ and s_τ are estimated from a sample by

$$m_\tau = \frac{1}{N} \sum_{p=1}^{N} X_{p,\tau} \quad , \qquad 3.2$$

and

$$s_\tau = \frac{1}{N}\left[\sum_{p=1}^{N} (X_{p,\tau} - m_\tau)^2\right]^{1/2} . \qquad 3.3$$

For a small sample size N, a better, unbiased estimate of σ_τ can be computed by replacing $1/N$ in Eq. 3.3 by $1/(N-1)$. This $\varepsilon_{p,\tau}$ series may also be regarded as a standardized series.

The second-order stationary monthly series as computed by Eq. 3.1 may be a sequentially time dependent or time independent series, which results from characteristics of processes producing each series. By fitting a proper sequentially dependent model to $\varepsilon_{p,\tau}$ - series as described in Chapter II, a sequentially independent time series $\delta_{p,\tau}$ can be computed from the $\varepsilon_{p,\tau}$ - series.

In this study the following notation for the different time series is used: $X_{p,\tau}(\cdot)$ is the original series, which in most cases is the non-stationary time series because of periodicity in parameters, with the dot in the parenthesis denoting the kind of data (for example, $X_{p,\tau}(P)$ is the value of the original series of precipitation for the month τ of the year p), $\varepsilon_{p,\tau}(\cdot)$ represents the second-order stationary series after periodicities are removed in μ and σ, and $\delta_{p,\tau}$ the series of residuals of the $\varepsilon_{p,\tau}(\cdot)$ after fitting a sequentially dependent model, with $\delta_{p,\tau}$ approximately an independent in sequence second-order stationary random variable.

3.1 Data for the Analysis of Precipitation Forecast

Precipitation. The West Coast region of the United States is divided into three areas as shown in Fig. 4. These areas, as proposed by Klein (1964), are topographically and meteorologically nearly homogeneous. The criterion used for data consistency of a precipitation station is that the changes of station location during the period of observation are less than one mile in the horizontal direction and less than 100 feet in elevation. Data of consistent monthly precipitation of 83 stations, uniformly distributed over the three coastal areas (17, 39, and 27 stations for coastal area 1, 2, and 3, respectively) are selected from "Climatological Data" published by the Weather Bureau, U. S. Department of Commerce. The locations of the selected stations are shown in Fig. 4 by dots. Their names and coordinates are given in Appendix B. The length of data is from January 1948 through September 1971.

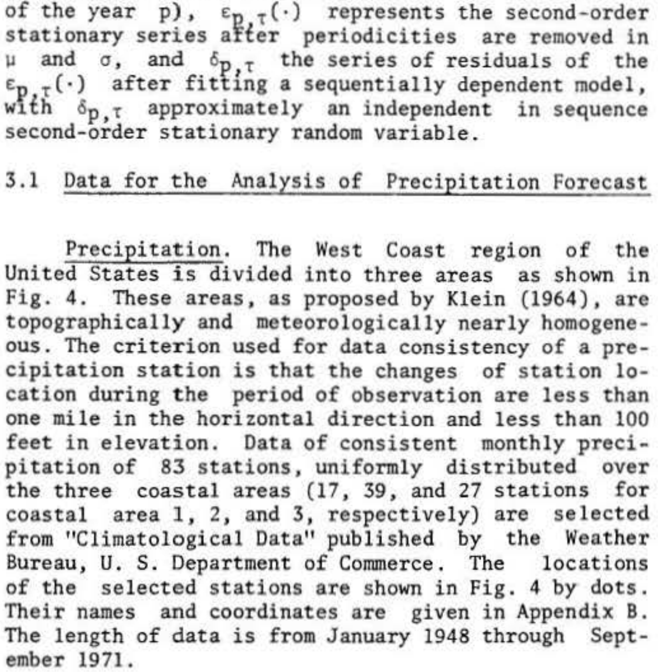
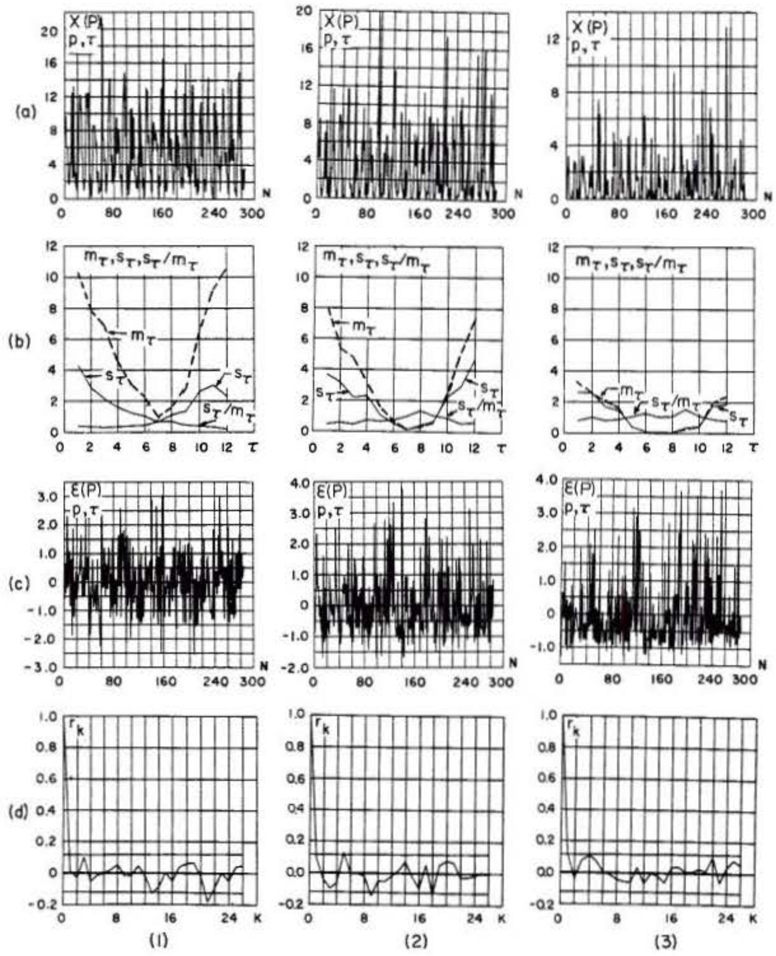


Fig. 4. U.S. coastal precipitation stations, and data points of sea surface temperature, used in this study for precipitation forecast.

A representative time series of monthly precipitation for each area is taken as a simple average of the monthly values of precipitation of all stations in the area. These periodic-stochastic time series, $X_{p,\tau}(P)$, for the three areas are shown in Fig. 5a.

The parameters m_τ and s_τ for $\tau = 1,2,\ldots,12$, are computed by using Eq. 3.2 and 3.3, respectively. These values are shown in Fig. 5b, together with the coefficient of variation s_τ/m_τ for each of the three areas. The twelve values of s_τ/m_τ for each of the



LEGEND

(1),(2),(3) The Three Coastal Regions of Fig. 4.
(a) The Original Series, $X_{p,\tau}(P)$, in Inches, N = Month Number.
(b) The Periodic Parameters m_τ and s_τ, in Inches, and the Approximate Constant Coefficient of Variation, s_τ/m_τ.
(c) The Independent Stochastic Second-Order Stationary Component, $\varepsilon_{p,\tau}(P)$.
(d) Correlogram of $\varepsilon_{p,\tau}(P)$-Series with 95% Confidence Limits of a Serially Uncorrelated Series.

Fig. 5. Coastal precipitation data.

three areas are not statistically significantly different from a constant. The mean annual precipitations are 66.9, 39.8, and 15.0 inches for areas number 1, 2, and 3, respectively.

The second-order stationary time series, $\varepsilon_{p,\tau}(P)$, for the three areas are computed by using Eq. 3.1, and are shown in Fig. 5c. The correlograms of $\varepsilon_{p,\tau}(P)$ series are shown in Fig. 5d, which indicate that all three $\varepsilon_{p,\tau}(P)$ series are practically independent in sequence time series.

The $\varepsilon_{p,\tau}(P)$ - components are fitted by a normal and a lognormal probability distribution with three parameters, and are tested for the goodness of fit by a chi-square test using ten equal probability classes. The results are shown in Table 1.

The $\varepsilon_{p,\tau}(P)$ series for area 1 is serially uncorrelated, and is standard normally distributed. For areas 2 and 3, $\varepsilon_{p,\tau}(P)$ series are also serially uncorrelated, but are lognormally distributed with three parameters. In other words, $\log_e[\varepsilon_{p,\tau}(P) + 1.710]$ of area 2 is normally distributed with mean 0.343 and standard deviation 0.699. Similarly, $\log_e[\varepsilon_{p,\tau}(P) + 1.288]$ of area 3 is normally distributed with mean -.040 and standard deviation 0.811.

The standard normal transform of $\varepsilon_{p,\tau}(P)$ of area 2, the $\varepsilon'_{p,\tau}(P)$ series, is computed by

$$\varepsilon'_{p,\tau}(P) = \{\log_e[\varepsilon_{p,\tau}(P) + 1.710] - 0.343\}/0.699. \qquad 3.4$$

TABLE 1

FITTING PROBABILITY FUNCTIONS TO PRECIPITATION DATA

Area Number	normal					lognormal						
	χ^2	$.95\chi^2 cr$	Result	Mean	Std Dev	χ^2	$.95\chi^2 cr$	Result	Lower Bound	Mean	Std Dev
I	6.39	15.5	Accept	0.0	1.0	52.0	14.1	Reject	-2.579	.846	.532
2	50.45	15.5	Reject	0.0	1.0	13.8	14.1	Accept	-1.710	.343	.699
3	110.1	5.5	Reject	0.0	1.0	10.5	14.1	Accept	-1.288	-.040	.811

For area 3, the $\varepsilon'_{p,\tau}(P)$ is computed by

$$\varepsilon'_{p,\tau}(P) = \{\log_e[\varepsilon_{p,\tau}(P) + 1.288] + 0.040\}/0.811 \quad . \qquad 3.5$$

For area 1, the $\varepsilon'_{p,\tau}(P)$ is the same as $\varepsilon_{p,\tau}(P)$.

Sea surface temperature of the Pacific Ocean. Variations of sea surface temperature depend on many factors such as insolation or exposure to the sun, evaporation from the sea, convective transfer of heat, mixing of deep and surface water, transport by currents, upwelling (the rising of water toward the surface from subsurface layers), and convergence and divergence of sea water. The exposure to the sun depends on the cloudiness of the atmosphere. Evaporation is controlled by the vapor pressure gradient of the layer of air near the sea surface and by wind velocities. The convective transfer of heat depends on the difference in the sea and air temperatures and on wind velocity. Deviations of sea surface temperature from the means are the indicators of heat surplus or deficit of the surface layer of the sea. They are strongly related to the mix-layer depths, e.g., the depth of relatively constant temperature extending from surface to the top of the thermocline. This is the reason for the relative persistence of large-scale deviations through winter, during which the mixed-layer depth is much greater than during the other season. According to Laevastu and Hubert (1970), the sea surface temperature deviations are relatively persistent through any given winter or summer season, but can change rapidly in spring and fall. The deviations are of the order of 1.5° to 2.5° C with an extreme of 4.5° C observed during late summer.

Because long records of data are not available, the areal coverage of the sea surface temperature of the Pacific Ocean, used in this study, is limited to the area east of 175° W longitude, between 20° N to 56° N latitude, as shown in Fig. 4. Two sources of data are used. The monthly data for the period January 1949 through December 1962 was obtained from the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. This set of data was originally prepared by Dr. Sette's group at the Bureau of Commercial Fisheries from records of sea surface temperature of ships operating in the area. More than two million observations were used, and an intensive editing procedure was applied to data. The procedure is explained in Circular 258 of the Bureau of Commercial Fisheries. The data are finally reduced to values at grid points of the two degree square latitude and longitude over the area. However, the data obtained from NCAR are at the grid points of a rectangular array. Formulas for computing the latitude and longitude of the grid points of the array were given.

The sea surface temperature data, in degrees centigrade at grid points of two degrees latitude by five degrees longitude, are computed from the data at the grid points of rectangular array by simple interpolation. The locations of the 2° x 5° grid points are shown as crosses in Fig. 4. The time series of sea surface temperature for the period from January 1949 through December 1962 at these grid points are used as basic data in this study.

A second period of data from January 1963 through October 1971 was obtained from the monthly publication "Fishing Information" of the Fishery-Oceanography Center, NOAA, United States Department of Commerce. The monthly values, in degrees Fahrenheit, at the same 2° x 5° grid points as used in the first period of data are read from the publication.

These two sources of data provide the basic surface temperature data for the period January 1949 through October 1971.

The surface of the Pacific Ocean under investigation is divided into 28 grid areas that are 10 degrees longitude by 6 degrees latitude, and one that is 10 degrees longitude by 4 degrees latitude, as shown in Fig. 6. A representative value for a particular grid area is computed for each month from all the data points in the area. Each datum point is considered to be representative of the area of a rectangle having sides at distances halfway between two data points. As shown in Fig. 6, the value at datum point 12 represents the values within the dashed area.

The representative values of temperature for each of the 28 grid areas are computed. Using area number 17 as an example, the representative value is computed as

$$X_{p,\tau}(T) = \frac{1}{6}\,[\tfrac{1}{4}(X^1_{p,\tau}(T)+X^3_{p,\tau}(T)+X^6_{p,\tau}(T)+X^8_{p,\tau}(T))+ + \tfrac{1}{2}(X^2_{p,\tau}(T)+X^4_{p,\tau}(T)+X^5_{p,\tau}(T)+X^7_{p,\tau}(T)+X^9_{p,\tau}(T) + + (X^{10}_{p,\tau}(T)) + (X^{11}_{p,\tau}(T)+X^{12}_{p,\tau}(T))] \quad . \qquad 3.6$$

Similarly, for area number 2 this value is

$$X_{p,\tau}(T) = \frac{1}{4}\,[\tfrac{1}{4}(X^a_{p,\tau}(T)+X^c_{p,\tau}(T)+X^e_{p,\tau}(T)+X^g_{p,\tau}(T)) + + \tfrac{1}{2}(X^b_{p,\tau}(T)+X^d_{p,\tau}(T)+X^f_{p,\tau}(T)+X^h_{p,\tau}(T)) + +X^i_{p,\tau}(T)] \quad , \qquad 3.7$$

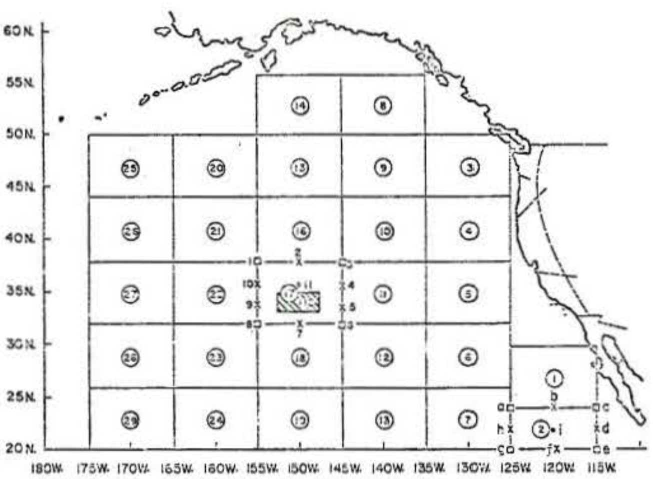


Fig. 6. Sea surface temperature areas used in this study, showing points for defining the formula for the computation of a representative value of the temperature of an area.

where $X^{j}_{p,\tau}(T)$ is the temperature for the month τ of the year p at the grid point j.

The values of m_τ and s_τ are computed for all 29 areas by using Eqs. 3.2 and 3.3, which are shown in Fig. 7, together with the coefficients of variation, s_τ/m_τ as they change along τ. Note that the areas shown in each row in the figure are at the same latitude. The range of variation of the twelve monthly mean temperatures is as low as 4° C at the low latitudes, and this range increases with latitude to become as high as 8° C for areas of high latitudes. The standard deviations are small compared to the means, resulting in the low and relatively constant values of s_τ/m_τ.

Correlograms of the $\varepsilon_{p,\tau}(T)$ - series, computed by Eq. 3.1 for each of the 29 areas, are shown in Fig. 8a, again the areas in each row are at the same latitude. These correlograms show that the $\varepsilon_{p,\tau}$ - series of all 29 areas are highly dependent in sequence. The areas at low latitudes have higher autocorrelation coefficients and longer lag times than the areas at high latitudes. Also, the areas closer to the coast have somewhat longer "memory" than the areas farther from the coast.

The first-order Markov model is fitted to the $\varepsilon_{p,\tau}(T)$ - series, and the series of the residuals of the model as the $\delta_{p,\tau}(T)$ series are computed. Correlograms of the $\delta_{p,\tau}(T)$ - series are shown in Fig. 8b. They indicate these series to be practically sequentially independent time series for all areas. Therefore, the standardized series of deviations of sea surface temperature are sequentially dependent time series with the dependence approximated by the first-order Markov linear model.

Normal probability distribution functions are fitted to all $\delta_{p,\tau}(T)$ - series by using the same technique as for the $\varepsilon_{p,\tau}(P)$ - series. The results are shown in Table 2. They indicate the $\delta_{p,\tau}(T)$ - series to be all normally distributed, with their means and variances given in that table.

3.2 Data for the Analysis of Snowmelt Runoff Forecast

Snowmelt runoff. Monthly mean discharges for 30 years at the three gaging stations, shown in Fig. 2 by dots and described in Table 3, from the water year 1939-40 through the water year 1968-69, are obtained from the U. S. Geological Survey Water Supply Papers. The monthly values of South Fork Flathead River near Columbia Falls and Flathead River at Columbia Falls are adjusted for the changes in content of the Hungry Horse reservoir. Based on the period of data used, the characteristics of the runoffs of the three stations are shown in Table 4.

The seasonal flow in Table 4 is the summation of the monthly mean values of April through July, inclusive. The mean seasonal flow for each station accounted for nearly 80 percent of its mean annual flow. The first- and the second-order autocorrelation coefficients for all three stations are not significantly different from those of a sequentially independent time series.

Monthly base flows of each gaging station are estimated by

$$Q_i = Q_o e^{-kt} \quad , \qquad 3.8$$

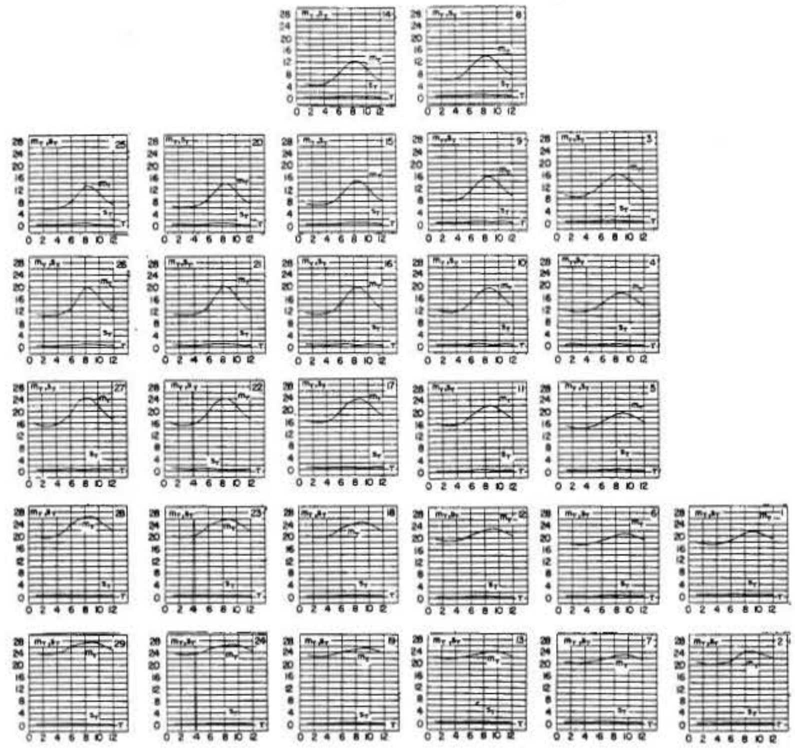
in which Q_i is an estimated base flow of the month i, Q_o is the base flow of the month o which is t months before the month o, k is a recession constant and e is the natural base of logarithm.

Using Eq. 3.8, total volume of base flows during the period of April through July are estimated for the three gaging stations and shown in Table 4. The estimated volume of baseflow during the snowmelt season is very small compared to the volume of the seasonal flow. Therefore, no adjustment for the baseflow is made, and the observed flow during the snowmelt season is used as the dependent variable in this study.

The sample cumulative distribution function of the 30 values of seasonal runoff for each station is computed by using the plotting position method m/(N + 1). These distribution functions for the three stations are plotted on normal probability paper, Fig. 9. Based on Smirnov-Kolmogorov test, the distribution functions at the three stations are not significantly different from the normal probability distribution at 95 percent level of confidence.

Therefore, the time series of seasonal runoff of the three stations are sequentially independent normally distributed processes, with the estimated means and standard deviation as shown in Table 4.

Method of computation of indices of snowmelt runoff. Most of the indices used in the correlation analysis for the forecast of snowmelt runoff are computed from the observed values at different times of the season. Two steps are usually used in computing the indices. For each month the effective monthly values are computed as the weighted average of data at the locations selected. Then the indices are computed from the obtained effective monthly values as the weighted average of all months of the season. Many criteria are used in assigning the weights. The station weights may be assigned proportionally to the Thiessen area of each station or proportionally to the variance of the data observed at each station. Sometimes, station weights are assigned according to the correlation between the data at each station and the seasonal runoff. Since the observed snow water equivalent highly depends on the elevation of the snow course, the elevation of each course is usually considered in assigning weights to snow courses. Work,



NOTE

Area number is shown at the upper right corner; area in each row are at the same latitude.

Fig. 7. Monthly means and monthly standard deviations of sea surface temperature data, in °C.

TABLE 2

FITTING PROBABILITY FUNCTIONS TO SEA SURFACE TEMPERATURE DATA

Area No.	χ^2	Normal Distribution $.95\chi^2$cr	Accept	Mean	Std Dev	Variance	3rd Moment
1	6.53	15.5	Yes	0	0.6584	0.433	- .037
2	7.72	15.5	Yes	0	0.7010	0.491	0.070
3.	7.21	15.5	Yes	0	0.6617	0.438	-0.009
4	4.41	15.5	Yes	0	0.6656	0.443	-0.046
5	10.85	15.5	Yes	0	0.6377	0.406	-0.049
6	8.50	15.5	Yes	0	0.6344	0.402	0.047
7	4.48	15.5	Yes	0	0.5926	0.351	0.049
8	9.18	15.5	Yes	0	0.6399	0.409	0.032
9	4.561	15.5	Yes	0	0.7058	0.498	-0.003
10	13.35	15.5	Yes	0	0.6932	0.480	0.059
11	4.11	15.5	Yes	0	0.6481	0.420	-0.022
12	5.47	15.5	Yes	0	0.6148	0.378	0.026
13	4.86	15.5	Yes	0	0.6202	0.385	0.052
14	7.74	15.5	Yes	0	0.7493	0.561	-0.026
15	8.88	15.5	Yes	0	0.6519	0.425	-0.004
16	3.42	15.5	Yes	0	0.7094	0.503	-0.034
17	6.08	15.5	Yes	0	0.6728	0.453	0.030
18	6.83	15.5	Yes	0	0.6895	0.475	0.030
19	7.67	15.5	Yes	0	0.5679	0.322	0.003
20	10.02	15.5	Yes	0	0.6923	0.429	0.019
21	4.86	15.5	Yes	0	0.6704	0.449	-0.006
22	8.35	15.5	Yes	0	0.7553	0.570	-0.055
23	4.18	15.5	Yes	0	0.7802	0.609	-0.152
24	9.41	15.5	Yes	0	0.6995	0.489	0.030
25	14.56	15.5	Yes	0	0.7499	0.562	0.105
26	9.86	15.5	Yes	0	0.6869	0.472	0.068
27	6.08	15.5	Yes	0	0.7427	0.552	0.020
28	11.76	15.5	Yes	0	0.8279	0.685	0.090
29	4.11	15.5	Yes	0	0.7251	0.526	-0.006

Remarks: Number of classes is 10. Data from January 1949 through December 1970, 22 years or 264 monthly values.

TABLE 3

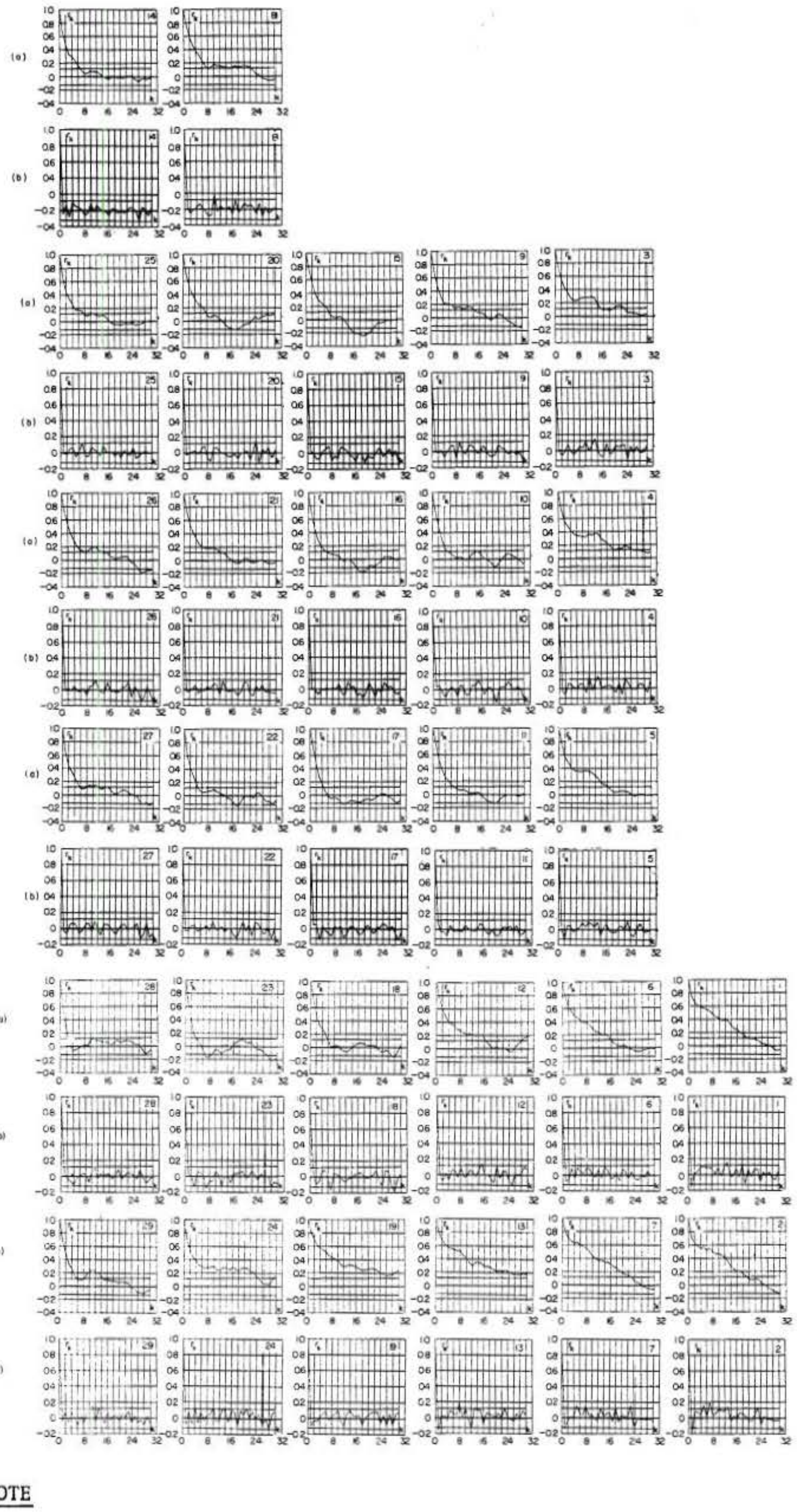
STREAM GAGING STATIONS

Station Number	Name of Station	Location	Drainage Area, sq mi
3585	The Middle Fork Flathead River near West Glacier, Mont.	48° 29' 43" N - 114° 00' 33" W	1128
3625	The South Fork Flathead River near Columbia Falls, Mont.	48° 21' 24" N - 114° 02' 12" W	1663
3630	The Flathead River at Columbia Falls, Mont.	48° 21' 43" N - 114° 11' 02" W	4464

TABLE 4

CHARACTERISTICS OF STREAMFLOW DATA

Station Number	Annual Flow, 10^5 AF		Seasonal Flow, 10^5 AF		Autocorrelation		Estimated Base Flow % of Mean Annual Flow
	Mean	Std Dev	Mean	Std Dev	1 st	2 nd	
3585	21.122	4.414	16.645	3.614	0.105	-0.038	2.3
3625	26.261	5.790	20.914	4.678	0.026	-0.078	2.9
3630	71.609	15.086	56.061	12.057	0.150	-0.038	3.2



NOTE

Area number is shown at the upper right corner; area in each row are at the same latitude.

Fig. 8. Correlograms of series of sea surface temperatures for 29 areas used in this study; (a) $\varepsilon_{p,\tau}$(T)-series, (b) $\delta_{p,\tau}$(T)-series.

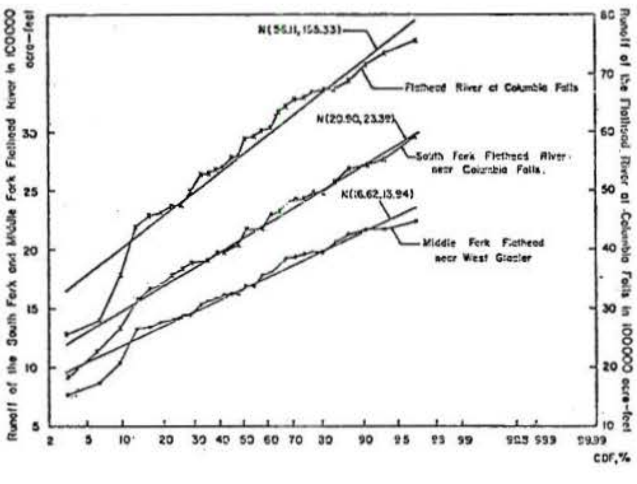


Fig. 9. Fittings of normal probability distribution functions to frequency distributions of snowmelt runoff data.

Beaumont, and Davis (1962) found that a more accurate forecast is obtained by using different forecast relations for different ratios of the snow water equivalent observed at high and low elevation. The monthly weights for indices used are usually assigned according to the estimated relative effects of each month on the runoff.

The method currently used in assigning weights for computing an index, say the index of winter precipitation, is to perform a multiple correlation analysis between the seasonal runoff and the precipitation in winter months at the selected stations. The stations weights are then assigned by using the multiple regression coefficients obtained, and taking into account other considerations as previously described. The effective monthly value for each month is then computed as the weighted average of values at the stations for that month. Using the effective precipitation for all of the winter months as independent variables, a multiple correlation analysis is again applied, with the seasonal runoff as a dependent variable. The weight for each month is assigned by using the multiple regression coefficients. The winter precipitation index is then computed as the weighted average of the effective precipitation of the winter months. There is no set rule regarding the magnitude of weights, but it is customary to make the sum of station weights as well as of monthly weights equal to unity. Normally, the highest weighting factor is not greater than three times the lowest.

The indices of forecast used in this study are similar to those used by the Water Management Subcommittee, Columbia Basin Inter-Agency Committee, U.S. Bureau of Reclamation (1964). The report describes the forecasting procedure for inflow into the Hungry Horse Reservoir. Both the station weights and monthly weights of these indices used in this study are mostly those given in the report. The exception is that they are scaled in such a way that the sums of station weights as well as of monthly weights are unities. Although the weights given in the report were derived especially for the South Fork Flathead River as the indicators of the potential of runoff from the basin, it is shown in Chapter IV that they may be used effectively as indices of the potential runoff of adjacent basins as well.

Fall and winter precipitation index. Locations of the five precipitation stations used in this study are shown in Fig. 2 as circles. The information about the stations is given in Table 5. These stations were selected because of their long records and a good correlation with the runoff of the South Fork Flathead River Basin.

Monthly total precipitation series of the five stations from January 1939 through September 1971 are obtained from the Climatological Data published by the U.S. Weather Bureau.

TABLE 5

PRECIPITATION STATIONS

Station Number	Station Name	Location	Elevation Ft	Station Weight
4328	Hungry Horse Dam	48° 21' N - 114° 00' W	3160	0.28
6302	Ovando	47° 01' N - 113° 09' W	4100	0.14
7448	Seeley Lake Ranger Sta.	47° 13' N - 113° 31' W	4100	0.32
7978	Summit	48° 19' N - 113° 21' W	5213	0.12
8809	West Glacier	48° 30' N - 113° 59' W	3154	0.14

For fall precipitation indices, the monthly weights for August, September, and October are 0.21, 0.32, and 0.47, respectively. Station weights are first assigned proportionally to the variance of the data of each individual station, then a trial-and-error procedure in adjusting the weights is made to obtain the best correlation between the fall precipitation index and the runoff. The station weights are shown in Table 5.

For winter precipitation indices, the monthly weights for November, December, January, February, and March are the same and equal to 0.20. The station weights are the same as those of the fall precipitation index.

Using the values of the weights as described, 30 years of fall precipitation indices, August-October for 1939 through 1968, and 30 years of winter precipitation indices, November-March for 1940 through 1969, are computed. The sample cumulative distribution functions of both indices are plotted on normal probability paper, Fig. 10. Based on the Smirnov-Kolmogorov test at 95 percent level of confidence, the distribution of fall precipitation index is not significantly different from the normal probability function, with the mean 1.972 and the variance 0.627, while the distribution of the winter precipitation index is not significantly different from the normal probability function, with the mean 2.166 and the variance 0.258.

Snow water equivalent index. Out of all the snow courses in and near the basin, five were selected for computing the snow water equivalent index for forecasting the seasonal inflow into Hungry Horse Reservoir. These snow courses are chosen based on five criteria of desirable features: length of record, good individual plots against runoff, good correlation in multiple correlation analysis with runoff, consistent double mass plots and good areal distribution. One more snow course near the Canadian border, Kishenehn, is added in this study to obtain a better areal coverage of the basin. The data of snow water equivalents as of April 1 for the six now courses used are obtained from the Water Management Subcommittee Report and from the publication "Water Supply Outlook and

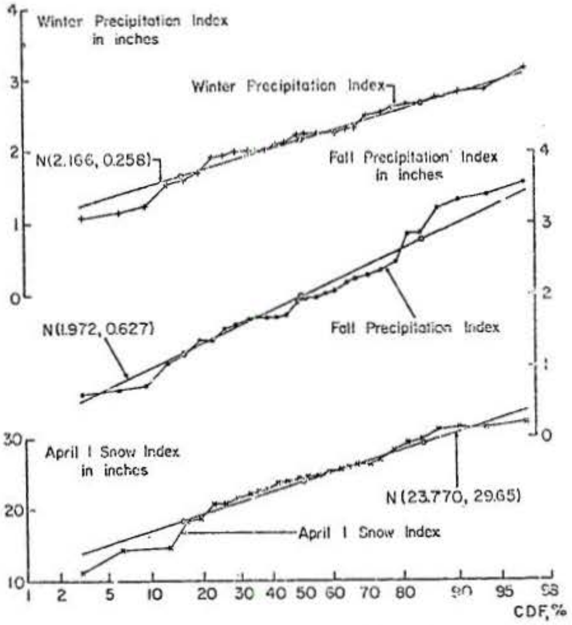


Fig. 10. Fittings of normal probability distribution function to frequency distributions of precipitation and snow index.

Federal - State - Private Cooperative Snow Surveys for Montana" of the Soil Conservation Service. The locations of the six snow courses are shown in Fig. 2 by squares. Some salient features of the snow courses are given in Table 6.

Weights of the five courses used by the Water Management Subcommittee are assigned proportionally to the variance of the observed data at each course. They were modified in such a way that the best correlation between the computed April 1 snow water equivalent index and the seasonal runoff was obtained. The weight of the added Kishenehn is assigned to be equal to the minimum weight of the five courses due to its small variance and low elevation. Weights of the six snow courses are given in Table 6.

TABLE 6

SNOW COURSES

ID No.	Name	Location	Elevation Ft	Weight
12B7	Goat Mountain	47° 39' N - 112° 55' W	7000	0.17
13A2	Desert Mountain	48° 26' N - 113° 58' W	5600	0.19
13A5M	Marias Pass	48° 19.5' N - 113° 21.5' W	5250	0.15
13B3	Big Creek	47° 40.5' N - 113° 57.5' W	6750	0.15
13B7	North Fork Jocko	47° 15.5' N - 113° 46' W	6330	0.19
14A6	Kishenehn	48° 58' N - 114° 25' W	3886	0.15

A sample cumulative distribution function of 30 values of the April 1 snow water equivalent index, from 1940 through 1969, are plotted on normal probability paper, Fig. 10. Again, based on the Smirnov-Kolmogorov test, at 95 percent confidence level, the distribution of the snow water equivalent index is not significantly different from a normal distribution, with the mean 23.77 and the variance 29.65.

Chapter IV

APPLICATION OF CANONICAL CORRELATION

This chapter presents results of correlation analyses of historical data, assembled as described in Chapter III. Using these results, the canonical correlation analysis is applied in the two examples of long-range forecasts.

4.1 Results of Analyses of Historical Data

Since the main purpose of this study is to demonstrate the potential of application of the canonical correlation analysis for hydrologic problems, only the results of correlation analyses for selecting the propper set of independent variables for each example of the long-range prediction are presented. The detailed discussion of the correlation analyses are presented elsewhere, Torranin (1972).

Coastal precipitation forecast. The monthly $\varepsilon'_{p,\tau}(P)$ - series of the three coastal areas are used as dependent variables, the independent variables being the monthly $\delta_{p,\tau}(T)$ - series of the 29 sea surface temperature areas shown in Fig. 6. The data of the $\delta_{p,\tau}(T)$ - series are for the period January 1949 through December 1969, or 21 years. The sample size for each season of precipitation is therefore 21 x 3 = 63.

Results of correlation analysis, Torranin (1972), show that the use of sea surface temperature for forecasting the coastal precipitation by the lag cross correlation method results in a border case of significance. For the purpose of demonstrating the application of canonical analysis, the forecast of the summer precipitation is used as an example, because practically all of the multiple correlation coefficients for summer precipitation proved to be significantly different from zero. For summer precipitation, the group of independent variables consists of the sea surface temperature at areas 16, 1, 27, 18, 28, 23, 4, 3, and 9 with time lag of one month and at areas 21, 27, 8, 14, and 12 with time lag of four months, a total of 14 independent variables.

From the canonical analysis between the group of three precipitation series and 14 sea surface temperature series, the vectors α and γ of coefficients of Eq. 2.21 are:

$$\alpha = [\alpha_1\ \alpha_2\ \alpha_3] = \begin{bmatrix} -.73963 & -.14096 & .77197 \\ -.51061 & -.13355 & -1.32588 \\ .27174 & -1.05020 & .22882 \end{bmatrix}, \quad 4.1$$

$$\gamma = [\gamma_1\ \gamma_2\ \gamma_3] = \begin{bmatrix} .20796 & 1.89924 & 1.44752 \\ -.86164 & -1.54182 & -.70774 \\ .18382 & -1.17142 & .20635 \\ -.16396 & .86917 & -.05202 \\ .11914 & .30801 & .18512 \\ 1.05531 & -.17980 & .22506 \\ -.52306 & .59146 & .11974 \\ .61735 & -.55442 & .07682 \\ -.13884 & .32808 & -.19386 \\ .43695 & .14030 & -.51693 \\ -.52511 & -.09499 & .01119 \\ -.72409 & -.12054 & .40572 \\ .73442 & .10392 & .87328 \\ -.35567 & .12199 & .02133 \end{bmatrix} . \quad 4.2$$

The linear correlation analysis between each of the three pairs of the canonical variables, computed by Eqs. 2.24 and 2.25, gives the following results. The first pair of canonical variables gives R_c=0.778, and

$$U_1 = -0.062 + 0.775\ V_1 . \qquad 4.3$$

The unbiased standard error of estimate is 0.628, the mean of U_1 is -0.038, the variance of U_1 is 1.000, the mean of V_1 is 0.031 and the variance of V_1 is 1.000.

The second pair of canonical variables gives R_c = 0.752, and

$$U_2 = -0.026 + 0.7516\ V_2 . \qquad 4.4$$

The unbiased standard error of estimate is 0.660, the mean of U_2 is -0.014, the variance of U_2 is 1.000, the mean of V_2 is 0.0156 and the variance of V_2 is 1.000.

The third pair of canonical variables gives R_c = 0.409, and

$$U_3 = -0.069 + 0.4095\ V_3 . \qquad 4.5$$

The unbiased standard error of estimate is 0.912, the mean of U_3 is -.069, the variance of U_3 is 1.000, the mean of V_3 is 0.001 and the variance of V_3 is 1.000.

From Fig. 3b, the canonical correlation coefficient of the third pair of canonical variables is not significantly different from zero. Figures 11 a, b, and c show the linear relations and the 80 percent confidence limits for a single forecast. Equations 4.3 through 4.5 are used in the later part of this chapter to demonstrate the application of canonical correlation analysis for monthly coastal precipitation.

The next section presents results of correlation analyses for the purpose of the forecast of snowmelt runoff at the three gaging stations as shown in Fig. 2.

Snowmelt runoff forecast. The correlation analysis in this part of study consists mainly of computing the canonical variables and of using the correlation analysis of each pair of canonical variables. The group of "independent" variables consists only of indices that can be computed from the observed original data as of April 1. These variables include the fall precipitation index, the snow water equivalent index as of April 1, and the winter precipitation index. This group of independent variables, with their definitions and methods of computation given in Chapter III, is used in the canonical analysis with a group of seasonal snowmelt runoffs at the three gaging stations. A linear correlation analysis between the pairs of derived canonical variables gives the basic forecasting equations of the runoff.

For this canonical analysis, the group of dependent variables, $X^{(1)}$ of Eq. 2.9, consists of seasonal runoff at the gaging stations 3585, 3625, and 3630,

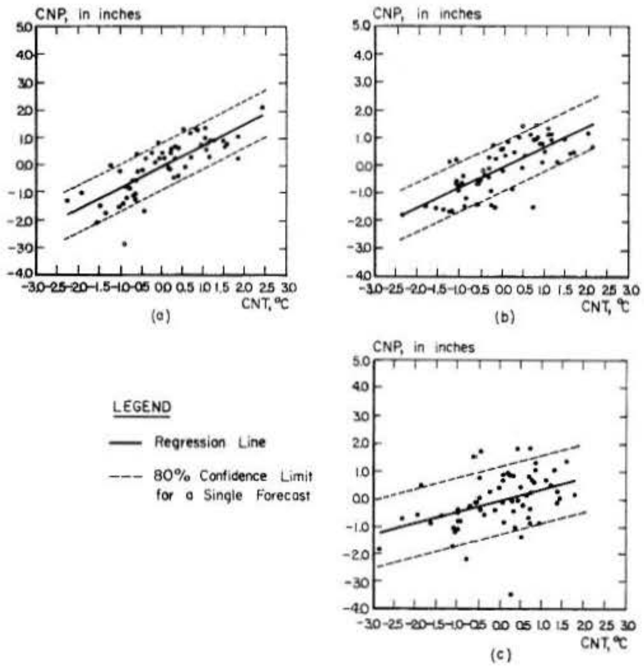


Fig. 11 Linear relation between canonical variables of precipitation, CNP, and of sea surface temperature, CNT, for (a) the first pair, (b) the second pair, and (c) the third pair.

while the group of independent variables, $X^{(2)}$ of Eq. 2.9, consists of the fall precipitation index, the snow water equivalent index, and the winter precipitation index, namely

$$X^{(1)} = \begin{bmatrix} x(1) \\ x(2) \\ x(3) \end{bmatrix} , \qquad 4.6$$

and

$$X^{(2)} = \begin{bmatrix} x(4) \\ x(5) \\ x(6) \end{bmatrix} , \qquad 4.7$$

in which

$x(1)$ = the seasonal runoff at gaging station 3585,
$x(2)$ = the seasonal runoff at gaging station 3625,
$x(3)$ = the seasonal runoff at gaging station 3630,
$x(4)$ = the fall precipitation index,
$x(5)$ = the April snow water equivalent index, and
$x(6)$ = the winter precipitation index.

The correlation matrix of x(1) through x(6) is as follows.

	x(1)	x(2)	x(3)	x(4)	x(5)	x(6)
x(1)	1.000	0.961	0.985	0.097	0.874	0.700
x(2)		1.000	0.973	0.160	0.916	0.692
x(3)			1.000	0.200	0.897	0.699
x(4)				1.000	0.239	0.073
x(5)					1.000	0.798
x(6)						1.000

From this correlation matrix, it is evident that the seasonal flows at the three stations are highly correlated among themselves, as expected. For the independent variables, the snow water equivalent and the winter precipitation index are also highly correlated. The correlations between the snow water equivalent index and each of the seasonal flows are of the same order of magnitude. This is also true for the winter precipitation index, but the correlations of the flow with the winter precipitation index are somwhat lower. Therefore, it is justified to use the indices derived for the flow of the South Fork Flathead River as the indices for the flow at the other two gaging stations of the adjacent river basins.

By using 30 years of data as discussed in Chapter III in the canonical analysis, the matrix α and γ of Eq. 2.21 are obtained as:

$$\alpha = [\alpha_1\ \alpha_2\ \alpha_3] = \begin{bmatrix} .09428 & 1.48235 & -.48860 \\ -.20967 & .24715 & .84349 \\ -.02657 & -.53209 & -.19342 \end{bmatrix} \qquad 4.8$$

$$\gamma = [\gamma_1\ \gamma_2\ \gamma_3] = \begin{bmatrix} .06128 & -1.29588 & -\ .28553 \\ -.20733 & .03498 & .24046 \\ .30034 & .05539 & -3.31876 \end{bmatrix} \qquad 4.9$$

The canonical correlation coefficients between the first, the second and the third pair of canonical variables, or λ_1, λ_2, and λ_3 of Eq. 2.20, are 0.9229, 0.6108, and 0.2059, respectively. From Fig. 3d and the 95 percent level of confidence, only the canonical correlation coefficient of the third pair of canonical variables is not significantly greater than zero.

Using the value of α and γ in Eqs. 4.8 and 4.9 the series of U_i and V_i, $i = 1, 2, 3$ are computed by using Eqs. 2.24 and 2.25. The linear correlation analysis between U_1 and V_1, and between U_2 and V_2, are performed with the following results:

$$U_1 = -0.47040 + 0.92288\ V_1 , \qquad 4.10$$

in which the canonical correlation coefficient $R_c = 0.9229$, the unbiased standard error of estimate is 0.38532, the mean of U_1 is -4.30675, the variance of U_1 is 1.000, the mean of V_1 is -4.15693, the variance of V_1 is 1.000; and

$$U_2 = 0.90555 + 0.61078\ V_2 , \qquad 4.11$$

with the canonical correlation coefficient R_c=0.6108, the unbiased standard error of estimate is 0.79228, the mean of U_2 is -0.5960, the variance of U_2 is 1.000, the mean of V_2 is -1.5802, and the variance of V_2 is 1.000.

Linear relations of Eqs. 4.10 and 4.11 are shown in Fig. 12 with the data points used. The values of α and γ of Eqs. 4.8 and 4.9 and Eqs. 4.10 and 4.11 are used in the forecast with Eqs. 2.34 and 2.38, as described in the later part of this chapter.

Next sections show how the results of the canonical analysis as obtained can be applied in hydrologic forecasts, specifically in the coastal precipitation forecasts and in the snowmelt runoff forecast.

4.2 Examples of Forecast by Using Canonical Correlation Analysis

In general, the main steps to be used in a forecast by the technique of canonical correlation analysis are:

(1) The canonical variables of the set of dependent variables (such as the monthly precipitation for the example of coastal precipitation forecast) are predicted by the canonical variables of the set of observed

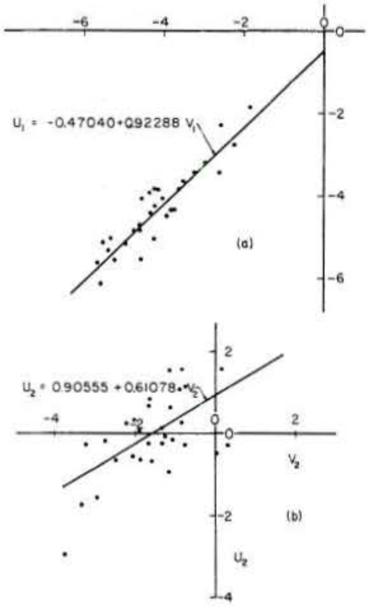


Fig. 12 Linear correlation between
(a) the first pair of canonical variables,
(b) the second pair of canonical variables,
of the snowmelt runoff, U, and the indices, V.

independent variables (such as the monthly sea surface temperature for the example of coastal precipitation forecast).
(2) The predicted canonical variables of dependent variables are transformed back to the predicted value of dependent variables.
(3) A confidence region for the predicted values of dependent variables is then constructed.

It should be noted that the data of the set of independent variables used in each of the following two examples of forecast are not used in the analysis previously described.

Coastal precipitation forecast. The total precipitation for June 1970 at the three coastal areas, as shown in Fig. 4, are forecast simultaneously as an example of previous developments. The group of independent variables to be used for the forecast represent the residuals of the first-order Markov model of standardized deviations of the sea surface temperature for May 1970, a time lag of one month, at areas 16, 1, 27, 18, 28, 23, 4, 3, and 9, as well as the residuals of February 1970, or a time lag of four months, at areas 21, 27, 8, 14, and 12. These 14 variables are arranged in a column vector of independent variables as

$$X^{(2)} = \begin{bmatrix} .6373 \\ -\ .4686 \\ -\ .3455 \\ .2410 \\ .2960 \\ .6253 \\ -\ .4071 \\ -\ .4061 \\ -\ .9917 \\ -\ .5487 \\ -1.1218 \\ .4894 \\ 1.4784 \\ -\ .6738 \end{bmatrix} . \qquad 4.12$$

Introducing the value of γ of Eq. 4.2 into Eq. 2.25, the canonical variables of the sea surface temperature are obtained as

$$V = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} = \begin{bmatrix} 1.48424 \\ -0.13905 \\ -1.04813 \end{bmatrix} . \qquad 4.13$$

The values of U_1 and U_2 are estimated by Eq. 4.3 and 4.4 as:

$$\hat{U}_1 = 1.09337, \text{ and } \hat{U}_2 = 0.13053 .$$

Since the canonical correlation coefficient of the third pair of canonical variables is not statistically significant, the mean value of U_3 is obtained as $\hat{U}_3 = -0.06900$. Therefore, the predicted canonical variables of precipitation for the month of June 1970 are

$$U = \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \hat{U}_3 \end{bmatrix} = \begin{bmatrix} 1.09337 \\ -0.13053 \\ -0.06900 \end{bmatrix} . \qquad 4.14$$

In order to transform the canonical variables back for predicting precipitation the inverse of the matrix α^T of the matrix α in Eq. 4.1 is computed, or

$$[\alpha^T]^{-1} = \begin{bmatrix} -0.9308 & -0.1592 & 0.3745 \\ -0.5092 & -0.2479 & -0.5331 \\ 0.1897 & -0.8993 & 0.0175 \end{bmatrix} . \qquad 4.15$$

From the matrix of Eqs. 4.14 and 4.15, the predicted precipitation at the three areas are obtained from Eq. 2.34 as

$$\hat{X}^{(1)} = \begin{bmatrix} -1.0288 \\ -0.4876 \\ 0.3236 \end{bmatrix} . \qquad 4.16$$

It should be noted that $\hat{X}^{(1)}$ are the predicted values of the normal transforms of standardized deviations of precipitation at the three areas. Equations 3.4 and 3.5 are then used to transform these three values into the standardized deviations at each of the three areas as

$$\hat{\varepsilon} = \begin{bmatrix} \hat{\varepsilon}(1) \\ \hat{\varepsilon}(2) \\ \hat{\varepsilon}(3) \end{bmatrix} = \begin{bmatrix} -1.0228 \\ -0.7080 \\ -0.0387 \end{bmatrix} . \qquad 4.17$$

The predicted total precipitation at each of the three areas is computed by Eq. 3.1, by using the means and standard deviations for the month of June. The predicted total precipitation for June 1970 is then

$$\hat{X}(P) = \begin{bmatrix} \hat{X}_1(P) \\ \hat{X}_2(P) \\ \hat{X}_3(P) \end{bmatrix} = \begin{bmatrix} 1.238 \\ 0.324 \\ 0.065 \end{bmatrix} . \qquad 4.18$$

The observed standardized deviations and the total precipitation for June 1970 are:

$$\tilde{\varepsilon} = \begin{bmatrix} -1.280 \\ .782 \\ -\ .135 \end{bmatrix} , \qquad 4.19$$

and

$$\tilde{X}(P) = \begin{bmatrix} 0.968 \\ 1.113 \\ 0.056 \end{bmatrix} . \qquad 4.20$$

The forecast errors at areas 1, 2, and 3 are 28, 71, and 16 percent of the observed values, respectively.

To construct a confidence region of the predicted precipitation, the matrices U* and E* of Eq.2.37 are computed from the matrices U, $(\alpha^T)^{-1}$, E, and $\{(\alpha^T)^{-1}\}^T$ by

$$U^* = (\alpha^T)^{-1} \hat{U} , \qquad 4.21$$

and

$$E^* = (\alpha^T)^{-1} E\{(\alpha^T)^{-1}\}^T . \qquad 4.22$$

The matrix E consists of the variance of a single forecast made by using the linear regression equation between U_i and V_j, i = j. The error of a single forecast by using a linear regression equation consists of the sampling error of regression and the error resulting from the variation of an actual value of dependent variable around the regression value. For the forecast of U_1, the variance of a single forecast ${e_1}^2$, for $V_1 = 1.48424$, is computed by

$${e_1}^2 = \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(V_1 - \overline{V}_1)}{\sum_{i=1}^{n} \{V_1(i) - \overline{V}_1\}^2} \right] , \qquad 4.23$$

in which $\hat{\sigma}$ is the unbiased standard error of estimate of U_1 by V_1, $\overline{V}_1$ is the mean of V_1 and n is the sample size used for the linear correlation analysis.

Using Eq. 4.23, ${e_1}^2$ is computed to be 0.41405 for $V_1 = 1.48424$. Similarly, ${e_2}^2$ is computed to be 0.44251 for $V_2 = -.13905$. Since the mean of U_3 is forecast, ${e_2}^2$ is the variance of U_3 itself. Therefore, for the forecast of precipitation of June 1970 the matrix E is

$$E = \begin{bmatrix} 0.41405 & 0 & 0 \\ 0 & 0.44251 & 0 \\ 0 & 0 & 1.000 \end{bmatrix} . \qquad 4.24$$

By using Eqs. 4.21 and 4.22 the matrices U* and E* are computed as

$$U^* = \begin{bmatrix} U^*(1) \\ U^*(2) \\ U^*(3) \end{bmatrix} = \begin{bmatrix} -1.0228 \\ -0.4876 \\ 0.3236 \end{bmatrix} , \qquad 4.25$$

and

$$E^* = \begin{bmatrix} 0.5080 & 0.0172 & -0.0033 \\ 0.0172 & 0.4142 & 0.0495 \\ -0.0033 & 0.0495 & 0.3731 \end{bmatrix} \qquad 4.26$$

Note that U* is the forecast precipitation. The inverse of E* is computed as

$$E^{*-1} = \begin{bmatrix} e'_{11} & e'_{12} & e'_{13} \\ e'_{21} & e'_{22} & e'_{23} \\ e'_{31} & e'_{32} & e'_{33} \end{bmatrix} = \begin{bmatrix} 1.9718 & -0.0855 & 0.0286 \\ -0.0855 & 2.4565 & -0.3265 \\ 0.0286 & -0.3265 & 2.7240 \end{bmatrix} \quad 4.27$$

From Eq. 2.38,

$$Q^*(X) = (X^{(1)} - U^*)^T E^{*-1} (X^{(1)} - U^*) \sim \chi^2(3)$$

or

$$\begin{aligned} e'_{11}(x(1)-U^*(1))^2 &+ e'_{22}(x(2)-U^*(2))^2 + e'_{33}(x(3)-U^*(3))^2 \\ &+ 2e'_{12}(x(1)-U^*(1))(x(2)-U^*(2)) \\ &+ 2e'_{13}(x(1)-U^*(1))(x(3)-U^*(3)) \\ &+ 2e'_{23}(x(2)-U^*(2))(x(3)-U^*(3)) \sim \chi^2(3), \end{aligned} \qquad 4.28$$

in which $\chi^2(3)$ is a chi-square distribution with three degrees of freedom. The confidence region for x(1), x(2), and x(3) at 80 percent level of confidence is obtained by Eq. 4.28 as

$$\begin{aligned} e'_{11}(x(1)-U^*(1))^2 &+ e'_{22}(x(2)-U^*(2))^2 + e'_{33}(x(3)-U^*(3))^2 \\ &+ 2e'_{12}(x(1)-U^*(1))(x(2)-U^*(2)) \\ &+ 2e'_{13}(x(1)-U^*(1))(x(3)-U^*(3)) \\ &+ 2e'_{23}(x(2)-U^*(2))(x(3)-U^*(3) \le 4.64, \end{aligned} \qquad 4.29$$

in which 4.64 is the 80th percentile value of the $\chi^2(3)$ distribution, and the e' are those of Eq. 4.27, and U* those of Eq. 4.25.

The interpretation of the confidence region defined by Eq. 4.29 is that there is 80 percent probability that each of x values to be observed for June 1970 will vary around the predicted values U* in such a manner that Eq. 4.29 is satisfied. Equation 4.29 represents an ellipsoid as shown in Fig. 13. Figure 13 shows the ellipses in $\varepsilon'(2) - \varepsilon'(3)$ plane, which are the results of passing a plane through the ellipsoid at $\varepsilon'(1)$ of -.2788, of -1.0228, and of -1.767. Note that -1.0228 is the value of the forecast $\varepsilon'(1)$, while -.2788 and -1.767 are the forecasts of $\varepsilon'(1)$ which have an estimated 70th and 30th percentile error associated with them, respectively. This ellipsoid shows that the errors of precipitation forecasts of the three areas are interrelated. If the precipitation for one area is predicted with a large error, the other two will be predicted with small error, such that there is a 80 percent probability for the observed precipitation (normal transforms of the standardized deviations) of these three areas to be in this ellipsoid.

Snowmelt runoff forecast. For this example, forecasts of snowmelt runoff (April through July runoff) at the three gaging stations, as shown in Fig. 2, for the year 1970 are made. Using the observed values of monthly totals of precipitation in the fall and winter at the five stations, and using the observed April 1 snow water equivalent at the six snow courses, the fall and winter precipitation indices and the snow water equivalent index are computed by applying the weights given in Chapter III. The three indices so computed represent the observed values of a set of idedent variables $X^{(2)}$ of Eq. 4.7, or

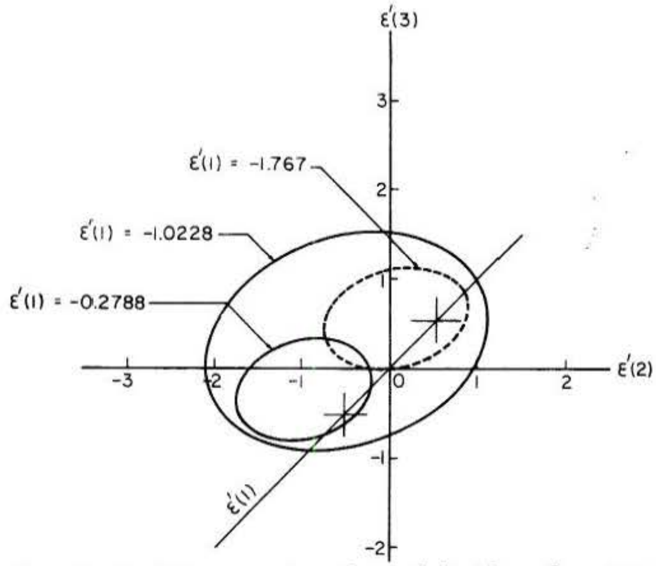


Fig. 13 Confidence region of precipitation forecasts at 80 percent confidence level.

$$X^{(2)} = \begin{bmatrix} 1.501 \\ 26.227 \\ 2.530 \end{bmatrix} . \qquad 4.30$$

The value of γ of Eq. 4.9 is used to compute the canonical variable of these indices by using Eq. 2.25, namely

$$V_1 = 0.06128 \times 1.501 - 0.20733 \times 26.227$$
$$+ 0.30034 \times 2.530 = -4.5858$$

and

$$V_2 = 1.29588 \times 1.501 + 0.03598 \times 26.227$$
$$+ 0.05539 \times 2.530 = -0.86133 .$$

Substituting V_1 and V_2 into Eqs. 4.10 and 4.11, the forecast values of U_1 and U_2 are $\hat{U}_1$=-4.70254, and $\hat{U}_2$=0.37947.

Since the canonical correlation coefficient of the third pair of canonical variables is not statistically significant, the mean value of U_3 is used, namely $\hat{U}_3$=-1.34037. Therefore, the predicted canonical variables of runoff for 1970 are

$$\hat{U} = \begin{bmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \hat{U}_3 \end{bmatrix} = \begin{bmatrix} -4.70254 \\ 0.37947 \\ -1.34037 \end{bmatrix} . \qquad 4.31$$

To transform the canonical variables back for predicting runoff, the inverse of the matrix α^T of the matrix α of Eq. 4.8 is computed as

$$[\alpha^T]^{-1} = \begin{bmatrix} -3.54096 & 0.55600 & -1.04300 \\ -4.82738 & 0.27566 & -0.09518 \\ -12.10702 & -0.20239 & -2.95019 \end{bmatrix} . \qquad 4.32$$

From the matrix $\hat{U}$ of Eq. 4.31 and $[\alpha^T]^{-1}$ of Eq. 4.32, the predicted flows are computed by using Eq. 2.34 as

$$\hat{X}^{(1)} = \begin{bmatrix} \hat{x}(1) \\ \hat{x}(2) \\ \hat{x}(3) \end{bmatrix} = [\alpha^T]^{-1} [\hat{U}] = \begin{bmatrix} 18.26062 \\ 22.93316 \\ 60.81130 \end{bmatrix} . \qquad 4.33$$

The canonical variable $\hat{U}_1$ can be predicted with the smallest error while U_3 has the largest error in the above matrix multiplication of Eq. 4.33. Therefore, the error associated with x(1), x(2) and x(3) are small if the contribution of U_1 to each x is large as compared with the contribution of U_3, or if the ratio of the magnitude of the element of the first column to the element of the third column of all rows of the matrix $(\alpha^T)^{-1}$ in Eq. 4.32 is large. A study of Eq. 4.32 reveals that such is the case for this particular problem, so that a small error in x should be expected.

The observed values of the snowmelt runoff at the three stations for 1970 are

$$\tilde{X}^{(1)} = \begin{bmatrix} 17.54 \\ 21.28 \\ 54.64 \end{bmatrix} \qquad 4.34$$

Therefore, the errors of runoff forecast at stations 3585, 3625, and 3630 are only 4.10, 7.75, and 11.4 percent of observed values, respectively.

The National Weather Service, NOAA also routinely issues forecasts of the snowmelt runoff at these three stations; the flows are forecast for the period April through September. The errors of forecast made by the agency for 1970 snowmelt runoff forecasts at stations 3585, 3625, and 3630 are 3.2, -6.4, and -5.3 percent of the observed values, respectively.

Although the number of the indices used in the forecast involving canonical correlation analysis is small, and despite the fact that it utilized observations only up to April 1, the error in this forecast exceeded only slightly that of the agency forecast. Discounting the Weather Service's long experience and abundance of available indices, it would seem that a forecast based on canonical correlation analysis would be at least equally accurate and probably less expensive.

Variances of a single forecast of U_1 and U_2, e_1^2 and e_2^2, respectively are computed by using equations similar to Eq. 4.23, as e_1^2 = 0.15968 and e_2^2 = 0.68256. Since the mean of $\hat{U}_3$ is predicted, e_3^2 is the variance of U_3 itself. Therefore, the matrix E of Eq. 2.31 for the forecasts of the snowmelt runoff for the year 1970 is

$$E = \begin{bmatrix} 0.15968 & 0 & 0 \\ 0 & 0.68256 & 0 \\ 0 & 0 & 1.000 \end{bmatrix} . \qquad 4.35$$

Using Eqs. 4.21 and 4.22, the matrices U* and E* of Eq. 2.37 are computed as

$$U^* = \begin{bmatrix} U^*(1) \\ U^*(2) \\ U^*(3) \end{bmatrix} = \begin{bmatrix} 18.26062 \\ 22.93316 \\ 60.81130 \end{bmatrix} , \qquad 4.36$$

and

$$E^* = \begin{bmatrix} 3.2649 & 2.9301 & 9.7436 \\ 2.9301 & 3.7818 & 9.5659 \\ 9.7436 & 9.5659 & 31.8476 \end{bmatrix} . \qquad 4.37$$

The forecast flows U* are the same as $\hat{X}^{(1)}$ of Eq. 4.33. The inverse of E* is

$$E^{*-1} = \begin{bmatrix} e'_{11} & e'_{12} & e'_{13} \\ e'_{21} & e'_{22} & e'_{23} \\ e'_{31} & e'_{32} & e'_{33} \end{bmatrix} = \begin{bmatrix} 3.52192 & -0.01338 & -1.07349 \\ -0.01338 & 1.10079 & -0.32655 \\ -1.03749 & -0.32655 & 0.45791 \end{bmatrix} . \qquad 4.38$$

From Eq. 4.29, the confidence region for x(1), x(2), and x(3) at 80 percent level of confidence is

$$3.52192(x(1)-18.26062)^2 + 1.10079(x(2)-22.93316)^2 + 0.45791(x(3)-60.81130)^2 + 2(-0.01338)(x(1)-18.26062)(x(2)-22.93316) + 2(-1.07349)(x(1)-18.26062)(x(3)-60.81130) + 2(-0.32655)(x(2)-22.93316)(x(3)-60.81130) \leq 4.64. \qquad 4.39$$

Equation 4.39 represents an ellipsoid in the space x(1) - x(2) - x(3), with a center located at the predicted values [x(1), x(2), x(3)]. The ellipsoid of Eq. 4.39 is shown in Fig. 14 by three ellipses which are the intersections of the plane x(1) = 17.49, x(1) = 18.26, and x(1) = 18.77 with the ellipsoid. The values of x(1) of 17.49, 18.26 and 18.77 are the predicted values of x(1) with 40th percentile error, no error, and 60th percentile error associated with it, respectively.

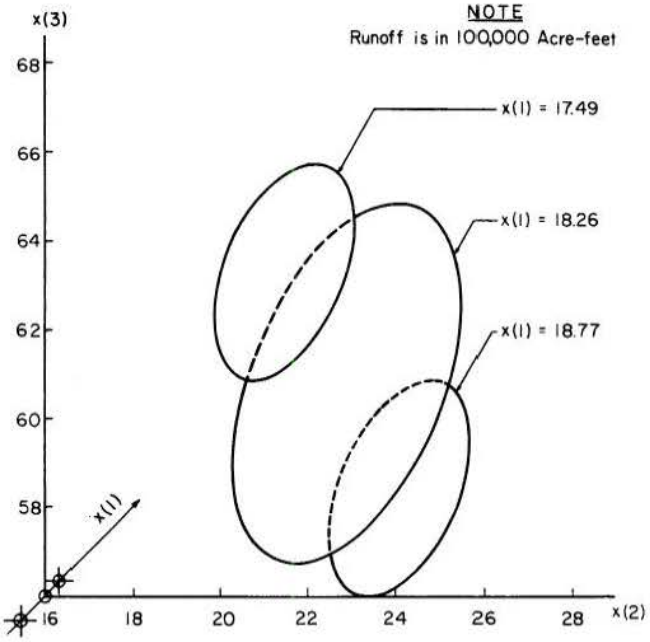


Fig. 14. Confidence region of streamflow forecasts at 80 percent confidence level.

The interpretation of the confidence region of Eq. 4.39, as represented by the ellipsoid of Fig. 14, is that there is an 80 percent probability that the observed values of the flows at stations 3585, 3625 and 3630 for the year 1970, as represented by x(1), x(2), and x(3), respectively, will vary around the forecast values, 18.26, 22.93, and 60.81 in such a way that Eq. 4.39 is satisfied. In other words, the observed flows at the three stations are represented by a point in the three-dimensional space within the ellipsoid of Fig. 14. The largest ellipse of Fig. 14 is the result of the assumption that there is no error in the forecast of x(1). When an error is assumed for x(1), the ellipse becomes smaller. The area of these ellipses or the volume of the ellipsoid, give the forecaster some idea about the variations in the predicted values to be expected for each set of forecasts. Intuitively, one would like to have the forecasts with the confidence region having as small a volume of the ellipsoid as possible, because in that case the overall error of forecast can be expected to be relatively small. The ellipsoidal confidence region for the forecast of 1970 snowmelt runoff is quite small in comparison with the magnitude of predicted values.

In the case of forecast of regional variables, such as in the two problems investigated, it is unlikely any predicted value will be equal to the observed value. The expected variations of predicted values about the observed values at the predicting times are useful information, as far as the overall regional forecast is concerned. The variations of individual predicted values in a region should not be considered separately, since they are correlated. The canonical analysis may be used effectively to obtain this information about the joint variation of predicted values, as shown in this study.

CHAPTER V

CONCLUSIONS

From the results of the investigation of two problems in long-range hydrologic prediction, used to demonstrate the potential for applying canonical correlation analysis to hydrologic problems, the following conclusions concern mainly the application of the canonical correlation analysis, the characteristics of the time series of the variables investigated and the feasibilities of the two long-range prediction problems.

(1) The problem of regional simultaneous forecast of mutually correlated dependent variables of area locations may be solved effectively by using the canonical correlation analysis, especially in constructing a confidence region for these forecasts. The confidence region gives overall information about the joint variation of predicted values. Other advantages observed concern the significance testing of linear correlation between the sets of dependent and independent variables and the saving in analysis by doing only one canonical correlation analysis instead of three separate analyses for each problem.

(2) While the mutual correlation usually observed in a set of time series representing a three-dimensional hydrologic process causes other techniques for the correlation analysis, such as the multiple correlation analysis, to be unsuitable for use with hydrologic data, canonical correlation analysis can be used effectively to investigate linear correlation between two or more hydrologic processes. The technique is very suitable for the investigation of linear relationships between two sets of variables, whose variables are mutually correlated in each set, in addition to a relatively high correlation between the two sets.

(3) The monthly periodic-stochastic time series of the coastal precipitation, after the periodicities in the mean and the standard deviation are removed, produce a standardized residual series that is close to a serially uncorrelated stationary time series. The probability distribution of the residuals is approximately normal for the uppermost coastal area, with mean annual precipitation of 66.9 inches. The distribution of the residuals of the two lower coastal areas are approximately lognormal, with mean annual precipitations of 39.8 and 15.0 inches, respectively.

(4) The monthly sea surface temperature of areas of the Pacific Ocean are used as a set of independent variables in the example of coastal precipitation forecast. After removing the periodicity in the mean and standard deviation in time series of the sea surface temperature, the resulting standardized stochastic time series are shown to be highly serially correlated, approximately of the first-order Markov linear model. The independent (residual) component computed from the Markov model, $\delta_{p,\tau}(T)$-series, is normally distributed.

(5) The contribution of the river base flow to the total snowmelt runoff during the snowmelt measured at each of the three gaging stations is small compared to the snowmelt runoff. The time series of snowmelt runoff, the fall and winter precipitation indices, and the snow water equivalent index are serially uncorrelated time series, with all of them having a normal distribution.

(6) The snowmelt runoff from the river basins has the largest correlation with the snow water equivalent index of all the three indices investigated for the snowmelt runoff forecast. Though the winter precipitation index is highly correlated with the snow water equivalent index, the runoff has a smaller correlation with the winter precipitation index than with snow water equivalent index. The canonical correlation coefficients between the set of the runoff dependent variables and the set of indices as independent variables are 0.923, 0.611, and 0.206; only the third canonical correlation coefficient may be considered as not being statistically significant.

(7) General results of the two examples of forecast by canonical correlation analysis are that the coastal precipitation forecast is not reliable, as indicated by a large percentage error, while the forecast of snowmelt runoff is reliable. The error of prediction at each gaging station in snowmelt runoff forecast is approximately of the same order of magnitude as the error in measuring the runoff itself.

BIBLIOGRAPHY

Anderson, T. W., 1958, An introduction to multivariate statistical analysis, John Wiley & Sons, Inc., New York.

Anderson, H. W., and Westl, A. J., 1965, Snow accumulation and melt in relation to terrain in wet and dry years, Proceedings of the 33rd Annual Meeting Western Snow Conference, Colorado Springs, Colorado, pp. 73-82.

Corps of Engineers, U.S. Army, 1956, Snow hydrology, Summary Reports of the Snow Investigations, pp. 371-405.

Dawdy, D. R., and Feth, J. H., 1967, Application of factor analysis in study of chemistry of the ground water quality, Mojave River Valley, California, Water Resources Research, V. 3, No. 2, pp. 505-510.

Diaz, G., Sewell, J. I., and Shelton, C. H., 1968, An application of principal components analysis and factor analysis in the study of water yield, Water Resources Research, V. 4, No. 2, pp. 299-306.

Dixon, W. J., 1970, BMD Biomedical computer programs, University of California Press, Berkeley, Los Angeles.

Eber, L. E., Saur, J. F. T., and Sette, O. E., 1968, Monthly mean charts sea surface temperature North Pacific Ocean 1949-62, Circular 258, Bureau of Commercial Fisheries, Washington D. C.

Eiselstein, L. M., 1967, A Principal component analysis of surface runoff data from a New Zealand Alpine watershed, Proceedings of the International Hydrology Symposium, Fort Collins, Colorado. V. 1, pp. 479-489.

Kaiser, H. F., 1958, The Varimax criterion for analytic rotation in factor analysis, Psychometrika, V. 23, No. 3, pp. 187-200.

Kendall, M. G., 1957, A course in multivariate analysis, Hafner Publishing Company, Inc., New York.

Klein, W. H., 1964, Application of synoptic climatology and existing numerical prediction to median range forecasting, WMO Technical Note No. 66, pp. 103-125.

Laevastu, T., and Herbert, W. E., 1970, The nature of sea surface temperature anomalies and their effects on weather, Technical Note No. 55, Fleet Numerical Weather Central, Monterey, California.

Marsden, M. A., and Davis, R. T., 1968, Regression on principal components as a tool in water supply forecasting, Proceedings 36th Annual Meeting Western Snow Conference, Lake Tahoe, Nevada, pp. 33-40.

Nimmannit, V., and Morel-Seytoux, H. J., 1969, Regional discrimination of change in runoff, Colorado State University Hydrology Papers, No. 14, pp. 1-65.

Rice, R. M., 1967, Multivariate methods useful in hydrology, Proceedings the International Hydrology Symposium, Fort Collins, Colorado, V. 1, pp. 471-478.

Rice, R. M., 1969, Storm runoff from chaparral watersheds, Ph.D. Dissertation, Colorado State University.

Roesner, L. A. and Yevjevich, V. M., 1966, Mathematical models for time series of monthly precipitation and monthly runoff, Colorado State University Hydrology Papers, No. 15.

Snyder, W. N., 1962, Some possibilities for multivariate analysis in hydrologic studies, Journal of Geophysical Research, V. 67, No. 2, pp. 721-729.

Torranin, P., 1972, Application of canonical correlation in hydrologic predictions, Ph.D. Dissertation, Colorado State University.

Veitch, L. G., and Shepherd, K. J., 1971, A statistical method for flow prediction, River Murray example, Water Resources Research, V. 7, No. 6, pp. 1469-1484.

Water Management Subcommittee, Columbia Basin Inter-Agency Committee, U.S. Bureau of Reclamation, 1964, Derivation of procedures for forecasting inflow to Hungry Horse Reservoir, Montana, U.S. Government Report.

Work, R. A., Beaumont, R. T., and Davis, R. T., 1962, Snowpack ratio in runoff forecasting, Proceedings 30th Annual Meetings Western Snow Conference, Cheyenne, Wyoming, pp. 60-69.

Appendix A

CANONICAL CORRELATION ANALYSIS

This appendix summarizes the mathematical background information for canonical correlation analysis, stressing some particular properties pertinent to the application of this study. Except for the derivation and computation of marginal cumulative distributions of the square of the canonical correlation coefficient, most of the information is extracted from Anderson (1958), and is given in this appendix in a summarized form for the purpose of rapid reference.

Basic Derivation Related to Canonical Correlation Analsis

Let X be a matrix of random variables with p components and with a covariance matrix Σ_{pxp}. For the sake of simplicity, let its vector of mean, Ex, be zero.

Let tne matrix X be partitioned into the two subvectors of p_1 and p_2 components each, or as

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} \quad , \quad p_1 \leq p_2 \quad . \qquad \text{A-1}$$

The covariance matrix is partitioned similarly into p_1 and p_2 rows and columns, as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \quad . \qquad \text{A-2}$$

Let U or V be an arbitrary linear combination of $X^{(1)}$, or of $X^{(2)}$, respectively,

$$U = \alpha^T X^{(1)} \quad , \qquad \text{A-3}$$

and

$$V = \gamma^T X^{(2)} \quad , \qquad \text{A-4}$$

in which α and γ are p_1x1 and p_2x1 column vectors, respectively.

The linear combinations, U or V, having the maximum correlation, are required in canonical correlation analysis. Since this linear correlation between a multiple of U and a multiple of V is the same as the correlation between U and V variables themselves, therefore an arbitrary normalization of α and γ can be made. Let α and γ be selected such that both U and V have unit variances, or

$$EU^2 = E\alpha^T X^{(1)} X^{(1)^T} \alpha = \alpha^T E X^{(1)} X^{(1)^T} \alpha = \alpha^T \Sigma_{11} \alpha = 1 \quad , \qquad \text{A-5}$$

and

$$EV^2 = E\gamma^T X^{(2)} X^{(2)^T} \gamma = \gamma^T E X^{(2)} X^{(2)^T} \gamma = \gamma^T \Sigma_{22} \gamma = 1 \quad .$$

The covariance, or in this case the correlation coefficient, between U and V is

$$EUV = E\alpha^T X^{(1)} X^{(2)} \gamma = \gamma^T \Sigma_{12} \gamma \quad . \qquad \text{A-7}$$

The required variables U and V in the canonical correlation analysis are obtained by maximizing Eq. A-7 subject to Eqs. A-5 and A-6.

Put further

$$\psi = \alpha^T \Sigma_{12} \gamma - \frac{1}{2}\lambda(\alpha^T \Sigma_{11} \alpha - 1) - \frac{1}{2}\mu(\gamma^T \Sigma_{22} \gamma - 1) \quad , \qquad \text{A-8}$$

in which γ and μ are the Lagrangian multipliers, The maximization is obtained by equating to zero the partial derivative of ψ, with respect to the vectors α and γ ,

$$\frac{\partial\psi}{\partial\alpha} = \Sigma_{12}\gamma - \lambda\Sigma_{11}\alpha = 0 \quad . \qquad \text{A-9}$$

and

$$\frac{\partial\psi}{\partial\gamma} = \Sigma^T_{12} - \mu\Sigma_{22}\gamma = 0 \quad . \qquad \text{A-10}$$

Multiplication of Eq. A-9 by α^T and of Eq. A-10 by γ^T gives

$$\alpha^T \Sigma_{12} \gamma - \lambda\alpha^T \Sigma_{11} \alpha = 0 \quad , \qquad \text{A-11}$$

and

$$\gamma^T \Sigma^T_{12} \alpha - \mu\gamma^T \gamma_{22} = 0 \quad . \qquad \text{A-12}$$

Since $\alpha^T \Sigma_{11} \alpha = \gamma^T \Sigma_{22} \gamma = 1$, and $\alpha^T \Sigma_{12} \gamma =$ EUV = EVU $= \gamma^T \Sigma^T_{12} \alpha$, then from Eqs. A-11 and A-12 it is seen that $\lambda = \mu = \alpha^T \Sigma_{12} \alpha$. Therefore, Eqs. A-9 and A-10 may be written as

$$-\lambda\Sigma_{11}\alpha + \Sigma_{12}\gamma = 0 \quad , \qquad \text{A-13}$$

$$\Sigma_{21}\alpha - \lambda\Sigma_{22}\gamma = 0 \quad , \qquad \text{A-14}$$

or in matrix form

$$\begin{bmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} = 0 \quad . \qquad \text{A-15}$$

In order for a nontrivial solution to exist, the matrix on the left side of Eq. A-15 must be singular, or

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} = 0 \quad . \qquad \text{A-16}$$

Equation A-16 is a polynomial equation of degree p with p roots, say with

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_p \quad .$$

From Eq. A-11 it is seen that $\lambda = \alpha^T\Sigma_{12}\gamma = EUV$, or that λ is the correlation coefficient between $U = \alpha^T X^{(1)}$ and $V = \gamma^T X^{(2)}$, when α and γ satisfy Eq. A-15 for some value of λ. Since the maximum correlation coefficient is required, $\lambda = \lambda_1$ is selected. Let a solution of Eq. A-15 for $\lambda = \lambda_1$ be $\alpha^{(1)}, \gamma^{(1)}$, with $U_1 = \alpha^{(1)^T} X^{(1)}$ and $V_1 = \gamma^{(1)^T} X^{(2)}$. Then U_1 and V_1 are the linear combinations of $X^{(1)}$ and $X^{(2)}$, respectively, with a maximum correlation coefficient.

A second linear combination of $X^{(1)}$ and second linear combination $X^{(2)}$ are sought next, such that, of all possible linear combinations, uncorrelated with U_1 and V_1, have the maximum correlation coefficient. This procedure is continued until the r-th step of linear combinations, or

$$U_1 = \alpha^{(1)^T} X^{(1)} \ , \ V_1 = \gamma^{(1)^T} X^{(2)} \ , \ \dots ,$$

$$U_r = \alpha^{(r)^T} X^{(1)} \ , \ V_r = \gamma^{(r)^T} X^{(2)}$$

until the corresponding correlation coefficients $\lambda^{(1)} = \lambda_1, \dots, \lambda^{(r)} = \lambda_r$ of Eq. A-16 are obtained. The next step is to find the linear combinations $U=\alpha^T X^{(1)}$, and $V=\gamma^T X^{(2)}$ which have the maximum correlation coefficient between them as compared to all the linear combinations uncorrelated with (U_1,V_1), (U_2,V_2), ..., (U_r,V_r). The conditions that U be uncorrelated with U_i and V_i, and V uncorrelated with V_i and U_i, $i = 1,2, \dots r$, are:

$$\alpha^T\Sigma_{11}\alpha^{(i)} = 0 \ , \qquad \text{A-17}$$

$$\alpha^T\Sigma_{12}\gamma^{(i)} = 0 \ , \qquad \text{A-18}$$

$$\gamma^T\Sigma_{22}\gamma^{(i)} = 0 \ , \qquad \text{A-19}$$

$$\gamma^T\Sigma_{21}\alpha^{(i)} = 0 \ . \qquad \text{A-20}$$

The correlation between U_{r+1} and V_{r+1} or $EU_{r+1} V_{r+1}$ is to be maximized subject to Eqs. A-5, A-6, A-17, and A-19, for $i = 1,2,3$, and ..., r. Let

$$\psi_{r+1} = \alpha^T\Sigma_{12}\gamma - \tfrac{1}{2}\lambda(\alpha^T\Sigma_{11}\alpha-1) - \tfrac{1}{2}\mu(\gamma^T\Sigma_{22}\gamma-1)$$

$$+ \sum_{i=1}^{r} \nu_i\alpha^T\Sigma_{11}\alpha^{(i)} + \sum_{i=1}^{r} \theta_i\gamma^T\Sigma_{22}\gamma^{(i)} \ , \qquad \text{A-21}$$

in which $\lambda, \mu, \nu_1, \dots, \nu_r, \theta_1, \dots, \theta_r$ are Lagrange multipliers. By taking partial derivatives of ψ_{r+1}, with respect to α and γ and equating them to zero, it can be shown that the maximized ψ_{r+1} is obtained when all the ν and θ multipliers are zero. The maximum correlation coefficient is obtained for the solution of Eq. A-16, say λ_{r+1}, when the values of α and γ come from solution of Eq. A-15 for $\lambda = \lambda_{r+1}$, $\alpha = \alpha^{(r+1)}$, $\gamma = \gamma^{(r+1)}$. Therefore, the (r+1)-th combination of $X^{(1)}$ and $X^{(2)}$ are $U_{r+1} = \alpha^{(r+1)^T} X^{(1)}$ and $V_{r+1} = \alpha^{(r+1)^T} X^{(2)}$, respectively. The total number of pairs of combinations is then p_1.

The results of derivation of the canonical correlation may be summarized as follows: The r-th pair of canonical variables are the linear combinations $U_r = \alpha^{(r)^T} X^{(1)}$ and $V_r = \gamma^{(r)^T} X^{(2)}$, each with the unit variance and uncorrelated with the first (r-1) pairs of canonical variables and having the maximum correlation of all the linear combinations uncorrelated with the first (r-1) - pairs. This correlation gives r-th canonical correlation coefficient which is the r-th largest root of Eq. A-16. The values of $\alpha^{(r)}$ and $\gamma^{(r)}$ are the solution of Eq. A-15 which corresponds to the value of the r-th largest root of Eq. A-16. It should be noted that, in the derivation of the canonical correlation so far, no assumption is made regarding the probability distribution function of the matrix X of random variables.

Probability Distribution of a Quadratic Form

This part of Appendix A is related to the construction of the confidence region of forecasts as given in the previous text.

A quadratic form is defined as

$$Y^T AY = \sum_{1,j=1}^{p} a_{ij}Y_iY_j \ , \qquad \text{A-22}$$

in which $Y^T = (Y_1,Y_2 \dots Y_p)$, and A is a symmetric matrix, $A = (a_{ij})$. The matrix A and the quadratic form are called positive definite if $Y^T AY>0$ for all $Y \neq 0$.

It is to be shown in this appendix that if $X_{px1} \sim N(\mu,\Sigma)$, or

$$f(X_1, X_2, \dots, X_p) = (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu)} \qquad \text{A-23}$$

with Σ positive definite, then

$$Q(X) = (X-\mu)^T \Sigma^{-1} (X-\mu) \sim \chi^2(p) \ . \qquad \text{A-24}$$

Proof.

If Σ is positive definite, there exists a nonsingular matrix B such that $B^T \Sigma B = I$. [Corollary 4, p. 339 of Anderson (1958)]. Therefore

$$\Sigma = (B^T)^{-1} I B^{-1} = (B^T)^{-1} B^{-1} \ ,$$

and

$$\Sigma^{-1} = B B^T \ . \qquad \text{A-25}$$

Let

$$Z = B^T (X-\mu) \ , \qquad \text{A-26}$$

Then

$$\begin{aligned} EZ &= 0 \ , \\ EZZ^T &= EB^T (X-\mu)(X-\mu)^T B \\ &= B^T[E(X-\mu)(X-\mu)^T]B \\ &= B^T\Sigma B \\ &= I \ . \end{aligned}$$

Since a linear transformation of a normal random variable is also normally distributed, therefore

$$Z \sim N\ (0,I) \qquad \text{A-27}$$

$$Z^T Z = \sum_{i=1}^{p} (Z_i)^2 \ , \qquad \text{A-28}$$

$$\sim \chi^2(p) \quad . \qquad \text{A-29}$$

Equation A-29 can be written because the summation of p squares of a standard normal random variable is distributed as chi-square distribution with p degree of freedom.

Consider the product $Z^T Z$,

$$Z^T Z = (X-\mu)^T BB^T (X-\mu) \ ,$$

substituting BB^T from Eq. A-25,

$$Z^T Z = (X-\mu)^T \textstyle\sum^{-1} (X-\mu) \ . \qquad \text{A-30}$$

Therefore $Q(X) = (X-\mu)^T \Sigma^{-1} (X-\mu) \sim \chi^2(p)$, and Eq. A-24 is proved.

Marginal Cumulative Distribution of Square of the Sample Canonical Correlation Coefficients

Let D_i for $i = 1, 2, \ldots, p_1$ be the squares of the sample canonical correlation coefficients of the two sets, $X^{(1)}$ of p_1 components and $X^{(2)}$ of p_2 components. For the case that $X^{(1)}$ is of a multivariate normal distribution and $X^{(1)}$ and $X^{(2)}$ are independently distributed, Anderson (1958) presented the joint probability distribution of D_i to be:

$$g(D_1,D_2\ ,\ldots,\ D_{p_1}) =$$

$$\pi^{\frac{1}{2}p_1} \prod_{i=1}^{p_1} \frac{\Gamma[\frac{1}{2}\ (N-i)]}{\Gamma[\frac{1}{2}\ (N-p_2-i)]\cdot\Gamma[\frac{1}{2}\ (p_1+1-i)]\cdot\Gamma[\frac{1}{2}\ (p_2+1-i)]}$$

$$\prod_{i=1}^{p_1} \left\{ D_i^{\frac{1}{2}\ (p_2-p_1-1)} (1-D_i)^{\frac{1}{2}\ (N-p_2-p_1-2)} \right\} \prod_{i<j}^{p_1} (D_i-D_j) \ , \qquad \text{A-31}$$

in which N is the sample size, and $p_1 \leq p_2$.

For the case $p_1 = 3$, Eq. A-31 becomes

$$g(D_1,D_2,D_3) = k_1 \prod_{i=1}^{3} \left\{ D_i^{k_2}\ (1-D_i)^{k_3} \right\} \prod_{i<j}^{3} (D_i-D_j) \ , \qquad \text{A-32}$$

in which

$$k_1 = \pi^{3/2} \prod_{i=1}^{3} \frac{\Gamma[\frac{1}{2}\ (N-i)]}{\Gamma[\frac{1}{2}\ (N-p_2-i)]\cdot\Gamma[\frac{1}{2}\ (4-i)]\cdot\Gamma[\frac{1}{2}\ (p_2+1-i)]} \qquad \text{A-33}$$

$$k_2 = \frac{1}{2}\ (p_2-4) \ , \qquad \text{A-34}$$

and

$$k_3 = \frac{1}{2}\ (N-p_2-5) \ . \qquad \text{A-35}$$

From the joint probability distribution given by Eq. A-32, the marginal distribution of D_1, D_2 and D_3 can be obtained by integrating out D_2 and D_3, D_1 and D_3, and D_1 and D_2, respectively, as follows:

$$g_1(D_1) = \int_0^1\int_0^1 \{g(D_1,D_2,D_3)dD_2dD_3\} I_{(D_3,D_1)}(D_2)$$

$$\cdot\ I_{(0,D_2)}(D_3) \ , \qquad \text{A-36}$$

in which $g_1(D_1)$ is the marginal distribution of D_1, and $I_{(A,B)}(X)$ is the indicator function of X,

$$I_{(A,B)}(X) = 1 \quad \text{for} \quad A \leq X \leq B$$

$$= 0 \quad \text{otherwise.}$$

Substitute $g(D_1,D_2,D_3)$ from Eq. A-32 in Eq. A-36;

$$g_1(D) = k_1 \int_o^1 \left[\prod_{i=1}^{2} \left\{ D_i^{k_2} (1-D_i)^{k_3} \right\} (D_1-D_2) \int_o^1 \right.$$

$$\left\{ D_3^{k_2} (1-D_3)^{k_3} (D_1-D_3)(D_2-D_3)dD_3 \cdot I_{(0,D_2)}(D_3) \right\}$$

$$\left. \cdot\ dD_2 \right] \cdot I_{(D_3,D_1)}(D_2) \ ,$$

$$= k_1 \left[\int_o^1 \prod_{i=1}^{2} \left\{ D_i^{k_2} (1-D_i)^{k_3} \right\} (D_1-D_2) F_{12}(D_1,D_2) dD_2 \right]$$

$$\cdot\ I_{(D_3,D_1)}(D_2), = k_1 D_1^{k_2} (1-D_1)^{k_3} F_1(D_1) \ , \qquad \text{A-37}$$

in which

$$F_{12}(D_1,D_2) = \int_o^1 \left\{ D_3^{k_2} (1-D_3)^{k_3} (1-D_3)(D_2-D_3) dD_3 \right\} I_{(0,D_2)}(D_3)$$

$$.\ I_{(D_3,D_1)}(D_2) \ , \qquad \text{A-38}$$

and

$$F_1(D_1) = \int_o^1 \left\{ D_2^{k_2} (1-D_2)^{k_3} (D_1-D_2) F_{12}(D_1-D_2) dD_2 \right\} I_{(D_3,D_1)}(D_2) \qquad \text{A-39}$$

By using similar integration techniques, the following are obtained.

$$g_2(D_2) = k_1 D_2^{k_2} (1-D_2)^{k_3} F_2(D_2) \ , \qquad \text{A-40}$$

in which $g_2(D_2)$ is the marginal distribution of D_2, and

$$F_2(D_2) = \int_0^1 D_1^{k_2}(1-D_1)^{k_3}(D_1-D_2)F_{12}(D_1,D_2)dD_1 I_{(D_2,1)}(D_1) \ , \qquad \text{A-41}$$

$$F_{12}(D_1,D_2) = \int_0^1 D_3^{k_2}(1-D_3)^{k_3}(D_1-D_3)(D_2-D_3)dD_3 I_{(0,D_2)}(D_3) \cdot I_{(D_3,D_1)}(D_2) \ . \qquad \text{A-42}$$

$$g_3(D_3) = k_1 D_3^{k_3}(1-D_3)^{k_3} F_3(D_3) \ , \qquad \text{A-43}$$

in which $g_3(D_3)$ is the marginal distribution of D_3, and

$$F_3(D_3) = \int_0^1 D_1^{k_2}(1-D_1)^{k_3}(D_1-D_3)F_{13}(D_1,D_3)dD_1 I_{(D_2,1)}(D_1) \ , \qquad \text{A-44}$$

$$F_{13}(D_1,D_3) = \int_0^1 D_2^{k_2}(1-D_2)^{k_3}(D_1-D_2)(D_2-D_3) \cdot dD_2 \cdot I_{(D_3,D_1)}(D_2) \cdot I_{(D_2,1)}(D_1) \ . \qquad \text{A-45}$$

Numerical integration is used to compute the cumulative distributions of D_1, D_2, D_3 by using Eqs. A-37, A-38, A-39; A-40, A-41, A-42; and A-43, A-44, A-45, respectively. The cumulative distributions for different values of p_2 and N are given in Figure 3.

Appendix B

PRECIPITATION STATIONS SELECTED

Coastal area 1

Sequence Number	Station Number	Name	Latitude	Longitude
1	45.0176	Anacortes	48.52	122.62
2	35.0318	Astor Experimental Sta.	46.17	123.82
3	45.0872	Bremerton	47.57	122.67
4	45.0945	Buckley 1 NE	47.17	122.00
5	45.1233	Cedar Lake	47.25	121.44
6	35.1552	Cherry Grove 2 S	45.42	123.25
7	45.1496	Clearwater	47.58	124.30
8	45.1679	Concrete	48.55	121.77
9	35.1817	Cottage Grove 1 S	43.47	123.04
10	35.2345	Disston 1 NE layng Cr	43.72	122.75
11	35.2673	Estacada 2 SE	45.16	122.19
12	35.4721	Lang Lois	42.93	124.45
13	45.4769	Longview	46.10	122.55
14	45.5880	New Halem	48.41	121.15
15	45.7507	Sedro Woolley	48.30	122.13
16	45.7548	Shelton	47.12	123.06
17	35.8481	Tide Water	44.42	123.91

Coastal area 2

Sequence Number	Station Number	Name	Latitude	Longitude
1	4.0227	Antioach Fibreboard Ml.	38.01	121.46
2	4.0383	Auburn	38.54	121.04
3	4.0693	Berkeley	37.52	122.15
4	4.1018	Bowman Dam	39.27	120.40
5	4.1112	Brooks Farnham Ranch	38.77	122.15
6	35.1055	Brookings	42.05	124.28
7	4.1214	Burney	40.88	121.67
8	4.1277	Calaveras Big Trees	38.28	120.32
9	4.1700	Chester	40.18	121.13
10	4.1715	Chico Experiment Sta.	39.42	121.47
11	4.1784	Clarksburg	38.42	121.53
12	35.1946	Crater Lake NP HQ	42.90	122.13
13	4.2147	Cresent City 1 N	41.77	124.20
14	4.2500	Downieville Ranger Sta.	39.57	120.83
15	4.2910	Euraka WB City	40.80	124.17
16	4.3134	Foresthill Ranger Sta.	39.02	120.82
17	4.3136	Fort Bragg	39.57	123.48
18	35.3455	Grants Pass	42.26	123.19
19	4.3191	Fort Ross	38.31	123.15
20	4.3761	Happy Camp	41.80	123.38
21	4.5188	Los Banos	37.05	120.85
22	4.5346	Mariposa	37.48	119.23
23	4.5449	Mc Cloud	41.16	122.08
24	4.6252	North Fork Ranger Sta.	37.23	119.50
25	4.7109	Potter Valley PH	39.37	123.13
26	35.6907	Prospect 2 SW	42.44	122.31
27	4.7292	Red Bluff WB Airport	40.15	122.25
28	4.7296	Redding Fire Sta No 2	40.58	122.40
29	4.8025	Sawyers Bar Ranger Sta.	41.30	123.13
30	4.8045	Scottia	40.29	124.06
31	4.8353	Sonora	37.59	120.23
32	4.8587	Stony Goerge Reservoir	39.58	122.53
33	4.8928	Tiger Creek PH	38.45	120.48
34	4.9035	Tulelake	41.97	121.47
35	4.9105	Twin Lakes	38.70	120.05
36	4.9490	Weaverville Ranger Sta.	40.73	122.93
37	4.9699	Willows	39.32	122.12
38	4.9814	Wrights	38.08	121.57
39	4.9855	Yosemite Park Headqtrs.	37.75	119.58

Coastal area 3

Sequence Number	Station Number	Name	Latitude	Longitude
1	4.0606	Beaumont	33.56	116.59
2	4.0790	Big Sur State Park	36.15	121.47
3	4.1864	Coalinga	36.15	120.35
4	4.2236	Cuyama	34.93	119.62
5	4.2239	Cuyamaca	32.59	116.35
6	4.2346	Delano	35.78	119.25
7	4.2516	Dry Canyon Reservoir	34.48	118.53
8	4.4022	Hollisler	36.51	121.24
9	4.4204	Idria	36.41	120.67
10	4.5107	Los Alamos	34.75	120.28
11	4.5215	Lytle Creek Ranger Sta.	34.20	117.45
12	4.5756	Mojave	35.05	118.17
13	4.6006	Mount Wilson FC 338 B	34.23	118.07
14	4.6175	Newport Beach Harber	33.60	117.88
15	4.6399	Ojai	34.27	119.15
16	4.6703	Parkfield	35.88	120.43
17	4.7077	Potterville	36.04	119.01
18	4.7253	Randburg	35.37	117.65
19	4.7306	Redlands	34.05	117.18
20	4.7470	Riverside Fire Sta. No 3	33.57	117.24
21	4.7672	Salinas Dam	35.33	120.50
22	4.7740	San Diego NB Airport	32.44	117.10
23	4.7851	San Luis Dam	35.30	120.67
24	4.8839	Tejon Rancho	35.03	118.75
25	4.8967	Topanja Patrol Sta. Fc 6	34.08	118.60
26	4.9087	Tustin Irvin Ranch	33.73	117.78
27	4.9552	Wasco	35.36	119.20

Appendix C

LIST OF SELECTED SYMBOLS

Symbol	Definition
α_i	Column vector of coefficients for the i-th canonical variable of the set of dependent variable
CNP	Canonical variable of precipitation
CNT	Canonical variable of sea surface temperature
D_i	Square of the sample canonical correlation coefficient
ε	Second-order stationary component of a time series
λ_i	The i-th canonical correlation coefficient
k	Recession constant
m_τ	Sample mean of a hydrologic variable for the month τ
μ_τ	Population mean of a hydrologic variable for the month τ
P	Monthly precipitation
ρ_τ	Population autocorrelation coefficient for τ months time lag
Q_i	River base flow of the i-th month
$Q(\cdot)$	Quadratic form
$\Gamma(\cdot)$	Gamma function
γ_i	Column vector of coefficients for the i-th canonical variable of the set of independent variables
R_c	Canonical correlation coefficient
r_k	Sample estimate of ρ_k
s_τ	Sample standard deviation of a hydrologic variable for the month τ
Σ	Covariance matrix
σ_τ	Population standard deviation of a hydrologic variable for the month τ
T	Sea surface temperature
t	Time
τ	Time lag
U_i	The i-th canonical variable of the set of dependent variables
V_i	The i-th canonical variable of the set of independent variables
δ	A sequentially independent stochastic component of second-order stationary time series

KEY WORDS: Multivariate analysis, canonical correlation analysis, long-range precipitation forecast, snowmelt runoff forecast, confidence region of forecasts, precipitation, snowmelt runoff.

ABSTRACT: The potential for application of canonical correlation analysis to hydrologic problems is demonstrated by two problems in long-range hydrologic prediction: 1) forecast of monthly precipitation of three large areas of the West Coast of the United States, and 2) forecast of seasonal snowmelt runoff for three gaging stations in the Flathead River Basin in Montana.

Canonical correlation analysis is found to be effective in investigating linear correlation between two or more three-dimensional hydrologic processes, in which the set of time series of each process are mutually correlated, in addition to a relatively high correlation between the processes themselves. The main advantages of using

KEY WORDS: Multivariate analysis, canonical correlation analysis, long-range precipitation forecast, snowmelt runoff forecast, confidence region of forecasts, precipitation, snowmelt runoff.

ABSTRACT: The potential for application of canonical correlation analysis to hydrologic problems is demonstrated by two problems in long-range hydrologic prediction: 1) forecast of monthly precipitation of three large areas of the West Coast of the United States, and 2) forecast of seasonal snowmelt runoff for three gaging stations in the Flathead River Basin in Montana.

Canonical correlation analysis is found to be effective in investigating linear correlation between two or more three-dimensional hydrologic processes, in which the set of time series of each process are mutually correlated, in addition to a relatively high correlation between the processes themselves. The main advantages of using

KEY WORDS: Multivariate analysis, canonical correlation analysis, long-range precipitation forecast, snowmelt runoff forecast, confidence region of forecasts, precipitation, snowmelt runoff.

ABSTRACT: The potential for application of canonical correlation analysis to hydrologic problems is demonstrated by two problems in long-range hydrologic prediction: 1) forecast of monthly precipitation of three large areas of the West Coast of the United States, and 2) forecast of seasonal snowmelt runoff for three gaging stations in the Flathead River Basin in Montana.

Canonical correlation analysis is found to be effective in investigating linear correlation between two or more three-dimensional hydrologic processes, in which the set of time series of each process are mutually correlated, in addition to a relatively high correlation between the processes themselves. The main advantages of using

KEY WORDS: Multivariate analysis, canonical correlation analysis, long-range precipitation forecast, snowmelt runoff forecast, confidence region of forecasts, precipitation, snowmelt runoff.

ABSTRACT: The potential for application of canonical correlation analysis to hydrologic problems is demonstrated by two problems in long-range hydrologic prediction: 1) forecast of monthly precipitation of three large areas of the West Coast of the United States, and 2) forecast of seasonal snowmelt runoff for three gaging stations in the Flathead River Basin in Montana.

Canonical correlation analysis is found to be effective in investigating linear correlation between two or more three-dimensional hydrologic processes, in which the set of time series of each process are mutually correlated, in addition to a relatively high correlation between the processes themselves. The main advantages of using

this technique concern the significance testing of the linear correlation between the processes, the reduced effort in the correlation analysis, and particularly for the prediction problem as it concerns the construction of a confidence region of the simulataneous predicted values. Though not demonstrated in the examples, canonical correlation analysis can also be used for selecting significant data observation stations for use in the correlation analysis.

A set of forecasts is made for each prediction problem by using the canonical correlation analysis of the historical data. Results of these forecasts indicate that the precipitation prediction is not reliable, while the runoff due to seasonal snowmelt can be well predicted.

Applicability of Canonical Correlations in
Hydrology - Padoong Torranin
Hydrology Paper #58

this technique concern the significance testing of the linear correlation between the processes, the reduced effort in the correlation analysis, and particularly for the prediction problem as it concerns the construction of a confidence region of the simulataneous predicted values. Though not demonstrated in the examples, canonical correlation analysis can also be used for selecting significant data observation stations for use in the correlation analysis.

A set of forecasts is made for each prediction problem by using the canonical correlation analysis of the historical data. Results of these forecasts indicate that the precipitation prediction is not reliable, while the runoff due to seasonal snowmelt can be well predicted.

Applicability of Canonical Correlations in
Hydrology - Padoong Torranin
Hydrology Paper #58

this technique concern the significance testing of the linear correlation between the processes, the reduced effort in the correlation analysis, and particularly for the prediction problem as it concerns the construction of a confidence region of the simulataneous predicted values. Though not demonstrated in the examples, canonical correlation analysis can also be used for selecting significant data observation stations for use in the correlation analysis.

A set of forecasts is made for each prediction problem by using the canonical correlation analysis of the historical data. Results of these forecasts indicate that the precipitation prediction is not reliable, while the runoff due to seasonal snowmelt can be well predicted.

Applicability of Canonical Correlations in
Hydrology - Padoong Torranin
Hydrology Paper #58

this technique concern the significance testing of the linear correlation between the processes, the reduced effort in the correlation analysis, and particularly for the prediction problem as it concerns the construction of a confidence region of the simulataneous predicted values. Though not demonstrated in the examples, canonical correlation analysis can also be used for selecting significant data observation stations for use in the correlation analysis.

A set of forecasts is made for each prediction problem by using the canonical correlation analysis of the historical data. Results of these forecasts indicate that the precipitation prediction is not reliable, while the runoff due to seasonal snowmelt can be well predicted.

Applicability of Canonical Correlations in
Hydrology - Padoong Torranin
Hydrology Paper #58

LIST OF PREVIOUS 25 PAPERS

No. 33 Runs of Precipitation Series, by Jose Llamas and M. M. Siddiqui, May 1969.

No. 34 Statistical Discrimination of Change in Daily Runoff, by Andre J. Dumas and Hubert J. Morel-Seytoux, August 1969.

No. 35 Stochastic Process of Precipitation, by P. Todorovic and V. Yevjevich, September 1969.

No. 36 Suitability of the Upper Colorado River Basin for Precipitation Management, by Hiroshi Nakamichi and H. J. Morel-Seytoux, October 1969.

No. 37 Regional Discrimination of Change in Runoff, by Viboon Nimmanit and Hubert J. Morel-Seytoux, November 1969.

No. 38 Evaluation of the Effect of Impoundment on Water Quality in Cheney Reservoir, by J. C. Ward and S. Karaki, March 1970.

No. 39 The Kinematic Cascade as a Hydrologic Model, by David F. Kibler and David A. Woolhiser, February 1970.

No. 40 Application of Run-Lengths to Hydrologic Series, by Jaime Saldarriaga and Vujica Yevjevich, April 1970.

No. 41 Numerical Simulation of Dispersion in Groundwater Aquifers, by Donald Lee Reddell and Daniel K. Sunada, June 1970.

No. 42 Theoretical Probability Distribution for Flood Peaks, by Emir Zelenhasic, December 1970.

No. 43 Flood Routing Through Storm Drains, Part I, Solution of Problems of Unsteady Free Surface Flow in a Storm Drain, by V. Yevjevich and A. H. Barnes, November 1970.

No. 44 Flood Routing Through Storm Drains, Part II, Physical Facilities and Experiments, by V. Yevjevich and A. H. Barnes, November 1970.

No. 45 Flood Routing Through Storm Drains, Part III, Evaluation of Geometric and Hydraulic Parameters, by V. Yevjevich and A. H. Barnes, November 1970.

No. 46 Flood Routing Through Storm Drains, Part IV, Numerical Computer Methods of Solution, by V. Yevjevich and A. H. Barnes, November 1970.

No. 47 Mathematical Simulation of Infiltrating Watersheds, by Roger E. Smith and David A. Woolhiser, January 1971.

No. 48. Models for Subsurface Drainage, by W. E. Hedstrom, A. T. Corey and H. R. Duke, February 1971.

No. 49 Infiltration Affected by Flow of Air, by David B. McWhorter, May 1971.

No. 50 Probabilities of Observed Droughts, by Jaime Millan and Vujica Yevjevich, June 1971.

No. 51 Amplification Criterion of Gradually Varied, Single Peaked Waves, by John Peter Jolly and Vujica Yevjevich, December 1971.

No. 52 Stochastic Structure of Water Use Time Series, by Jose D. Salas-La Cruz and Vujica Yevjevich, June 1972.

No. 53 Agricultural Response to Hydrologic Drought, by V. J. Bidwell, July 1972.

No.54 Loss of Information by Discretizing Hydrologic Series, by Mogens Dyhr-Nielsen, October 1972.

No.55 Drought Impact on Regional Economy, by Jaime Millan, October 1972.

No.56 Structural Analysis of Hydrologic Time Series, by Vujica Yevjevich, November 1972.

No.57 Range Analysis for Storage Problems of Periodic-Stochastic Processes, by Jose Salas La-Cruz, December 1972.