MULTIVARIATE FUNCTIONS APPLIED TO LONG-RANGE WEATHER FORECASTING

by

DAVID REGINALD THOMAS

A THESIS

submitted to

OREGON STATE UNIVERSITY

in partial fulfillment of
the requirements for the
degree of

MASTER OF SCIENCE

June 1962

APPROVED:

Redacted for privacy

Professor of Statistics

In Charge of Major

Redacted for privacy

Chairman of Department of Statistics

Redacted for privacy

Chairman of Graduate School Committee

Redacted for privacy

Dean of Graduate School

Date thesis is presented May 15, 1962

Typed by Carol Baker

ACKNOWLEDGEMENTS

Appreciation is extended to Dr. Lyle Calvin for his help and guidance in this study. Thanks are also conveyed to Dr. Richard Link for his contributions to the project.

The author wishes to acknowledge the research grant from the Bonneville Power Administration which made this study possible.

Appreciation is also extended to the Western Data Processing Center for the use of their computing facilities. The author is further indebted to Mr. David Oatey and Mr. William Anderson of the Western Data Processing Center for their tutoring in computer programming.

TABLE OF CONTENTS

MULTIVARIATE FUNCTIONS APPLIED TO LONG-RANGE WEATHER FORECASTING

CHAPTER 1

INTRODUCTION

In recent years our business and industrial economy has begun to make effective use of weather forecasting. In a talk given to the October 1957 meeting of the American Meteorology Society Thomas F. Malone (6, p. D-21) reported on a survey that was made several years ago. The study was an attempt to place a monetary value on applied meteorology in the United States. The survey was made on business, industry, and agriculture and reported a value of approximately one billion dollars savings per year. These savings were realized by utilization of daily reports, forecasts, storm warnings, and climatological data prepared by the U.S. Weather Bureau. With the incentive for making even larger savings which result from better and longer forecasts, considerable emphasis has been placed on meteorological research.

The general problem of forecasting weather may be divided into two parts. First, it is necessary to predict future atmospheric circulation and temperature patterns; and second, it is necessary to interpret these circulation and temperature patterns in terms of the resulting weather conditions which will exist at various localities.

Effective interpretation of circulation and temperature patterns in terms of local weather conditions is a subjective procedure that takes years of meteorological training and experience. However for the prediction of circulation and temperature patterns, both subjective and objective methods are currently being used.

This thesis will be devoted to an objective method for the construction and evaluation of mathematical models which can be used for predicting circulation and temperature patterns. Statistical methods will be used to construct mathematical models which will attempt to connect future weather with past events. The statistical methods used in this thesis have been known to statisticians for considerable time as parts of the topics of multivariate and regression analysis, but only in recent years have these methods been used for forecasting weather.

Gilman (1957) made use of multivariate analysis for the construction of spatial patterns for 30-day mean values of surface pressure and surface temperature. He made use of regression analysis for constructing prediction equations based on these spatial patterns. Gilman reported that prediction equations constructed for predicting the 30-day mean surface temperature patterns over the United States concurrently and one month in advance from the 30-day surface pressure patterns over the Northern Hemisphere

were to some extent successful (6, p. 87 - 96).

The methods used in this thesis for constructing spatial patterns and prediction equations based on these patterns are basically the same methods that Gilman used. This study differs from Gilman's mainly in that this study is concerned with the use of upper-air data, 700 millibar height, 700 millibar temperature, in addition to surface pressure. This study will also be concerned with predictions for all 12 months of the year in contrast to Gilman's study of only the winter months, December, January, February, and March. The analysis for this thesis was programmed by the author and was run on the IBM 7090 computer at the Western Data Processing Center in Los Angeles, California.

CHAPTER 2

WEATHER DATA

Source of Weather Data

Data were obtained from the National Weather Records Center for the following three weather elements: surface pressure, 700 millibar height, and 700 millibar temperature. Measurements of these three elements at 108 grid point intersections over the Northern Hemisphere were available. The grid point intersections used were at every 10^{o} latitude, 30^{o} N - 80^{o} N, and every 20^{o} longitude, 000^{o} - 340^{o}. Measurements of surface pressure and 700 millibar height on every other day from March 1, 1948 through June 31, 1960 were available. The first measurements for 700 millibar temperature began one year later, March 2, 1949. March 1 or March 2 was used as beginning days for each of the 360-day "weather years".

The data for the last few days of February were not used in order that the 30-day means would represent the same part of each year. Although the second 30-day mean in the "weather year" may contain an observation from March, it still will be referred to as the April mean. The same terminology holds for the other 30-day means. In other words, the 30-day means referred to as monthly means are approximate terminology used for the sake of

convenience.

The data, with corresponding coding information as to location and time, were available in punched-card form. When the data were checked, a number of omissions and errors were found, and corrections were made from the Weather Bureau Historical Series Data and Maps. Where it was possible to fill in missing data directly from these records it was done. When it was not possible to obtain data for particular grid points at particular times, interpolations were made. Interpolations were made across geographical locations when possible. When it was not possible, they were made across time. For example, to obtain a missing value at one location, the average of the values on each side were used. If these values were not available, the average of the values at the same location on days immediately before and after the missing value day were used. By using one of these two methods it was possible to correct all errors detected.

Thirty-Day Mean Deviations

Thirty-day means (over 15 observations) were taken for each of the three elements at each of the 108 grid points. Deviations from the 9-year monthly means, March 1949-February 1958, were calculated for each 30-day period. This can be written in symbolic

form as

$$(2.1) \qquad \theta^{o}_{ik} = \theta^{oo}_{ik} - \overline{\theta}^{oo}_{i'k} .$$

θ^{o}_{ik} represents the ith 30-day mean deviation ordered in time at the kth grid point for the weather element θ. θ^{oo}_{ik} and $\overline{\theta}^{oo}_{i'k}$ represent respectively the original 30-day mean and the 9-year mean for the same ith month. For example, if i represented April 1954, then then θ^{o}_{ik} would be the difference between the mean for April 1954 and the mean for the 9 Aprils from 1949-1957 for the weather variable θ at the kth grid point.

Breakdown of Data into Development and Test Samples

The year was divided into the three seasons listed in the following table.

Table 1

Breakdown of the Year into Seasons

Spring	Summer	Winter
March	July	November
April	August	December
May	September	January
June	October	February

The data for each of the nine combinations of weather elements and seasons were separated into a "development sample" and a "test sample". The development samples contain the 30-day mean deviations for the 9 years from March 1949 through February 1958, and the test samples contain the remainder of the 30-day mean deviations. The development samples are used to construct spatial patterns and prediction equations based on these spatial patterns. The test samples are used to evaluate the accuracy of the prediction equations when they are applied to a portion of the population which did not enter into their construction.

CHAPTER 3

SPATIAL WEATHER PATTERNS

Principal Component Analysis

Principal component analysis is a branch of multivariate analysis, i. e. , analysis of two or more interrelated variables. Principal components which have special properties in terms of variances and covariances are linear combinations of these interrelated variables. These linear combinations can be written in matrix form as,

(3. 1a) $$X = \Theta M,$$

or in expanded matrix form as,

(3. 1b) $$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1s} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2s} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{is} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{ns} \end{bmatrix} =$$

$$\begin{bmatrix} \theta_{11} & \theta_{12} \cdots \theta_{1k} \cdots \theta_{1s} \\ \theta_{21} & \theta_{22} \cdots \theta_{2k} \cdots \theta_{2s} \\ \vdots & \vdots \\ \theta_{i1} & \theta_{i2} \cdots \theta_{ik} \cdots \theta_{is} \\ \vdots & \vdots \\ \theta_{n1} & \theta_{n2} \cdots \theta_{nk} \cdots \theta_{ns} \end{bmatrix} \begin{bmatrix} m_{11} & m_{12} \cdots m_{1j} \cdots m_{1s} \\ m_{21} & m_{22} \cdots m_{2j} \cdots m_{2s} \\ \vdots & \vdots \\ m_{k1} & m_{k2} \cdots m_{kj} \cdots m_{ks} \\ \vdots & \vdots \\ m_{s1} & m_{s2} \cdots m_{sj} \cdots m_{ss} \end{bmatrix} .$$

The general terms, θ_{ik}, m_{kj}, x_{ij}, of the three matrices above are defined as follows: θ_{ik} is the value for the ith observation on the kth variable; m_{kj} is the value for the jth multiplier associated with the kth variable; x_{ij} is the ith value for the jth principal component.

In order for the columns of X to fulfill the requirements for principal components, the following two conditions must hold:

(3. 2i) $M' = M^{-1}$; that is, M must be an orthogonal matrix, and

(3. 2ii) the elements for a column of the M matrix must be found such that $\sum_{i=1}^{n} x_{ij}^2$ will have a maximum value under the condition that $\sum_{i=1}^{n} x_{ij}x_{ih} = 0$ for $h = 1, 2, \ldots j-1$, when $j \geq 2$. The columns, $M_{(j)}$, of M are found in the order $j = 1, 2, \ldots s$.

By multiplying X on the left by its transpose, we have

(3. 3) $$X'X = [\Theta M]'[\Theta M]$$

$$= M'[\Theta'\Theta] M$$

$$= M'RM,$$

where

(3. 4) $$R = \Theta'\Theta.$$

In order to satisfy conditions (3. 2i) and (3. 2ii), the columns of M must be chosen in such a way as to be orthogonal to one another and to make X'X a diagonal matrix. A derivation of the reduction of a symmetric matrix to its diagonal form, using conditions (3. 2i) and (3. 2ii), is given by T. W. Anderson (1, p. 273-277). The procedure for finding the columns of M so that conditions (3. 2i) and (3. 2ii) hold is known as diagonalization of a symmetric matrix. This procedure is also known as finding the characteristic roots, or eigenvalues, and the corresponding characteristic vectors, or eigenvectors. The eigenvalues are the diagonal elements of X'X and the corresponding eigenvectors are the columns of M. Iterative methods for finding the eigenvalues and corresponding eigenvectors of a symmetric matrix have been programmed for computers. For this study Jacobi's method for diagonalization was used. Even with the facility of a high speed computer, diagonalization of a large matrix is a time-consuming procedure. Without the facility of a computer the diagonalization of a large matrix would be virtually impossible.

Once the values for the eigenvectors have been found, the values for the principal components are obtained by the matrix multiplication shown in (3.1b).

A direct consequence of (3.2ii) is that

$$\sum_{j=1}^{s}\sum_{i=1}^{n} x_{ij}^2 = \sum_{k=1}^{s}\sum_{i=1}^{n} \theta_{ik}^2 . \tag{3.5}$$

In other words, an orthogonal transformation preserves the sums of squares of the original matrix. A proof of (3.5) is given by T. W. Anderson (1, p. 277).

By using expression (3.1b) the mean of the jth principal component can be expressed as

$$\bar{x}_{\cdot j} = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{3.6}$$

$$= \frac{\sum_{k=1}^{s} m_{kj} \sum_{i=1}^{n} \theta_{ik}}{n}$$

$$= \sum_{k=1}^{s} m_{kj} \bar{\theta}_{\cdot k} .$$

If the s variables all have sample means equal to zero, i.e.,

$$\bar{\theta}_{\cdot k} = 0, \text{ for } k = 1, 2, \ldots s, \tag{3.7}$$

then it follows from (3.6) that

$$\bar{x}_{\cdot j} = 0, \text{ for } j = 1, 2, \ldots s . \tag{3.8}$$

If (3.7) is satisfied, and consequently (3.8), then

$$\sum_{j=1}^{s} v(x_{(j)}) = \frac{\sum_{j=1}^{s}\sum_{i=1}^{n}(x_{ij} - \bar{x}_{\cdot j})^2}{n-1} \tag{3.9}$$

$$= \sum_{j=1}^{s}\sum_{i=1}^{n} x_{ij}^2 \Big/ (n-1),$$

where $v(x_{(j)})$ represents the sample variance for the jth component. Also from (3.7) we have

$$\sum_{k=1}^{s} v(\theta_{(k)}) = \frac{\sum_{k=1}^{s}\sum_{i=1}^{n}(\theta_{ik} - \bar{\theta}_{\cdot k})^2}{n-1} \tag{3.10}$$

$$= \frac{\sum_{k=1}^{s}\sum_{i=1}^{n}\theta_{ik}^2}{n-1} .$$

If (3.7) is satisfied then from (3.5), (3.9), and (3.10) we have

$$\sum_{j=1}^{s} v(x_{(j)}) = \sum_{k=1}^{s} v(\theta_{(k)}) . \tag{3.11}$$

In other words, the sum of variances of s principal components is equal to the sum of variances of the original s variables.

Principal Components and Their Corresponding Spatial Weather Patterns

In meteorological terms, the multipliers in a column of the M matrix may be thought of as values for a spatial weather pattern. Curves may be drawn on a map of the Northern Hemisphere through geographical locations with equal multipliers. The pattern that these curves form is referred to as a spatial weather pattern. Since the columns of the M matrix are orthogonal, the corresponding patterns may be referred to as orthogonal spatial patterns.

The s interrelated variables, $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(s)}$ take on the values of the "standardized" 30-day mean deviations that occur at the s grid points. A "standardized" 30-day mean is defined to be

$$\theta_{ik} = \frac{\theta^{o}_{ik}}{w_k} , \tag{3.12}$$

where

$$w_k = \sqrt{\sum_{i=1}^{n} \theta^{o2}_{ik}} . \tag{3.13a}$$

From expression (2.1) it follows that

$$\bar{\theta}^{o}_{\cdot k} = 0, \tag{3.14}$$

so that expression (3.13a) may also be written as

(3.13b)
$$w_k = \sqrt{\sum_{i=1}^{n} (\theta^{o}_{ik} - \bar{\theta}^{o}_{.k})^2}$$

$$= \sqrt{(n-1)\, v(\theta^{o}_{(k)})}.$$

In other words, the 30-day mean deviations are weighted by the reciprocal of the product of the constant $\sqrt{n-1}$ and the development sample standard deviations at the respective grid points. If the 30-day means were not standardized, it can be seen from (3.2ii) and (3.11) that the grid points at which the 30-day means had large variances would dominate those at which the variances were small in the selection of orthogonal spatial patterns. From (3.12) and (3.13a) it follows that the variances of the 30-day means at all the grid points are the same; that is,

(3.15)
$$v(\theta_{(k)}) = \frac{\sum_{i=1}^{n} (\theta_{ik} - \bar{\theta}_{.k})^2}{n-1}$$

$$= \frac{\sum_{i=1}^{n} \theta^{2}_{ik}}{n-1}$$

$$= \frac{1}{n-1} \left(\frac{\sum_{i=1}^{n} \theta^{o2}_{ik}}{\sum_{i=1}^{n} \theta^{o2}_{ik}} \right)$$

$$= \frac{1}{n-1} .$$

Because $w_k = \sqrt{\sum_{i=1}^{n} \theta_{ik}^{o2}}$ and $\bar{\theta}_{.k}^{o} = 0$, R in expression (3.4) is the s x s correlation matrix of 30-day mean deviations among the s grid points.

Written in expanded matrix form,

(3.16)
$$R = \begin{bmatrix} r_{11} & r_{12} \cdots r_{1k} \cdots r_{1s} \\ r_{21} & r_{22} \cdots r_{2k} \cdots r_{2s} \\ \vdots & \vdots \quad \vdots \quad \vdots \\ r_{g1} & r_{g2} \cdots r_{gk} \cdots r_{gs} \\ \vdots & \vdots \quad \vdots \quad \vdots \\ r_{s1} & r_{s2} \cdots r_{sk} \cdots r_{ss} \end{bmatrix},$$

where

(3.17)
$$r_{gk} = \sum_{i=1}^{n} \theta_{ig} \, \theta_{ik}$$

$$= \frac{\sum_{i=1}^{n} \theta_{ig}^{o} \, \theta_{ik}^{o}}{\sqrt{\sum_{i=1}^{n} \theta_{ig}^{o2}} \sqrt{\sum_{i=1}^{n} \theta_{ik}^{o2}}}$$

is the correlation between the 30-day mean deviations at the grid points g and k. In other words, the correlation matrix of the 30-day mean deviations among the s grid points is diagonalized in order to find the orthogonal spatial weather patterns.

As Lorenz (5, p. 2) indicates, atmospheric variations are predictable because they are governed by physical laws which presumably do not change with respect to time. These laws tell us that past and future weather are in some way related.

As Gilman (3, p. 21) indicates, experience shows us that the principal components which have small variances are those of smaller physical scale. The small variances result from small-scale and random local effects.

In light of these discussions by Lorenz and Gilman, we shall use the orthogonal spatial patterns in an attempt to filter the predictable atmospheric variations (those governed by underlying physical laws) from the unpredictable atmospheric variations (those due to small-scale and random local effects) of the 30-day mean deviations.

The question that now arises is, how many of these orthogonal spatial patterns are useful in accounting for the predictable atmospheric variations of the 30-day mean deviations. Chapter 5 contains a study of this question, but for the present discussion the number of patterns of interest is that number which accounts for approximately 90 percent of the total variance over the development sample of the standardized 30-day means. Using (3.9), (3.11), and (3.15) we have

$$(3.19) \qquad \sum_{j=1}^{s} \left(\sum_{i=1}^{n} x_{ij}^2 \right) = s.$$

The number of spatial patterns which account for approximately 90 percent of the total variance is the number r, chosen so that

$$(3.20) \qquad \sum_{j=1}^{r} \left(\sum_{i=1}^{n} x_{ij}^2 \right) \cong (.90)\, s.$$

The number r also satisfies the following relations:

$$(3.21) \qquad r \leq s$$

$$r \leq n - d;$$

that is, the maximum number of principal components needed to account for all of the variance is the smaller of the numbers s or n-d, where d is the number of linear dependencies among the rows of X (2, p. 115). For this study, at most 32 principal components explain all of the variance among the standardized 30-day means, because there are 4 linear dependencies among the rows of the X matrix.

Predictors and Predictants

The first r principal components which have the largest variances and account for approximately 90% of the total variance for the development sample will be used as the predictors

(independent variables) in forming the prediction equations. Discarding the remaining s - r columns in the X and M matrices, (3.1a) is rewritten as

$$X_{(\theta)} = \Theta M_{(\theta)} . \tag{3.22}$$

The subscript θ is used to identify the weather element that the predictors are obtained from. For this study there will be nine expressions of the form (3.22); the nine expressions result from the different combinations of weather elements and seasons.

The matrix of the values for the predictants is defined by the product matrix,

$$Y_{(\phi, L)} = \Phi_{(L)} M_{(\phi)}, \tag{3.23}$$

the subscript ϕ denotes the weather element for which the 30-day mean deviations are to be predicted L months in advance. If prediction equations are to be formed for predicting 30-day means one month in advance, then $L = 1$. With respect to weights, w_k's, and orthogonal spatial patterns, $m_{(j)}$'s, the 30-day mean deviations are treated as though they were moved back in time L months. For example, if predictants are formed for the spring season, with $L = 1$, the months of April, May, June, and July are weighted by the weights obtained from the months of

March, April, May, and June. The orthogonal spatial patterns, $M_{(\phi)}$, formed from standardized 30-day means for the months of March, April, May, and June are used in (3. 23). This means that the 30-day weighted means in the columns of $\Phi_{(L)}$ may not all have equal variances such as the standardized 30-day means in the columns of Θ did (see 3.15), but for small L (L = 1, 2) the variances should be nearly equal. In other words, some grid points may carry slightly more weight, in terms of variance, than other grid points for determining values in the $Y_{(\phi, L)}$ matrix.

CHAPTER 4

PREDICTION EQUATIONS

Linear least-squares prediction equations are formed from the x's (predictors) and the y's (predictants) that were found in Chapter 3. The ith term for the kth predictant can be written as the following linear combination of the ith terms for the k predictors:

$$(4.1) \quad y_{ik} = b_{1k}x_{i1} + b_{2k}x_{i2} + \ldots + b_{jk}x_{ij} + \ldots + b_{rk}x_{ir} + e_{ik}(r) = \hat{y}_{ik} + e_{ik}(r),$$

where

$$(4.2) \quad \hat{y}_{ik} = b_{1k}x_{i1} + b_{2k}x_{i2} + \ldots + b_{jk}x_{ij} + \ldots + b_{rk}x_{ir}.$$

The b's are the prediction coefficients and $e_{ik}(r)$ is a residual term which depends on the choice of r. $\hat{y}_{ik}$ is the estimated value of y_{ik} which is obtained by using the prediction equations in (4.2).

The b's are found so as to minimize the sum of squares of the residual terms over the development sample. In other words, the prediction coefficients are chosen such that

$$(4.3)\quad \sum_{i=1}^{n} e_{ik}^2(r) = \sum_{i=1}^{n} (y_{ik} - b_{1k}x_{i1} - b_{2k}x_{i2} - \ldots - b_{jk}x_{ij} - \ldots - b_{rk}x_{ir})^2$$

is a minimum.

In order to find the b's which minimize (4.3), the following equations are solved for the b's:

$$(4.4)\quad \frac{\partial(\sum_{i=1}^{n} e_{ik}^2(r))}{\partial b_{1k}} = 2\sum_{i=1}^{n}(y_{ik} - (\sum_{j=1}^{r} b_{jk}x_{ij})(-x_{i1})) = 0$$

$$\frac{\partial(\sum_{i=1}^{n} e_{ik}^2(r))}{\partial b_{2k}} = 2\sum_{i=1}^{n}(y_{ik} - (\sum_{j=1}^{r} b_{jk}x_{ij})(-x_{i2})) = 0$$

$$\vdots$$

$$\frac{\partial(\sum_{i=1}^{n} e_{ik}^2(r))}{\partial b_{jk}} = 2\sum_{i=1}^{n}(y_{ik} - (\sum_{j=1}^{r} b_{jk}x_{ij})(-x_{ij})) = 0$$

$$\vdots$$

$$\frac{\partial(\sum_{i=1}^{n} e_{ik}^2(r))}{\partial b_{rk}} = 2\sum_{i=1}^{n}(y_{ik} - (\sum_{j=1}^{r} b_{jk}x_{ij})(-x_{ir})) = 0 .$$

Equations (4.4) can be written in matrix notation as

$$(4.5)\quad X'Y_{(k)} = [X'X]\, B_{(k)},$$

where $Y_{(k)}$ and $B_{(k)}$ respectively are the column matrices of the values for the kth predictant and the corresponding kth set of prediction coefficients. By letting k = 1, 2, . . . q, when q is the number of predictants, we have

$$(4.6) \qquad X'Y = [X'X]\ B.$$

By multiplying each side of the equality in (4.6) by $[X'X]^{-1}$, we have

$$(4.7a) \qquad B = [X'X]^{-1}[X'Y],$$

or written in expanded form we have

$$(4.7b) \qquad \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1k} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2k} & \cdots & b_{2q} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{j1} & b_{j2} & \cdots & b_{jk} & \cdots & b_{jq} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rk} & \cdots & b_{rq} \end{bmatrix} =$$

$$\begin{bmatrix} \frac{1}{\Sigma x_{i1}^2} & 0 \cdots & 0 \cdots & 0 \\ 0 & \frac{1}{\Sigma x_{i2}^2} \cdots & 0 \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 \cdots & 0 \cdots & \frac{1}{\Sigma x_{ij}^2} \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 \cdots & 0 \cdots & 0 \cdots & \frac{1}{\Sigma x_{ir}^2} \end{bmatrix} \begin{bmatrix} \Sigma x_{i1}y_{i1} & \Sigma x_{i1}y_{i2} & \cdots & \Sigma x_{i1}y_{ik} & \cdots & \Sigma x_{i1}y_{iq} \\ \Sigma x_{i2}y_{i1} & \Sigma x_{i2}y_{i2} & \cdots & \Sigma x_{i2}y_{ik} & \cdots & \Sigma x_{i2}y_{iq} \\ \vdots & \vdots & & \vdots & & \vdots \\ \Sigma x_{ij}y_{i1} & \Sigma x_{ij}y_{i2} & \cdots & \Sigma x_{ij}y_{ik} & \cdots & \Sigma x_{ij}y_{iq} \\ \vdots & \vdots & & \vdots & & \vdots \\ \Sigma x_{ir}y_{i1} & \Sigma x_{ir}y_{i2} & \cdots & \Sigma x_{ir}y_{ik} & \cdots & \Sigma x_{ir}y_{iq} \end{bmatrix},$$

where all summations are over the subscript i. $[X'X]$ is a diagonal matrix (see condition (3.2ii)), so it follows from the definition of an inverse matrix that $[X'X]^{-1}$ will be a diagonal matrix which has the reciprocal of the corresponding diagonal elements of the $[X'X]$ matrix on the diagonal. Because $[X'X]^{-1}$ is a diagonal matrix, the elements in the B matrix reduce to the form

(4. 7c)
$$b_{jk} = \frac{\sum_{i=1}^{n} x_{ij} y_{ik}}{\sum_{i=1}^{n} x_{ij}^2} .$$

Because the coefficient b_{jk} only depends on the jth predictor, terms on the right side of the prediction equation (4. 2) can be added or omitted without affecting the other prediction coefficients.

The equations of the type (4. 2) can be written in matrix form as

(4. 8a)
$$\hat{Y}_{(\phi,\theta)} = X_{(\phi)} B(\phi, \theta, L)$$

and in expanded form as,

(4. 8b)
$$\begin{bmatrix} \hat{y}_{11} & \hat{y}_{12} & \cdots & \hat{y}_{1k} & \cdots & \hat{y}_{1q} \\ \hat{y}_{21} & \hat{y}_{22} & \cdots & \hat{y}_{2k} & \cdots & \hat{y}_{2q} \\ \vdots & \vdots & & \vdots & & \vdots \\ \hat{y}_{i1} & \hat{y}_{i2} & \cdots & \hat{y}_{ik} & \cdots & \hat{y}_{iq} \\ \vdots & \vdots & & \vdots & & \vdots \\ \hat{y}_{n1} & \hat{y}_{n2} & \cdots & \hat{y}_{nk} & \cdots & \hat{y}_{nq} \end{bmatrix} = \qquad (\phi, \theta, L)$$

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1r} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2r} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ir} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{nr} \end{bmatrix}_{(\theta)} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1k} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2k} & \cdots & b_{2q} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{j1} & b_{j2} & \cdots & b_{jk} & \cdots & b_{jq} \\ \vdots & \vdots & & \vdots & & \vdots \\ b_{r1} & b_{r2} & \cdots & b_{rk} & \cdots & b_{rq} \end{bmatrix}_{(\phi, \theta, L)}.$$

The subscript (ϕ, θ, L) attached to a matrix refers to the predictants for the weather element ϕ, predicted from the predictors for the weather element θ at lag L months.

The predicted values for the 30-day means at the s grid points are obtained from

$$\hat{\Phi}_{(\phi, \theta, L)} = \hat{Y}_{(\phi, \theta, L)} M'_{(\phi)}, \tag{4.9}$$

where $M'_{(\phi)}$ is the transpose of $M_{(\phi)}$ in expression (3.23). The general term in (4.9) can be expressed as

$$\hat{\phi}_{ik} = \hat{y}_{i1} m_{k1} + \hat{y}_{i2} m_{k2} + \cdots + \hat{y}_{ij} m_{kj} + \cdots + \hat{y}_{iq} m_{kq}. \tag{4.10}$$

In other words, the prediction for the ith standardized 30-day mean deviation at the kth grid point is a linear combination of estimates for the ith terms of the q predictants and the corresponding q multipliers that are associated with the weather element ϕ at the kth grid point.

CHAPTER 5

EVALUATION OF PREDICTION MODELS

Development Sample Evaluation

The accuracy of the predicted standardized 30-day mean deviations at each of the s grid points can be evaluated by the following expression:

$$\text{(5.1a)} \qquad \text{reduction of error} = 1 - \frac{\sum_{i=1}^{n} (\phi_{ik} - \hat{\phi}_{ik})^2}{\sum_{i=1}^{n} \phi_{ik}^2} .$$

Expression (5.1a) is known as the reduction of error for n predictions made at the kth grid point. If the n predictions, $\hat{\theta}_{ik}$'s, are equal to the corresponding n true values, θ_{ik}'s, the reduction of error at the kth grid point is equal to 1. If the n predictions were made equal to the 9 year mean of the 30-day mean deviations, i.e., $\hat{\phi}_{ik} = 0$ for i = 1, 2, ... n, the reduction of error would be equal to zero. Predictions made equal to the mean over the development sample are termed climatology predictions.

$$(5.1b)\quad \text{reduction of error} = \frac{\sum_{i=1}^{n}\phi_{ik}^2 - \sum_{i=1}^{n}(\phi_{ik}-\hat{\phi}_{ik})^2}{\sum_{i=1}^{n}\phi_{ik}^2}$$

$$= \frac{\sum_{i=1}^{n}\phi_{ik}^2 - (\sum_{i=1}^{n}\phi_{ik}^2 - 2\sum_{i=1}^{n}\phi_{ik}\hat{\phi}_{ik} + \sum_{i=1}^{n}\hat{\phi}_{ik}^2)^2}{\sum_{i=1}^{n}\phi_{ik}^2}$$

$$= \frac{2\sum_{i=1}^{n}\phi_{ik}\hat{\phi}_{ik}}{\sum_{i=1}^{n}\phi_{ik}^2} - \frac{\sum_{i=1}^{n}\hat{\phi}_{ik}^2}{\sum_{i=1}^{n}\phi_{ik}^2}.$$

If the n predicted and true values were not correlated, i. e. ,

$\sum_{i=1}^{n}\phi_{ik}\hat{\phi}_{ik} = 0$, then the reduction of error would be equal to

$$(5.3)\quad \text{chance reduction of error} = -\frac{\sum_{i=1}^{n}\hat{\phi}_{ik}^2}{\sum_{i=1}^{n}\phi_{ik}^2}.$$

The chance reduction of error is used as a measure of comparison for the ordinary reduction of error defined by expression (5.1a). The accuracy of the predicted values for the standardized 30-day

means over the development sample can be measured at each of the s grid points by using expression (5. la) and (5. 3). The values for expressions (5. la) and (5. 3) may change for the different prediction models that result from using different numbers of predictors, r, and predictants, q.

Test Sample Evaluation

The prediction models developed in Chapters 3 and 4 are tested on data which did not enter into their construction. The test sample is used to simulate the actual use of the prediction models for predicting circulation and temperature patterns. The prediction models constructed from the data in the development samples are used to generate predictions for the corresponding test samples. The predictions are generated by the following sequence of calculations:

(5. 4a) $$X_{(\theta)} = \Theta M_{(\theta)}$$

where the general term of the Θ matrix is obtained from

$$\theta_{ik} = \frac{\theta^{o}_{ik}}{w_k}$$

$$= \frac{\theta^{oo}_{ik} - \overline{\theta}^{oo}_{ik}}{w_k} ,$$

$$\hat{Y}_{(\phi,\theta,L)} = X_{(\theta)} \underline{B_{(\phi,\theta,L)}}, \tag{5.4b}$$

$$\hat{\Phi}_{(\phi,\theta,L)} = \hat{Y}_{(\phi,\theta,L)} \underline{M'_{(\phi)}}. \tag{5.4c}$$

The quantities underlined in the above expressions are those which were obtained from a development sample. These quantities are all defined in Chapters 2, 3, and 4. The matrices that are not underlined all have n' rows, where n' is the number of months in the test sample for which 30-day mean predictions are made. Θ is the n' by s matrix of standardized 30-day mean deviations for the weather element θ at the s grid points, and $X_{(\theta)}$ is the corresponding n' by r matrix of predictors. $\hat{Y}_{(\phi,\theta,L)}$ is the n' by q matrix of predicted values for the q predictants, and $\hat{\Phi}$ is the corresponding n' by s matrix of predicted standardized 30-day mean deviations for the weather element ϕ over the n' months at the s grid points.

The evaluation measures defined by expressions (5.1a) and (5.3) for the development sample are also those used for the test sample evaluation with n replaced by n'. The values for these evaluation measures may change for the different prediction models that result from using different numbers of predictors, r, and predictants, q.

CHAPTER 6

RESULTS

Intermediate Results

Results are shown for steps made in the construction of several of the prediction models. These results are presented to show the variation of the 30-day mean deviations over the Northern Hemisphere and the proportion of total variance of the standardized 30-day means accounted for by the principal components.

Shown in Figures 1-3 of the Appendix are values for standard deviations of 30-day mean deviations over the 9-year development samples. These standard deviations are plotted on maps of the Northern Hemisphere for each of the three weather elements for the winter season. The standard deviations are products of the constant $1/\sqrt{35}$ and the weights, $w_{(k)}$'s, which were discussed in Chapter 3. The standard deviations, and consequently the weights, can be seen to vary greatly with respect to their grid points locations. In fact, the ratio of the largest to the smallest standard deviation in Figure 5 is nearly 10:1.

Proportions of the total variance of 30-day mean deviations accounted for by the first 20 principal components are given in

Table 2

PROPORTION OF TOTAL VARIANCE OF 700 MILLIBAR HEIGHT ACCOUNTED FOR BY PRINCIPAL COMPONENTS DURING THE WINTER SEASON

Principal Component Number	Proportion of Total Variance	Cumulative Proportion of Total Variance
1	.162	.162
2	.135	.297
3	.113	.411
4	.086	.497
5	.075	.571
6	.063	.634
7	.054	.688
8	.039	.728
9	.037	.765
10	.034	.799
11	.027	.826
12	.024	.850
13	.021	.871
14	.020	.891
15	.016	.907
16	.015	.922
17	.012	.934
18	.009	.943
19	.009	.952
20	.007	.959

Table 3

PROPORTION OF TOTAL VARIANCE OF SURFACE PRESSURE ACCOUNTED FOR BY PRINCIPAL COMPONENTS DURING THE WINTER SEASON

Principal Component Number	Proportion of Total Variance	Cumulative Proportion of Total Variance
1	.185	.185
2	.142	.328
3	.116	.443
4	.082	.525
5	.075	.601
6	.064	.665
7	.047	.712
8	.041	.752
9	.031	.783
10	.028	.812
11	.027	.838
12	.021	.859
13	.020	.879
14	.015	.894
15	.014	.908
16	.012	.921
17	.012	.932
18	.010	.942
19	.009	.951
20	.007	.958

Table 4

PROPORTION OF TOTAL VARIANCE OF 700 MILLIBAR TEMPERATURE ACCOUNTED FOR BY PRINCIPAL COMPONENTS DURING THE WINTER SEASON

Principal Component Number	Proportion of Total Variance	Cumulative Proportion of Total Variance
1	.152	.152
2	.130	.282
3	.102	.384
4	.091	.475
5	.075	.550
6	.054	.604
7	.048	.652
8	.046	.697
9	.037	.735
10	.034	.769
11	.028	.796
12	.024	.821
13	.022	.843
14	.019	.862
15	.018	.880
16	.015	.894
17	.015	.909
18	.014	.923
19	.011	.934
20	.010	.944

Tables 2-4. These proportions of variance are tabled for the winter development samples of the three weather elements. It can be seen from these tables that 15 principal components for 700 millibar height and surface pressure and 17 principal components for 700 millibar temperature were needed to account for at least 90% of the total variance. For the spring and summer seasons two or three additional principal components were needed for each of the three weather elements in order to account for 90% of the total variance. Maps of the spatial patterns for the corresponding first three principal components in Tables 2-4 are given in Figures 4 - 12 of the Appendix.

Evaluation Results

Evaluation results for the specifications and the predictions of one month in advance are presented in this section. The reduction of error with respect to climatology is one measure of evaluation that will be used. The mean of this measure over the 108 grid points can be obtained from (5.1a) and written as the following expression:

(6.1)

$$\text{mean reduction of error with respect to climatology} = \frac{\sum_{k=1}^{108} \left(1 - \frac{\sum_{i=1}^{n} (\phi_{ik} - \hat{\phi}_{ik})^2}{\sum_{i=1}^{n} \phi_{ik}^2} \right)}{108} ,$$

where $\hat{\phi}_{ik}$ is the ith prediction made at the kth grid point, and ϕ_{ik} is the corresponding true value. Another measure used for evaluation is the following:

(6.2)

$$\text{mean reduction of error with respect to chance} = \frac{\sum_{i=1}^{n} \phi_{ik} \hat{\phi}_{ik}}{\sum_{i=1}^{n} \phi_{ik}^2} .$$

Expression (6.2) is the difference between expressions (5.1b) and (5.3) divided by two. From (6.2) we have that the

(6.3)

$$\text{mean reduction of error with respect to chance} = \frac{\sum_{k=1}^{108} \left(\frac{\sum_{i=1}^{n} \phi_{ik} \hat{\phi}_{ik}}{\sum_{i=1}^{n} \phi_{ik}^2} \right)}{108} .$$

Both expressions (6.1) and (6.3) would have values equal to 1 if the

n (108) predictions, $\hat{\phi}_{ik}$'s, were exactly equal to their corresponding true values, ϕ_{ik}'s.

The mean reductions of error with respect to climatology and chance are given in Tables 5 and 6 for the specification and the prediction of 700 millibar height from surface pressure over the test samples. From Table 5 it can be seen that the reductions of error with respect to climatology and chance increase as the number of predictors and predictants in the specification model increase. In addition to the specification of pressure height from surface pressure, specifications were also evaluated on the test sample for surface pressure from 700 millibar height, 700 millibar temperature from 700 millibar height, and 700 millibar height from 700 millibar temperature. For the specification of surface pressure from 700 millibar height the mean reductions of error with respect to climatology were found to be slightly lower than those corresponding measures shown in Table 5. The specifications of 700 millibar height and 700 millibar temperature made from each other were found to be only slightly better than those specifications made by climatology.

From Table 6 it can be seen that the reductions of error with respect to climatology are all negative quantities. In other words, more accurate results could have been obtained by using

Table 5

SPECIFICATION OF 700 MILLIBAR HEIGHT FROM SURFACE PRESSURE OVER THE TEST SAMPLE FOR THE WINTER SEASON

Number of predictors and predictants	Mean reduction of error with respect to climatology	Mean reduction of error with respect to chance
4	.33	.40
8	.40	.49
12	.45	.55
16	.52	.63
20	.55	.66

Table 6

PREDICTION OF 700 MILLIBAR HEIGHT FROM SURFACE PRESSURE OVER THE TEST SAMPLE FOR THE WINTER SEASON

Number of predictors and predictants	Mean reduction of error with respect to climatology	Mean reduction of error with respect to chance
4	-.09	.00
8	-.12	.03
12	-.21	.03
16	-.39	.03
20	-.60	.04

the 9-year mean values (climatology) at their respective grid points as predicted values rather than using the prediction models to obtain predictions. The reductions of error with respect to chance can be seen from Table 6 to be nearly equal to zero.

Table 7

PREDICTION EVALUATION OVER THE TEST SAMPLE USING 12 PREDICTORS AND 12 PREDICTANTS FROM THE SAME WEATHER ELEMENT

Season	Weather element	Mean reduction of error with respect to climatology	Mean reduction of error with respect to chance
Winter	700 m.b. height	-.22	.00
	surface pressure	-.31	.02
	700 m.b. temperature	-.20	-.01
Spring	700 m.b. height	-.47	-.01
	surface pressure	-.44	-.02
	700 m.b. temperature	-.27	.01
Summer	700 m.b. height	-.28	-.02
	surface pressure	-.28	-.04

Evaluation results using 12 predictors and 12 predictants from the same weather element are given in Table 7. These results were obtained from 30-day mean predictions made one month in advance over the test sample. The mean reductions of error with respect to climatology are all negative for the combinations of

weather elements and seasons. Evaluations were made over the summer development sample for predicting 700 millibar height one month in advance from 700 millibar height. The evaluations for this model which contains 12 predictors and 12 predictants were as follows: the reduction of error with respect to climatology was. 25, and the reduction of error with respect to chance was . 26. By comparing these reductions with corresponding reductions of -. 28 and -. 02 from Table 7, one can see the importance of using a test sample for evaluation.

The reductions of error for specification were found to be generally better in the regions closest to the North Pole. Figures 13 and 14 in the Appendix show the reductions of error with respect to climatology for locations over the Northern Hemisphere for winter test samples. Figure 13 is a map of the reductions for 700 millibar height specified from surface pressure. Figure 14 is a map of the reductions for the reconstruction of 700 millibar height. By reconstruction is meant the specification of 30-day means of a weather element from that same weather element. In other words, the true values replace the estimated values for the predictants in expression (4. 9).

Discussion

The results presented in the previous section would certainly not encourage one to use the particular prediction models developed in this study. However, before judgment is passed on the procedure for obtaining the prediction models, an attempt should be made to see how these prediction models might be improved.

An increase in the size of the development sample is one change that could be expected to increase the accuracy of the predictions for a test sample; that is, if relationships among the 30-day means are really predictable over time, an increase in the amount of information concerning these relationships should result in more accurate predictions. The difference in the size of the development samples for this study and Gilman's study (36 compared to 120) may account for the more accurate prediction results that Gilman obtained (3, p. 90).

By using the predictors from more than one weather variable in the prediction equations, improvements might also be made in the accuracy. Another possible improvement would be to include non-linear terms of the predictors in the prediction equations. There is also the possibility that relationships among means for shorter periods may be more predictable over time. For example, 15-day

means may be more predictable at two 15-day periods in advance than 30-day means predicted one month in advance.

BIBLIOGRAPHY

1. Anderson, T. W. An introduction to multivariate statistical analysis. New York, Wiley, 1958. 374 p.

2. Cramér, Harald. Mathematical methods of statistics. Rev. ed. Princeton, Princeton University Press, 1946. 575 p.

3. Gilman, Donald L. Empirical orthogonal functions applied to thirty-day forecasting. Cambridge, Massachusetts, Massachusetts Institute of Technology, Department of Meteorology, June 1957, 129 p. (Scientific Report No. 1, Statistical Forecasting Project) (AF 19(604)1283)

4. Kendall, M.G. A course in multivariate analysis. New York, Hafner, 1957. 185 p.

5. Lorenz, Edward N. Empirical orthogonal functions and statistical weather prediction. Cambridge, Massachusetts, Massachusetts Institute of Technology, Department of Meteorology, December 1956. 49 p. (Scientific Report No. 1, Statistical Forecasting Project) (AF 19(604)1566)

6. Malone, Thomas F. Whither applied meteorology in the United States? Proceedings of the First National Conference on Applied Meteorology. 1:D-20-D-27. October 1957.

7. Scheffé, Henry. The analysis of variance. New York, Wiley, 1959. 477 p.

APPENDIX

Figure 1. Standard Deviations of 30-Day Means for 700 Millibar Height Over the Winter Season

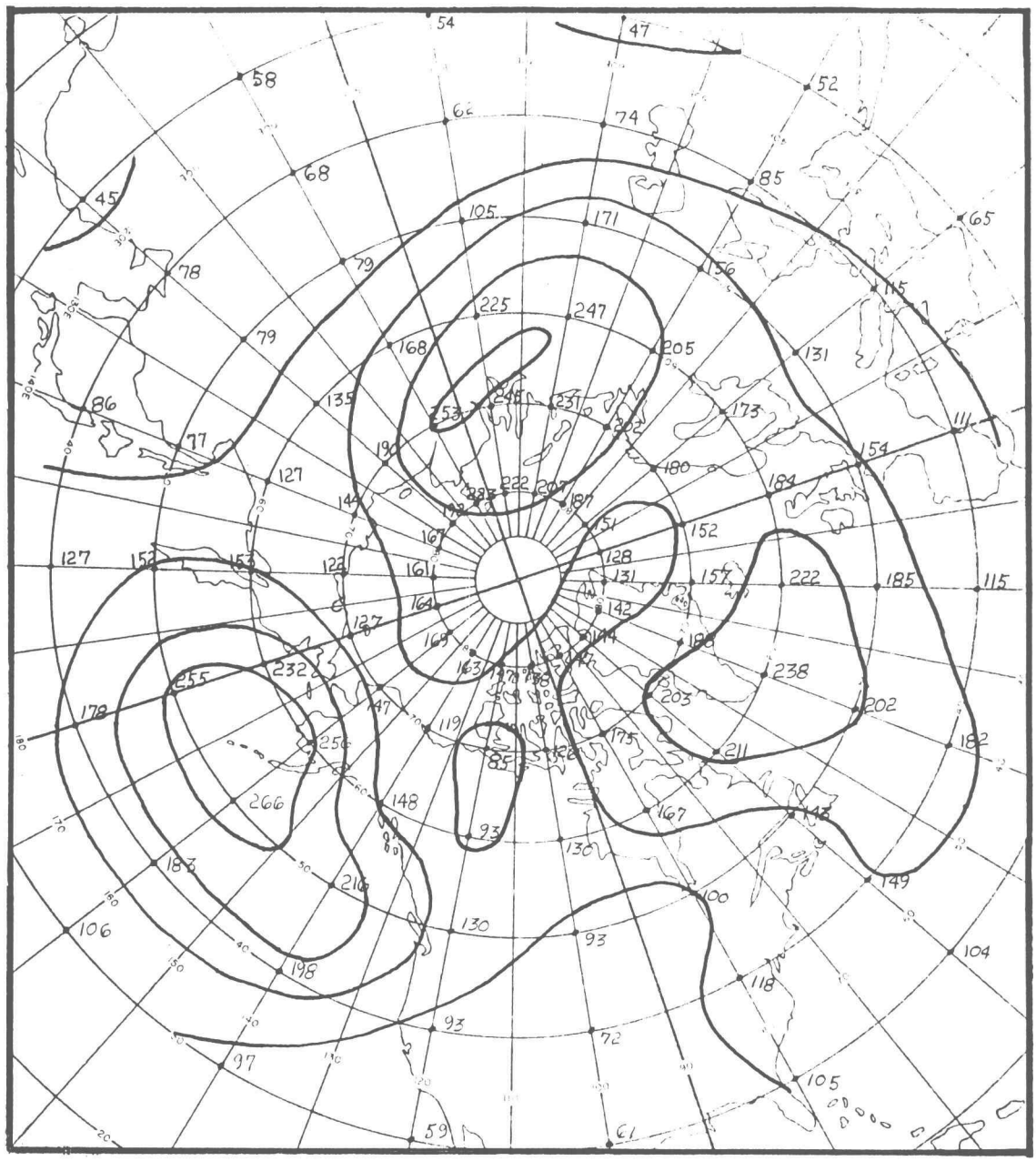

Figure 2. Standard Deviations of 30-Day Means for Surface Pressure Over the Winter Season

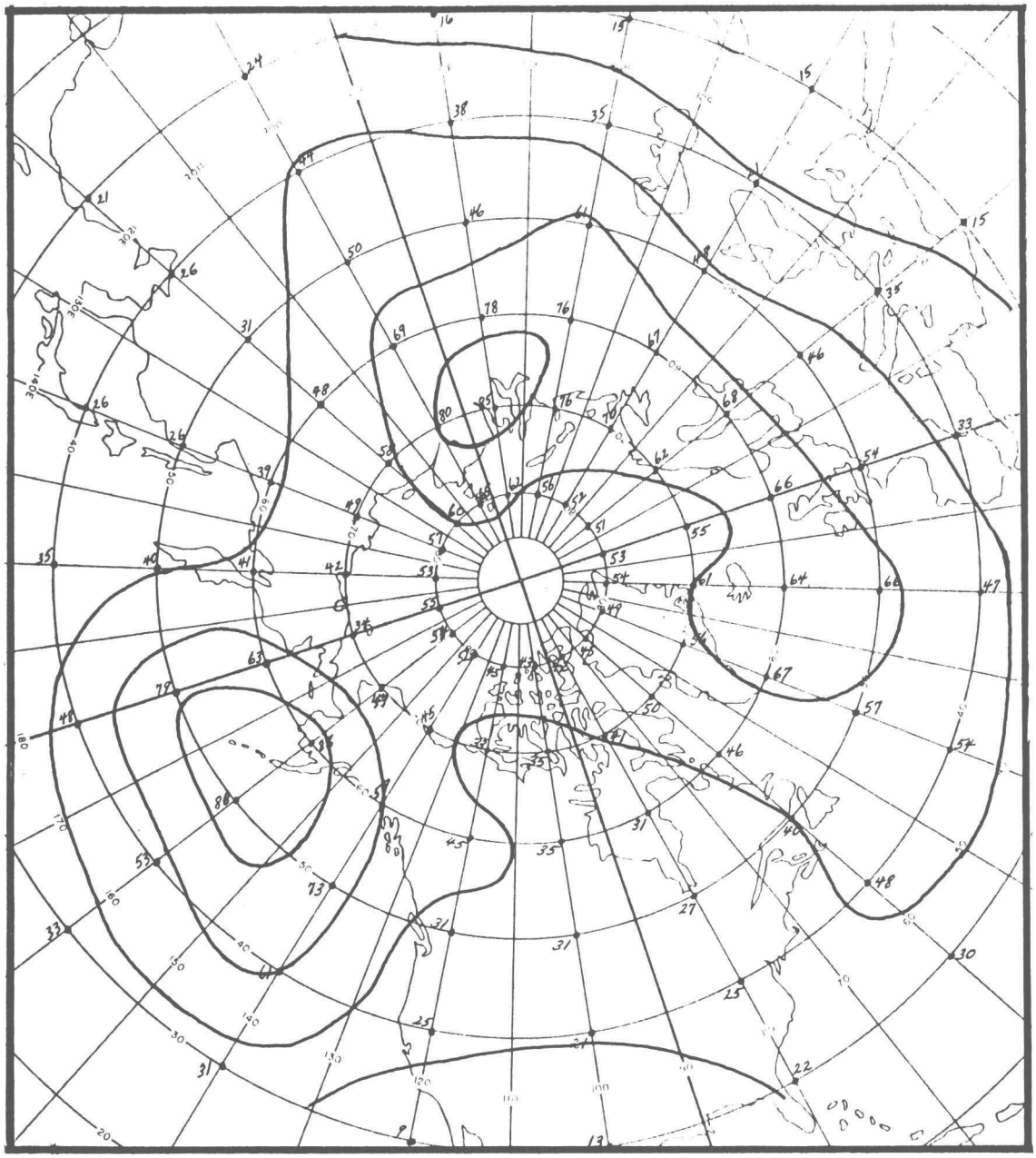

Figure 3. Standard Deviations of 30-Day Means for 700 Millibar Temperature Over the Winter Season

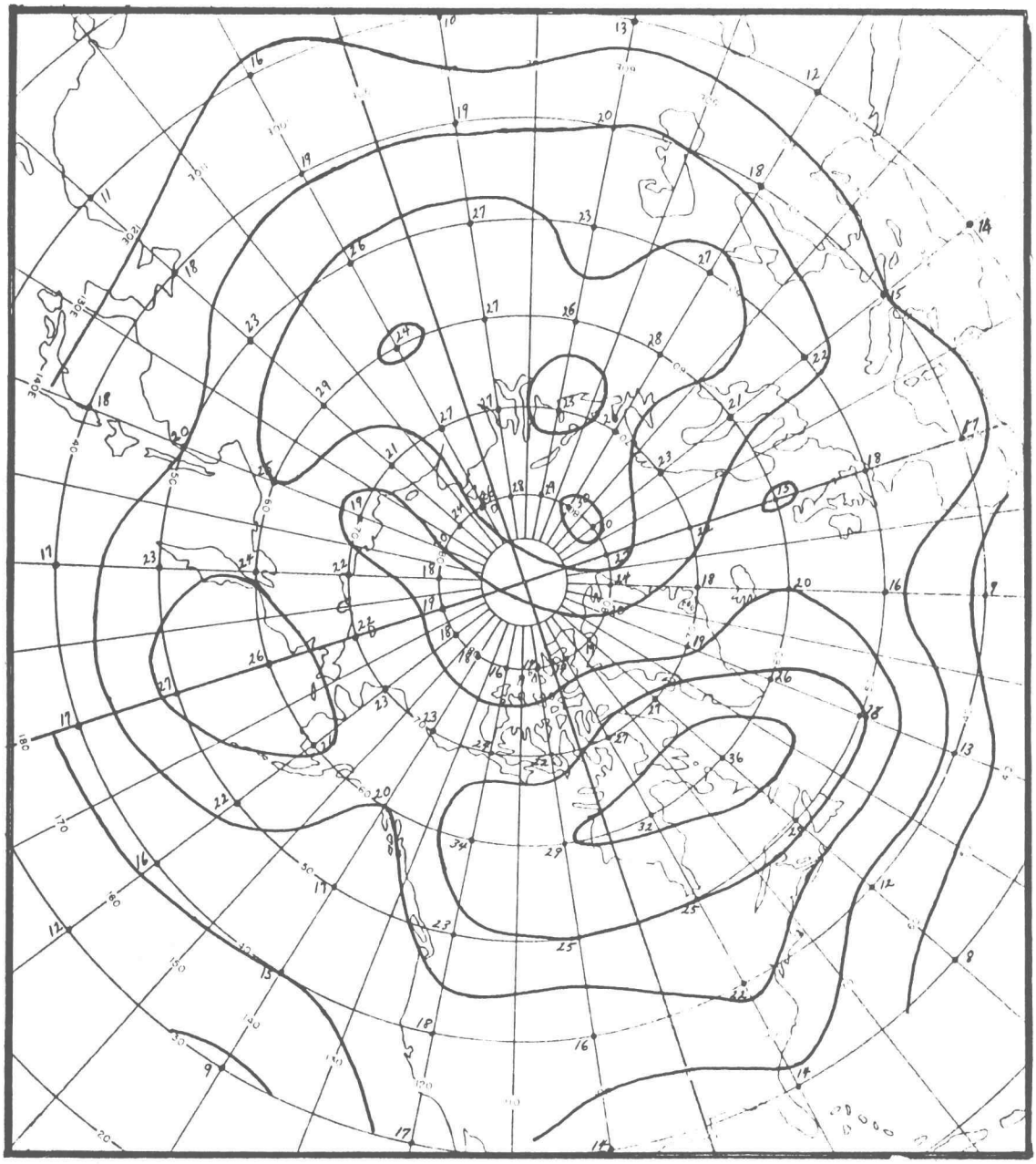

Figure 4. Spatial Pattern Number 1 for 700 Millibar Height Over the Winter Season

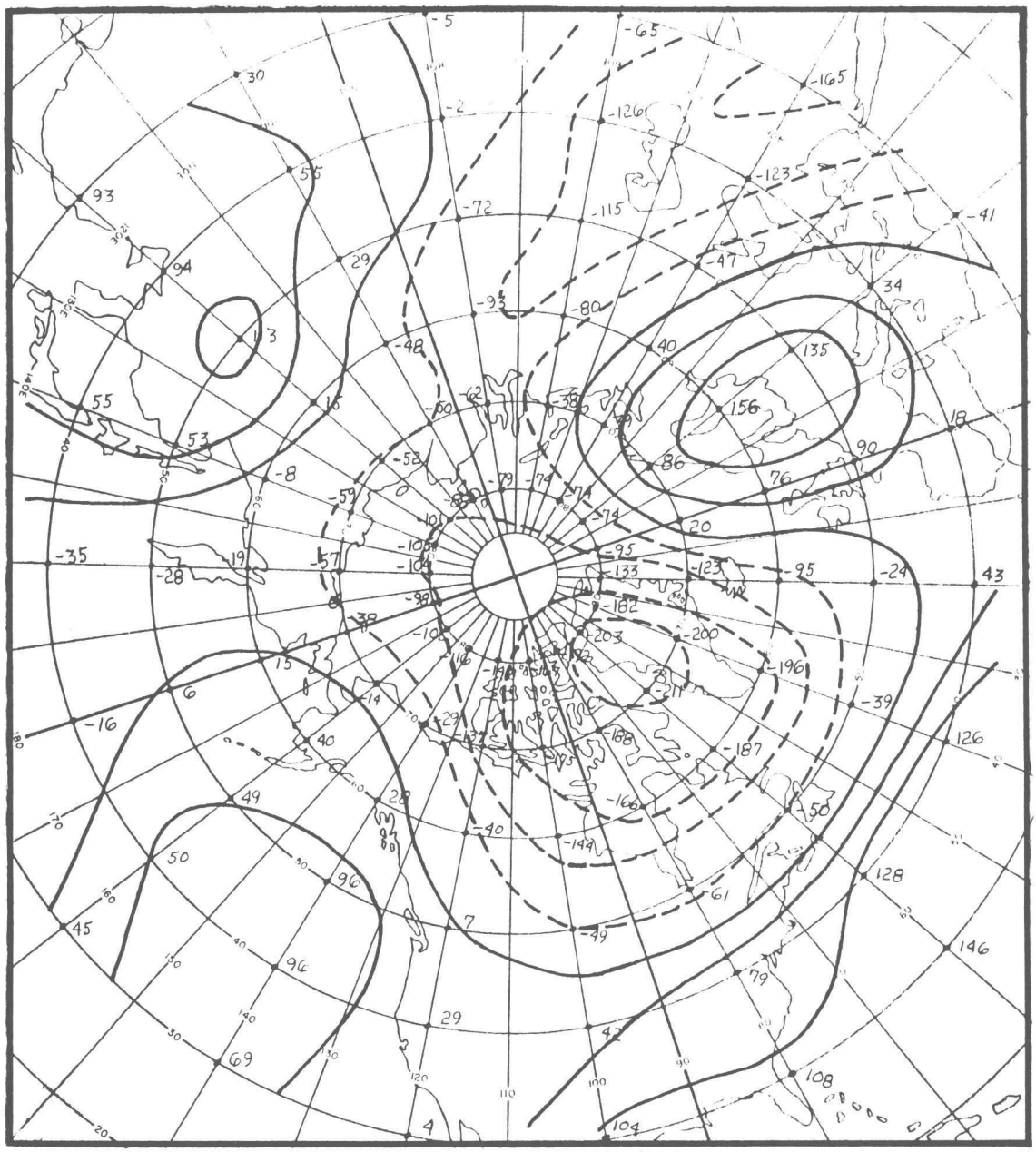

Figure 5. Spatial Pattern Number 2 for 700 Millibar Height Over the Winter Season

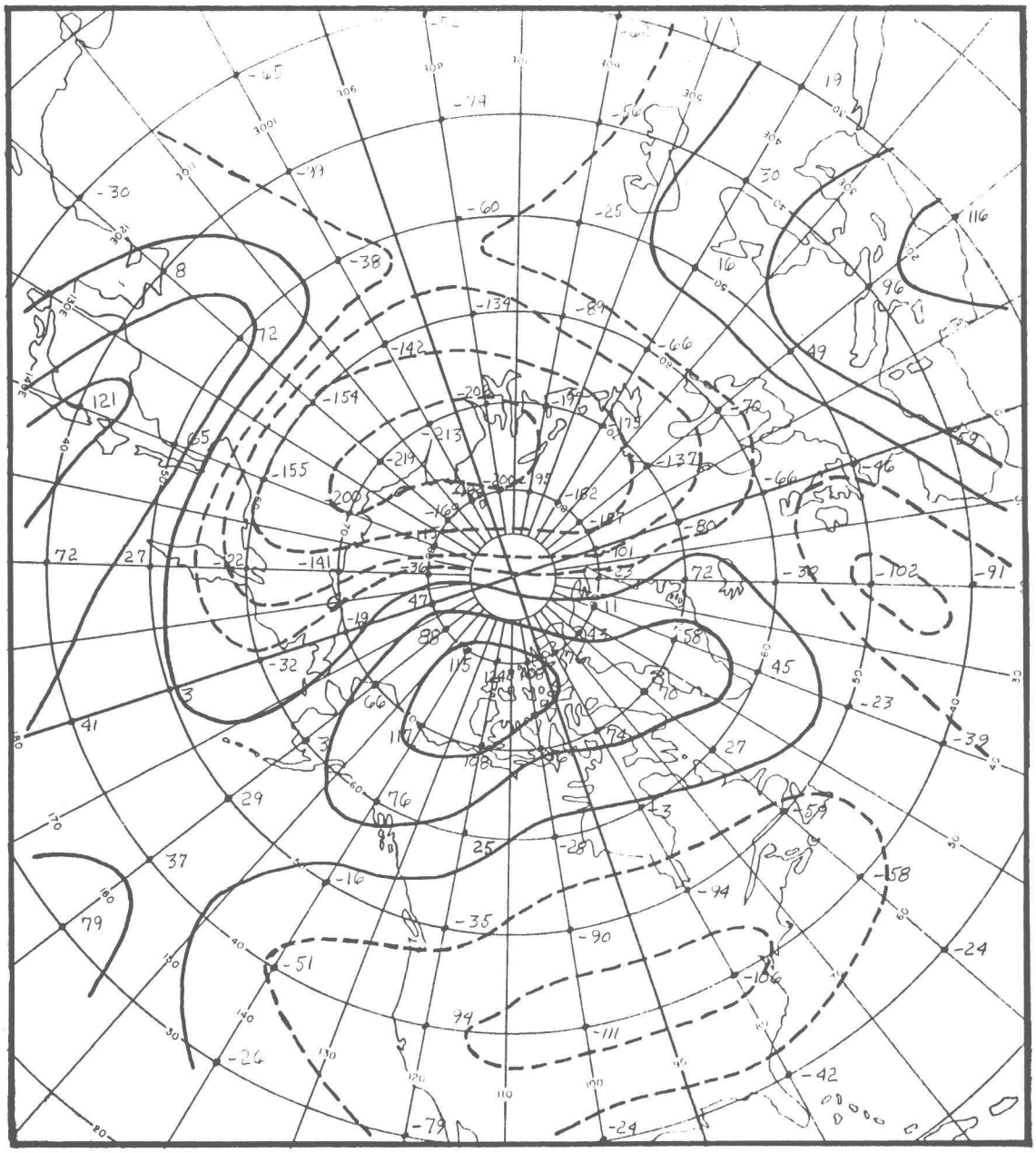

Figure 6. Spatial Pattern Number 3 for 700 Millibar Height Over the Winter Season

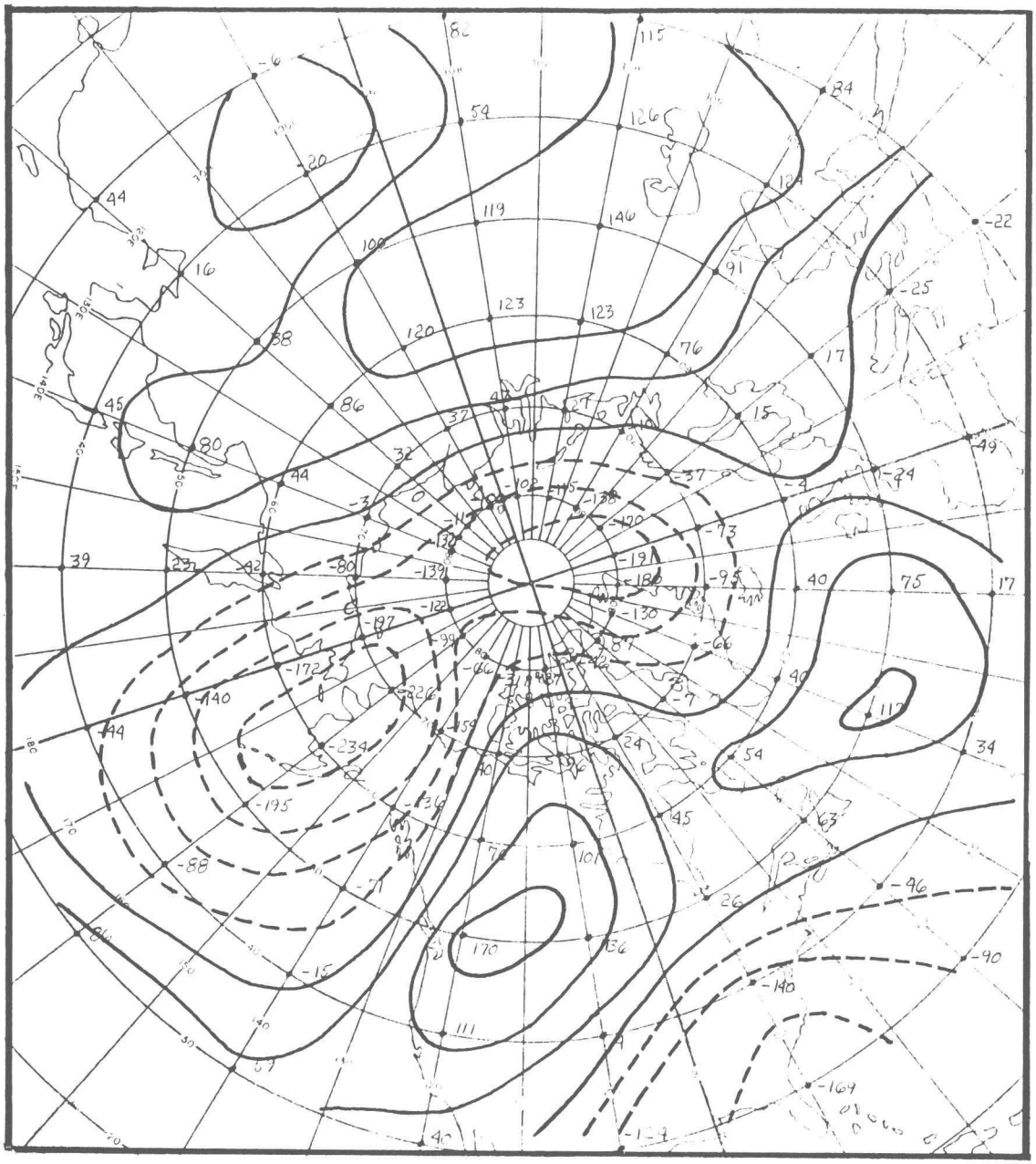

Figure 7. Spatial Pattern Number 1 for Surface Pressure Over the Winter Season

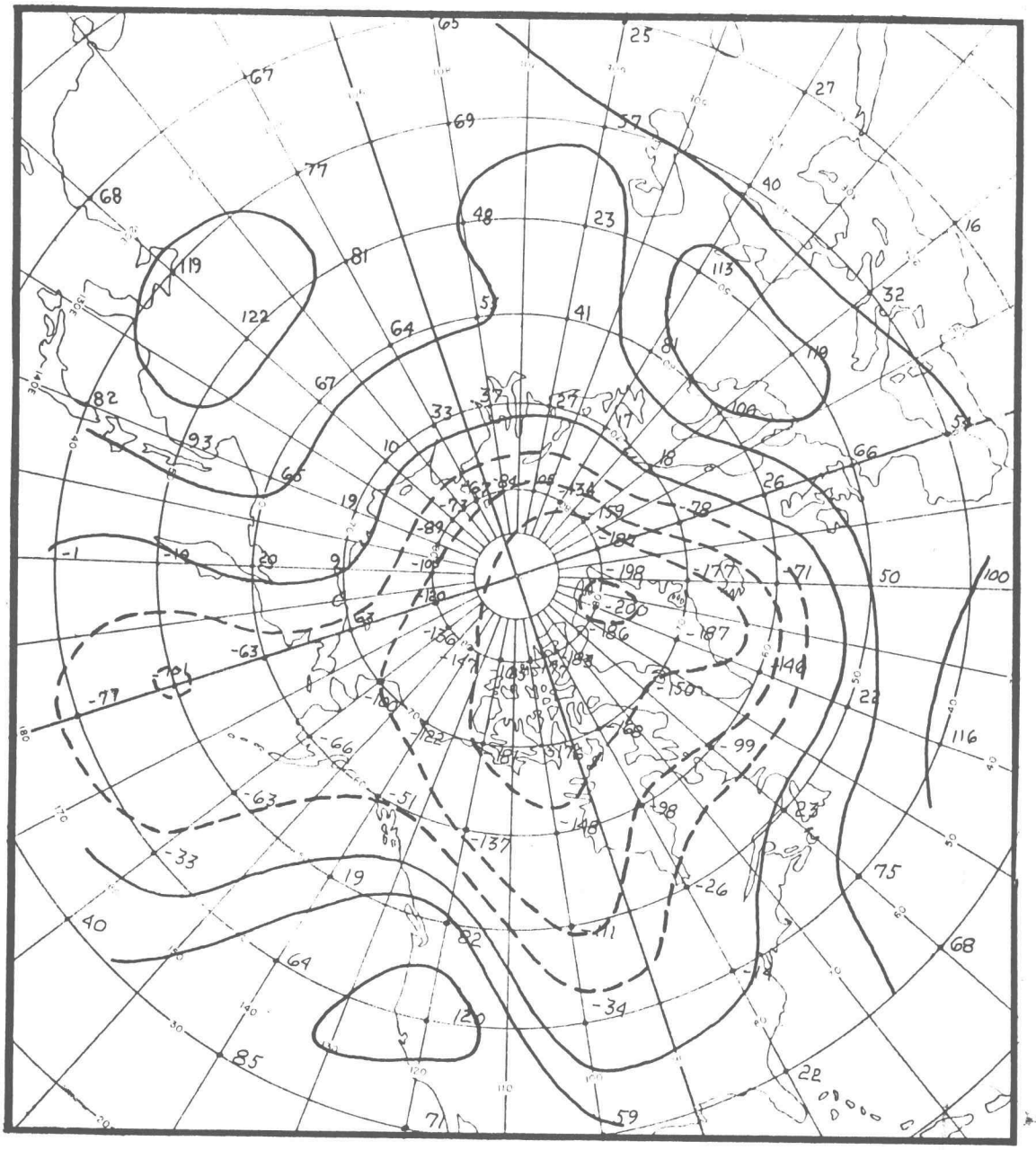

Figure 8. Spatial Pattern Number 2 for Surface Pressure Over the Winter Season

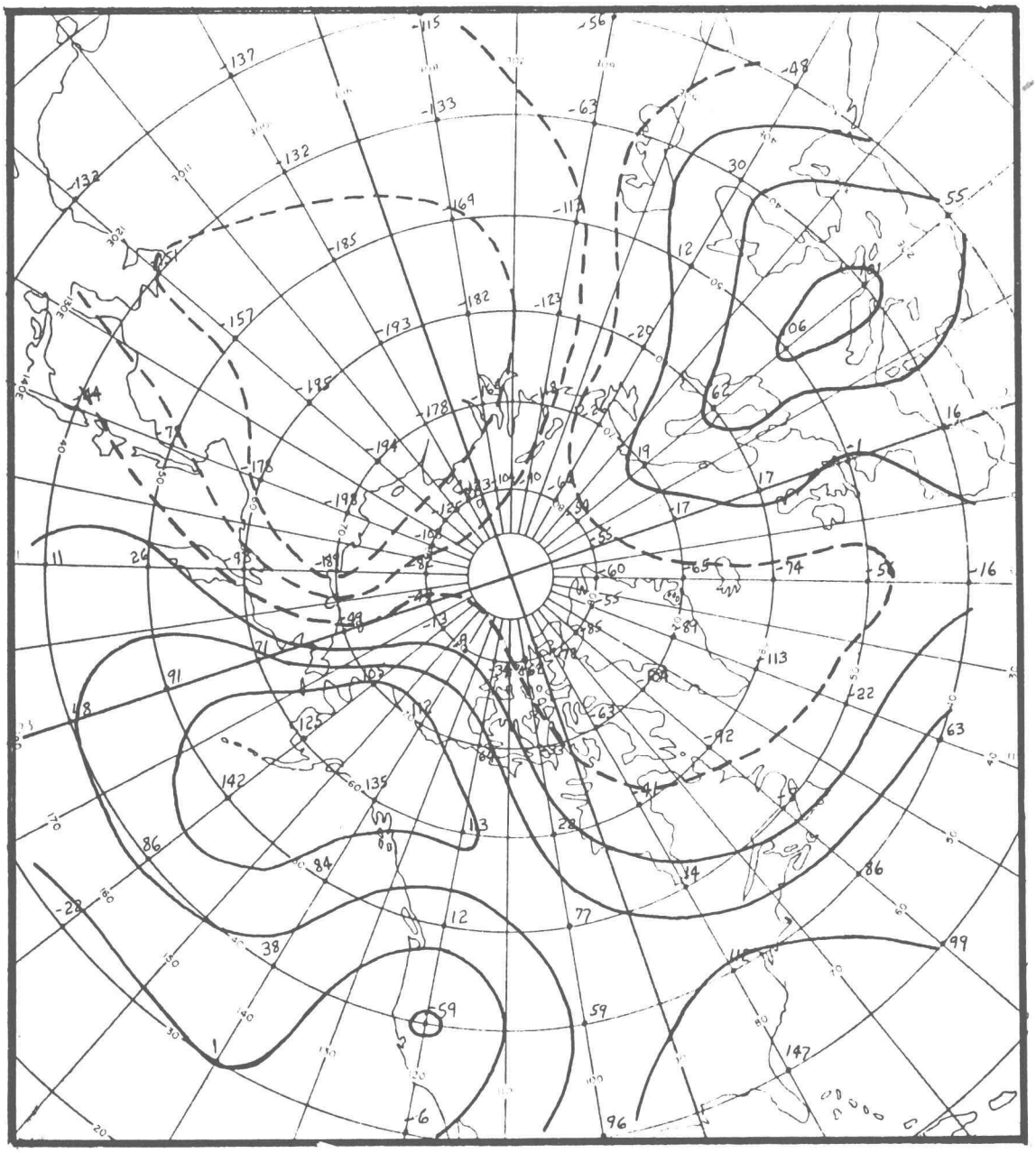

Figure 9. Spatial Pattern Number 3 for Surface Pressure Over the Winter Season

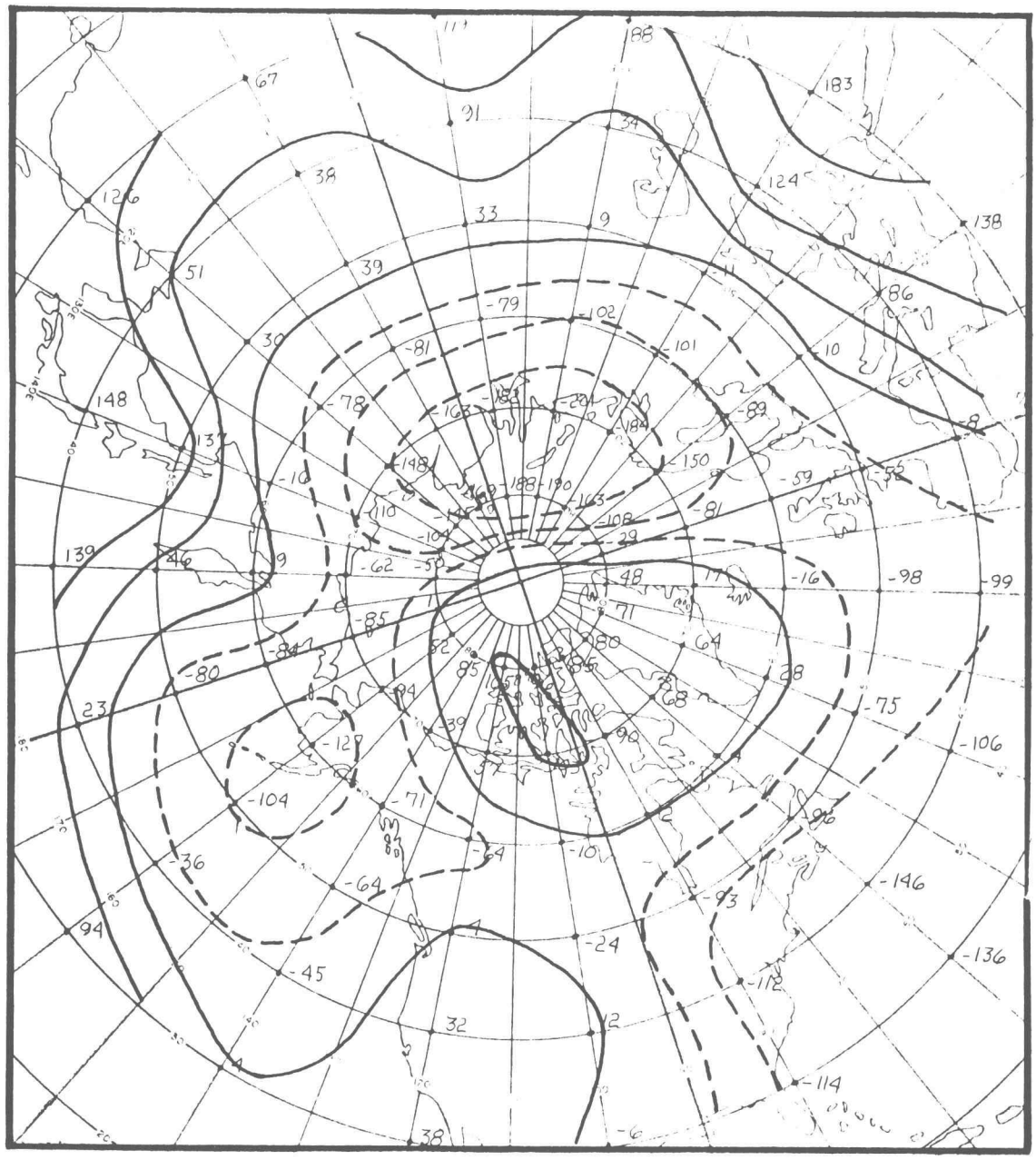

Figure 10. Spatial Pattern Number 1 for 700 Millibar Temperature Over the Winter Season

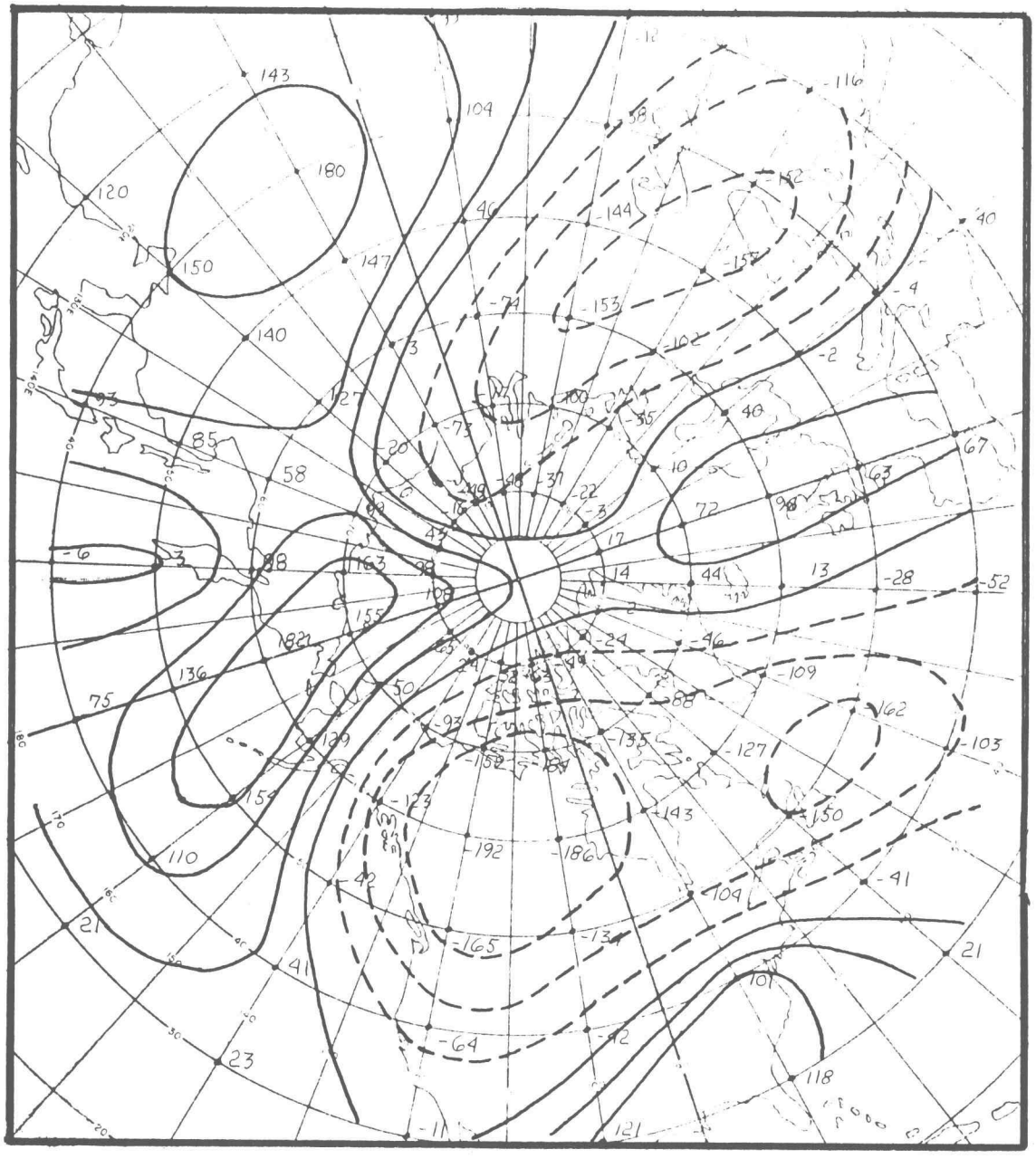

Figure 11. Spatial Pattern Number 2 for 700 Millibar Temperature Over the Winter Season

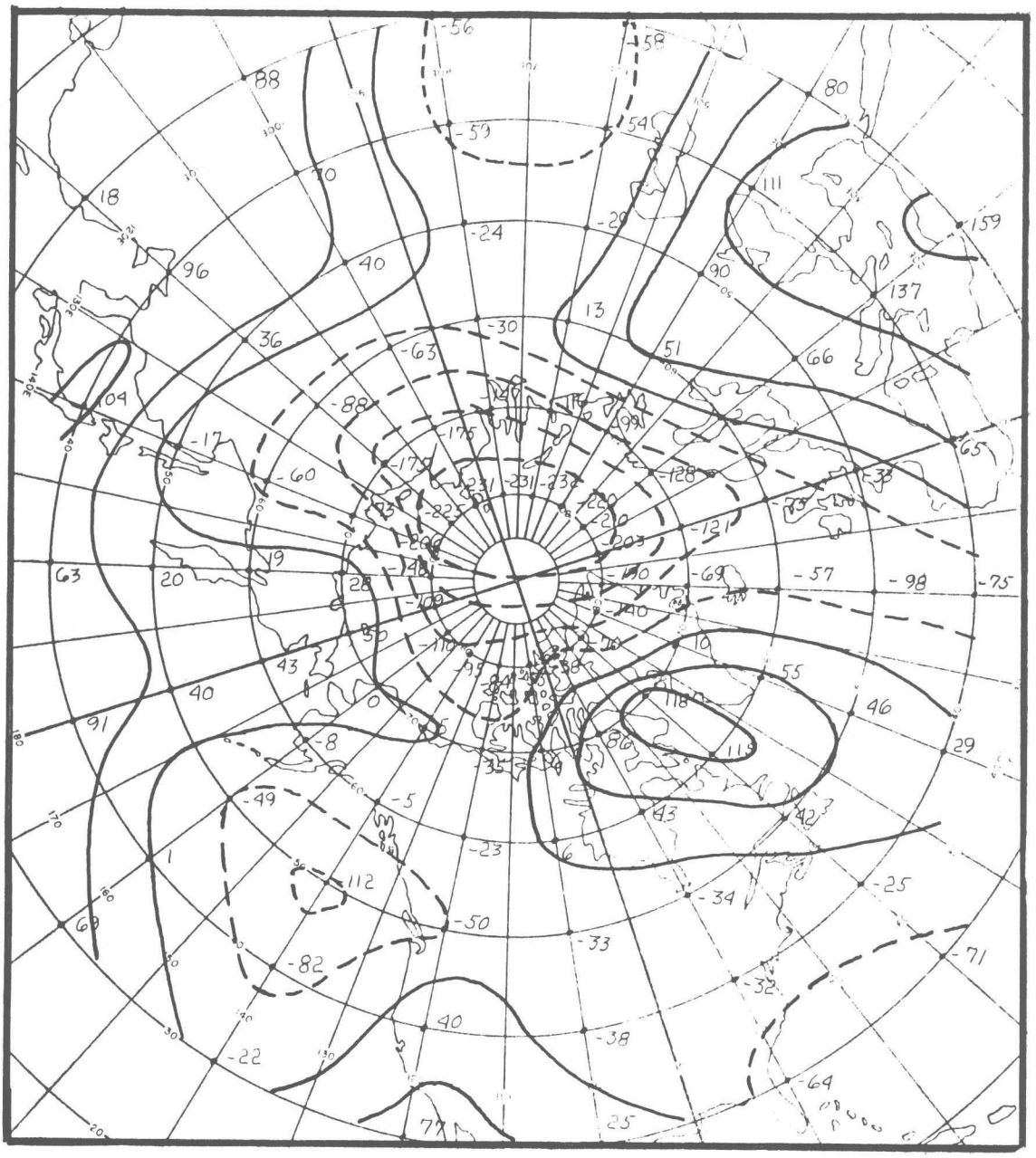

Figure 12. Spatial Pattern Number 3 for 700 Millibar Temperature Over the Winter Season

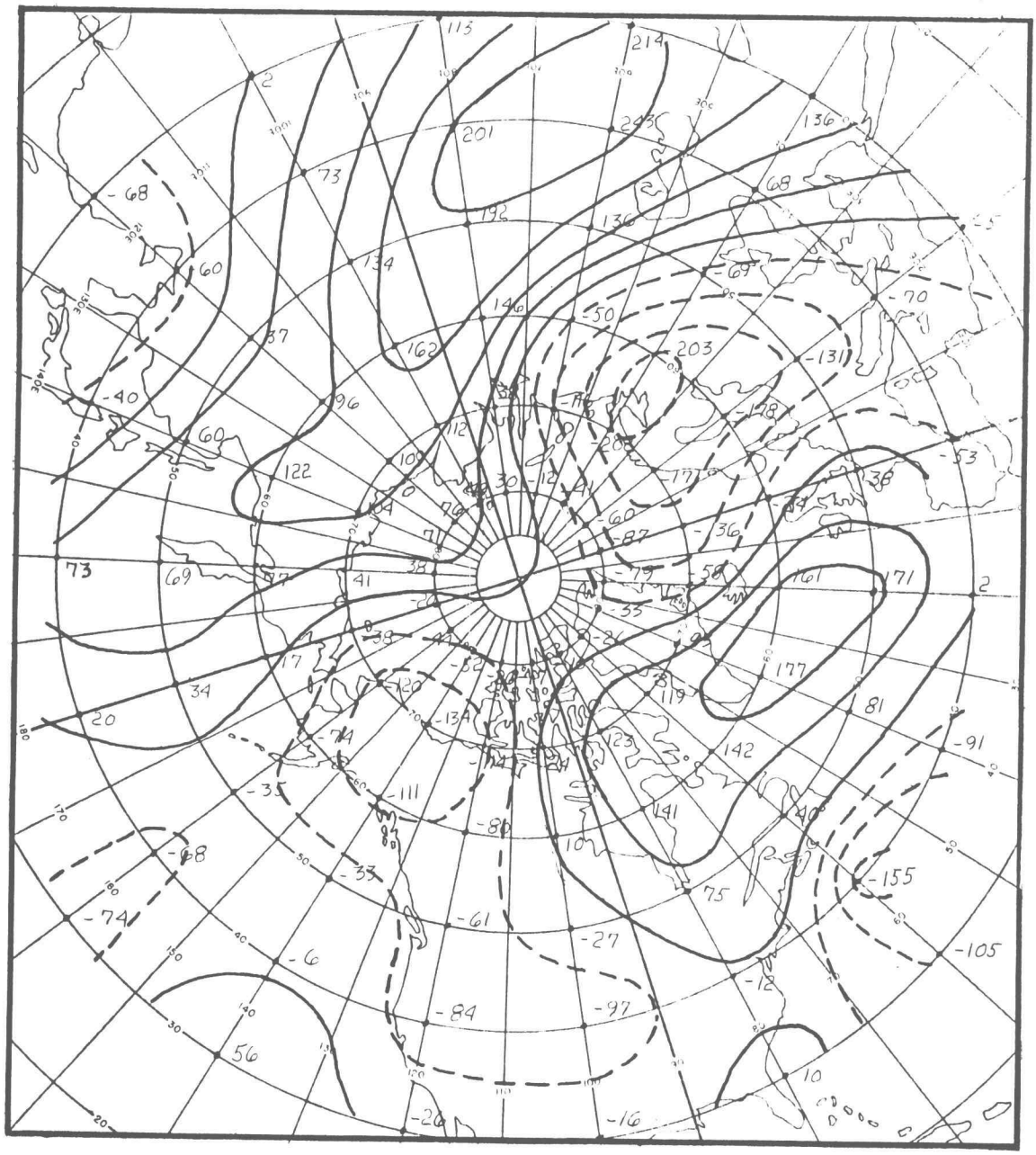

Figure 13. Reduction of Error with Respect to Climatology for the Specification of 700 Millibar Height from Surface Pressure Using 16 Predictors and 20 Predictants Over the Winter Season

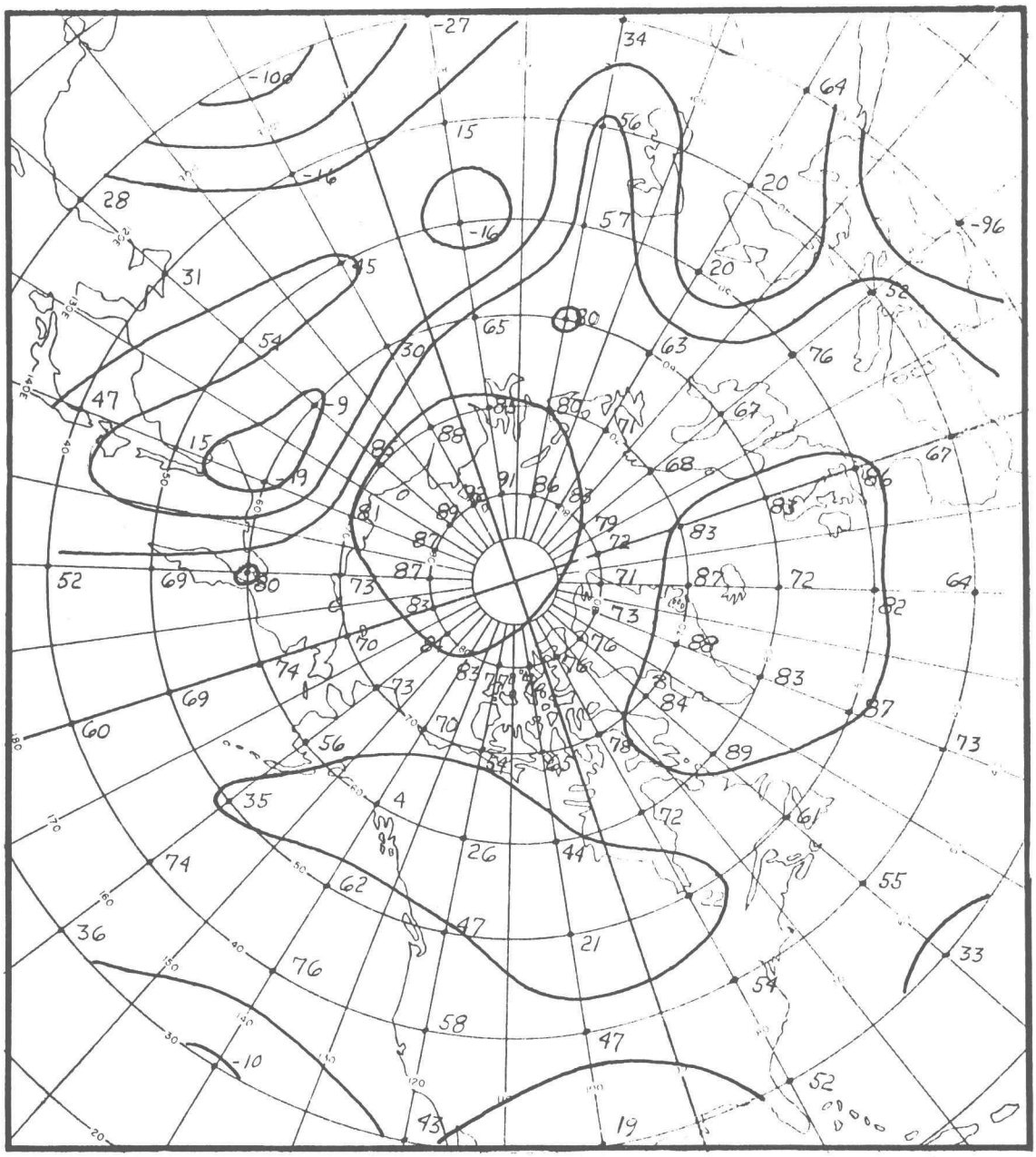

Figure 14. Reduction of Error with Respect to Climatology for the Reconstruction of 700 Millibar Height Using 12 Predictors and 20 Predictants Over the Winter Season