THE DEVELOPMENT OF EFFICIENT LINEAR STATISTICAL OPERATORS FOR THE PREDICTION OF SEA-LEVEL PRESSURE

By R. M. White, D. S. Cooley, R. C. Derby, and F. A. Seaver

Geophysics Research Directorate, Air Force Cambridge Research Center

(Original manuscript received 4 December 1957; revised manuscript received 28 April 1958)

ABSTRACT

The design of efficient linear statistical operators for the 24-hour prediction of the sea-level pressure distribution over the United States is considered. Factor analysis techniques for reduction and selection of independent variables in regression analysis are used as a means of obtaining efficient statistical forecasting equations. The effects of the variations in data density in time and space, and the extent of geographical coverage upon the explained variance of the sea-level pressure are examined.

1. Introduction

As part of a general program of experimentation with statistical approaches to meteorological prediction, the authors have recently been engaged in an inquiry into sea-level pressure forecasting. Among the problems encountered in developing statistical forecasting techniques are those of a mathematical and procedural nature as well as those which are more meteorological in character. The first group of problems is concerned with the development of mathematical procedures which enable one to deal with large numbers of correlated variables and to extract their informational and predictive content in an efficient manner. The problems of the latter kind which we have considered in this paper refer to the effects of variations in the sea-level-pressure data density and geographical coverage upon the reduction of the variance of the 24-hour sea-level pressures.

2. Scope

In this investigation a series of statistical experiments is conducted in which progressively larger numbers of independent variables are used in a linear regression analysis to explain the sea-level-pressure variance at 24-hr lag. This explained variance, also known as the per-cent reduction of the variance, will henceforth be designated simply as P.R. In all of these experiments, factor-analysis techniques are used as a means of extracting the informational and predictive content of the independent variables and reducing their number.

In all of the experiments, the dependent variables are the 24-hour sea-level pressures, located at 42 grid points, distributed over the United States as indicated by the solid circles in fig. 1. For the first statistical experiment, which we shall designate as E_1, the 42 sea-level pressures taken at initial time and located at the same coarse network as the 42 dependent variables, form the set of independent variables. In the second experiment, (E_2), the sea-level-pressure observations are extended geographically along the western and northern boundaries at initial time and their density is doubled over the United States. The number of independent variables is thus increased from 42 to 105, located as indicated by the complete grid in fig. 1. In the last experiment, (E_3), the dependent variables are related to the two-lag sequence of the surface-pressure distribution defined at the 105 grid points at time $t=0$ and $t=-12$ hr. This last experiment contains a total of 210 independent variables.

Comparisons are then made between the P.R.'s of each experiment, and information about the effects of varying data density and geographical coverage upon the explained variance of the 24-hour sea-level pressure distribution is obtained. By proper statistical tests, some information is also gained about the probable performance of the statistical operators upon new data.

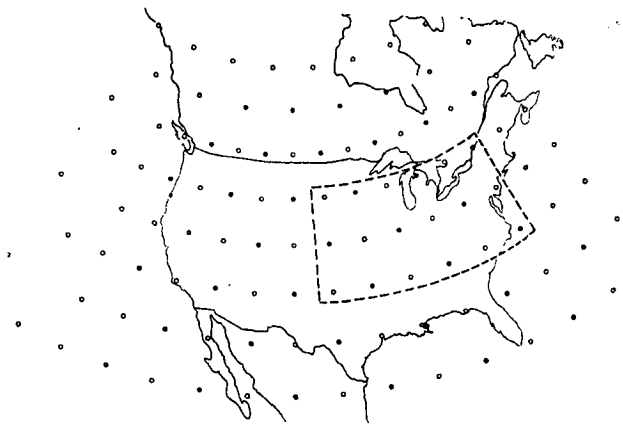

FIG. 1. Grid system used in statistical experiments. Dependent variables are located at points indicated by solid circles for all experiments.

3. Mathematical and computational procedures

Let the pressure deviation from the mean at some grid point i be designated as p_i, and its 24-hour predicted value as $\hat{p}_i$. We wish to express $\hat{p}_i$ as a linear function of some combination of pressures at initial time and at time $t=-12$ hours. Thus we can write

$$\hat{p}_i = \left(\sum_{j=1}^{n} a_{ij}p_j\right)_{t=0} + \left(\sum_{j=1}^{n} a_{ij}p_j\right)_{t=-12}, \quad (1)$$

where j indicates the location in space of the independent variable and i indicates the location of the dependent variable. For E_1, $n=42$ and the pressures are taken at 42 grid points at time $t=0$. For E_2, $n=105$ and the pressures are taken at 105 grid points at time $t=0$. In the last experiment (E_3), $n=105$ and the pressures are taken at all 105 grid points at the initial time and time $t=-12$ hours.

In matrix notation, equation (1) may be written for each of the three experiments

$$\hat{\mathbf{P}}_1=\mathbf{G}_1\mathbf{P}_1, \quad \hat{\mathbf{P}}_2=\mathbf{G}_2\mathbf{P}_2, \quad \hat{\mathbf{P}}_3=\mathbf{G}_3\mathbf{P}_3, \quad (2)$$

where the $\hat{\mathbf{P}}$'s are the vectors of the 42 predictands, the $\mathbf{G}$'s are in order 42×42, 42×105, and 42×210 matrices of coefficients a_{ij}, the $\mathbf{P}$'s are vectors of independent variables and the subscripts refer to the experiment.

The coefficient matrices $\mathbf{G}$ are determined by the standard least-squares process as described by White *et al* (1957) which minimizes the error sum of the squares of the linear estimates. This process leads to the necessity for solving the matrix equations

$$\mathbf{G}_1\mathbf{R}_1=\mathbf{A}_1, \quad \mathbf{G}_2\mathbf{R}_2=\mathbf{A}_2, \quad \mathbf{G}_3\mathbf{R}_3=\mathbf{A}_3, \quad (3)$$

where the $\mathbf{A}$'s are matrices of covariances between the independent variables and the predictands, and the $\mathbf{R}$'s are symmetrical covariance matrices of the independent variables. The solution for the regression coefficient matrices $\mathbf{G}$, may in principle be obtained by inversion of the matrices $\mathbf{R}$. Thus we can write

$$\mathbf{G}_1=\mathbf{A}_1\mathbf{R}_1^{-1}, \quad \mathbf{G}_2=\mathbf{A}_2\mathbf{R}_2^{-1}, \quad \mathbf{G}_3=\mathbf{A}_3\mathbf{R}_3^{-1}, \quad (4)$$

where $\mathbf{R}^{-1}$ is the inverse of $\mathbf{R}$.

Many difficulties arise in practice when the straightforward solution indicated by equation (4) is attempted.

a. It is necessary to constrain the number of independent variables to a very small fraction of the total number of degrees of freedom in the data sample. Unless this requirement is fulfilled, the stability and hence the usefulness of the statistical-prediction operators is greatly impaired.

b. The independent variables, which in this case are sea-level pressures, are highly correlated in both time and space rendering the matrices $\mathbf{R}$ nearly singular.

c. The order of the matrices $\mathbf{R}$ which is equal to the number of independent variables in any given experiment soon becomes too great for efficient processing by even the largest electronic computers available.

Various procedures have been proposed to overcome these difficulties. Recently, Lorenz (1956) has proposed that these difficulties may be overcome by an empirical orthogonal linear transformation of the independent variables such that a much smaller set of transformed variables is obtained which are also linearly independent in time. Linear orthogonal transformations of sets of independent variables as a means of reducing the total number of such variables have been extensively used in various types of statistical prediction schemes by G. P. Wadsworth (1948), R. G. Miller and T. F. Malone (1954), and the present authors (White *et al*, 1957). However, the Tchebyscheff transformations used in these studies do not result in variables which are independent in time and hence do not avoid the problem of nearly singular matrices. The procedure proposed by Lorenz is not essentially different from standard schemes for factor analysis (Thurstone, 1947). Its adaptation to meteorological problems promises to be fruitful.

Consider the procedure as applied to E_1. The set of independent variables p_i ($i=1\cdots42$) is to be expressed by a transformed set which has linear combinations of these variables. If we denote these new variables as Q_j ($j=1\cdots42$) we may write

$$Q_j = \sum_{i=1}^{42} m_{ij}p_i \quad (j = 1\cdots42), \quad (5)$$

where m_{ij} is the set of transformation coefficients.

In matrix notation we may write

$$\mathbf{Q}=\mathbf{M}\mathbf{P}, \quad (6)$$

where $\mathbf{Q}$ is now the vector of the transformed variables, $\mathbf{M}$ is the transformation matrix of the coefficients m_{ij}, and $\mathbf{P}$ is the vector of the independent variables. The problem is to determine the matrix $\mathbf{M}$ which will satisfy the following conditions:

a. The transformation vectors comprising the $\mathbf{M}$ matrix must be orthogonal. The orthogonality condition requires that the $\mathbf{M}$ matrix satisfy the condition that

$$\mathbf{M}\mathbf{M}^T=\mathbf{I}, \quad (7)$$

where $\mathbf{M}^T$ is the transpose of the matrix $\mathbf{M}$ and $\mathbf{I}$ is the identity matrix.

b. The transformed variables $\mathbf{Q}$ must be linearly independent in time. This condition is satisfied if

$$\overline{\mathbf{Q}\mathbf{Q}^T}=\mathbf{D}, \quad (8)$$

where the bar indicates a time average, $\overline{\mathbf{Q}\mathbf{Q}^T}$ takes

the form of the covariance matrix of the transformed variables, and **D** is a diagonal matrix. Equation (8) states that all covariances between the transformed variables vanish, which means that the new transformed variables are linearly independent.

If we now substitute in (8) from (6) we find

$$\mathbf{M}\overline{\mathbf{P}\mathbf{P}^T}\mathbf{M}^T=\mathbf{D}, \tag{9}$$

but we have already indicated that $\overline{\mathbf{P}\mathbf{P}^T}$ is the same as the matrix $\mathbf{R}_1$ of (3) and thus we may write

$$\mathbf{M}\mathbf{R}_1\mathbf{M}^T=\mathbf{D}. \tag{10}$$

The problem of obtaining **M** and **D** will be recognized as being identical with the problem of obtaining the characteristic vectors and roots of the matrix $\mathbf{R}_1$. A procedure for obtaining **M** and **D** which involves a series of elementary rotations of the matrix $\mathbf{R}_1$ has been devised by Lorenz (1956). Programs for obtaining **M** and **D** on the IBM 704 EDPM located at the General Electric Company in Lynn, Massachusetts, were prepared by the authors in collaboration with the programming staff at the General Electric Company. The General Electric machine having only 8192 words of high speed storage permitted the diagonalization of matrices only up to 63rd order in one pass.

We can now reformulate equation (2) for E_1 in terms of the transformed variables by using equations (6) and (7)

$$\hat{\mathbf{P}}_1=\mathbf{G}_1\mathbf{M}_1{}^T\mathbf{Q}_1=\mathbf{H}_1\mathbf{Q}_1 \tag{11}$$

where $\mathbf{H}_1$ is now a set of coefficients operating on the transformed variables. Instead of (3) and (4), we have by the use of (8) that

$$\mathbf{H}_1\overline{\mathbf{Q}_1\mathbf{Q}_1{}^T}=\mathbf{B}_1, \quad \mathbf{H}_1\mathbf{D}_1=\mathbf{B}_1, \quad \mathbf{H}_1=\mathbf{B}_1\mathbf{D}_1{}^{-1}, \tag{12}$$

where $\overline{\mathbf{Q}_1\mathbf{Q}_1{}^T}=\mathbf{D}$ is the covariance matrix of the transformed variables which is a diagonal matrix, $\mathbf{B}_1$ is a matrix of the covariances of the transformed variables and the predictands.

The reformulation of the problem in terms of the transformed variables as indicated in equations (12) has many advantages. As indicated previously, the problem of singular matrices is avoided because the Q's are linearly independent. The inversion of the matrix **D** consequently becomes a trivial problem. Furthermore, it is possible to specify any set of independent variables to any desired degree of average accuracy. This is because the sum of the variances of the original variables remains invariant under the diagonalization process, permitting the assignment of a specific fraction of the total original variance to each transformed variable. The time independence of the Q's also permits their efficient screening on the basis of their contributions to the explained variance of the dependent variables. This results from the fact that the total explained variance of the dependent variable is the simple sum of the contributions to this variance from the individual transformed variables.

4. Organization of the statistical experiments

To carry out the series of statistical investigations, we used a sample of sea-level pressure data for the months January and February, 1948 to 1955, which was kindly made available to us by the Extended Forecast Section of the U. S. Weather Bureau. These data were prepared twice daily on an operational basis and provided a total sample of 944 cases spaced 12 hr apart.

In E_1, the 42 sea-level pressures at time $t=0$ were transformed into a set of 13 Q_1's which explained 94 per cent of the variance of the original variables.

In E_2, the remaining 63 sea-level pressures at time $t=0$ indicated by the open circles in fig. 1 were transformed into a set of 25 Q's which explained 97 per cent of the variance of these 63 pressures. These Q's were combined with $19Q_1$'s which explained 97 per cent of the variance of the 42 sea-level pressures in E_1, into a new set of 44 variables. This set of 44 independent variables was transformed into a new set of 19 Q_2's which specified 97 per cent of the variance of the 44 variables. The Q_2's therefore, explain 94 per cent of the variance of the original 105 sea-level pressures at time $t=0$. The necessity for proceeding in this manner arises from the limitations of the machine capacity.

In E_3, the $19Q_2$'s taken at time $t=0$ and $t=-12$ hours were combined into a new set of 38 independent variables. In this case, the machine was instructed to provide a set of transformed variables, Q_3's, which accounted for 99 per cent of the total variance of the functions describing the 12-hour map sequence. This high degree of specification in the time domain was necessary to insure that pressure changes over 12-hour periods would be faithfully represented. It was possible to specify 99 per cent of the variance by retaining $30Q_3$'s.

5. Results

a. *Experiment No. 1.*—The efficiency of the transformation of the 42 independent variables in E_1 is indicated in table 1 where the per cent of the variance of the original variables explained by the first twenty transformed variables is given, together with the cumulative total of the explained variance. It can be seen that slightly more than 75 per cent of the variance can be accounted for by five, and 90 per cent by ten transformed variables.

The empirical orthogonal functions (characteristic vectors of $\mathbf{R}_1$) are given by the rows of the transformation matrix $\mathbf{M}$. The geographical configurations of the empirical orthogonal functions associated with first, second, fourth, seventh, twelfth, and thirteenth transformed variables are shown in fig. 2. The regularity and scale of the empirical-function patterns decrease with decreasing importance of the functions in specification of the variance of the pressure distribution. In many cases, the patterns of these functions are similar to certain persistent characteristics of sea-level circulation patterns and indeed can be interpreted in this manner. This is because a large positive or negative value of any Q_1 arises from a large positive or negative correlation between the associated functions and the pressure pattern. The functions may thus be regarded as empirical but specific components of the circulation patterns.

TABLE 1. Average per cent of variance of the sea-level pressures at the 42 grid points of E_1 specified by the transformed variables (Q_1's).

Q_1	Per cent variance	Cumulative per cent variance	Q_1	Per cent variance	Cumulative per cent variance
1	29.70	29.70	11	1.37	92.16
2	16.61	46.31	12	1.09	93.25
3	12.36	58.67	13	0.86	94.11
4	9.80	68.47	14	0.70	94.81
5	7.04	75.51	15	0.59	95.40
6	5.16	80.67	16	0.50	95.90
7	3.59	84.26	17	0.46	96.36
8	2.76	87.02	18	0.40	96.76
9	2.10	89.12	19	0.31	97.07
10	1.67	90.79	20	0.30	97.37

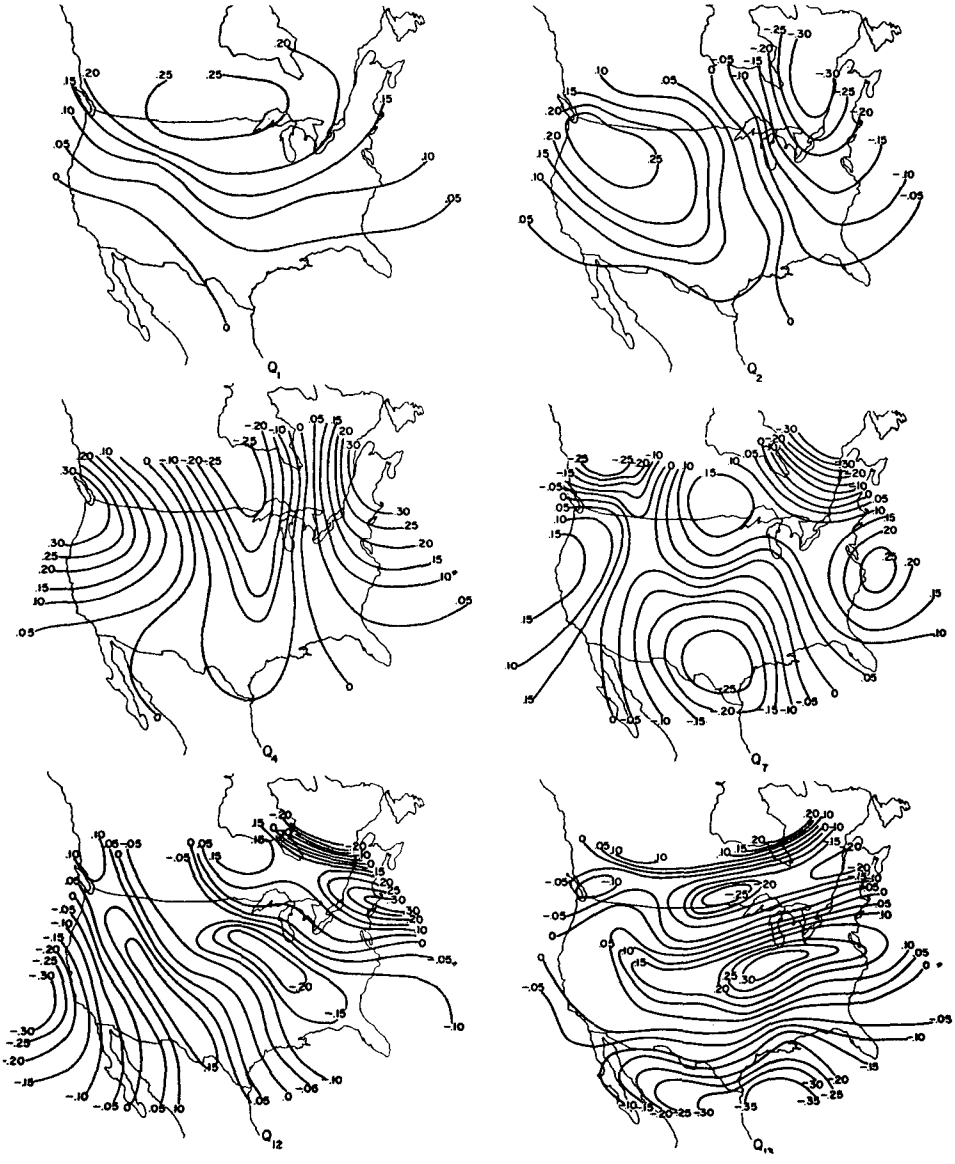

FIG. 2. The geographical patterns of the empirical orthogonal functions associated with the transformed variables of the first experiment as indicated.

The geographical distribution of the total explained variance (R^2) is shown at the top of fig. 3. The explained variance is generally in excess of 50 per cent, except along western and northern boundaries. The distribution of the standard deviation of the residual errors is shown at the bottom of fig. 3. For comparative purposes, the distribution of the standard deviation of the sea-level pressure is shown in fig. 4.

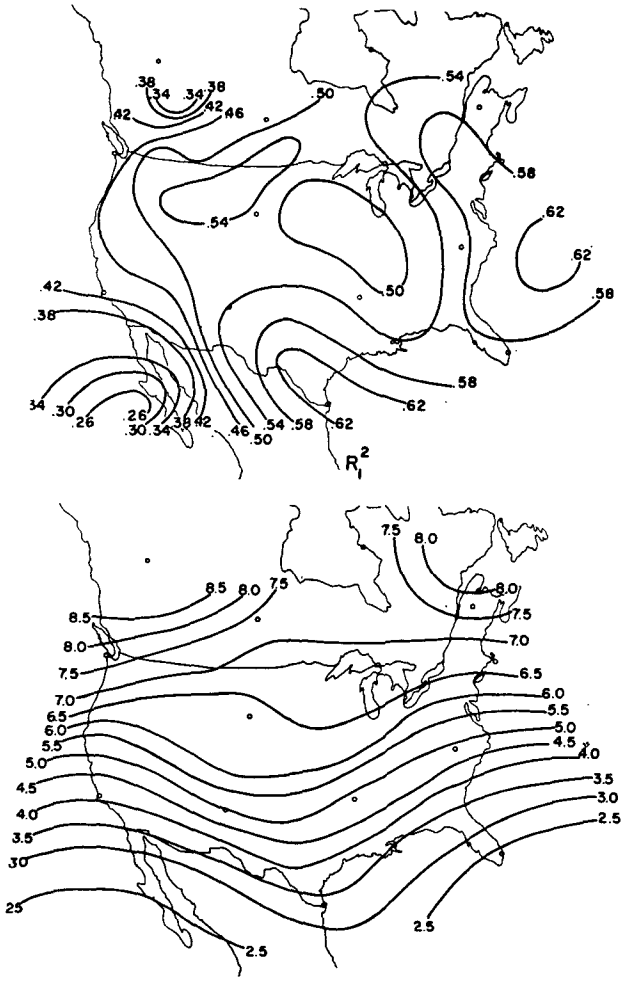

FIG. 3. The geographical distribution of the reduction of the variance of the 24-hr sea-level pressure (top), and the standard error of estimate in mb (bottom) for E_1.

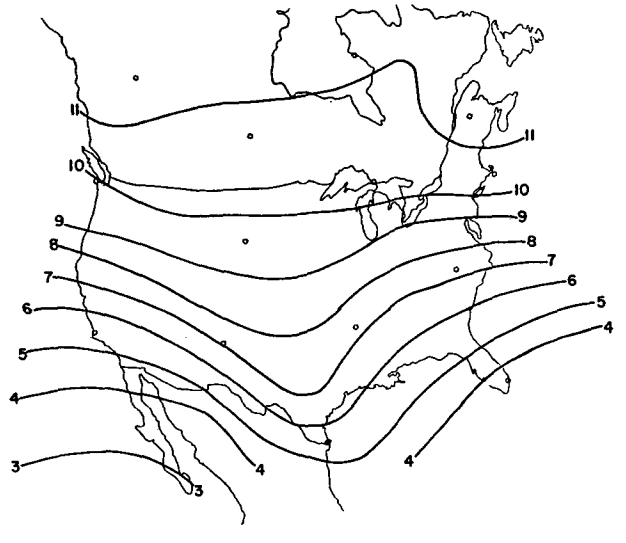

FIG. 4. The geographical distribution of the standard deviation of the sea-level pressure in mb in this winter sample.

We investigated the manner in which the individual Q_1's contributed to R_1^2. In general, there is an association between the importance of the Q_1's in specification of the variance of the independent variables and explanation of the variance of the dependent variables. Further, the geographical variation of R_1^2 is closely related to the geographical configuration of the empirical orthogonal functions. This is illustrated in fig. 5 where the fields of the coefficients of linear correlation (r) between the Q_1's of fig. 2 and the 42 predictands are plotted. The geographical distribution of these correlations reflects the normal progression of the circulation components represented by the empirical functions. Thus, the charts in the upper left corner of fig. 2 and 5 reflect the fact that when a high pressure area covers the central region of the U. S. it generally proceeds south-eastward. This motion accounts for the region of high positive correlation to the south and east of the maximum in the empirical function. Similar reasoning leads to an appreciation of the relations between the corresponding charts of fig. 2 and 5.

The selection of the Q_1's to this point has been made without any reference to the dependent variables. It is possible to rank each of the thirteen Q_1's according to their contribution to the explained variance of the dependent variables and perform an elimination of all those which do not contribute significantly. An illustrative sample of dependent variables has been chosen from the central region of the grid system. The nine variables chosen for display are those indicated by the solid circles within the dashed curve of fig. 1 and are numbered from north to south and east to west.

The screening analysis is shown in table 2 where the contributions of each of the 13 Q_1's to the explained variance of the dependent variables are listed.

If it were possible to specify the magnitude of the per cent reduction which could be considered statistically significant, it would be a simple matter to eliminate all those Q_1's accounting for less than this amount. Unfortunately, it is not possible to assign such a statistical significance criterion rigorously because the number of degrees of freedom is not exactly known. The authors have arbitrarily considered that an explained variance by one independent variable of less than 1 per cent is not significant. This criterion corresponds to about 385 deg of freedom in the sample at the 5 per cent significance level. This is not an unreasonable number considering that

the total sample in E_1 consists of 928 cases. If only those variables which explain more than 1 per cent of the variance of the dependent variable are retained, it is possible to reduce the number of independent variables to as few as four in one case. The largest number of variables which would be retained would be nine. The total variance explained by the fully screened system is shown in next to the last line of table 2.

b. *Experiment No. 2.*—The independent-variable data density was doubled and the geographical coverage extended along the western and northern boundaries for the second experiment. This leads to the use of the sea-level pressures at the 105 points of fig. 1 as the independent variables. This distribution of sea-level pressure was represented by a set of 19 transformed variables, Q_2's in the manner described in Section 3.

In both E_1 and E_2 we require that the set of transformed variables specify 94 per cent of the total variance. In E_1 this required $13Q_1$'s while in E_2 this required $19Q_2$'s. This means that the additional information added in E_2 over E_1 could be represented by only six more independent variables.

The geographical distribution of R_2^2 for E_2 and the standard deviation of the residual errors is shown in fig. 6. Comparison with similar results for E_1 (fig. 3) shows only slight differences, except in western and northern regions where the additional boundary information has added considerably to the explained

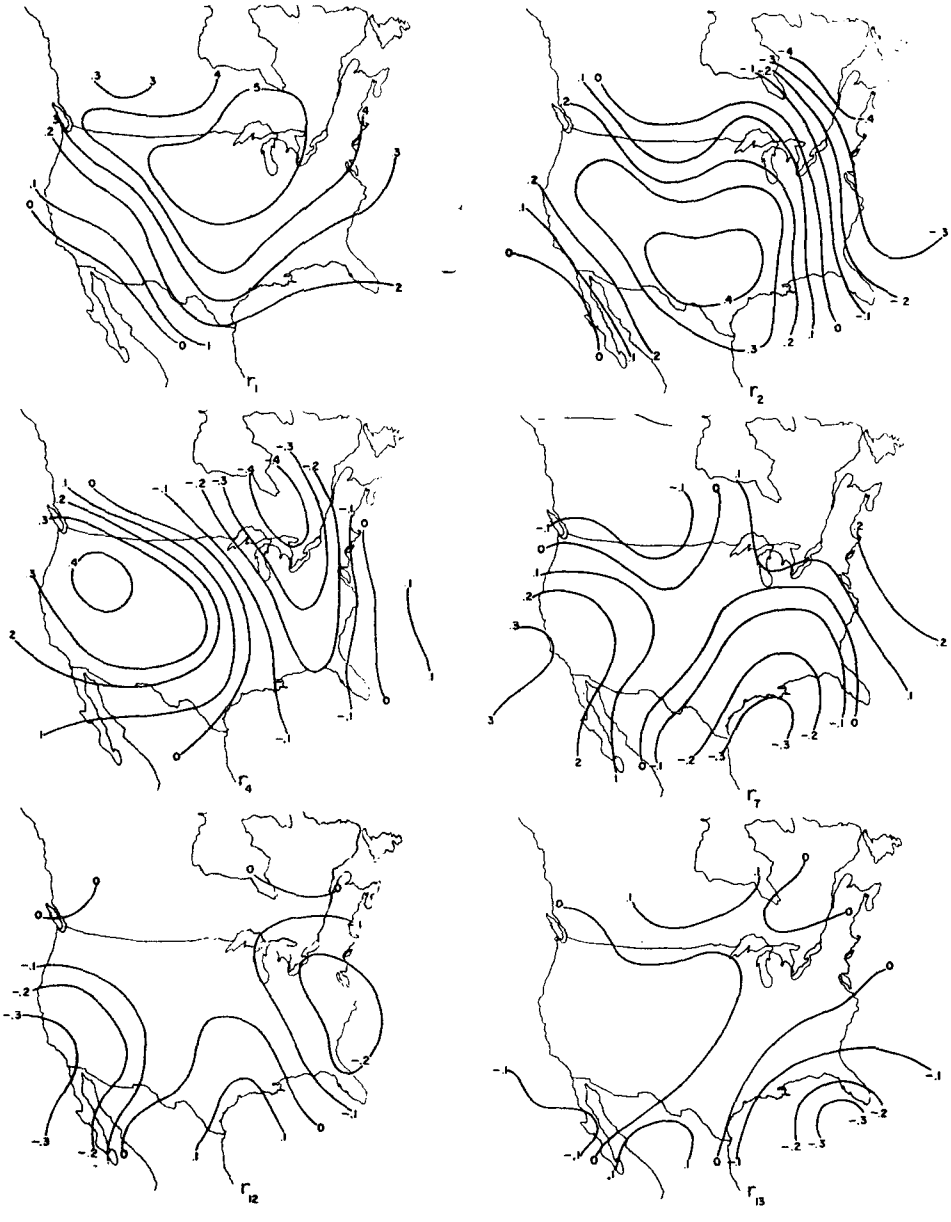

FIG. 5. The geographical distribution of the field of linear correlation between each of the transformed variables associated with the empirical functions of fig. 2 and the 24-hr sea-level pressure.

variance. Doubling the data density in the manner done here does not appear to add significantly to the prediction accuracy.

Because of this result the authors conducted two subsidiary experiments to determine whether there existed a critical data density above which the

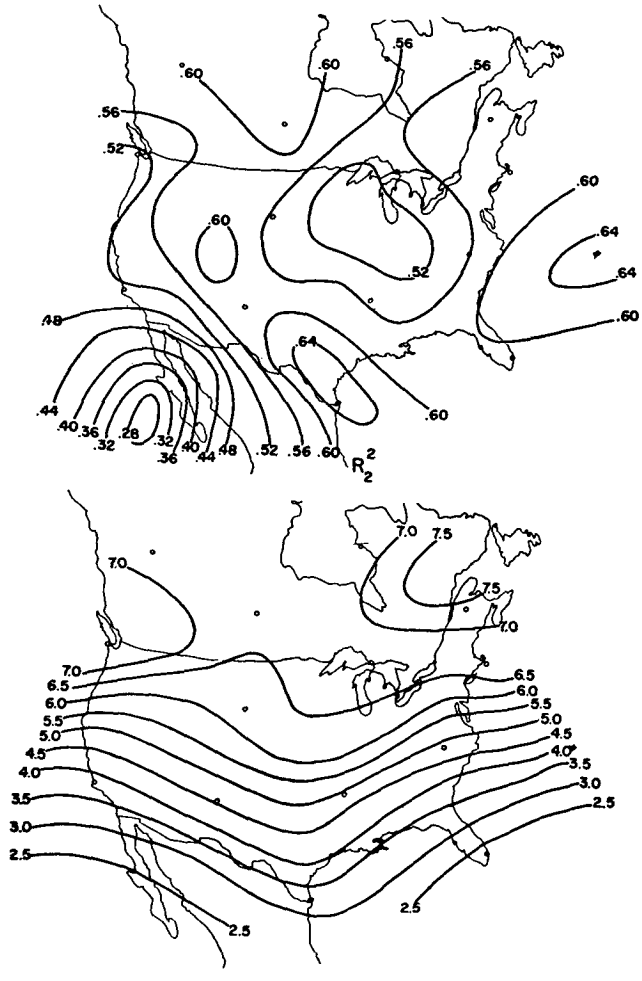

FIG. 6. The geographical distribution of the reduction of the variance of the 24-hr sea-level pressure (top) and the standard error of estimate in mb (bottom) for E_2.

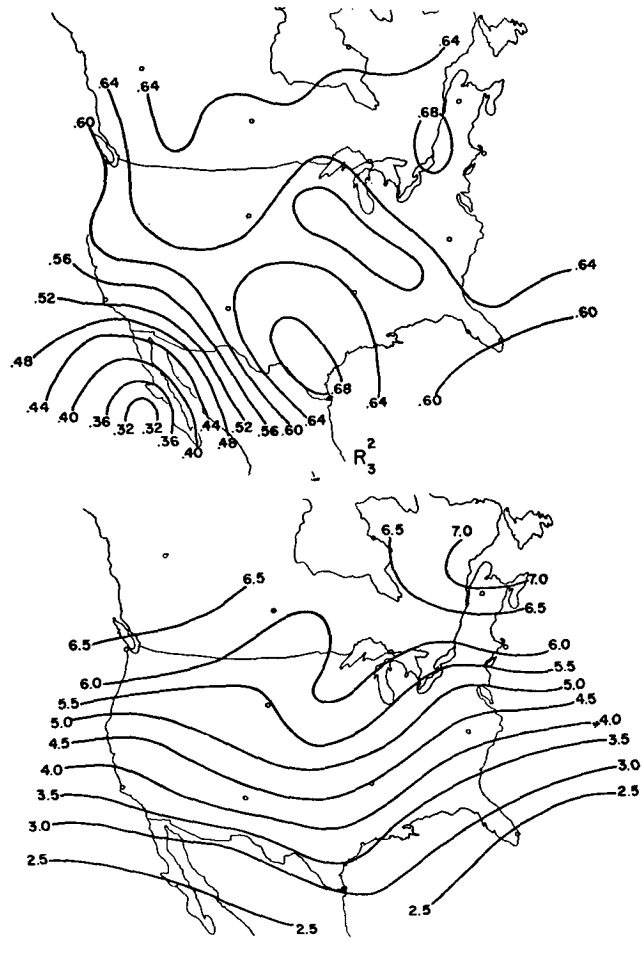

FIG. 7. The geographical distribution of the reduction of the variance of the 24-hr sea-level pressure (top) and the standard error of estimate mb (bottom) for E_3.

TABLE 2. Screening analysis of the independent variables (Q_1's) of E_1, for a set of dependent variables in northeastern United States. Q_1's are listed in order of decreasing contribution to the explained variance of the dependent variables together with their corresponding ranks in specifying the variance of the sea-level pressure distribution. r^2 expressed in per cent.

	Predictands																	
	1		2		3		4		5		6		7		8		9	
	Q_1 rank	r^2	Q_1 rank	r^2	Q_1 rank	r^2	Q_1 rank	r^2	Q_1 rank	r^2	Q_1 rank	r^2	Q_1 rank	r^2	Q_1 rank	r^2	Q_1 rank	r^2
1	1	22.8	5	18.0	1	18.4	1	25.8	5	14.5	1	25.9	1	29.9	1	24.0	1	27.4
2	4	9.9	1	10.7	4	10.7	4	13.8	1	13.5	2	11.8	3	7.1	2	18.0	2	10.4
3	2	6.1	2	9.6	5	7.1	8	5.0	3	7.8	8	3.4	2	6.4	6	6.8	4	10.0
4	3	6.0	6	7.0	3	6.9	12	2.7	9	4.7	6	3.3	6	1.6	3	3.8	3	1.5
5	6	5.3	12	5.7	12	4.8	2	1.7	4	4.7	5	1.1	8	0.9	7	1.0	11	1.3
6	12	2.5	9	2.8	6	3.3	7	1.4	7	2.2	4	1.0	5	0.8	4	0.9	6	0.3
7	7	2.5	7	2.8	11	1.7	5	0.9	12	1.6	3	0.9	12	0.5	8	0.7	13	0.2
8	5	1.2	3	2.2	9	0.7	11	0.8	11	1.3	12	0.4	4	0.2	11	0.5	12	0.2
9	9	0.9	8	1.1	8	0.5	10	0.6	6	0.6	9	0.3	11	0.1	5	0.3	7	0.1
10	11	0.8	4	0.6	13	0.5	6	0.3	13	0.3	13	0.1	10	0.0	13	0.0	10	0.1
11	10	0.8	11	0.0	10	0.3	3	0.3	2	0.3	7	0.1					8	0.0
12	8	0.3			2	0.1	9	0.0	10	0.0	11	0.1						
13	13	0.1			7	0.0					10	0.0						
Sum ($r^2>1\%$)		56.3		59.9		52.9		50.3		50.3		46.5		45.0		53.6		50.6
R^2		59.2		60.5		54.9		53.4		51.8		48.5		47.5		56.1		51.6

explained variance of the dependent variables is not substantially improved. In their subsidary experiments, independent variables covering the same geographical areas as in E_1 were selected at 18 and 9 evenly spaced grid points. For dependent variables located at the 9 grid points indicated by the solid circles inclosed by the dashed line of fig. 1, average P.R.'s of 46 per cent, 52 per cent, and 54 per cent were obtained when 9, 18, and 42 independent variables were used. This is a rather interesting result indicating that there is a data-density limit beyond which improved reduction of variance is almost negligible when this type of linear statistics is used.

A screening analysis of the Q_2's has been undertaken in a manner analogous to that shown in table 2 with essentially similar results.

c. *Experiment No. 3.*—Having expressed the sea-level circulation as defined at the 105 points of fig. 1 in terms of $19Q_2$'s, we investigated the effects of introducing the circulation information at time $t=-12$ hours. The $19Q_2$'s at time $t=-12$ hours with those at time $t=0$ form a new set of 38 independent variables. A set of $30Q_3$'s, which retained 99 per cent of the variance of the 38 Q_2's was obtained.

For this third experiment the geographical distribution of R_3^2 and the standard deviation of the residual errors are shown in fig. 7. Rather marked increases in the explained variance are evident when comparison is made with the results of E_2 in fig. 6. The largest differences appear in an east-west belt through the northern United States where they exceed 10 per cent. This improvement in R_3^2 is not due to the larger number of independent variables in E_3. The R_3^2's accounted for by the 19 best Q_3's were computed, yielding results which were not significantly different from those shown in fig. 7.

Since the difference $R_3^2-R_2^2$ is not due to the increased number of independent variables in E_3, and since the conditions of the experiments E_2 and E_3 are the same except for the introduction of sea-level pressure information at time $t=-12$ hours in E_3, it is reasonable to ascribe some reality to the difference. Aside from actual tests upon independent data, however, there is no way of making rigorous significance statements. Treatment of the multiple correlations as simple correlations between the linear estimates and the observed values of the dependent variables (Fisher, 1946) permits the application of standard tests of statistical significance which must be considered as necessary rather than sufficient conditions. Such tests were applied with the result that the differences across the northern regions of the

TABLE 3. Screening analysis of the independent variables (Q_3's) of E_3, for a set of dependent variables in northeastern United States. Q_3's are listed in order of decreasing contribution to the explained variance of the dependent variables together with their corresponding ranks in specifying the variance of the sea-level pressure distribution. r^2 expressed in per cent.

	Predictands															
	1		2		3		4		5		6		7		8		9	
	Q_3 rank	r^2	Q_3 rank	r^2	Q_3 rank	r^2	Q_3 rank	r^2	Q_3 rank	r^2	Q_3 rank	r^2	Q_3 rank	r^2	Q_3 rank	r^2	Q_3 rank	r^2
1	1	13.6	6	14.6	1	12.3	1	16.4	3	11.4	1	20.8	1	24.7	1	21.6	1	25.4
2	3	9.5	1	8.0	3	12.2	20	8.2	1	10.4	24	7.2	6	6.1	6	9.3	8	8.0
3	6	6.9	10	5.6	8	6.6	3	7.1	24	6.7	20	6.2	23	5.6	24	7.0	6	5.3
4	20	5.5	2	5.6	20	6.2	4	6.4	5	4.3	6	6.0	4	5.0	3	6.6	24	3.7
5	8	5.2	12	5.2	24	3.8	27	5.8	10	3.4	3	5.9	20	3.9	2	3.6	2	3.2
6	2	4.3	7	3.6	27	3.5	8	5.4	20	2.9	5	2.2	26	2.5	5	3.3	12	3.0
7	12	4.3	29	3.5	6	3.2	23	3.4	22	2.8	15	1.8	24	2.4	20	3.2	20	1.9
8	27	2.9	21	2.1	29	2.5	24	3.1	8	2.7	2	1.7	8	1.5	15	2.0	27	1.8
9	7	2.4	22	2.1	11	2.2	6	1.5	6	2.3	22	1.2	12	1.3	22	1.8	23	1.6
10	28	2.3	23	2.1	4	2.0	7	1.4	21	2.3	23	1.2	5	1.1	8	1.5	30	1.2
11	29	2.0	8	1.7	12	2.0	11	1.2	11	2.2	4	1.0	11	0.7	16	0.9	29	1.1
12	4	1.6	25	1.6	22	1.6	9	1.0	29	1.8	9	0.7	2	0.6	13	0.8	26	1.1
13	5	1.0	5	1.5	21	1.4	22	0.8	15	1.5	16	0.6	15	0.6	30	0.8	25	0.9
14	26	0.9	30	1.2	30	1.3	16	0.8	30	1.5	30	0.5	7	0.6	28	0.7	7	0.8
15	11	0.9	3	1.1	25	0.7	12	0.4	14	1.1	11	0.4	25	0.6	27	0.6	9	0.7
16	22	0.9	11	1.0	2	0.6	15	0.4	4	0.8	27	0.4	28	0.4	4	0.5	5	0.5
17	30	0.8	18	1.0	10	0.6	14	0.3	7	0.5	26	0.4	3	0.4	23	0.3	3	0.4
18	21	0.7	20	0.9	18	0.3	21	0.2	25	0.2	19	0.2	30	0.4	26	0.3	22	0.4
19	24	0.7	24	0.9	5	0.3	13	0.2	27	0.2	13	0.2	9	0.4	18	0.2	16	0.4
20	10	0.6	27	0.8	26	0.2	30	0.2	2	0.1	29	0.1	16	0.3	25	0.2	19	0.3
21	14	0.3	14	0.7	28	0.2	28	0.2	12	0.1	28	0.1	13	0.3	29	0.1	15	0.2
22	16	0.3	4	0.7	9	0.1	10	0.1	23	0.1	14	0.1	29	0.2	7	0.0	29	0.2
23	15	0.2	28	0.6	7	0.1	18	0.1	13	0.1	21	0.1	14	0.2			11	0.1
24	17	0.1	17	0.4	15	0.1	2	0.0	19	0.1	7	0.1	10	0.1			4	0.1
25	19	0.1	9	0.1	13	0.1			28	0.0	8	0.0	27	0.0			13	0.1
26	9	0.1	15	0.0	19	0.0											14	0.1
27	25	0.1															10	0.1
28	18	0.1															18	0.0
29	13	0.0																
Sum ($r^2>1\%$)		61.5		61.5		60.8		60.9		57.0		55.2		54.1		59.9		57.3
R^2		68.5		66.7		64.2		64.7		59.7		59.0		59.8		65.1		62.4

United States are large enough to satisfy these statistical significance requirements at the 5 per cent level.

Analysis of the manner in which the Q_3's contribute to the explained variance of the selected predictands is shown in table 3. For E_3, many more independent variables now contribute more than 1 per cent of the explained variance than in E_1. Generally, the largest P.R.'s associated with individual Q's in E_3 are smaller than those in E_1. Furthermore, many of the very high order Q's in E_3 associated with the small-scale circulation features in the space-time domain contribute very importantly to the P.R.'s. This is much less true in E_1 where the small-scale features in the space domain are generally of secondary importance in explaining the variance of the dependent variables.

The latter phenomenon is associated with the fact that over short time periods of the order of 12 hr, the large-scale circulation features do not undergo marked changes. This is brought out in table 4 where the 12-hr lag correlations of the Q's of E_2 are shown. It can be seen that the low order Q_2's tend to have higher lag correlations than the higher order Q_2's. This indicates that the smaller scale components of the circulation are more sensitive to and hence contain considerable information about the way in which the circulation is changing. That this information is of importance in explaining the variance of the 24-hr sea-level pressure is not unexpected.

The former phenomenon reflects the correlation between the pressure distributions at time $t=0$ and $t=-12$ hours and the allocation of the total predictive information among the larger number of independent variables in E_3.

6. Summary

The increase in sea-level-pressure data density in space beyond certain critical limits does not significantly add to the explained variance of the 24-hr sea-level-pressure distributions. Addition of information in the time domain does appear to add significantly to the P.R. as indicated by the results of E_3. As discussed by White *et al* (1957), addition of such information should lead to improved P. R.'s on physical grounds.

TABLE 4. Twelve hour lag correlations of Q_2's.

Q_2 Rank	12-hr lag r	Q_2 Rank	12-hr lag r
1	0.93	11	0.66
2	0.90	12	0.40
3	0.79	13	0.55
4	0.87	14	0.29
5	0.85	15	0.73
6	0.71	16	0.56
7	0.76	17	0.57
8	0.62	18	0.62
9	0.80	19	0.55
10	0.78		

The question arises as to whether it might not be possible to achieve equivalent P.R.'s by a variety of combinations of time and space densities of weather information. This would imply an ability to substitute time for space information, and vice-versa, and might lead to better schemes for arranging observation networks and the scheduling of their output.

The techniques of factor analysis used in this paper to reduce the number of independent variables in the regression system and to select those which contribute significantly to the P.R., appear to be efficient and useful. Some information about the usefulness of the statistical operators for actual prediction is given by the magnitude of the explained variance, although actual tests should be made on independent data before rigorous statements about the utility of the operators can be made.

REFERENCES

1. Fisher, R. A., 1946: *Statistical methods for research workers.* Edinburgh, Oliver and Boyd, Ltd., 354 pp.
2. Lorenz, E. N., 1956: *Empirical orthogonal functions and statistical weather prediction.* Sci. Rep. No. 1, Statis. Fcst. Proj., Contract No. AF19(604)-1566, Mass. Inst. Tech.
3. Malone, T. F., and R. G. Miller, 1956: *Studies in synoptic climatology.* Final Rep., Off. Nav. Res. No. N5ori-07883, Mass. Inst. Tech.
4. Thurstone, L. L., 1947: *Multiple-factor analysis.* Chicago, Univ. of Chicago Press.
5. Wadsworth, G. P., 1948: *Short range and extended forecasting by statistical methods.* Tech. Rep. No. 105–37, Air Wea. Serv., Mass. Inst. Tech.
6. White, R. M., R. C. Derby, D. S. Cooley, and F. A. Seaver, 1957: Hemispherical forecasting by statistical techniques. *J. Meteor.*, **14**, 448–457.