

AD-A034 379

AIR FORCE GEOPHYSICS LAB HANSCOM AFB MASS  
MULTI-PREDICTOR CONDITIONAL PROBABILITIES.(U)  
OCT 76 I I GRINGORTEN

F/G 4/2

UNCLASSIFIED

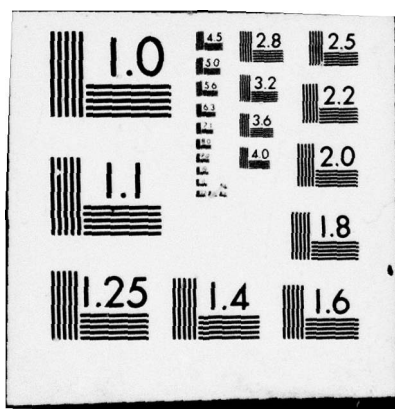
AFGL-TR-76-0239

NL

1 OF 1  
AD-A  
034 379



END  
DATE  
FILMED  
2-15-77  
NTIS



U.S. DEPARTMENT OF COMMERCE  
National Technical Information Service

AD-A034 379

MULTI-PREDICTOR CONDITIONAL PROBABILITIES

AIR FORCE GEOPHYSICS LABORATORY  
HANSCOM AIR FORCE BASE, MASSACHUSETTS

6 OCTOBER 1976

ADA034379

018057

AFGL-TR-76-0239

AIR FORCE SURVEYS IN GEOPHYSICS, NO. 354



## Multi-Predictor Conditional Probabilities

IRVING I. GRINGORTEN

6 October 1976

Approved for public release; distribution unlimited.

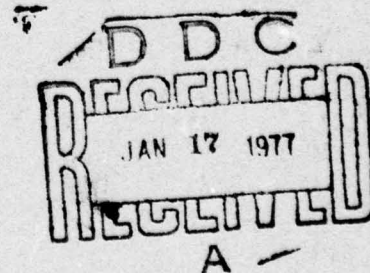
METEOROLOGY DIVISION PROJECT 8624

**AIR FORCE GEOPHYSICS LABORATORY**

HANSCOM AFB, MASSACHUSETTS 01731

**AIR FORCE SYSTEMS COMMAND, USAF**

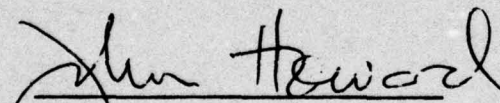
REPRODUCED BY  
NATIONAL TECHNICAL  
INFORMATION SERVICE  
U. S. DEPARTMENT OF COMMERCE  
SPRINGFIELD, VA. 22161





This technical report has been reviewed and  
is approved for publication.

FOR THE COMMANDER:

  
\_\_\_\_\_  
Chief Scientist

Qualified requestors may obtain additional copies from the Defense  
Documentation Center. All others should apply to the National  
Technical Information Service.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFGL-TR-76-0239	2. GOVT ACCESSION NO.	3. PERFORMER'S CATALOG NUMBER
4. TITLE (and Subtitle) MULTI-PREDICTOR CONDITIONAL PROBABILITIES		5. TYPE OF REPORT & PERIOD COVERED Scientific, Interim
		6. PERFORMING ORG. REPORT NUMBER AFSG No. 354
7. AUTHOR(s) Irving I. Gringorten		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Air Force Geophysics Laboratory Hanscom AFB Massachusetts 01731		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS P. E. 62101F 86240205
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Geophysics Laboratory Hanscom AFB Massachusetts 01731		12. REPORT DATE 6 October 1976
		13. NUMBER OF PAGES 24
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES Tech, Other		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Stochastic modelling Climatology Conditional probability Predictand Multi-Predictor		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) A predictand's probability distribution is modified by information on one or more of its predictors. If linear dependence is assumed between the predictand and the predictors transformed into normal Gaussian variates, then a model algorithm is possible for the conditional probability of the predictand. It is given as the probability that a Gaussian variable ( $\eta$ ) will equal or exceed a threshold value ( $\eta_c$ ) where $\eta_c$ is expressed linearly in terms of specific normalized values of the predictors. The predictor		

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)

20.

coefficients, known as partial regression coefficients, are functions of the correlations between predictors and the correlations between each predictor and the predictand.

This stochastic model was tested on regular 3-hourly observations of precipitation-produced radar echoes at five widely scattered stations in the eastern half of the United States. The results revealed strong evidence of the validity of the probability estimates, but more importantly revealed that the model can yield sharp estimates of the conditional probability with as many as seven predictors.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE(When Data Entered)



## Preface

The subject of this report in its broadest aspects, has been the topic of many conferences between personnel of the Design Climatology Branch, AFGL; its contractor, St. Louis University; and the Air Weather Service. The latter had expressed requirements, since 1969, for conditional climatology for military operations aimed at short-range predictions of 1 to 6 hours in advance, and more recently for military planning of worldwide Air Force missions. This writer's discussions with Capt. Albert R. Boehm, AWS DNT, have been most valuable. Capt. Boehm used an alternative method to obtain the model algorithm from the basic equations of T. W. Anderson. His knowledge and expertise in the statistical literature, freely shared, are gratefully acknowledged.

ACCESSION for	
DTIC	White Section <input checked="" type="checkbox"/>
DOC	Buff Section <input type="checkbox"/>
UNANNOUNCED	<input type="checkbox"/>
JUSTIFICATION	
BY	
DISTRIBUTION/AVAILABILITY CODES	
Dist.	AVAIL. AND/OR SPECIAL
A	

## Contents

1. INTRODUCTION	7
2. HYPOTHESIS	8
3. ANALYSIS	8
3.1 Single Predictor	11
3.2 Two Predictors	11
3.3 Three or More Predictors	11
4. SAMPLE APPLICATION	12
4.1 Data Collection	13
4.2 Finding the Correlations	14
4.3 Finding the Partial Regression Coefficients	17
4.4 Estimating Conditional Probability	18
5. SUMMARY AND CONCLUSIONS	22
REFERENCES	24

## Illustrations

1. Sample of a PPI Radar-scope Picture of Storms Within 125 nmi of the Radar Station	13
2. Illustration of the Eight Equal Areas, Each Obtained by the Grouping of Eight Cells (Figure 1)	13



## Tables

1. Sample Data Collection of Radar Coverage of Precipitation	14
2. The Frequency of Fractional Cover, in Oktas	15
3. The Cumulative Relative Frequency of Coverage Equal to or Greater Than L Oktas	15
4. The Values of the Transnormalized Variables Corresponding to the Relative Frequencies of Table 3	16
5. The Cross-Correlation Coefficients ( $\rho_{ij}$ ) Between the Transnormalized Values of the Fractional Cover or Okta Count in Groups 1 to 8	16
6. The cc of the Predictand With Each Predictor	17
7. The Partial Regression Coefficients of Eq. (11) as Determined for Key West, Florida, January, for Time Lags 0, 3, 6, 9, 12 hr Between Predictors and Predictand	17
8. The Coefficients $a_i$ Divided by b, Determined for Key West, Florida, for Time Lag of 6 hr, in January, April, July, and October	18
9. The Number of Times That the Conditional Probability of Fractional Coverage of 1/8, 1/4, 1/2, or Full Coverage was Estimated in the Range 0.0 to 0.099, 0.1 to 0.199, ---.	19
10. The Frequency (%) of the Estimates of Conditional Probability of Fractional Coverage (Including 8, 16, 32, and 64 Oktas) Among the Cases When the Conditional Probability was Estimated $\geq 0.10$	20
11. The Skill Scores Which Range From 0 for "No Skill" to 1.0 for Perfectly Correct Forecasts	22

## Multi-Predictor Conditional Probabilities

### 1. INTRODUCTION

Beginning with the climatic frequency of a weather condition or event, it is possible to estimate the modified probability of its occurrence, or to give its conditional probability following certain antecedents. Symbolically, if  $Y$  is the predictand, and  $X_1, \dots, X_n$  are  $n$  predictors, then it is desired to find

$$P(Y \geq Y_c \mid X_1, \dots, X_n)$$

where  $Y_c$  is a threshold value of the predictand, and  $P$  is the symbol for probability.

It need not be assumed that  $Y$  and  $X_i$  ( $i = 1, n$ ) are described in categories only. Instead, in this report, all weather elements are assumed transformable into continuous variables. If  $X$  or  $Y$  is transformed into a normal Gaussian variable, then the new variable,  $x$  or  $y$ , is conveniently called the transnormalized variable, a term first suggested to this writer by Capt. Albert Boehm, Air Weather Service. The predictors,  $X_1, \dots, X_n$  and predictand  $Y$  are considered specific values such as temperature (for example,  $12^\circ\text{C}$ ) or cloud cover (for example,  $3/10$ ). Each transnormalized counterpart,  $x_1, \dots, x_n$  or  $y$  has mean zero and variance 1.0 and Gaussian distribution.

---

(Received for publication 5 October 1976)

While often times there are analytical formulas for the transformation of  $X$  to  $x$ , it is always possible to do this transformation and obtain a one-to-one correspondence, through the cumulative distribution of  $X$  or by plotting  $X$  on normal probability paper<sup>1</sup>. Essentially, the transnormalization is accomplished through the probabilities. Thus:

$$P(x \geq x_c) = P(X \geq X_c) \quad (1)$$

where the subscript (c) is used to denote an assigned threshold value. In this report, for convenience, the symbol for probability is sometimes written as  $P(\xi)$  to denote the probability of equalling or exceeding the bracketed quantity. Thus, Eq. (1) is also written as

$$P(x_c) = P(X_c) .$$

A difficulty arises when the minimum of  $X$  must be categorized, such as "no rain". When this happens, arbitrarily, half of the relative frequency of no rain ( $X_0$ ) can be added to the frequency of rain to give  $P(X \geq X_0)$ .

## 2. HYPOTHESIS

The analytical model for conditional probability, or algorithm, is based on the assumption of linear dependence of the transnormalized predictand on transnormalized predictors. Thus:

$$y = a_1 x_1 + \dots + a_n x_n + b \eta \quad (2)$$

where  $\eta$  is random and normally distributed,  $a_1, \dots, a_n$  are partial regression coefficients, and  $b$  is a coefficient with magnitude such that the normality of  $y$  is preserved.

## 3. ANALYSIS

Because of normality, correlation coefficients (hereafter abbreviated as cc) are given by:

1. Gringorten, I.I. (1972) Conditional probability for an exact noncategorized initial condition, *Monthly Weather Review* 100:796-798.

$$\rho_{ij} = E x_i x_j$$

$$\rho_i = E y x_i$$

where  $\rho_{ij}$  is the symbol for cc between predictors  $x_i$  and  $x_j$ , and  $\rho_i$  is the symbol for cc between predictand ( $y$ ) and the  $i$ th predictor ( $x_i$ ).

From Eq. (2), after squaring both sides of the equation and taking expected values,

$$1 = \sum_{i=1}^n a_i^2 + 2 \sum_{i \neq j} a_i a_j \rho_{ij} + b^2. \quad (3)$$

By multiplying both sides of (2) by  $x_i$  and taking expected values,

$$\rho_i = a_i + \sum_{j \neq i} a_j \rho_{ij}. \quad (4)$$

In matrix form, where

$$R = \begin{pmatrix} \rho_i \\ - \\ - \\ - \\ \rho_n \end{pmatrix} \quad (5)$$

$$C = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1n} \\ - & - & - & - \\ - & - & \rho_{ij} & - \\ - & - & - & - \\ \rho_{n1} & - & - & - & 1 \end{pmatrix} \quad (6)$$

$$A = \begin{pmatrix} a_1 \\ - \\ - \\ - \\ a_n \end{pmatrix} \quad (7)$$

then Eq. (4) becomes  $R = CA$ .

By premultiplying both sides of this last equation by the inverse of  $C$ ,

$$A = C^{-1}R = \hat{C}R / |C| \quad (8)$$



where  $|C|$  is the determinant corresponding to  $C$ ;  $\hat{C}$  is the adjoint matrix of  $C$ , or the matrix of cofactors ( $p_{ij}$ );  $p_{ij} = (-1)^{i+j} |M_{ij}|$  where  $|M_{ij}|$  is the determinant of the submatrix of order  $(n-1)$  obtained by deleting the  $i$ th row and  $j$ th column from  $C$ .<sup>2</sup>

From Eq. (3)

$$b^2 = 1 - |A^T C A| = 1 - |A^T R|$$

or

$$b = \sqrt{1 - a_1 \rho_1 - \dots - a_n \rho_n} \quad (9)$$

From Eq. (2) for specific values of  $x_1, \dots, x_n$  the value of  $\eta$  will exceed  $\eta_c$  as frequently as  $y$  exceeds an assigned minimum  $y_c$  or

$$P(\eta \geq \eta_c) = P(y \geq y_c | x_1, \dots, x_n) \quad (10)$$

Or, the probability that  $\eta$  will exceed  $\eta_c$  is the conditional probability that  $y$  will exceed  $y_c$ . Its solution, therefore, is given by the solution of  $\eta_c$  in

$$\eta_c = (y_c - a_1 x_1 - \dots - a_n x_n) / b \quad (11)$$

where  $a_i$  ( $i = 1, n$ ) and  $b$  are given by Eqs. (8) and (9) supported by Eqs. (5) and (6). Since  $\eta$  has a Gaussian distribution, the probability of it exceeding  $\eta_c$  is given in tables of almost any text in statistics or approximately by algorithm (see below).

Capt. Albert R. Boehm, in personal correspondence, revealed that the results in Eqs. (10) and (11) are obtainable from the deductions of T.W. Anderson.<sup>3</sup> Anderson's theorem bears promise of further enhancement of the results by giving solutions for the conditional probability of several predictands in combination, as well as predictors in combination. This possibility, however, has not yet been examined.

2. Hohn, Franz E. (1964) Elementary Matrix Algebra, The MacMillan Company, New York.
3. Anderson, T. W. (1958) An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, New York.



### 3.1 Single Predictor

For a single predictor  $R = \rho$ ,  $C = 1$ , and consequently  $a_1 = \rho$  whence

$$\eta_c = (y_c - \rho x) / \sqrt{1 - \rho^2} \quad (12)$$

a result previously obtained<sup>1</sup>.

### 3.2 Two Predictors

In this case, the solution for  $\eta_c$  in Eq. (11) becomes

$$\eta_c = (y_c - a_1 x_1 - a_2 x_2) / \sqrt{(1 - a_1 \rho_1 - a_2 \rho_2)} \quad (13)$$

$$\begin{aligned} a_1 &= (\rho_1 - \rho_2 \rho_{12}) / (1 - \rho_{12}^2) \\ \text{where} \\ a_2 &= (\rho_2 - \rho_1 \rho_{12}) / (1 - \rho_{12}^2) \end{aligned}$$

### 3.3 Three or More Predictors

In terms of the cc's ( $\rho_i$  and  $\rho_{ij}$ ), the expressions for  $a_1, \dots, a_n$  become too long and complicated for hand-written analysis. But solutions are feasible if a computer technique will provide numerical answers for A in Eq. (8), using Eqs. (5) and (6) as input. One such computer program was used as a subroutine in the example described below. It is known as the "CDC-6600 Subroutine MATRIX". The method of this subroutine is basically the classical Gauss-Jordan Method.<sup>4</sup> Once the vector A, or the partial regression coefficients ( $a_1, \dots, a_n$ ), are determined, then b is determined from Eq. (9).

Now, the values of the transnormalized predictors,  $x_1, \dots, x_n$  and predictand y may be given. Alternatively, each  $x_i$  may need to be found from the given probability of  $x_i$  being equalled or exceeded. Where  $P(x) = p$ ,

$$\frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\xi^2/2} d\xi = p, \quad (14)$$

4. Dixon, W.J. (1965) BMD Biomedical Computer Programs, University of California at Los Angeles, Health Sciences Facility, Los Angeles.

the value of  $x$  can be approximated to a high degree of accuracy by a convenient algorithm:<sup>5</sup>

$$x = k \left[ t - \frac{2.30753 + 0.27061t}{1 + 0.99229t + 0.04481t^2} \right] \quad (15)$$

where

$$k = 1, \quad t = \sqrt{\ln(1/p^2)} \quad \text{for } p \leq 0.5, \text{ and}$$

$$k = -1, \quad t = \sqrt{\ln\{1/(1-p)^2\}} \quad \text{for } p > 0.5.$$

Once  $x_i$  and  $y_c$  are known, then  $\eta_c$  is determined by Eq. (11), and the conditional probability follows by Eq. (10). Again a useful approximation is provided by an algorithm:<sup>5</sup>

$$P(\eta_c) = l + m \left[ 2 \left( 1 + c_1 \eta_c + c_2 \eta_c^2 + c_3 \eta_c^3 + c_4 \eta_c^4 \right)^4 \right]^{-1} \quad (16)$$

where

$$c_1 = 0.196854,$$

$$c_2 = 0.115194,$$

$$c_3 = 0.000344,$$

$$c_4 = 0.019527,$$

$$l = 0, \quad m = 1 \text{ for } \eta \geq 0, \text{ and}$$

$$l = 1, \quad m = -1 \text{ for } \eta < 0.$$

#### 4. SAMPLE APPLICATION

For the following illustrative application, some readily available data were chosen from another recent study. Records of WSR-57 radar pictures at five stations in the eastern half of the United States, in the midseasonal months of January, April, July and October, taken every 3 hours, that is, eight times per day, were used. The five stations were: (1) Minneapolis, Minnesota, (2) Key West, Florida, (3) Wichita, Kansas, (4) Cape Hatteras, North Carolina, (5) Evansville, Indiana.

5. National Bureau of Standards (1964) Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, Applied Mathematics Series No. 55, for sale by Supt. of Documents, U. S. Govt. Printing Office, Washington, D. C.

Figure 1 is an example of a radar picture. The observer usually sees the negative of pictures like Figure 1, so that radar echoes from rainstorms appear as white clouds on a black background surrounding the radar station. In Figure 1, the outermost circle has a radius of 125 nmi, and the second circle 100 nmi. A large storm is shown centered approximately 85 nmi northeast of the radar station.

#### 4.1 Data Collection

To study each picture, a transparent template was superimposed over it – as shown in Figure 1 – to divide the area between the 25-nmi circle and the 100-nmi circle into 64 cells, each of  $460 \text{ nmi}^2$ . Each of the 64 cells was examined to estimate what fraction of the cell, in eights or oktas<sup>6</sup> was covered by precipitation echo. The coverage was recorded as zero for no coverage, one okta for  $1/8$  coverage, ---, eight oktas for full coverage of the cell. For this exercise the cells were grouped into eight equal-area groups, each group consisting of a ring or part of a ring, and numbered 1 to 8 (Figure 2). A record of the number of oktas (from 0 to 64) in each group was made for each 3-hourly picture. As an example, in Figure 1, the number of oktas in Groups 1 to 8 were estimated, by eye, as 3, 6, 6, 1, 17, 6, 3, 0, respectively.

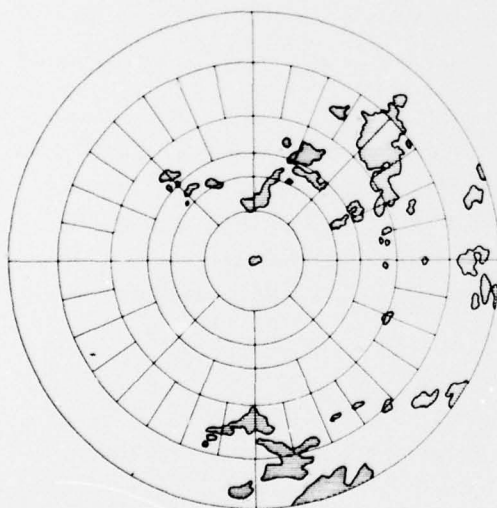


Figure 1. Sample of a PPI Radar-scope Picture of Storms Within 125 nmi of the Radar Station

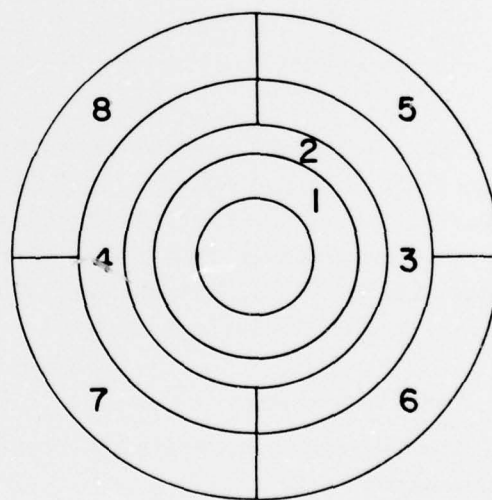


Figure 2. Illustration of the Eight Equal Areas, Each Obtained by the Grouping of Eight Cells (Figure 1)

6. McIntosh, D.H. (1972) Meteorological Glossary, Chemical Publishing, New York.



Table 1 shows an example of the entries of the okta counts in groups 1 to 8. The okta count in group 1 was chosen as the predictand. The okta counts in groups 2 to 8 were chosen as the seven predictors, and were matched with okta counts in group 1 at time lags of zero (for simultaneous coverage), 3 hours, 6 hours, 9 hours, and 12 hours later. Thus, in the example of Table 1, the numbers on the first line (17, 18, 35, 4, 26, 30, 9) were tested as predictors of the okta count (11) on the first line, then of the okta count (40) on the second line, then (12) on the third line, then (7) on the fourth line, and lastly (2) on the fifth line.

Table 1. Sample Data Collection of Radar Coverage of Precipitation. The example is for Minneapolis, Minnesota, 26 October 1970. The entries are the okta counts

Hour	Group							
	1	2	3	4	5	6	7	8
0100	11	17	18	35	4	26	30	9
0400	40	31	35	9	12	18	7	8
0700	12	13	9	9	30	3	1	8
1000	7	7	8	8	19	9	3	18
1300	2	4	17	0	10	3	0	0
1600	0	0	0	0	4	1	0	0
1900	0	0	0	0	0	0	0	0
2200	0	0	0	0	0	0	1	0

For each station the number of radar pictures, in one midseasonal month, was equal to the number of years (6) times the number of days per month (30 to 31) times the number of 3-hourly periods per day (8) for a total of 1440 or 1488 pictures, less a generally small number of pictures missing from the record. Most of the pictures, by far, showed no precipitation echoes, revealing that there was little or no precipitation on most days.

## 1.2 Finding the Correlations

For each station and midseason month, a computer program yielded the following:

- (1) The sample size ( $\leq 1488$ ).
- (2) The number of missing observations, which together with the sample size added up to 1440 or 1488.
- (3) The sample size, in which the seven predictors (groups 2 to 8) of one picture could be matched to the predictand (group 1) at lag times 0, 3, 6, 9, or 12 hours later.

(4) A table of the frequency of the areal coverage in each group (1 to 8) starting with full coverage (64 oktas), then 63 oktas, --- 2 oktas, 1 okta, and lastly none (for example, Table 2).

Table 2. The Frequency of Fractional Cover, in Oktas. The coverage is by radar echoes as in Figure 1. The case is for Minneapolis, Minnesota, January 1969-1974.

Okta Count (L)	Group (i)							
	1	2	3	4	5	6	7	8
64 (full cover)	1	0	0	0	0	0	0	0
63	0	0	0	0	0	0	0	0
62	2	1	0	0	0	0	0	0
4	10	16	8	11	3	4	2	1
3	11	12	14	16	3	5	9	4
2	14	17	28	17	6	8	6	11
1	22	32	31	31	13	12	18	10
0	1217	1231	1288	1298	1395	1388	1384	1390
Total	1428	1428	1428	1428	1428	1428	1428	1428

(5) A table of the cumulative relative frequency of coverage equal to or greater than L oktas in each group, for L = 64, 63, ---, 1, 0 (for example, Table 3). The Blom formula<sup>7</sup> was used to calculate relative frequency.

Table 3. The Cumulative Relative Frequency of Coverage Equal to or Greater Than L Oktas. The example is for Minneapolis, Minnesota, January

Okta Count (L)	Group (i)							
	1	2	3	4	5	6	7	8
64	0.0004	-	-	-	-	-	-	-
63	0.0004	-	-	-	-	-	-	-
62	0.0018	0.0004	-	-	-	-	-	-
4	0.115	0.095	0.047	0.046	0.0074	0.0102	0.0074	0.0088
3	0.122	0.103	0.056	0.057	0.0095	0.0137	0.0137	0.0116
2	0.132	0.115	0.0761	0.0691	0.0137	0.0193	0.0179	0.0193
1	0.148	0.138	0.0978	0.0908	0.0228	0.0277	0.0305	0.0263
*0	0.575	0.569	0.549	0.546	0.512	0.514	0.515	0.514

\*This relative frequency is obtained by adding half of the frequency of no echo to the total number of cases with echoes, and dividing by the overall sample size.

7. Blom, Gunnar (1958) Statistical Estimates and Transformed Beta-Variables, John Wiley and Sons, New York.



(6) A table of the transnormalized variable ( $x_i$ ) corresponding to the okta count (L) in the  $i$ th group. This was obtained by substituting relative frequency for the probability (p) in Eq. (15) (for example, Table 4).

Table 4. The Values of the Transnormalized Variables Corresponding to the Relative Frequencies of Table 3

Okta Count (L)	Group (i)							
	1	2	3	4	5	6	7	8
64	3.33	-	-	-	-	-	-	-
63	3.33	-	-	-	-	-	-	-
62	2.91	3.33	-	-	-	-	-	-
4	1.20	1.31	1.68	1.68	2.44	2.32	2.44	2.37
3	1.16	1.26	1.58	1.58	2.34	2.21	2.21	2.27
2	1.12	1.20	1.43	1.48	2.21	2.07	2.10	2.07
1	1.04	1.09	1.29	1.34	2.00	1.92	1.87	1.94
0	-0.19	-0.17	-0.12	-0.11	-0.03	-0.04	-0.04	-0.03

(7) The matrix of the correlation coefficients ( $\rho_{ij}$ ) between the predictors (for example, Table 5).

Table 5. The Cross-Correlation Coefficients ( $\rho_{ij}$ ) Between the Transnormalized Values of the Fractional Cover or Okta Count in Groups 1 to 8. The example is for Minneapolis, Minnesota, January

Group (j)	Group (i)							
	1	2	3	4	5	6	7	8
1	1.00	0.90	0.72	0.68	0.32	0.44	0.43	0.29
2		1.00	0.81	0.77	0.39	0.49	0.49	0.36
3			1.00	0.63	0.51	0.59	0.42	0.37
4				1.00	0.38	0.41	0.62	0.54
5					1.00	0.28	0.24	0.51
6						1.00	0.44	0.21
7							1.00	0.26
8								1.00

(8) A matrix of correlation coefficients, each row of which consisted of the cc ( $\rho_i$ ) between the predictand (y) and each predictor  $x_i$  ( $i = 2, 8$ ). Table 6 is an example. In each successive row, the predictand is the transnormalized value of the okta count at time 0, 3, 6, 9, 12 hours later.

Table 6. The cc of the Predictand With Each Predictor. The First Line is for the group 1 predictand occurring at the same time as the predictors. The second line is for the predictand occurring 3 hr later, ---, the fifth line is for the predictand occurring 12 hr later. The column headed "1" gives the incidental information of the correlations of group 1 of one time with group 1 of 0, 3, 6, 9, 12 hr later

Predictand Group 1 LAG	Predictor Group (i)							
	1	2	3	4	5	6	7	8
0 hours	1.00	0.90	0.72	0.68	0.32	0.44	0.43	0.29
3	0.67	0.70	0.51	0.67	0.19	0.28	0.40	0.34
6	0.46	0.48	0.35	0.45	0.10	0.16	0.27	0.16
9	0.30	0.33	0.25	0.32	0.08	0.09	0.20	0.12
12	0.18	0.21	0.14	0.20	0.04	0.03	0.14	0.05

#### 4.3 Finding the Partial Regression Coefficients

Given the  $\rho_i$  and  $\rho_{ij}$  for one station, one month and a time lag, the computer solution of Eqs. (8) and (9) yielded values for  $a_2$ , ---,  $a_8$  and  $b$ . Table 7 shows the set of values of  $a_i$  ( $i = 2$ , ---, 8) and  $b$  for time lags 0 to 12 hours, for Key West, Florida, in January. It demonstrates decreasing importance, with time, of group 2 as a predictor of group 1 and the varying importance of the other predictors. Group 8, which is situated northwest of the radar station (Figure 2) is a better predictor of the weather of group 1 than group 2 for time lags 6, 9 and 12 hours. This is easily understood since the weather tends to migrate eastward.

Table 7. The Partial Regression Coefficients of Eq. (11) as Determined for Key West, Florida, January, for Time Lags 0, 3, 6, 9, 12 hr Between Predictors and Predictand

Time Lag (Hours)	$a_2$	$a_3$	$a_4$	$a_5$	$a_6$	$a_7$	$a_8$	$b$
0	0.751	0.071	-0.007	0.079	-0.023	0.031	0.033	0.474
3	0.316	0.139	0.136	-0.038	0.004	0.120	0.197	0.675
6	0.118	0.076	0.102	-0.029	0.082	0.125	0.263	0.809
9	0.010	0.053	0.049	0.000	0.110	0.159	0.219	0.881
12	0.034	0.046	-0.001	-0.004	0.100	0.133	0.159	0.931

This last consideration prompts an examination of the coefficients at one station (Key West) to compare their values from season to season (Table 8). For Key West, Florida, at a time lag of 6 hours, they show a greater importance, in

Table 8. The Coefficients  $a_i$  Divided by  $b$ , Determined for Key West, Florida for Time Lag of 6 hrs, in January, April, July, and October

Midseason Month	$(a_i/b)$							
	i = 2	3	4	5	6	7	8	b
January	0.15	0.09	0.13	-0.04	0.10	0.15	0.32	0.809
April	0.19	0.09	0.15	0.14	-0.06	0.12	0.17	0.832
July	0.13	0.02	0.13	0.11	0.15	-0.03	0.19	0.887
October	0.13	0.18	0.17	0.09	0.12	0.24	0.11	0.770

winter of the condition 55 to 73 miles to the northwest (group 8) than the weather closer to the area of verification (group 2). This is true in summer also, although not as spectacularly. The month of October offers the best opportunity for prediction by conditional probability. In the fall there must be a more prevalent southwesterly flow (group 7) than northwesterly.

#### 4.4 Estimating Conditional Probability

Given coefficients  $a_2, \dots, a_8$  and  $b$ , the model of Eqs. (11) and (10) yields probability estimates. For example, on 3 January 1974, at 0100, at Key West, Florida, the seven predictor values, in oktas, were observed to be 2, 0, 2, 0, 0, 1, 1 oktas, which are transnormalized into  $x = 1.219, -0.453, 1.223, -0.255, -0.371, 0.924, 1.449$ , respectively. The predictand values, at time lags 3, 6, 9, 12 hours later were 1, 18, 25 and 20 oktas respectively, which had (unconditional) frequencies of being equalled or exceeded: 0.16, 0.026, 0.017, 0.022 respectively, and therefore were transnormalized into  $y_0 = 0.977, 1.937, 2.122$  and  $2.006$ , respectively. Inserting these transnormalized values and the coefficients,  $a_i$  ( $i = 2, 8$ ) and  $b$  (Table 7) into Eq. (11), the values of  $\eta_c$  become 0.124, 1.52, 1.865, 1.791 and consequently the conditional probabilities, through Eq. (16), are  $P(\eta_c) = 0.451, 0.064, 0.031, 0.036$ , respectively. As to be expected, the increase of the conditional probabilities over the unconditional probability is demonstrated best on the 3-hour lag.

A priori, Eqs. (11) and (10) are applied to give the conditional probability that certain thresholds of areal coverage are exceeded. As an example, if the problem is to give the probability that group 1 will have 8 oktas or more coverage, of which the unconditional probability is 0.0413, then in the above example for Key West, 3 January 1974 the model estimates of conditional probabilities for 3-, 6-, 9-, 12-hour lags are 0.106, 0.102, 0.077, 0.066, respectively. These are all improved probabilities over the unconditional probability. In verification, there was an echo on all except the 3-hour lag.



A computer program was written to calculate the conditional probabilities of the fractional cover in group 1, successively equal to or greater than 8 oktas, 16 oktas, 32 oktas, and lastly 64 oktas (complete coverage), given the number of oktas in groups 2 to 8 at the initial hour. This was done for time lags of 0, 3, 6, 9, 12 hours, at each of the five stations for each midseasonal month and for each hour for which there was a recorded verification.

The following illustration is for Minneapolis in October. The climatic frequency of the predictand, equal to or exceeding the fractional cover, was:

Full cover (64 oktas) : 0.0011  
 Half cover (32 oktas) : 0.0340  
 One-quarter cover (16): 0.0764  
 One-eighth cover (8) : 0.1236  
 Average: 0.059

The total number of "forecasts", always less than the maximum possible number (4 x 1488), is shown in the second column (Table 9). The number of times that the conditional probability was computed to be: 0.00 to 0.099, 0.100 to 0.199, ---, 0.900 to 0.999 is shown in the next 10 columns. All eleven pieces of information are shown for the forecasts of time lags 0, 3, 6, 9, 12 hours.

Table 9. The Number of Times That the Conditional Probability of Fractional Coverage of 1/8, 1/4, 1/2 or Full Coverage was Estimated in the Range 0.0 to 0.09%, 0.1 to 0.19%, ---. The example is for Minneapolis, Minnesota, October

Time Lag (Hours)	Total No. Fcsts	Conditional Probability Range				
		0 to 0.099	0.1 to 0.199	0.2 to 0.299	0.3 to 0.399	0.4 to 0.499
0	5844	5247	123	87	51	52
3	5780	5169	169	124	66	73
6	5752	5078	312	137	101	61
9	5724	5049	395	164	79	31
12	5692	5071	413	152	47	9
		0.5 to 0.599	0.6 to 0.699	0.7 to 0.799	0.8 to 0.899	0.9 to 0.999
0	5844	50	42	34	51	107
3	5780	49	52	34	28	16
6	5752	36	21	6	0	0
9	5724	6	0	0	0	0
12	5692	0	0	0	0	0

Viewed as forecasts, the conditional probability estimates, by definition, are sharp when they are close to 1.0 or 0.0, implying a high degree of confidence by the forecaster. In actual fact, there are many low estimates, simply because the echo-producing rainstorms are so infrequent. But there are relatively few high conditional probabilities. Table 10 presents the overall relative frequency of prediction, for all five stations and all four midseasonal months. An estimate of 0.70 or higher was never made beyond the 3-hour lag. Even on the zero lag, the degree of certainty of radar echoes in group 1 — when there are echoes in groups 2 to 8 — is not generally high.

Table 10. The Frequency (%) of the Estimates of Conditional Probability of Fractional Coverage (Including 8, 16, 32 and 64 Oktas) Among the Cases When the Conditional Probability was Estimated  $\geq 0.10$ . The figures are averages for all five stations and midseasonal months

Time Lag (Hours)	Conditional Probability Estimate:								
	$\geq 0.10$	$\geq 0.20$	$\geq 0.30$	$\geq 0.40$	$\geq 0.50$	$\geq 0.60$	$\geq 0.70$	$\geq 0.80$	$\geq 0.90$
0	100%	72%	53%	43%	34%	26%	20%	14%	9%
3	100%	57%	37%	23%	14%	9%	4%	2%	1%
6	100%	41%	19%	8%	4%	1%	-	-	-
9	100%	26%	8%	2%	1%	-	-	-	-
12	100%	10%	2%	-	-	-	-	-	-

The sharpness of the conditional probabilities, as a forecast tool, was measured by a statistic related to Brier's score<sup>8</sup> as described below. Let

- (1)  $P(L) = P(\eta \geq \eta_c)$ , the conditional probability that the areal coverage (number of oktas) in group 1 will equal or exceed the number  $L$  ( $= 8, 16, 32, 64$ );
- (2)  $F(L)$  = Climatic frequency or the unconditional probability that the number of oktas in group 1 will equal or exceed the number  $L$ ;
- (3)  $V = 1$  when, on the hour of verification, the number of oktas equals or exceeds the number  $L$ ; and
- (4)  $V = 0$  when, on the hour of verification, the number of oktas is less than the number  $L$ .

8. Brier, G.W., and Allen, R.A. (1951) Verification of weather forecasts, Compendium of Meteorology, American Meteorological Society, Boston, MA., 925-996.



For this test of verification and scoring, L was made successively equal to 8, 16, 32, and 64 corresponding to fractional coverages of 1/8, 1/4, 1/2 and full coverage, respectively.

In the Brier scoring system, if the estimated conditional probabilities are correct, or perfectly valid, then the essence of the test is in the sharpness of the probability,  $P(L)$ . If  $P(L)$  is greater than the climatic frequency,  $F(L)$ , then the difference squared,

$$E_f = [P(L) - V]^2$$

would be smaller than

$$E_s = [F(L) - V]^2 \quad (18)$$

when  $V = 1$  denoting occurrence. Conversely, if  $P(L)$  is smaller, than  $F(L)$ , implying lowered confidence in the occurrence of the event (L), then  $E_f$  would be smaller when the event does not occur.

For all forecasts, totalled over all probability estimates, over all lags and for all times when the forecast is made, the total

$$T = \sum [P(L) - V]^2$$

is intended to be as small as possible. For completely sharp probability statements  $P(L)$  is either equal to 1 or 0, and if completely matched to the values of  $V$ , then  $T$  would be identically zero.

Skill, as apparently first proposed by Heidke,<sup>7</sup> is judged by the advantage of the forecasts over the information in the climatic frequencies. If the symbol  $C$  is used, instead of  $T$ , for the total when climatic frequencies are used as probabilities, then

$$C = \sum [F(L) - V]^2 .$$

The skill ( $S$ ) is ultimately judged by

$$S = (C - T) / C .$$

The statistic  $S$  has the property that it will be equal to 1.0 for perfectly sharp and accurate forecasts, and will be equal to 0.0 for conditional probability statements that do not differ profitably from the unconditional probabilities. It is possible for forecast probabilities to be so much in error that  $T$  will be greater than  $C$ , to

make  $S$  negative. Generally speaking, the conditional probability will be greater than the unconditional probability when there is an improved chance of the event occurring, and smaller for a reduced chance. Hence, the term  $E_f$  will generally be less than the corresponding  $E_c$ . All told, the score  $S$  is best when it is closest to 1.0. A value of 0.0 for  $S$  would imply that the conditional probabilities offer no information over that of the unconditional probabilities. Table 11 lists all of the skill scores for each of the five stations by midseasonal month and lag. As to be expected, the score decreases with time lag. At zero lag, or for simultaneous occurrences, the fractional covers over groups 2 to 8 collectively must be considered as fair to good estimates of the coverage over group 1. But, in 3 hours the predictability decreases rapidly. In 12 hours it almost vanishes, and in fact does at Cape Hatteras, North Carolina.

Table 11. The Skill Scores Which Range From 0 for "No Skill" to 1.0 for Perfectly Correct Forecasts. The stations are (1) Minneapolis, Minnesota, (2) Key West, Florida, (3) Wichita, Kansas, (4) Cape Hatteras, North Carolina, and (5) Evansville, Indiana

Time Lag (Hours)	January					April				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
0	0.63	0.65	0.66	0.76	0.73	0.71	0.54	0.58	0.65	0.70
3	0.35	0.40	0.43	0.50	0.40	0.41	0.33	0.32	0.40	0.35
6	0.14	0.19	0.20	0.31	0.25	0.20	0.11	0.13	0.18	0.11
9	0.07	0.08	0.10	0.14	0.14	0.11	0.06	0.03	0.10	0.05
12	0.03	0.03	0.04	0.00	0.08	0.07	0.02	0.02	0.00	0.03
	July					October				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
0	0.64	0.42	0.53	0.56	0.52	0.71	0.55	0.60	0.59	0.72
3	0.32	0.10	0.27	0.27	0.27	0.44	0.32	0.35	0.32	0.48
6	0.06	0.02	0.09	0.10	0.08	0.23	0.20	0.17	0.17	0.27
9	0.00	0.01	0.03	0.05	0.04	0.12	0.13	0.04	0.10	0.12
12	0.00	0.04	0.01	0.00	0.02	0.08	0.06	0.02	0.00	0.07

## 5. SUMMARY AND CONCLUSIONS

The conditional probability of  $Y$  equal to or greater than a threshold  $Y_c$ , given several or many predictors  $X_1, \dots, X_n$ , is determined from the transnormalized variables. Symbolically

$$P(Y \geq Y_c \mid X_1, \dots, X_n) = P(y \geq y_c \mid x_1, \dots, x_n) \\ = P(\eta \geq \eta_c)$$

where  $y_c$ ,  $x_1$ , ...,  $x_n$  are the transnormalized variables obtained from their respective climatic probabilities ( $p$ ) through Eq. (14) or its approximation Eq. (15).

The value of  $\eta_c$  is obtained through Eq. (11). If the partial regression coefficients  $a_i$  ( $i = 1, \dots, n$ ) and their function  $b$  are not provided, they are determinable in terms of the cross-correlation coefficients between the predictors ( $\rho_{ij}$ ) and the cc's between predictand and predictors ( $\rho_i$ ) through Eqs. (8) and (9). The approximation for the conditional probability, good to four decimals, is given by Eq. (16).

The model was tested on some readily available data, consisting of radar-echo coverage around each of 5 stations in the eastern half of the United States. The tests herein for verification provided no measure of the degree of validity of the conditional probability estimates. But there is strong evidence, in Table 11, that they are satisfactorily valid, because the skill scores fall to zero but do not fall below zero.

Assuming validity, the Brier scoring system provides a measure of sharpness. Table 10 displays a low measure, in general. But the model is capable of estimating high probabilities as evidenced by the frequencies in the first two lines of Table 10.



## References

1. Gringorten, I.I. (1972) Conditional probability for an exact noncategorized initial condition, Monthly Weather Review 100:796-798.
2. Hohn, Franz E. (1964) Elementary Matrix Algebra, The MacMillan Company, New York.
3. Anderson, T.W. (1958) An Introduction to Multivariate Statistical Analysis, John Wiley and Sons, New York.
4. Dixon, W.J. (1965) BMD Biomedical Computer Programs, University of California at Los Angeles, Health Sciences Facility, Los Angeles.
5. National Bureau of Standards (1964) Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, Applied Mathematics Series No. 55, for sale by Supt. of Documents, U. S. Govt. Printing Office, Washington, D.C.
6. McIntosh, D.H. (1972) Meteorological Glossary, Chemical Publishing, New York.
7. Blom, Gunnar (1958) Statistical Estimates and Transformed Beta-Variables, John Wiley and Sons, New York.
8. Brier, G.W., and Allen, R.A. (1951) Verification of weather forecasts, Compendium of Meteorology, American Meteorological Society, Boston, MA., 925-996.