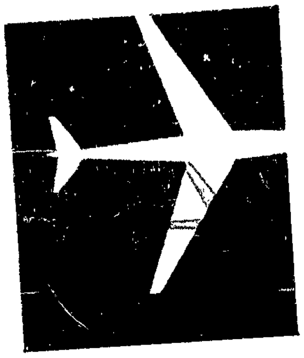
N-63-4-4

Contract FAA/BRD-363
Tech. Publication 16



This report has been approved for general distribution.

[illegible]

PRECIPITATION

THE TRAVELERS RESEARCH CENTER INC.

[illegible]

Contract FAA/BRD-363
Tech. Publication 16

OBJECTIVE ANALYSIS OF PRECIPITATION OCCURRENCE AND TYPE

This report has been approved for general distribution.

Frederick P. Ostby, Jr.

August 1962
7044-29

This report has been prepared by The Travelers Research Center, Inc., for the Aviation Research and Development Service, formerly the Bureau of Research and Development, Federal Aviation Agency, under Contract FAA/BRD-363. The content of this report reflects the views of the contractor, who is responsible for the facts and accuracy of the data presented herein. The content of this report does not necessarily reflect the official policy of the Federal Aviation Agency.

THE TRAVELERS RESEARCH CENTER INC.
650 MAIN STREET HARTFORD 3, CONNECTICUT

ABSTRACT

An objective technique was developed to analyze precipitation occurrence and type. A weighted-averaging method was used, in which the contribution of a station observation to the analysis is a function of its distance from the gridpoint. A square area centered on a gridpoint was searched to locate stations to be used in the analysis. Developmental testing on 49 hr of data from January 1961 and 72 hr from September 1960 over the eastern half of the United States showed that the method correctly specified whether precipitation was occurring for more than 90% of the analysis area.

TABLE OF CONTENTS

1.0 INTRODUCTION

Detailed analysis of precipitation occurrence requires a dense data network to depict the mesoscale character of precipitation. Although with conventional data it is not possible to capture individual precipitation cells over a given region, it is possible to produce a reasonable analysis of precipitation occurrence on a scale comparable to the hourly airways network. To go below this scale requires either the use of mesonet data or the incorporation of radar information. The problem of combining ground observations of precipitation and radar information has been discussed previously [2]. While such a synthesis is desirable, this report will deal with objective techniques based solely on ground observations.

An objective analysis of precipitation is useful for display purposes and as possible input to a prognosis of precipitation. Futhermore, although precipitation is not considered a critical in-flight weather problem, precipitation may imply related information, such as presence of middle clouds, icing, and turbulence, which are difficult to specify directly.

2.0 THE ANALYSIS TECHNIQUE

2.1 Analysis of Precipitation Occurrence

Determining the optimum distance between gridpoints is an important consideration in objective analysis procedures. Clearly, this is a function of both the scale of the phenomenon to be analyzed and the density of the data. The density of hourly airways data over the United States does not permit the analysis of individual precipitation cells. The grid used for the development of precipitation analysis was derived from a portion of the 1,977-point octagonal grid employed by the Joint Numerical Prediction Unit (JNWP). The grid spacing that appeared to be most consistent with the available data was one-quarter of the basic JNWP grid spacing (roughly 50 mi between gridpoints). This was the grid chosen for precipitation analysis and is shown in Fig. 1. Undoubtedly, the data density in the vicinity of major air terminals (such as New York and Washington) would warrant an even finer grid.

The analysis methods tested were weighted-averaging techniques, in which observations surrounding a gridpoint are weighted according to their distance from that gridpoint. The form of the observation for precipitation occurrence is obviously not on a continuous scale but rather takes on discrete values of "yes" or "no." It was decided to code observations as 100 for occurrence and 0 for non-occurrence.

In the first method, the analyzed value of precipitation occurrence, A, at a gridpoint is given by

$$A = \frac{\sum W\phi}{\sum W}, \tag{2-1}$$

where ϕ is the coded station observation of 100 or 0 and W is a distance weighting function:

$$W = [1 + (ad^2)^b]^{-1}. \tag{2-2}$$

Here, d is the distance in grid units between the station and the gridpoint, and a and

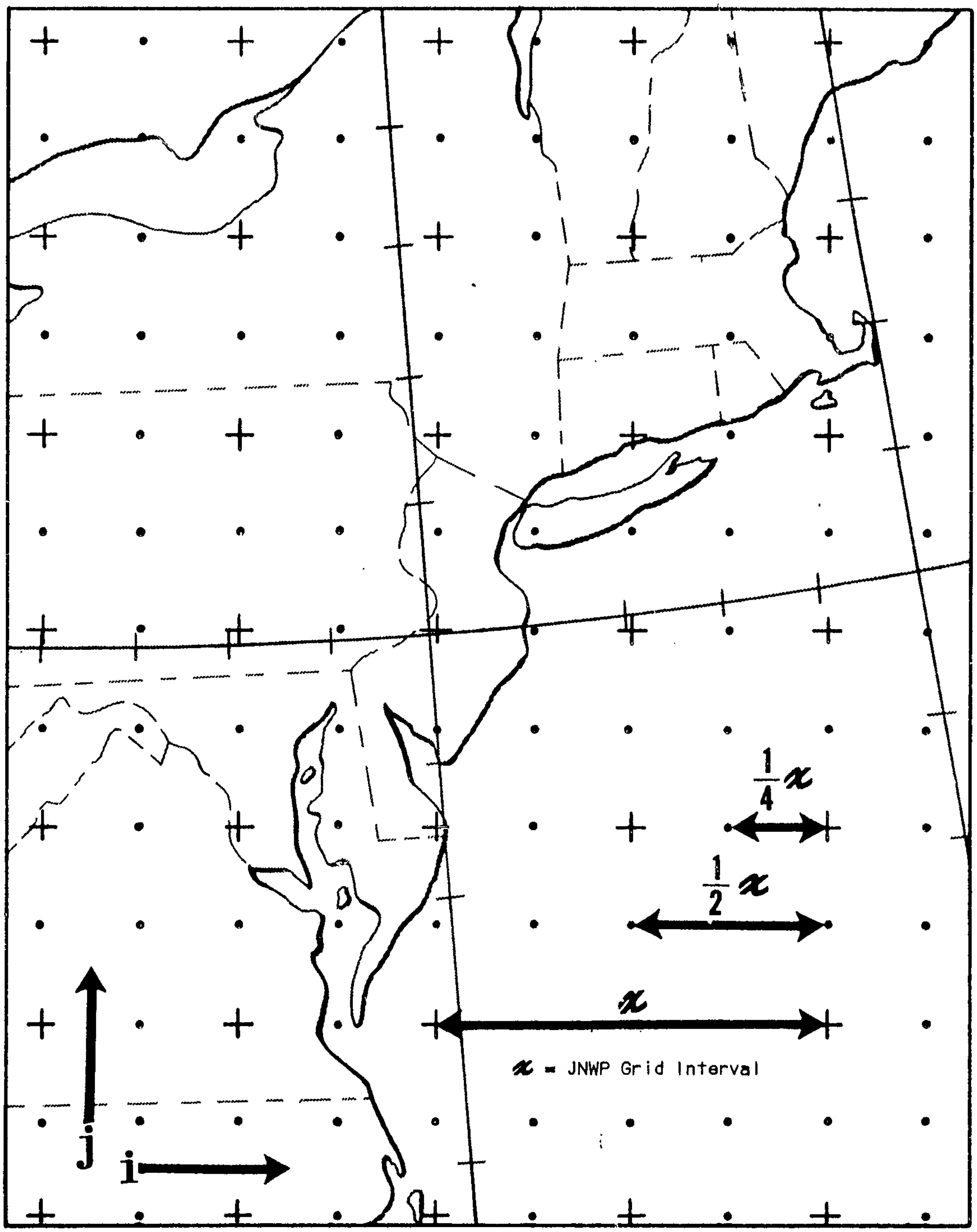


Fig. 1. Joint Numerical Weather Prediction Unit grid.

b are assigned constants. When $d = 0$ (the station is at the gridpoint), W is obviously equal to one. The manner in which W decreases with increasing d is controlled by the parameters a and b. The relationships among W, d, a, and b are illustrated in Fig. 2.

Solution of Eq. (2-1) results in analysis values at gridpoints where $0 \leq A \leq 100$. This kind of analysis yields at least two possible interpretations. One interpretation is that the dividing line between occurrence and non-occurrence of precipitation is at $A = 50$ and that all gridpoints having values of $A \geq 50$ are regarded as locations of precipitation occurrence.

The other interpretation results from taking a simple nonweighted average. This can be done by assigning the value 0 to either a or b in Eq. (2-2), which makes W constant. Equation (2-1) then becomes

$$A = \frac{\sum \phi}{N_s}, \qquad (2\text{-}3)$$

where N_s is the number of stations in an area surrounding the gridpoint. Since ϕ is either 100 or 0, it follows that for constant W,

$$A = 100 \frac{N_o}{N_s}, \qquad (2\text{-}4)$$

where N_o is the number of stations reporting precipitation in the area. Such an analysis gives the percentage of stations within an area that are reporting precipitation, and this could be interpreted as the percentage of an area in which precipitation is occurring.

In addition to the assignment of a pair of values for a and b, consideration must be given to the size of the area in which observations are allowed to affect the analysis at the gridpoint. Too large an influence area would tend to smooth out the analysis and would result in loss of detail, whereas too small an influence area could cause a loss of representativeness by not considering enough observations.

To take both of these factors into account, the concept of a variable search area was adopted. This procedure begins by searching for observations in a

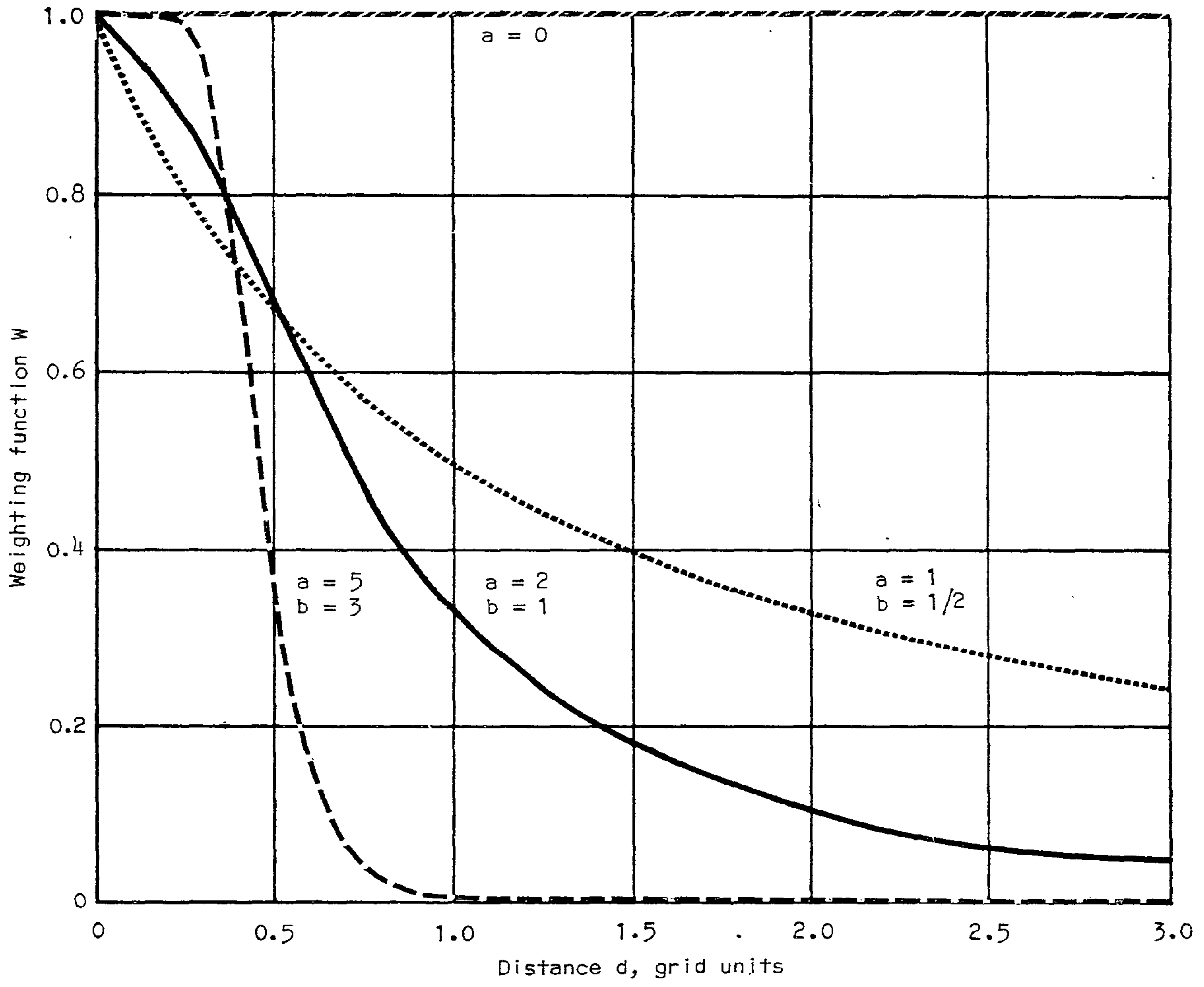


Fig. 2. Relationships among distance d, weighting function W, and constants a and b.

small square of L_1 grid units on a side, centered on the gridpoint for which an analysis value is being computed. If enough observations, N_{min}, fall within this square, an analyzed value is computed from Eq. (2-1). If there are fewer than N_{min} observations, the size of the search square is expanded to L_2 grid units on a side. If now at least N_{min} observations are found within the square, the analysis is performed using all of them. If the L_2-square contains too few observations, the final expansion (to L_3) is made and, if there are now any observations at all, the analysis is made. No search is made in an area larger than L_3. If no data are found, a generated value is assigned to the gridpoint. The use of a stepwise expansion of the search area permits the analysis value at a gridpoint in a dense-data area to be determined by only those observations in the immediate vicinity of the gridpoint.

In analyzing for a gridpoint, it is not uncommon to find some observations clustered in a small area while another observation also affecting the same gridpoint is relatively isolated. This kind of uneven distribution, if unaccounted for, can distort the analysis by allowing the observations in the dense-data region to unrealistically outweigh the sparser data.

The second analysis method, a modification of the first, was formulated to account for nonuniform observational distribution. This technique uses a density factor ρ, which is proportional to the number of stations in the neighborhood of each station. If each station is weighted according to this density approximation as well as to its distance from the gridpoint, Eq. (2-1) becomes

$$A = \frac{\sum W\,\phi\rho^{-1}}{\sum W\rho^{-1}}\,. \tag{2-5}$$

For this approach, the value of ρ is computed by preprocessing programs.

Another analysis problem concerns indeterminate gridpoints. These arise when no data surround a gridpoint, and, if no provision is made, they result in an analysis with missing gridpoint values. Since values at all gridpoints are necessary for verification and since the analysis may be used for prognostic or diagnostic programs, two methods for generating data for indeterminate gridpoints were formulated. One (space extrapolation) assigns a value to an indeterminate gridpoint by taking the average value of neighboring determinate gridpoints; the other (time persistence) assigns the value used in the previous analysis.

Analyses of precipitation occurrence are presented in Fig. 3.

2.2 Analysis of Precipitation Type

The analysis of precipitation type is like the analysis of precipitation occurrence except that the former excludes observations of no precipitation. The observation is coded as 100 if the precipitation is frozen and 0 if it is liquid. The solution of Eq. (2-1) and the stepwise expansion procedure are the same. The dividing line between frozen and liquid is at $A = 50$. Note that the analyses of occurrence and type are performed independently; however, the two analyses may be superimposed to depict the final occurrence and type analysis, as shown in Figs. 4 through 6.

2.3 Verification

Proper comparison of different analysis techniques requires an objective verification procedure that yields a representative error statistic. One suitable method is the areal-mean-error method of analysis verification [3]. In this method, some of the observations (generally 10%) are set aside for verification of the analysis. The verification is accomplished by comparing the observed value (at both analysis and withheld stations) with the corresponding analysis value (interpolated from the surrounding gridpoints). A verification statistic based solely on analysis-station errors is likely to exhibit a bias toward smaller errors. Since the analysis technique tends to force the analysis to fit the station observations, the errors in the vicinity of these stations are generally smaller than in regions between observations. To obtain a more realistic error estimate, it is necessary to sample the errors in these between-observation regions and

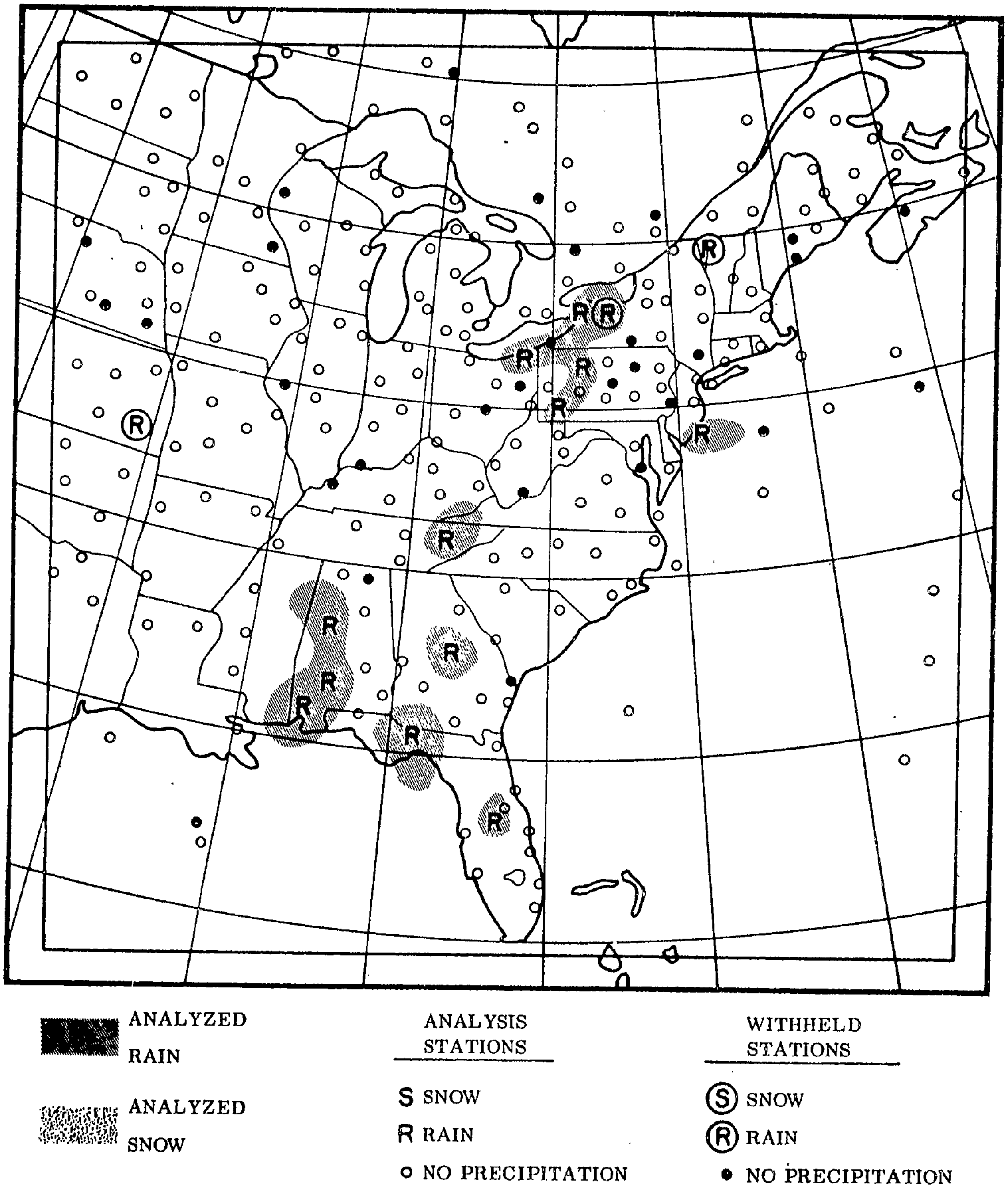


Fig. 3. Objective analysis of precipitation occurrence. 0000Z 16 September 1960 (test 13, Table I).

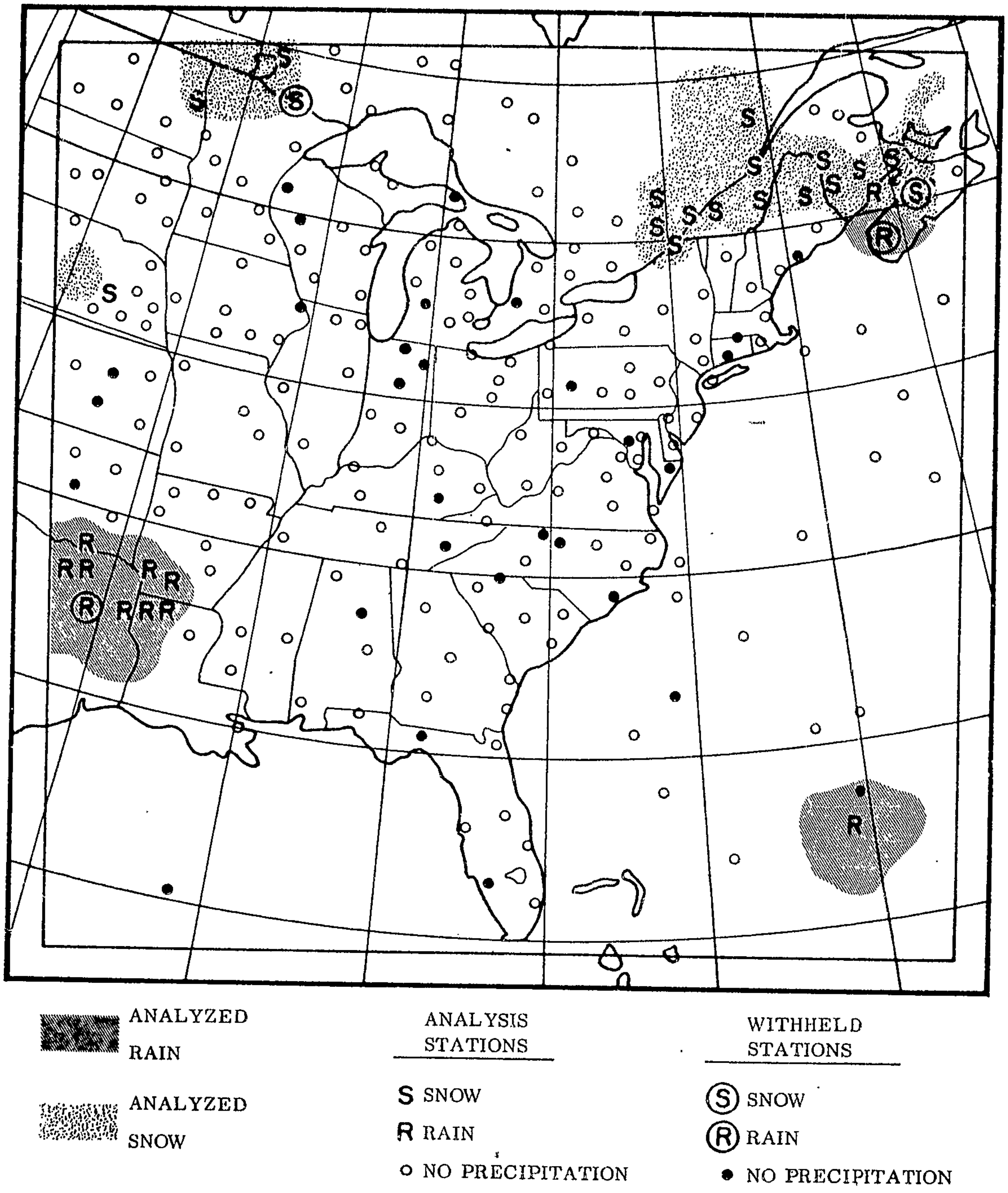


Fig. 4. Objective analysis of precipitation occurrence and type. 0000Z 7 January 1961 (test 10, Table I).

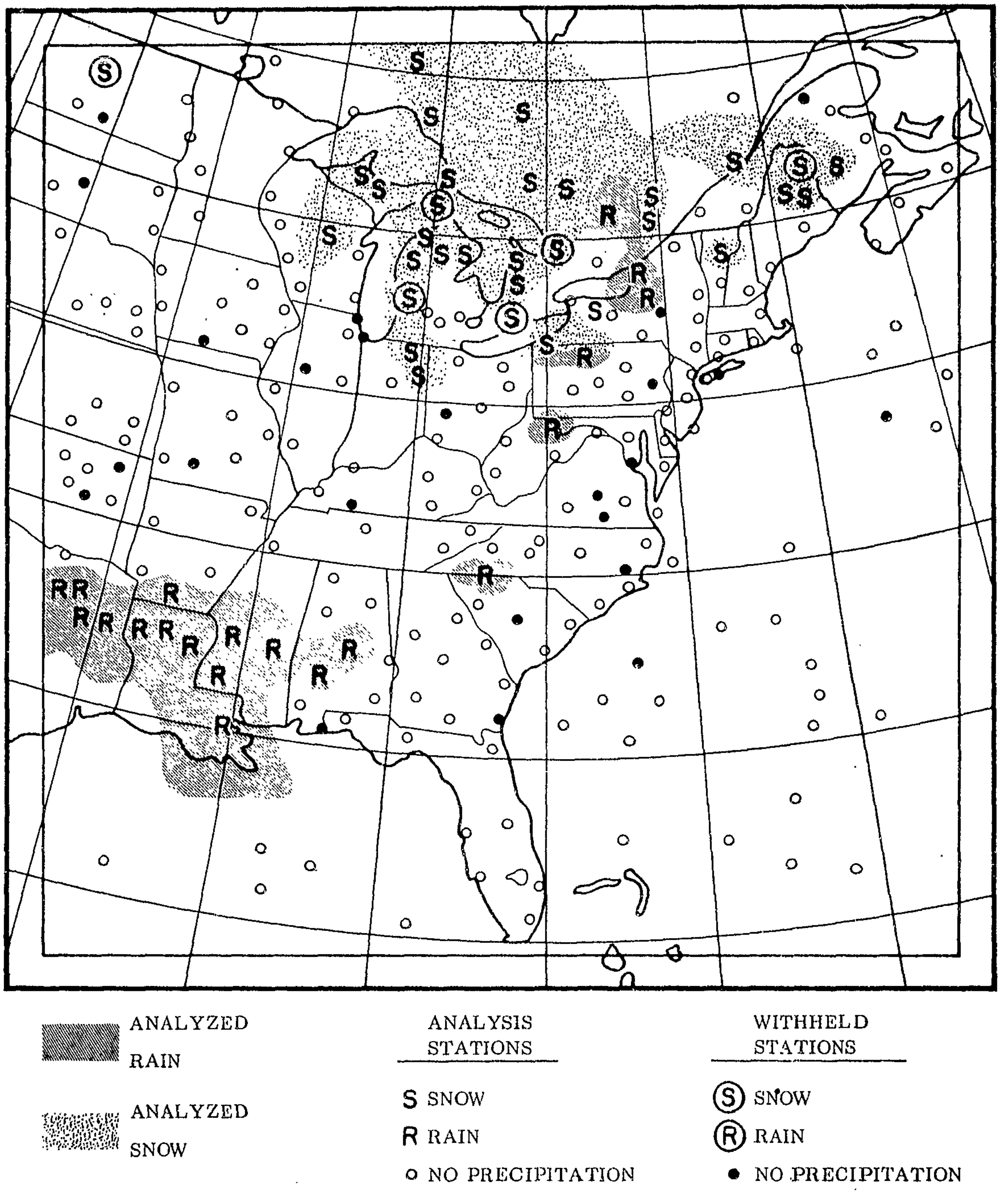


Fig. 5. Objective analysis of precipitation occurrence and type. 0000Z 8 January 1961 (test 10, Table I).

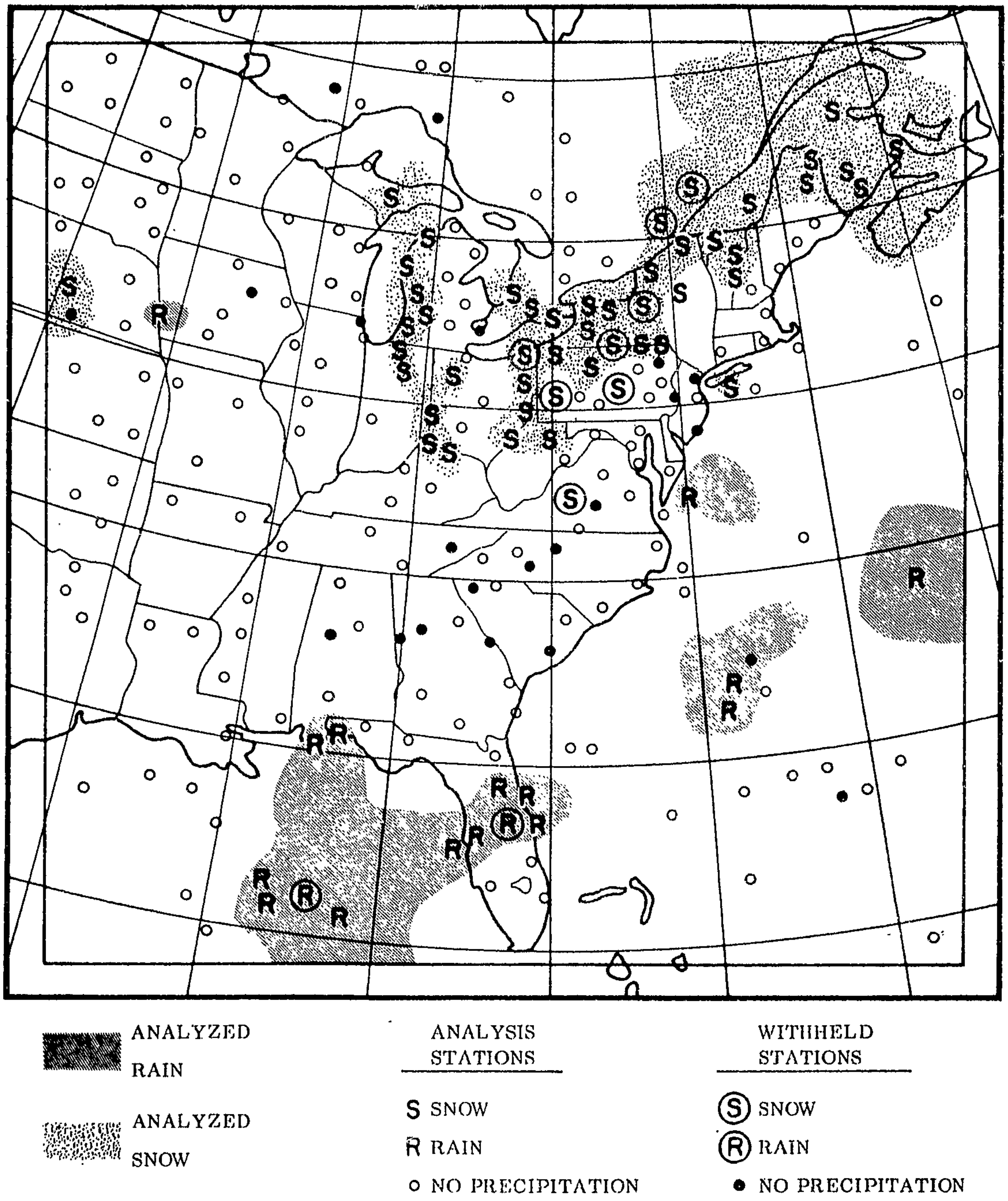


Fig. 6. Objective analysis of precipitation occurrence and type. 0000Z 9 January 1961 (test 10, Table I).

incorporate them in the verification statistic. The computation of errors at withheld stations and their combination with analysis-station errors was performed to provide a representative error estimate.

An unrepresentative verification statistic may also be produced by giving each station error equal weight. Variable data density causes the errors in the dense-data regions to contribute more to an over-all score than the errors from sparse-data regions. Since analysis errors tend to be smaller in dense-data regions, the result is a bias toward smaller errors. The areal-mean-error method compensates for variable data density by weighting each station error by an approximation of the area represented by that station. The appropriate area approximation is given by the relationship

$$\text{Area} = C\rho^{-1},$$

where C is a constant (assumed equal to 1) and ρ is a measure of the density of reporting stations, as in Eq. (2-5). Summation over all stations of the areal error estimate yields an estimate of the analysis error for the entire map, rather than for discrete points. Since this verification method deals in terms of area, the elements in the contingency tables that it generates show the percentage of the analysis area within each category. The appendix contains an example of a contingency table.

2.4 Analysis Data

Data for development testing came from two separate hourly airways data collections. These collections were for September 15 through 17, 1960 (72 hr), and January 7 through 9, 1961 (49 hr), and covered approximately the eastern half of the United States. The collections provided, on the average, about 300 observations per hour. Figures 3 through 6 illustrate analyses of the precipitation occurring during the times represented by these collections.

Raw data were processed from punched teletype paper tape by an input data-handling program [1] that sorted, edited, and converted the observations into a fixed format compatible with high-speed computations. Further preprocessing followed to tailor the data to the specific analysis methods used.

3.0 RESULTS

Developmental tests were made to determine the following.

(1) best combination of a and b to use in Eq. (2-2),

(2) best values for N_{min}, the desired minimum number of stations to compute a gridpoint value,

(3) best set of values L_1, L_2, and L_3, the search area,

(4) possible improvement by taking into account variable station distribution (data density), and

(5) best method of data generation for indeterminate values.

Some of the more important results are shown in Table 1. A discussion of the effect of the individual variables involved in an analysis follows.

3.1 The Weighting Function W

Many combinations of a and b were tested on the 49- and 72-hr samples, a few of which are shown in Table 1 (tests 1, 2, 4, 5, 6, 13, 14, and 15). For both data samples, the best results were obtained with $a = 5$ and $b = 3$. The poorest results came with $a = 0$ (i.e., all observations were given equal weight, as in test 1).

The net difference, however, was not very striking. On the 49-hr sample, for example, $a = 5$ and $b = 3$ gave a score of 93% correct, and $a = 0$ gave a score of 90%. For the 72-hr sample, $a = 5$ and $b = 3$ gave a score of 96% and $a = 0$ gave 94%.

3.2 Minimum Number of Observations

One, 2, 3, and 4 were tested as values of N_{min} in tests 6 through 9. Variation within the range $1 \le N_{min} \le 4$ did not cause any significant difference in results.

3.3 Size of the Search Area

Only two sets of values for the variable side length L were tested (tests 10 and 11). The first ($L_1 = 2$, $L_2 = 3$, and $L_3 = 4$ grid units) gave slightly better scores than the second ($L_1 = 1$, $L_2 = 3$, and $L_3 = 5$). Values of 2, 3, and 4 are roughly equivalent to side lengths of 100, 150, and 200 mi, respectively.

TABLE 1
DEVELOPMENTAL TEST RESULTS

Test	Data	Hours	Wt. function const.		N_{min}	Analysis score, %	
			a	b		Occurrence	Type
1	Jan 61	49	0	2	3	90.2	94.7
2	Jan 61	49	4	2	3	93.0	94.6
3*	Jan 61	49	4	2	3	93.0	94.8
4	Jan 61	49	1	0.5	3	91.6	—
5	Jan 61	49	6	0.5	3	92.0	—
6	Jan 61	49	5	3	4	93.1	—
7	Jan 61	49	5	3	3	93.1	—
8	Jan 61	49	5	3	2	93.1	—
9	Jan 61	49	5	3	1	92.8	—
10†	Jan 61	49	5	3	3	93.2	95.4
11†‡	Jan 61	49	5	3	3	93.0	94.0
12	Sep 60	72	5	3	3	95.8	—
13†	Sep 60	72	5	3	3	96.0	—
14	Sep 60	72	0	2	3	94.2	—
15	Sep 60	72	4	2	3	95.8	—

*Test 3 used Eq. (2-5); all others used Eq. (2-1).

†Tests 10, 11, and 13 used time persistence at indeterminate gridpoints; all others used space extrapolation.

‡Test 11 used a search area of 1, 3, then 5 grid units on a side; all others used 2, 3, then 4.

3.4 Variable Station Density

The variable-station-density factor, although it improved the analysis at a few gridpoints, did not have a profound effect on the map as a whole. Comparison of tests 2 and 3 shows a negligible difference in over-all percent-correct scores. However, this factor may be useful if the arrangement of observing stations is critical.

3.5 Data Generation for Indeterminate Values

When analyzing on a 1-hr cycle, testing showed time persistence to be slightly better than space extrapolation for generating data at indeterminate gridpoints (tests 7, 10, 12, and 13). Presumably, as the time cycle between analyses increases, the advantage of time persistence over space extrapolation disappears. The tests conducted did not determine at what point this occurs.

4.0 SUMMARY

Many parameters were tested in computing analyses, but no significantly best set appeared. The set that yielded the highest scores consisted of a = 5 and b = 3 in Eq. (2-2); $N_{min} = 3$; $L_1 = 2$, $L_2 = 3$, and $L_3 = 4$ grid units; variable station density omitted; and time persistence used at indeterminate gridpoints. Since these values are based on only 121 hr of data, testing on a much larger set of data should precede operational use.

5.0 ACKNOWLEDGEMENT

Required computer programming was done by the United Aircraft Research Laboratories. In particular, the author wishes to acknowledge the work of Mr. Peter Vaughn, who programmed the analysis technique.

6.0 REFERENCES

1. Marx, H. P., and R. M. Shroyer, The Development of Computer Input Data Handling Systems for the Processing of Hourly and Other Operational Weather Data, Tech. Memo, 6, Travelers Research Center, Inc., Hartford, July 1961.

2. Reed, R. J., A Plan for the Integration of Radar and Routine Data in the Objective Analysis of Cloud, Precipitation, and Convective Activity, TRC-20, Travelers Research Center, Inc., Hartford, December 1961.

3. Thomasell, A. L., The Areal-mean-error Method of Analysis Verification, Tech. Rpt. 7083-12, Travelers Research Center, Inc., Hartford, June 1962.

APPENDIX, EXAMPLE OF POOLED CONTINGENCY TABLES

Contingency tables are prepared for each analysis (hour) with a pooled table computed to summarize the complete test. The pooled table is an average of the individual hourly tables and provides one statistic for a series of analyses. Table A-1 is the pooled contingency table for 49 hr of data from January 1961, and Table A-2 is for 72 hr from September 1960. The row and column totals are 100, which refers to 100% of the analysis area. The sum of the main diagonal is the percentage of the analysis area correctly specified as to the occurrence or nonoccurrence (or type) of precipitation. The analyses of occurrence and type that these tables represent were computed by Eq. (2-1); $a = 5$ and $b = 3$ in Eq. (2-2); $N_{min} = 3$; search area equal to approximately 100, 150, then 200 mi; and time persistence used at indeterminate gridpoints. These conditions correspond to tests 10 and 13 in Table 1.

TABLE A-1
CONTINGENCY TABLE*

Analyzed	Observed precipitation occurrence, %		Total, %
	No precip.	Precip.	
No precip.	77.11	3.20	80.31
Precip.	3.57	16.12	19.69
Total	80.68	19.32	100.00
Percent correct = 93.23			

Analyzed	Observed precipitation type, %		Total, %
	Liquid	Frozen	
Liquid	33.86	1.88	35.74
Frozen	2.75	61.51	64.26
Total	36.61	63.39	100.00
Percent correct = 95.37			

*Test 10, Table 1. January 7/0000Z through 9/0000Z, 1961.

TABLE A-2
CONTINGENCY TABLE*

Analyzed	Observed precipitation occurrence, %		Total, %
	No precip.	Precip.	
No precip.	90.31	2.32	92.63
Precip.	1.68	5.69	7.37
Total	91.99	8.01	100.00
Percent correct = 96.00			

*Test 13, Table 1. September 15/0000Z through 17/2300Z, 1960.