

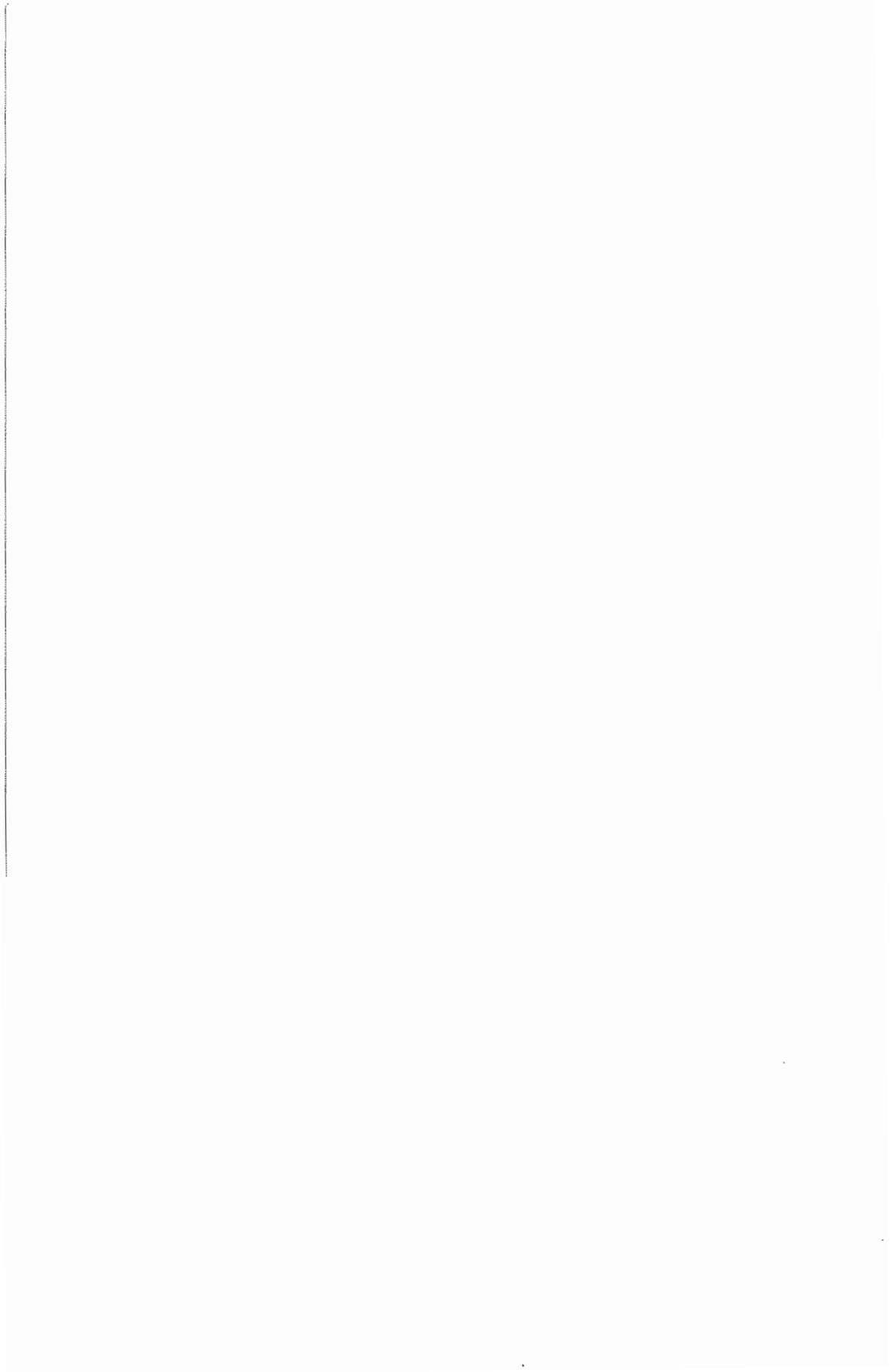
U. S. DEPARTMENT OF COMMERCE
Environmental Science Services Administration
Weather Bureau

ESSA Technical Memorandum WBTM SR-48

ON COMPARATIVE RATING OF FORECASTERS

SOUTHERN REGION HEADQUARTERS
SCIENTIFIC SERVICES DIVISION
FORT WORTH, TEXAS
October 1969





ON COMPARATIVE RATING OF FORECASTERS

by

George T. Gregg

WBFO, Albuquerque, New Mexico

For as long as weather forecasts have been made, verification systems have been proposed, tried, touted, disparaged, and, in most cases, abandoned. Inside the framework of these various systems attempts have as frequently been made to comparatively rate individual forecasters, i.e., to show that forecaster Adams is demonstrably superior in his results to forecaster Brown, or Brown to Clarke, etc. Most of these efforts have been indecisive at best.

This was particularly true in the field service of the Weather Bureau in the era preceding the mid-1960's, at least to this writer's observation. For both the district (public) forecasts and aviation terminal forecast verification systems, scores for individuals firmly established only one positive fact: who worked the greatest number of bad-weather situations.

The Bureau-wide advent of precipitation probability forecasting around 1966 offered hope for a more precise scoring system in all respects. Certainly, it seemed, the skill of a probability forecast of precipitation could be more sharply measured and evaluated than a simple categorical declaration of rain or no rain. At about the same time a Bureau-wide policy of evaluating temperature forecasts on an "absolute error" basis with well defined time periods was instituted in marked contrast to the previous overlapping categories of "Warmer", "Colder", and "No Change" with seasonally varying definitions of the categories. Both of these departurés, as well as other desirable features were incorporated in the "FP/NMC verification program". (1)

Preliminary indications from this verification system have been described by several writers, notably Roberts (2) for the conterminous 48 states, and Hughes (3), Dickey (5), and others for intra-region features. Hughes has also described conclusions which he draws for individual forecasters. (4)

At the Albuquerque WBFO rather complete records have been maintained and regularly reviewed since the official beginning of precipitation probability forecasting, December 1965, and the beginning of the FP/NMC comparison (precipitation and temperature) in March 1966. These records through October 1968, with one exception, are the basis for this paper. Moreover, through this period the Albuquerque WBFO has been fortunate in maintaining a relatively stable cadre of experienced, well-trained forecasters in the intraoffice unit responsible for area (i.e., state or general) forecasts. Personnel flux through the unit has been slight with the result that for most of the period examined eight individual have created 88% of the data for verification. They will be referred to below only as Forecaster A, Forecaster B, etc. with the order of letters having no significance beyond convenience and aimless selection. The five verification stations were Albuquerque, Clayton, and Roswell, New Mexico, and Flagstaff and Phoenix, Arizona.

2.

Obviously, several factors must be considered in evolving ratings equitable to each individual, the most important being an assurance that either meteorological difficulties confronting each individual are reduced to a common base or that difficulty and score are considered together. As Roberts (2) and others have shown, precipitation skill scores increase regularly with precipitation frequency and temperature forecast errors with temperature variability. For the Albuquerque forecast district, precipitation frequency is much greater in midsummer than through the cooler and drier seasons; and, conversely, day-to-day temperature variability is minimal through summer and greatest through winter. Furthermore, as we demonstrate below, rotational scheduling of individuals will equalize difficulties only over a very long time span, certainly longer than the approximate three years of data here examined. Even more obviously, geographical differences in difficulty must be considered in comparative rating of scores of individuals at different facilities and in disparate climatological regimes. Of course, this last consideration was not a factor in the data examined herein. Suggestions are offered below to control or evaluate the other variables...

PRECIPITATION FORECASTING RESULTS

The obvious basis for individual forecaster ranking is the average Brier score he achieved for the forecasts he prepared. (The score referred to here and throughout this paper is actually the "abbreviated" Brier score, one-half the total Brier score.) Hence, the eight men are ranked below according to their raw monthly scores averaged for the 34 months of data. The second column lists the percentage of the total number of forecasts made by the individual.

Defining a "forecast" as a probability specification for one station for one period and considering five stations and 34 months of data, there were in excess of 30,000 individual forecasts verified. Actually, it was considerably in excess of 30,000 since for about two-thirds of the months four editions of forecasts per day were verified (two editions per day for the last ten months). The data sample then was substantial, individually and collectively.

<u>Forecaster</u>	<u>Percentage</u>	<u>Brier Score</u>	<u>Frequency</u>
H	7.8	.063	.072
D	14.2	.067	.079
C	8.6	.070	.084
A	6.1	.070	.088
E	7.8	.077	.090
G	11.2	.078	.093
F	15.7	.080	.100
B	16.5	.080	.101
	<u>87.9</u>		

Thus, it is seen that average scores range from .063 for Forecaster H to .080 for Forecasters F and B. One might hastily conclude that the efforts of H were about 21% better than those of F and B, a not unreasonable figure. But this ranking took no account of difficulty, that is, of "threat". Attention is directed to the final column of the table, "Frequency". Entries in this column are the frequency of precipitation which was observed for the subset of forecasts made by each individual. It is no coincidence that this figure for frequency increases as the score increases, or, rather, that the magnitude of the score is directly proportional to the frequency. Hence, Forecasters B and F might, with reason, contend that they were confronted with situations 29% more difficult than Forecaster H; no wonder their scores were 21% worse!

In attempting to compensate for this direct relation between increasing rain frequency and increasing (worsening) scores, we have practiced the following procedure:

For each month Brier Scores and precipitation frequencies are computed for each of the five verification stations. These average scores and frequencies are treated as ordered pairs of numbers and plotted on coordinate axes (dots) as illustrated in Figure 1 and a least-squares regression line computed and drawn. Figure 1 illustrates data for May 1968, a reasonably typical example. The high correlation coefficient (.97) is not unusual. With rare exceptions values ranged above .8 for this relation and frequently above .9. (As demonstrated by Dickey (5), this relation is actually parabolic, not linear. However, the deviation from linearity for the range of values under consideration is assumed negligible.)

Two important and useful parameters are noted for the regression line for each month: the slope of the line (rate of rise of score with frequency of events) and the value of the y-intercept (the point at which the line crosses the zero-frequency value). This last value may be interpreted as a "caution error", related to the degree of threat the forecasters consider to exist even if precipitation never occurs.

To compensate for seasonal influences, a normalized score is computed from the regression parameters. This is arbitrarily defined as the score value corresponding to a precipitation frequency of .20. This value is designated \bar{BS} on Figure 1 and is beyond the actual range of scores for that particular month.

The regression line on the frequency-Brier score plane represents an average for the group effort for the period under consideration, usually one month. Any point located below the line would represent a superior forecast effort--Brier score less than the average score for the given frequency. Correspondingly, a point above the line would exhibit less successful forecasts than the average.

So, after the parameters of this relation have been ascertained for the equally separated group effort for the month, a new partition of the month's data is made, this time separating scores and frequencies for

4.

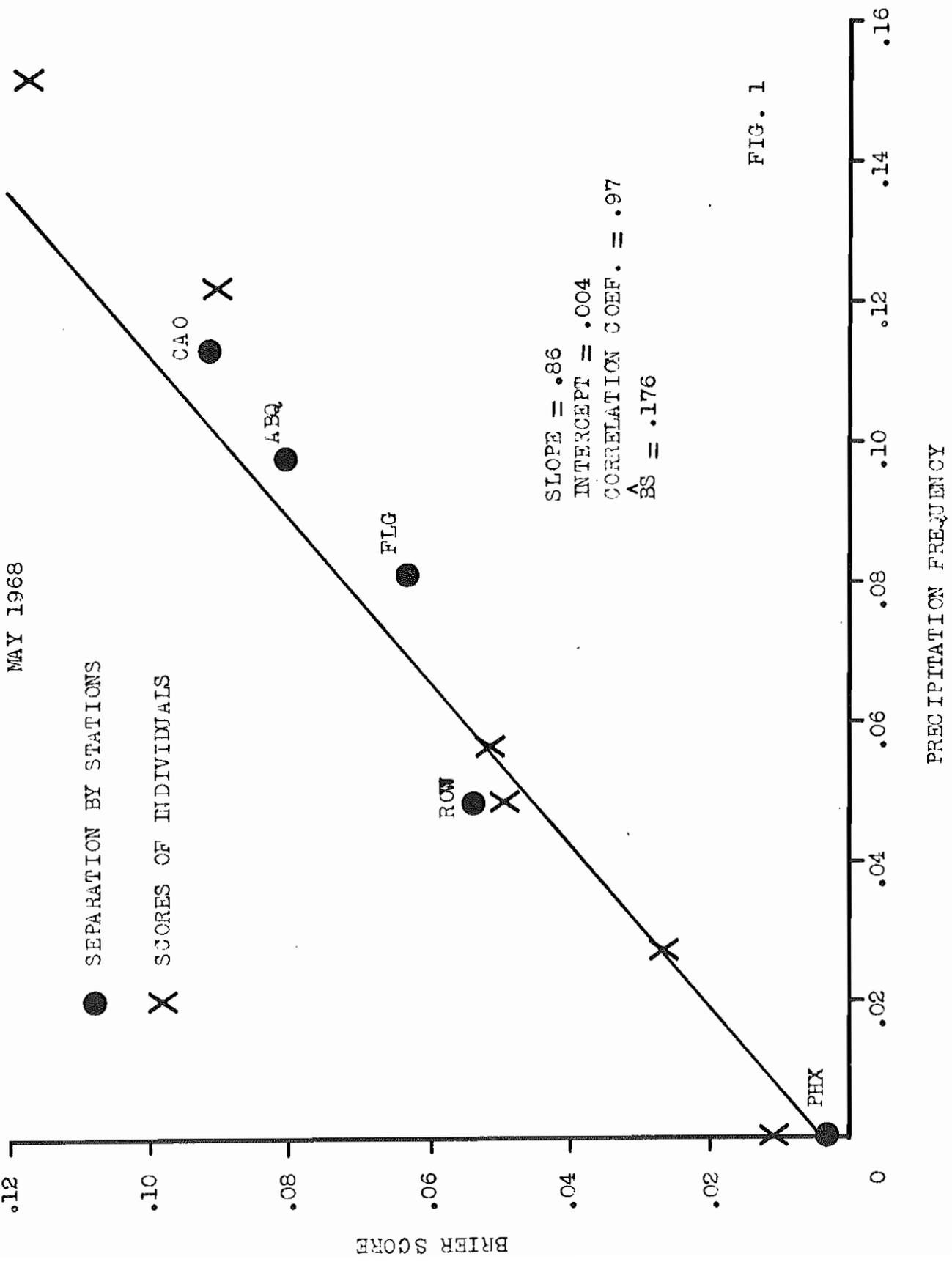


FIG. 1

individuals. These individual ordered pairs of numbers are then compared to the group effort. This comparison is visually indicated in Figure 1. Plotted circles represent the equal partition of the data--the separation from which the line was drawn. Crosses show average results achieved by each individual forecaster.

Individual scores (X's) in relation to the group average are then indicated by a position below (better efforts) the line or above (less skillful) and magnitudes are measured by vertical distances from the line to the individual's point. (In practice the cumbersome graphical methods described herein are not literally used; all results are obtained by two runs of the FP/NMC cards through an IBM/360 computer.)

At first glance, it might appear that scores thus obtained (a signed number, vertical distance from the individual's point to the line for his frequency) would be far superior to raw Brier scores for individuals, and probably they are. However, at least one major source of possible inequity still remains; there is no weighting of data for repeated exposure; the individual subsets are not numerically equal for a small data sample (month, year); and fortuitous handling of one or a few high-frequency precipitation situations may yield unrealistically good scores. The opposite situation is equally possible, of course. So, a weighting factor was added to the machine computations for individual scores. Each score departure (signed vertical distance) was multiplied by the percentage of the total number of forecasts (for the month) made by the individual forecaster. This should at least operate toward equalization since those forecasters with approximately equal numbers of forecasts would be only slightly affected and those with unrealistically small samples would have their absolute differences moved sharply toward a central value.

The final step was to obtain algebraic sums of the weighted score departures for the period of record--34 months. Clearly, if an individual could obtain a net positive score for the full period, he must either have stayed slightly below the average line consistently or have had at least a few comparatively very large positive departures to overbalance the zero and negative ones. The ranked results are:

Forecaster:	C	H	E	G	D	B	F	A
Dep. Sum ($\times 10^3$):	9.83	5.83	1.90	1.73	1.00	.08	-4.12	-5.91

It is seen that this is a substantially different ranking than that given by raw average Brier scores.

At face value this system of cumulative departure scores would appear more significant and equitable to individuals. However, we were vaguely uneasy about the ranges of these monthly and cumulative departures. Closer examination of month-to-month variations for individuals disclosed quite considerable departures--often from very high to abysmally low from one month to the next. Such gyrations could have little rational

6.

base and could easily be more fortuitous and related to the size of the numbers than to express actual evidence of skill. As a check we compared the ranking above with what would have been obtained from cumulative scores at the end of August 1967--about a year earlier but still encompassing a substantial amount of data. The change was appreciable. Two men slipped from the top four into the bottom, and half the individuals moved three full slots from their previous position. Hence, we judge the numerical values given here are too unstable for conclusive use.

Some value may still exist in the system to demonstrate to the individual forecaster how his efforts compare to those of the group for a specific sample of data. Probably conclusions regarding consistency of performance may safely be drawn by average and extreme positions. But we have to distrust numerical values from this system as long-term measures of individual skill.

So, another organization of the basic data was devised. Since the score vs frequency relation appears to be the most dependable feature of the probability scoring system, we took the ordered pair of numbers, frequency-Brier score, for each forecaster for each month for which he had verification data. From these data, a regression line was computed for the individual. Parameters of these individual lines are presented in the table below; it will be seen that the correlation coefficients are uniformly near-perfect, none below .95. Hence, the relations must be exceedingly reliable.

With this assurance, can we look more confidently at performance characteristics? What would be indications of superior performance? Clearly, a minimum slope would, for one. The more slowly scores increase as frequency increases, the better the forecasts must be. But this must be tempered to at least some extent by the individual's "caution factor", i.e., the average magnitude of his probability specification for situations from which no rain ensues. This factor is measured by the number in the "Intercept" column. This number represents the extension of the regression line to a frequency value of zero. It will be noted that these intercepts range from .006 for the apparently more confident forecasters (confident, at least, of "No rain" forecasts) to .018 for the man who appears to be most cautious.

In an attempt to combine these two indices, slope and caution factor, we computed a normalized score for each defined as the projection of his performance line to a frequency of .20. These are listed in the column labeled "BS_{.20}" and range from .135 to .161.

Incidentally, a rain frequency of .20 is roughly double that experienced over the Albuquerque forecast district on an annual average. Hence, for comparative purposes, a more realistic choice of frequency for normalization would be .10. This reduces score variations between individuals substantially but shuffles rankings only slightly.

In addition to the tabular presentation, these relations are displayed graphically in Figure 2. The performance curves form a rather close-knit family from which an appraiser has some difficulty discerning features for individuals. This observation has considerable import--we doubt that differences between these displayed curves are significant. Certainly for frequency values between .06 and .12, the range prevailing most of the time, there is little to determine significant differences between individual forecasters.

<u>Fcstr</u>	<u>Fqcy</u>	<u>BS</u>	<u>Cor.Coef.</u>	<u>Slope</u>	<u>Intcp.</u>	<u>BS</u>	<u>BS</u>
						<u>.20</u>	<u>.10</u>
H	.072	.063	.97	.77	.008	.161	.085
D	.079	.067	.97	.74	.008	.156	.082
C	.084	.070	.99	.76	.006	.158	.082
E	.090	.077	.97	.75	.009	.158	.084
F	.100	.080	.95	.60	.020	.140	.080
B	.101	.080	.96	.66	.014	.145	.080
G	.093	.078	.97	.66	.016	.148	.082
A	.088	.070	.98	.59	.018	.135	.077

And on this unhappily equivocal note, we leave precipitation forecasting verification for the moment and consider temperatures...

TEMPERATURE FORECASTING VERIFICATION

Records available for comparative, individual temperature forecasting scores are chronologically almost identical with those for precipitation. Data referenced below span the period March 1, 1966, through October 31, 1968. Individual percentages for portions of the total forecasts made by individual forecasters were not computed but would be approximately equal to those for precipitation. All data were extracted from the FP/NMC punch card decks.

As with precipitation, we are distrustful of raw "average absolute error" scores. For example, an individual might take a long mid-winter vacation or for any of many other reasons have a disproportionately high or low number of duty shifts during seasons of high temperature variability. Only very long-term data collections would smooth out such possible inequities, and during such long periods personnel complements are apt to change. Hence, we seek a measure of "average threat" to couple with "average error".

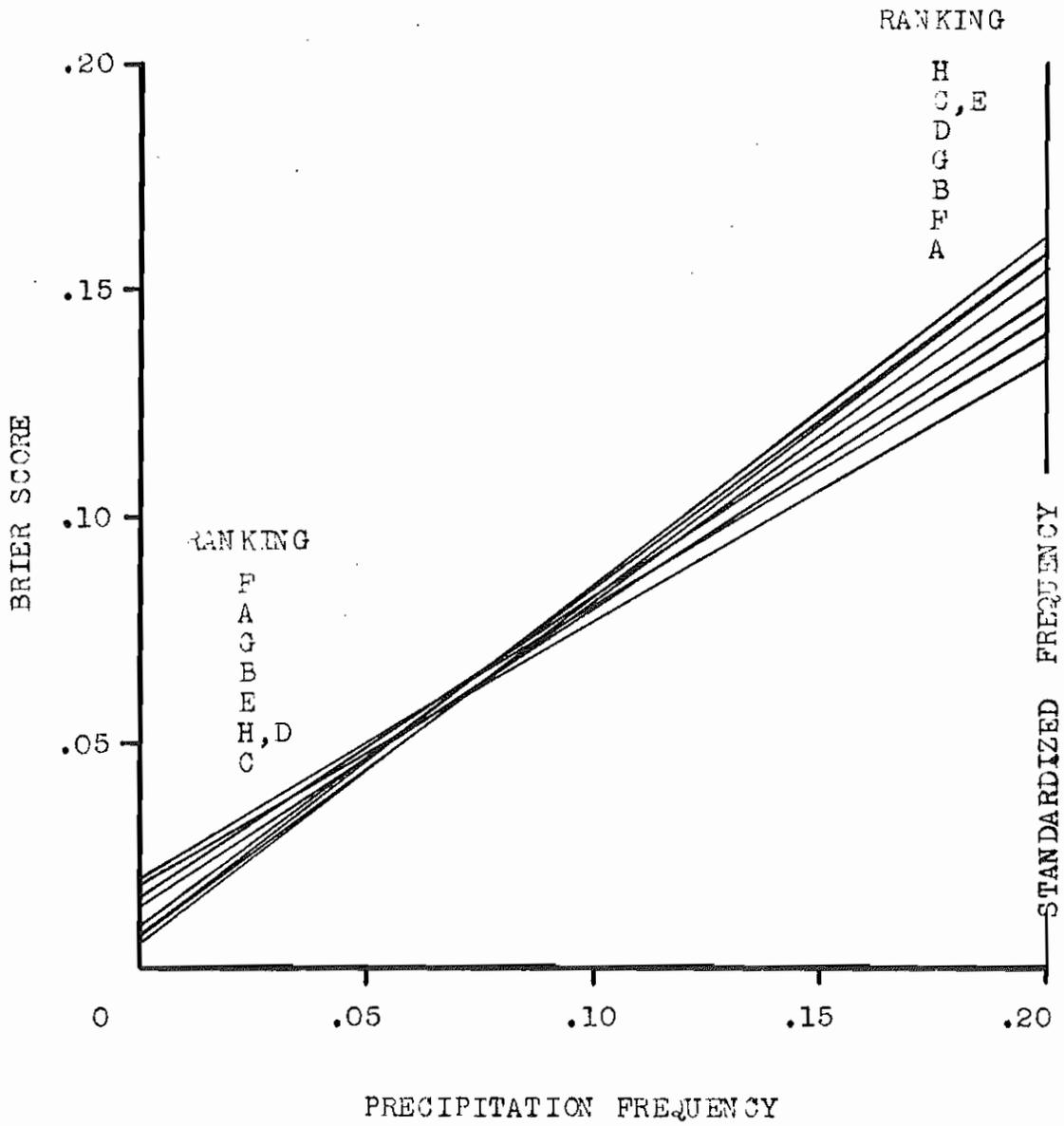


FIG. 2

The most obvious such measure is "variability", for which several definitions are available. Roberts (2) has used "interdiurnal variability", the average 24-hour change in maximum or minimum temperature for the period of record--usually a calendar month. This is compared to average absolute error of the temperature forecast. As a corresponding control we have, at the Albuquerque WBFO, long used standard deviation of maximum or minimum temperatures for each verification station as compared to the average absolute error in the forecasts. These parameters when plotted for the five verification stations usually yield a picture similar to Figure 3. It will be noted that this is decidedly comparable to the frequency-Brier score curves. However, the correlation coefficients are not usually as high as those for precipitation, i.e., the relation is less reliable.

However, either measure of the variability, average interdiurnal change or standard deviation, appears useful and probably adequate. For no very specific reason we use both at Albuquerque. A first run of the data cards separates errors and standard deviations for verification stations. This gives an indication of the staff effort for the month. A second run culls out average scores and corresponding interdiurnal variabilities for individual forecasters. This latter separation is the one discussed and illustrated below...

So, for a relatively stable cadre of personnel, we have comparative data for a continuous span of 32 months. (One man had data in all 32 months, two in 31 months, one in 28, one in 24, one in 19, one in 18, and one in 14...there was no seasonal pattern.) Data then should be reasonably comparable. Maximum temperature forecasts and minimum temperature forecasts were examined separately although, as will be seen, patterns and magnitudes were not significantly different.

The results of this close examination are not presented in tabular form. Rather, we feel that the graphical representations, Figures 4 and 5, convey the obvious conclusions much more readily. Plotted points (dots and X's--see below) represent unweighted averages for individual forecasters for all months in which they made verifying forecasts. The close cluster of points is indeed remarkable, at least to our expectation. Not only did threats turn out to be fairly close, less than one degree, but so did average errors. If any difference in average skill is demonstrated here it is exceedingly slight.

Dots represent data averaged per month and plotted for whatever period of record was available regardless of season. Since this evidenced little if any distinction between individuals (an ellipse encompassing the scatter points is almost circular), it was suspected the all-season average might have masked some real individual differences. Hence, we selected data for cold months only (high variability), November through April. These results for individuals are plotted as crosses on Figures 4 and 5. This new separation presents a slightly different depiction and more strongly displays the expected relation between threat and score.

DECEMBER 1967

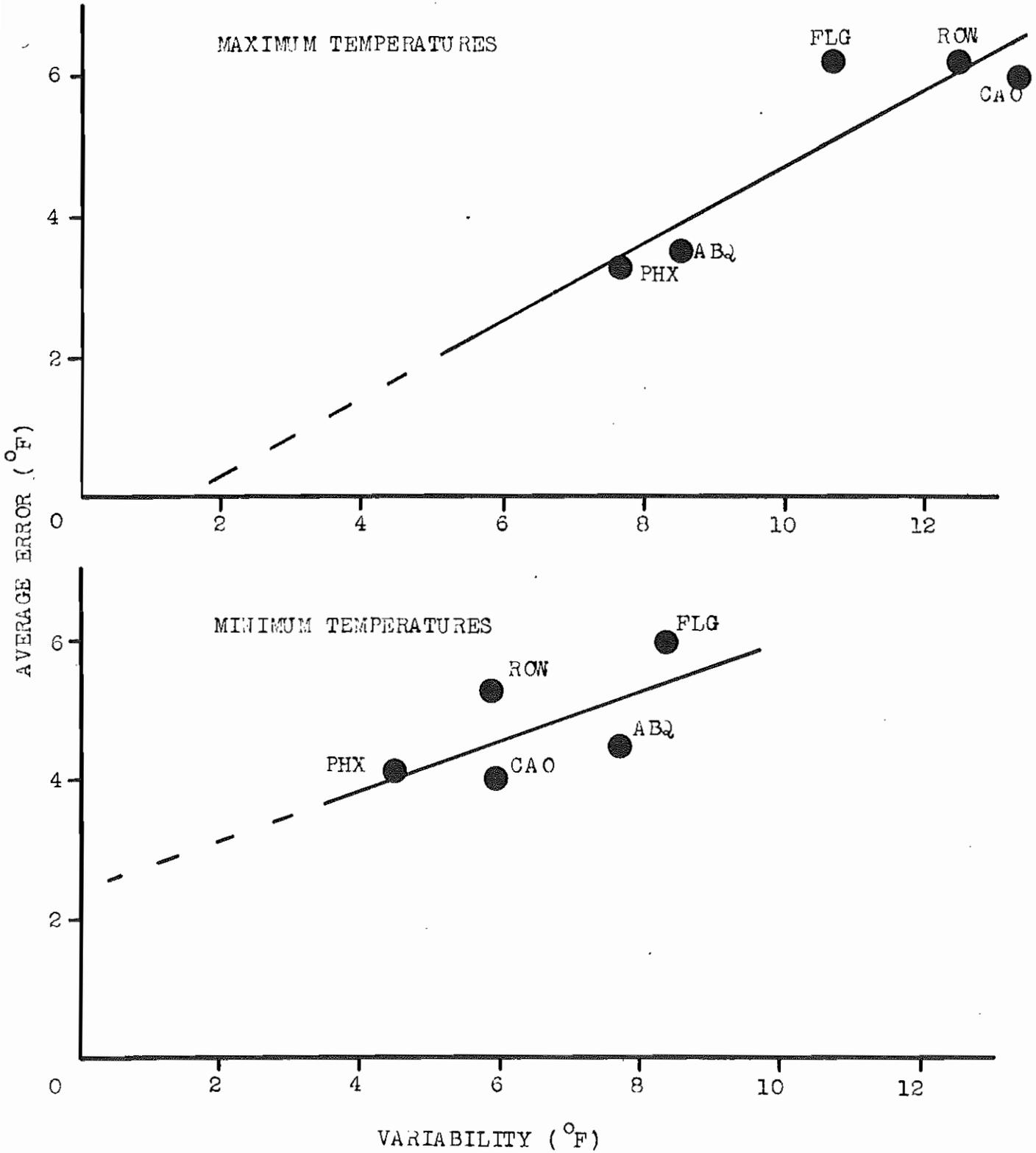
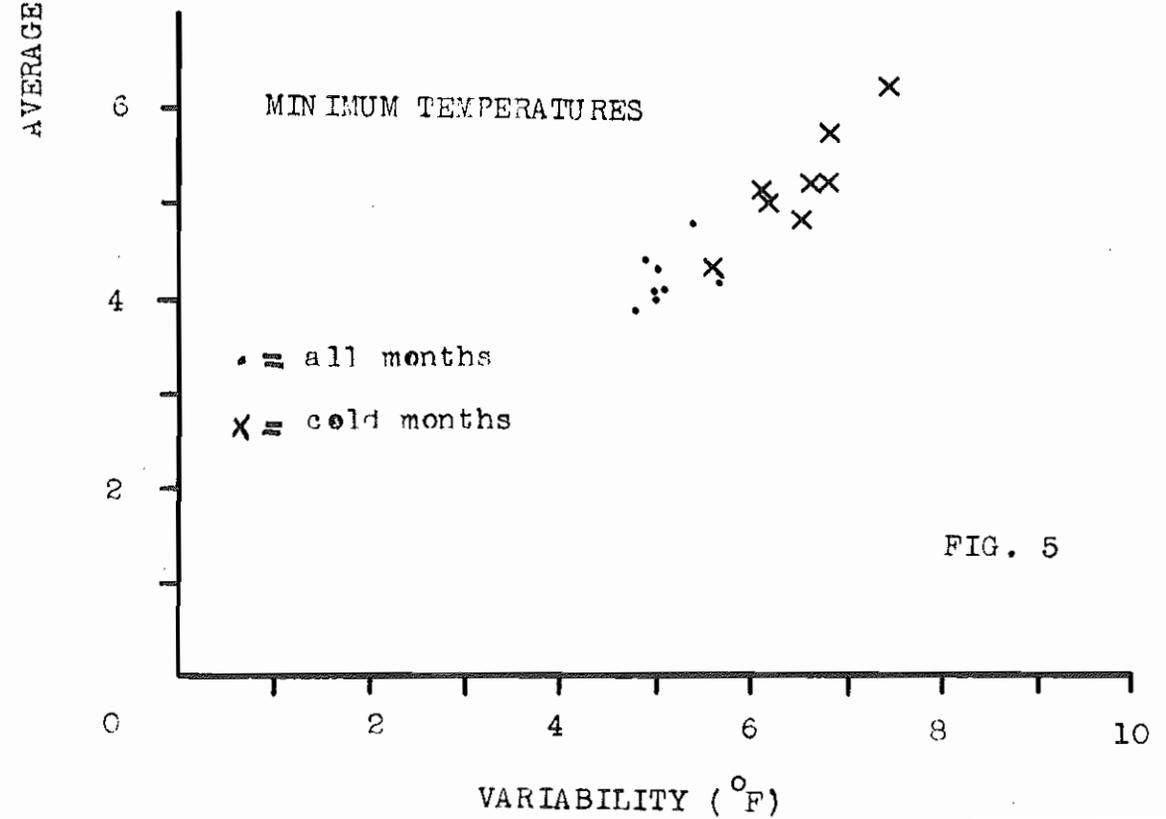
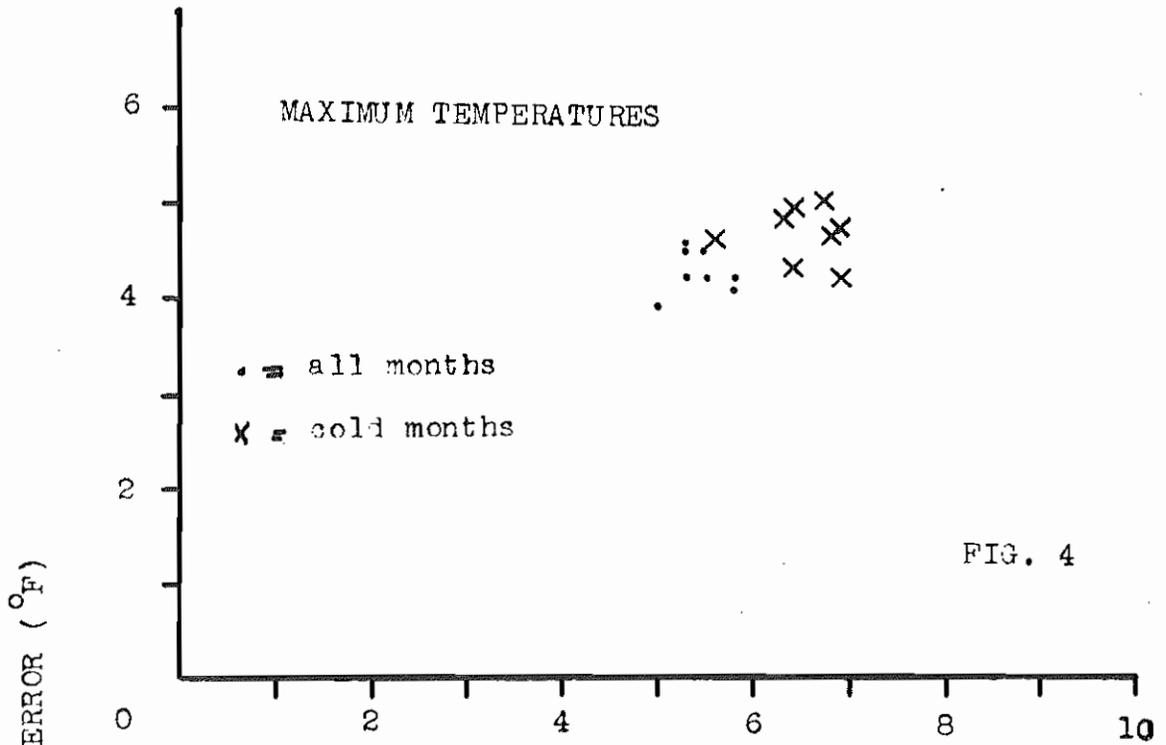


FIG. 3



Note particularly the ellipse enclosing minimum temperature data-- major and minor axes are strongly different and the major axis definitely points toward zero conditions. However, this strengthening of the relation is about all that is demonstrated by the second separation. Certainly it would be questionable and suspect to claim from this exhibit that any one forecaster's efforts were crowned with significantly more success than another's -- it only emphasizes which men had an unfortunately large share of rapidly changing temperature regimes...

DISCUSSION AND CONCLUSIONS

We were, quite candidly, rather surprised at the results of this close examination of individual verification data. We feel that only one reasonable conclusion can be drawn from this study: THAT for this particular set of forecasters, no outstanding individual strengths or weaknesses are apparent from objective verification data. After considering achieved scores along with threat, no significant trend toward superiority is evident.

Can such a conclusion be extended generally or are these results unique only for this group of men? While we have the highest regard for these individuals -- all colleagues and friends of many years standing and all eminently qualified both academically and by experience -- we feel that the group is not exceptional. A moderate extension of the above conclusion is probably valid with only slight reservations. In the present era of meteorology a certain amount of professional averaging is inherent in the system -- some by design, some by inadvertence, and most, probably, beneficial in the balance. The basic tools available to the forecaster are the products of the facsimile circuits, and these leave little room for individual improvement. He may, and for the most part probably does, critically review pertinent analyses for areas of particular concern to him. But for macro-scale projection into the future he is overwhelmingly dependent on the ubiquitous numerical models. They are by no means perfect, and the local forecaster does frequently improve on and sharpen their implications for his geographical area of specialization. But their basic objectiveness and large-scale accuracy and freedom from personal bias, mood, or prejudice pressure him away from frequent or significant deviation.

Some local supplementary aids are prepared and considered at most forecast centers, of course, and these, hopefully, sharpen and improve forecast products and perhaps allow some differences between facilities. But further averaging is deliberately encouraged and sought at local levels by reason of the well-proved concept of continuity: "If you don't have a strong and justifiable reason for disagreeing with the previous forecast or if the change you propose isn't significant, better forget it!" Most of us have considerable respect for the man on the preceding shift...

The individual forecaster, even if he accepts the conclusions asserted above (which is doubtful, forecasters being the individualists they are), need not be dismayed at a lack of opportunity to achieve distinction. All the data cited and examined refer only to scoring, to numerical verification of the products. This is so because errors in estimate of specific elements (e.g., temperature and precipitation) are all we can objectively analyze. Weather advices need not only to be accurate but also, to be of maximum value, need to be effective. This quality of effectiveness is an elusive one that so far has escaped numerical measure. Forecasts for the same period and area issued by two different men may score identically by numerical measure but can have vastly differing impacts -- and, hence, values -- from considerations of phrasing, presentation, user confidence and acceptance, timeliness, and other intangibles. A knowledge of and facility in communication should be as integral a part of a forecaster's training and capability as his familiarity with the geostrophic wind relation. If he cannot communicate effectively and convincingly, his technical competence and expertise is largely wasted. Hence, we must insist that a run-of-the-mill forecaster who can and does transfer his knowledge effectively is probably more valuable to the user than a brilliant one who masks and obscures his conclusions and expectations beneath a welter of trite phrases and stilted phraseology.

ACKNOWLEDGMENTS

Any worthwhile end product stems from many sources of input. To cite only a few: We are grateful to H. L. Jacobson for his sage advice, discerning criticism, and unflinching encouragement; to Ralph Pike for his unstinting assistance and valuable liaison with computational facilities; to Thomas May for graciously contributing his drafting competence; and to Mrs. Peggy Williams for her unstinting typographical efforts and cogent grammatical suggestions. Most especially, we are grateful for the patience and consideration of the doughty crew who created the basic data -- our esteemed colleagues of the Albuquerque Weather Bureau Forecast Office staff: Messrs. Beeler, Bussiere, Eldred, Erickson, Morris, Murray, and Ropar. (In actual alphabetic order this time!)

REFERENCES

1. ESSA, Weather Bureau, Operations Manual Letter 68-51, dated 11/17/68.
2. Roberts, C. F.; Porter, J.M.; Cobb, G. F.: "Report on the Forecast Performance of Selected Weather Bureau Offices for 1966-67", OML Technical Memorandum WBTM FCST-9.
3. Hughes, L. A., "Public Probability Forecasts", Technical Memorandum WBTM CR-11.
4. Hughes, L. A., "Probability Verification Results (24 months)", Technical Memorandum WBTM CR-19.
5. Dickey, W. W., "Verification of Operational Probability of Precipitation Forecasts, April 1966-March 1967", Technical Memorandum WBTM WR-25.