UNITED STATES DEPARTMENT OF COMMERCE
ENVIRONMENTAL SCIENCE SERVICES ADMINISTRATION
WEATHER BUREAU SOUTHERN REGION
Fort Worth, Texas

TECHNICAL MEMORANDUM NO. 10

A QUICK LOOK AT THE RESULTS OF ONE MONTH'S PRECIPITATION PROBABILITY FORECASTING

by

George T. Gregg
Weather Bureau State Forecast Center
Albuquerque, New Mexico

Scientific Services Division

January 1966

FOREWORD

This summary of one month's probability forecasts made by the Albuquerque Forecast Center is being distributed not so much for its informational content (while interesting and informative), but as an excellent example of the type of analysis that can be carried out at any forecast center, or local station, to attempt to discover individual and group biases or systematic errors in probability of precipitation forecasts.

W. W. Dickey
Chief, Scientific Services

Sure wish some one would volunteer JS

A QUICK LOOK AT THE RESULTS OF ONE MONTH'S PRECIPITATION PROBABILITY FORECASTING

With the completion of the first full month of official probability specification by the Albuquerque Area Forecast Center, the verifying data for these forecasts were rather eagerly examined. It was anticipated that such an examination would not only assure the staff that they do indeed possess some appreciable skill in classifying current and predicted synoptic patterns as to point probabilities of precipitation, but might also disclose at least a few correctable systematic individual and group errors of practice. Hence, this rather hasty and hurried survey.

The data examined in this analysis are for December of 1965. The ABQ FP-3 lists precipitation probabilities for six locations in Arizona and for five in New Mexico in accordance with codes and time intervals specified by the Salt Lake City Regional Office for Arizona and by the Fort Worth Regional Office for New Mexico. These rules are not contradictory although the SLC rules permit an additional probability class (2%) and add two additional time periods beyond those required by Southern Region (FTW) rules. Since the Albuquerque forecast district intersects both administrative regions, probability estimates published in the ABQ FP-3 follow the specifications of the administrative region in which the point is located for which the forecast is being made.

However, it was considered advisable for local verification purposes to keep records for internal information according to the more comprehensive system embracing, mainly, the extra 36 hours of forecast period. We wished, literally, to learn our own capabilities and limitations. Verifying data have, therefore, been maintained for FLG, INW and PHX in Arizona and for ABQ, CAO and ROW in New Mexico. These are the data examined in this report. It should also be mentioned that December 1965 was an unusually active month meteorologically over the western half of this district. Recurrent heavy rains and snows assaulted Arizona from about the 9th of the month on till near Christmas resulting from a series of short waves moving out of the southeastern Pacific. Flagstaff, with nearly 6 inches of moisture, had the greatest monthly precipitation of record. Large-scale and important flooding occurred in the Phoenix area late in the month due to full reservoirs, snow melt and from fresh rain. On the other hand, eastern New Mexico was quite dry with only a few days of light to moderate precipitation. Thus a complete spectrum of activity greeted our first month's efforts in the precipitation probability specification field.

Data were basically examined for accuracy, measured by the abbreviated Brier Score $(F-O)^2$, and for reliability, the property of separating weather patterns into precipitation probability classes. Average Brier scores are presented in the table below.

	FLG	INW	PHX	ABQ	CAO	ROW	AVG
First three periods	.16	.14	.16	.11	.08	.09	.12
Five periods	.23	.19	.23	.14	.09	.10	.16

The general gradient of error (decreasing eastward) rather obviously corresponds to the gradient of precipitation frequency. Magnitude of errors increased significantly from the three-period summary to the five-period summary for the wetter stations. This is not surprising since forecasts for (and observations at) CAO and ROW were predominantly zero, e.g., there were for ROW in the first three periods 239 forecasts of zero probability, only five of which resulted in rain events, a frequency of .02. Scores for the wetter stations for the initial three periods were higher (worse forecasts) than those for all five periods for the drier stations. Indeed, scores for the initial period only at FLG (.10) were comparable to the five-period average at CAO and ROW.

A consensus forecast was made once daily coincident with the 0900M official FP-3 estimate. Participants in the consensus were most frequently one or both of the aviation forecasters on duty. On an "available" basis other personnel such as the fire-weather meteorologist, the Chief, meteorologists performing research or other non-routine duties, and occasionally an observer were members of the consensus. Apparently this heterogeneous mixture was effective, as witness Figure 1. This diagram presents an array of the six stations with the abbreviated Brier score as ordinate. Solid lines connect official (FP-3) scores while dashed lines connect consensus scores. Clearly consensus scores are rather consistently better than those of the official forecasts. This tendency has been noted elsewhere and diverse morals and conclusions drawn. We choose to only express mild surprise (and perhaps chagrin!), to note that the difference is slight, and to resolve to carefully examine future data to see if the tendency persists. Other features displayed by Figure 1 are about as might be expected: Scores are worse where precipitation frequency is highest and worsen as the interval between forecast preparation and verifying time lengthens.

Figure 2 presents Brier score data (as ordinate) for the six stations individually for the 0900M edition of forecasts. Abscissae in these panels are the five time periods of the forecasts. Curves are drawn for the official forecast, the consensus forecast, and, where data were available, for the forecasts possible with climatological averages. The advantage of the consensus forecast is again apparent as is the progressive and rapid deterioration of both subjective forecasts with increasing time. As was expected, through the third period (roughly 36 hours) subjective forecasts display an improvement over climatology; beyond the fourth period the advantage usually goes the other way. The marked dip between first and second period for FLG and INW and some of the variations at CAO and ROW are, to say the least, intriguing. Data for succeeding months will be carefully examined to see if these anomalies recur. (One possible explanation for the FLG and INW dip: overforecasting for the initial--in this case six to nine hours--period. This could likely be proved or disproved by a closer analysis of the types of errors involved in these means; time has not yet been available for such a close scrutiny.)

Figure 3 is a reliability curve for the Albuquerque effort. For this presentation the first three forecast periods only were summarized since we consider these periods to constitute our principal responsibility in

the guidance chain. The figure shows that our product was not by any means completely satisfactory; there is evident an appreciable degree of underforecasting of precipitation probabilities in the lower categories and a certain amount of overforecasting of rain in the .7, .8 and .9 brackets. Fortunately the 100% class verified perfectly--an appropriate termination. Average deviation from perfect reliability was .15. As can be noted from the "Popularity of numbers" bar graph, the staff forecasters displayed a singular reluctance to make use of the .02 and .05 categories; thus these data are somewhat unrepresentative. If we lump the few .02 forecasts into the .00 class and the .05 forecasts into the .1 class, the average deviation decreases to .13 and the curve is less irregular. We note without comment or suggested explanation the relative unpopularity of the .7 category.

We feel that this record is encouraging for an initial effort and does demonstrate a real ability by the forecast staff to categorize situations and flow patterns as to their probability of producing measurable precipitation at a specified point. We hope to attain a greater degree of reliability with more experience.

A considerable deterioration in reliability (as well as skill scores) occurs in the last two periods of the five-period set of forecasts. This is strikingly illustrated by Figure 4, a reliability curve for FLG with the solid line representing reliability of precipitation specifications for the initial three periods and the dashed line for the final two periods only. Underforecasting of precipitation is again illustrated for this phenomenally wet month for the first three periods for most classes but gross overforecasting appears for the final two periods in the higher classes. Obviously high probabilities should be used rarely if ever for times in excess of 36 hours and most probabilities for these longer-term forecasts should approach climatological averages.

Thus we come to these four main conclusions from this hasty and incomplete review of our first month's effort:

1. Definite skill is evident in estimating precipitation probabilities for periods up to 36-48 hours.

2. More skill and reliability would probably result from a conscious use of higher probability figures for initial periods when the situation seems to warrant assurance of rain in the area.

3. Similarly, probabilities for periods more than 24 hours should be consciously tempered, and beyond 48 hours should not depart too far from climatic averages.

4. More use should be made of the class(es) between 0 and .1.

We hope in succeeding months as reliability (hopefully) increases to be able to better resolve our forecasts. No analysis along this line has yet been ventured. We also hope, whenever and if time permits, to look at

other features contained in these data, e.g., departure of individual scores from consensus (as a control), reliability of individuals, diurnal influences, relative difficulties of initial six vs 12 hour periods, etc.

Data now being generated by the new NMC/FP comparison will yield added interesting information from temperature forecast scores and from further vertical comparisons along the chain of guidance. We enthusiastically support and approve of these critical self- and system-analyses and only hope we can adequately supplement, digest, and exploit their indications.

George T. Gregg
WBAS, Albuquerque, New Mexico

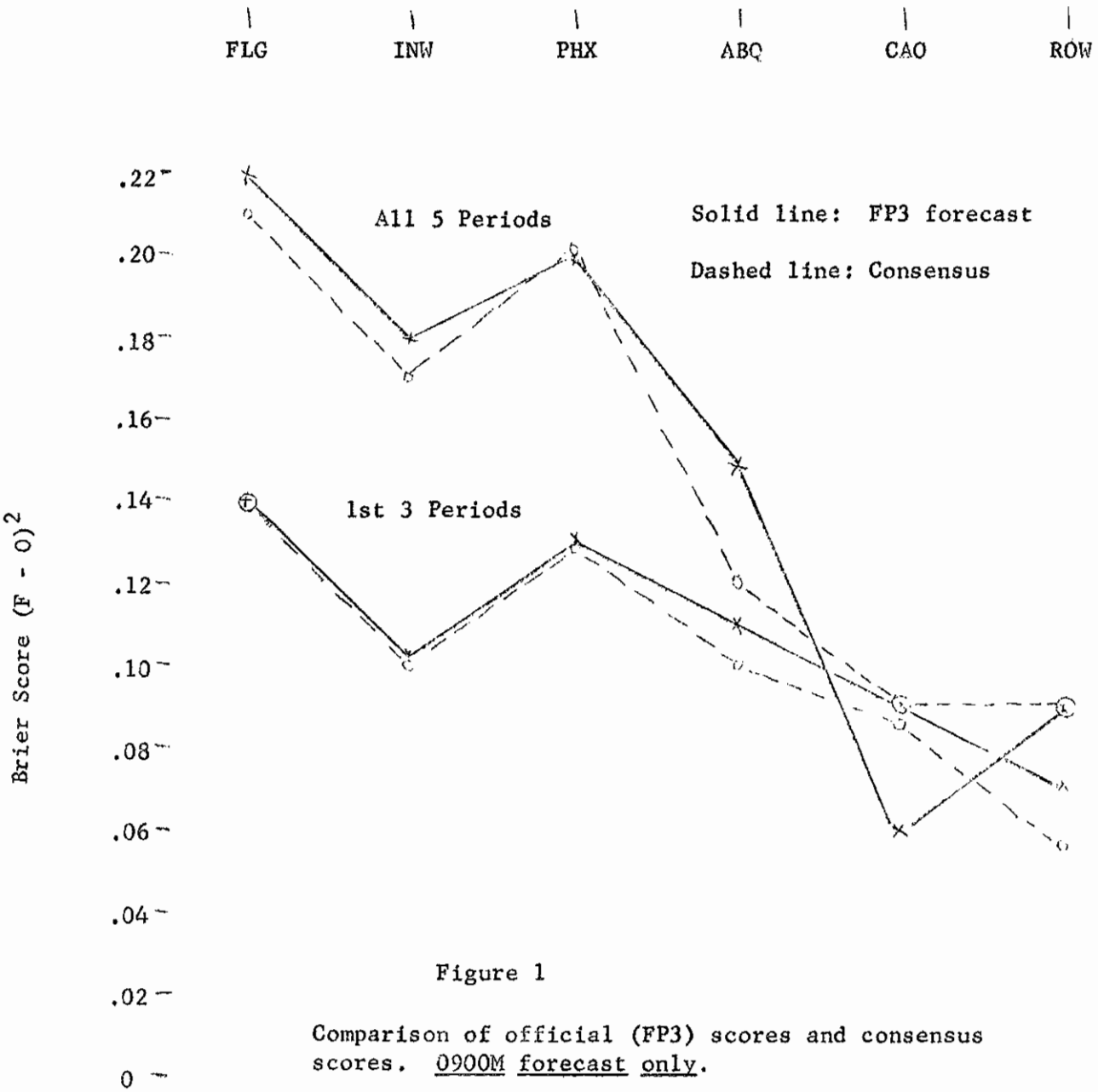

Figure 1

Comparison of official (FP3) scores and consensus scores. <u>0900M</u> <u>forecast</u> <u>only</u>.

December 1965

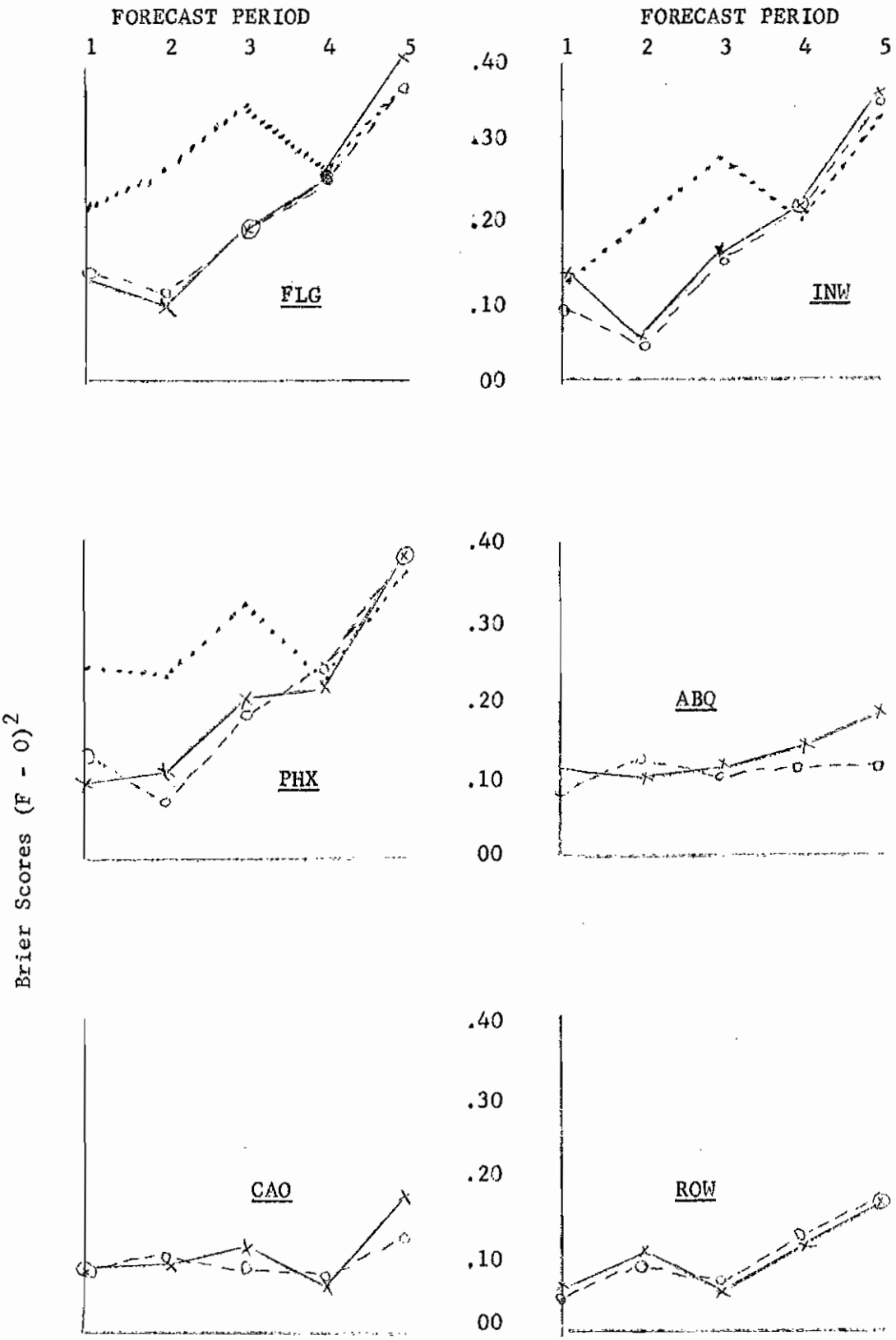

Figure 2

Comparative scores for individual stations.
0900M forecasts only. December 1965

Solid line: FP3 forecast
Dashed line: Consensus
Dotted line: Climat forecast

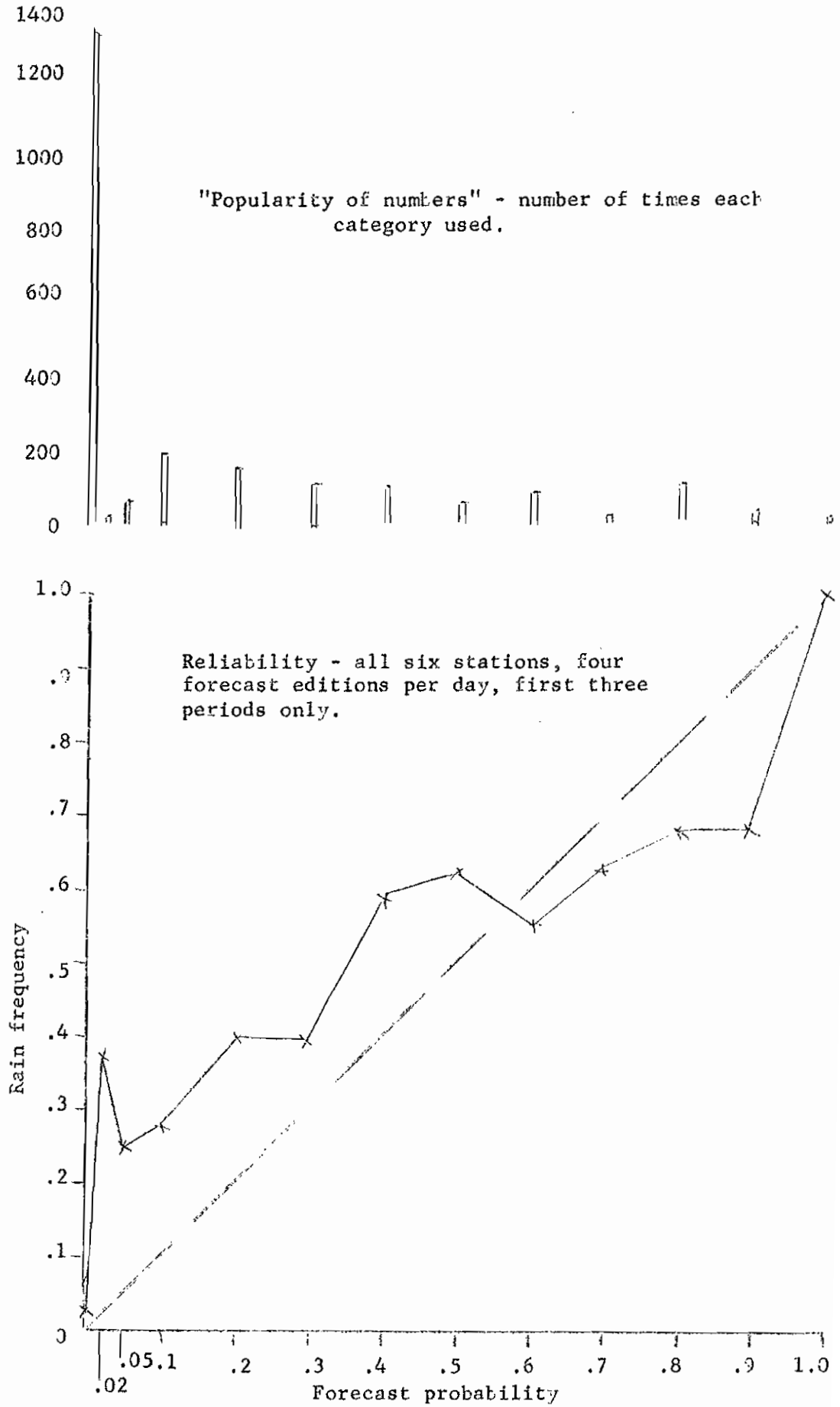
1400
1200
1000
800
600
400
200
0
"Popularity of numbers" - number of times each category used.
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
Rain frequency
Reliability - all six stations, four forecast editions per day, first three periods only.
.02
.05
.1
.2
.3
.4
.5
.6
.7
.8
.9
1.0
Forecast probability

Figure 3

FLAGSTAFF December 1965

Solid line: 1st three periods only
Dashed line: last two periods only

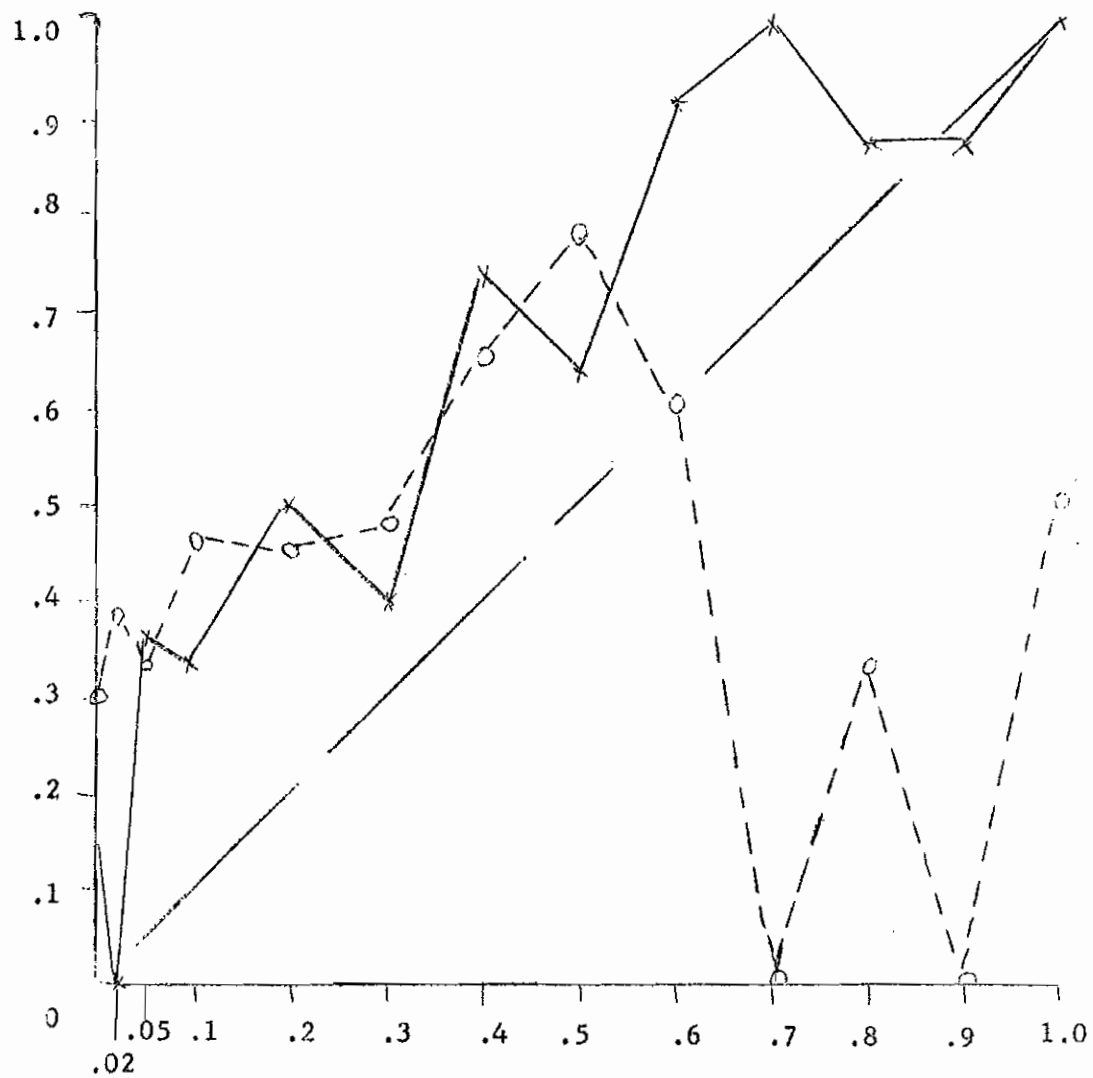

Figure 4