NOAA Technical Memorandum NWS SR-107

COMPARATIVE FORECAST VERIFICATION--A STATISTICAL APPROACH

Thomas M. Hicks
WSFO San Antonio, Texas

Scientific Services Division
Southern Region
Fort Worth, Texas
April 1983

**UNITED STATES
DEPARTMENT OF COMMERCE
Malcolm Baldrige, Secretary** /
**National Oceanic and
Atmospheric Administration**
John V. Byrne, Administrator /
National Weather
Service
Richard E. Hallgren, Director

COMPARATIVE FORECAST VERIFICATION--A STATISTICAL APPROACH

by
Thomas M. Hicks
WSFO San Antonio, Texas

1. INTRODUCTION

Forecast verification has been a controversial topic for about as long as forecasts have been issued. Perhaps most controversial have been those verification programs which attempt to compare the relative performance of forecasters. The reason for this is that most often the forecasters are judged on the basis of forecasts which simply cannot be compared. The degree of forecasting difficulty varies so much from one circumstance to another that a very large sample of forecasts is needed to assure that no forecaster has an unfair advantage over another (Panofsky and Brier, 1968).

The purpose of this paper is to offer an objective verification technique that minimizes the forecast difficulty factor and allows forecaster comparisons to be made--even for relative small samples of data. Fundamental statistical techniques are used to (1) analyze the relationships among the verification data, (2) develop an unbiased basis of forecaster comparison, and (3) test the significance of the forecaster scores. The technique currently evaluates only the temperature forecasts; however a similar technique is planned for precipitation forecasts.

2. PURPOSE OF VERIFICATION

Before any verification scheme is proposed or adopted, it is necessary to determine the primary purpose or purposes to be served by the verification. One purpose, and perhaps the most important, is to examine the forecast errors in order to determine their nature and possible cause (Panofsky and Brier, 1968). This hopefully will then lead to better forecasts. Many such verification programs are currently in use and should certainly be continued.

Another purpose, and the reason for this paper, is to objectively evaluate the overall performance of the forecasters. More specifically, this verification program seeks to answer the following questions:

(1) How has each of the forecasters performed within the past month and the past year?

(2) Is there any significant difference in forecaster performance? Or is the variation among forecasters simply random variation due to chance?

(3) Which of the forecasters (if any) have performe significantly above or below the station average during the pas month and the past year?

3. ANALYSIS OF VERIFICATION DATA

In order to answer the above questions, it is first necessary to select a basis of forecaster comparison that meets the following criteria:

(1) Must be proportional to forecaster skill.

(2) Must be free from external influences.

(3) Must be statistically suitable for significance tests.

Measures of Forecaster Skill

Perhaps the most common temperature verification statistic related to forecaster skill is mean absolute error (MAE). This statistic is a direct measure of the quality of the forecast, and has often been used--with controversy--as a basis of forecaster comparison.

An alternative verification statistic related to forecaster skill is percent improvement over guidance. This is essentially a quantitative evaluation of a forecaster's decision to follow, deviate from, or ignore guidance. A forecaster who consistentl improves over guidance has demonstrated skill and will have relatively high score. A forecaster who consistently follows guidance--or deviates in the wrong direction--has demonstrated no skill over guidance and will have a relatively low score.

External Influences

Roberts (1967) and others have shown that temperature variability is perhaps the most important external influence on the accuracy of temperature forecasts. Variability is a climatological measure of forecast difficulty, and can be expressed quantitatively as the net 24-hour change in temperature.

Another external influence is that of guidance. The overall quality of model output statistics (MOS) temperature forecasts has been quite good. But how might variations in quality of quidance affect forecaster comparisons?

Relationships Among Verification Statistics

To aid in the selection of a basis of forecaster comparison, all possible linear correlation coefficients were computed among the following verification statistics:

(1) Mean absolute forecaster error

(2) Mean absolute MOS error

(3) Mean absolute 24-hour temperature variability

(4) Percent improvement over MOS.

The data consisted of a random sample of 240 forecasts drawn from two full years of verification data. Since seasonal effects were of special concern, the data was stratified to ensure each month was equally represented. Each forecast actually consisted of the total scores for 18 separate forecasts (six verification stations--each for three periods).

The null hypothesis (population correlation coefficient = 0) was then tested to determine the significance of the correlation (Snedecor and Cochran, 1980). Those correlation coefficients determined to be highly significant indicate rejection of the null hypothesis at the 99% confidence level. The results are summarized in table 1.

TABLE 1. LINEAR CORRELATION OF VERIFICATION STATISTICS
(** denotes highly significant correlation)

Correlation (240 pairs)	r	Variation explained
FORECASTER MAE vs. VARIABILITY	+.794 **	63%
FORECASTER MAE vs. MOS MAE	+.848 **	72%
MOS MAE vs. VARIABILITY	+.717 **	51%
% IPVMT OVR MOS vs. FORECASTER MAE	-.416 **	17%
% IPVMT OVR MOS vs. VARIABILITY	-.267 **	7%
% IPVMT OVR MOS vs. MOS MAE	+.103	1%

The following conclusions can be drawn from the correlation analysis:

Mean Absolute Error vs. Variability
Forecaster mean absolute error is highly correlated with temperature variability--as previously determined by Roberts. The relationship indicates that temperature errors should be considerably higher during the winter months when variability is high, and lower during the summer months when variability is low.

Mean Absolute Error vs. Quality of Guidance
Forecaster mean absolute error is also highly correlated with the quality of guidance available (MOS MAE). The implication here is that a forecaster might put forth no effort whatsoever and yet still attain a comparatively low mean absolute error--by simply following guidance.

The apparent high correlation between forecast error and guidance is worthy of further discussion. It is evident from table 1 that both forecaster error and MOS error are highly related to variability. Since (for any forecast) variability is always the same for both the forecaster and guidance, is the correlation between forecaster error and guidance real? Or is it simply because of their common association with variability?

This question was answered by eliminating the effect o variability through partial correlation. This technique measures that part of the correlation between two variables that is not simply a reflection of their association with a third variable (Snedecor and Cochran, 1980). The resulting partial correlation coefficient, r = .658, was again determined to be highly significant with approximately 43% of the variation in forecaster error explained by variation in MOS error. More importantly, with equal variability, forecasters with more accurate guidance will tend to issure more accurate forecasts.

Mean Absolute Error: Unacceptable as Measure of Skill

At this point, it is apparent that mean absolute error is a very poor indicator of forecaster skill. The influence of variability and guidance are simply overwhelming. To attempt forecaster comparisons on the basis of mean absolute error could only lead to one inevitable fact: which of the forecasters had the most fortunate combination of easy forecasts and good guidance?

It should be pointed out that large samples would tend to equalize the influence of external factors among forecasters. However, Gregg (1969) determined that the adverse effect of variability was still strongly evident with almost three years of verification data. The effect of variability can also be minimized by normalization techniques. Such a technique was implemented by Gregg. However, the resulting variation in scores might still b the result of variation in guidance.

Another alternative would be to eliminate the effect of guidance, which would simultaneously reduce the effect of variability. However, the resulting difference between guidance and mean absolute error would still be higher in winter and lower in summer. This problem could be minimized by expressing the resulting difference in terms of percent. The resulting statistic is percent improvement over guidance!

Percent Improvement vs. Quality of Forecast

From the correlation analysis in table 1, it is evident that percent improvement over MOS is proportional to the quality of forecast issued. Higher improvement over MOS leads to lower mean absolute error. The reverse relationship is also true: lower mean absolute error leads to higher improvement over MOS.

Percent Improvement vs. Variability

Percent improvement over MOS is also related to temperature variability. However, the relationship is very weak, with only 7% of the variation in improvement explained by the correlation. The inverse relationship suggests that percent improvement should be slightly higher during the summer months when temperature variability is low, and slightly lower during the winter months when variability is high.

Percent Improvement vs. Quality of Guidance

Percent improvement over MOS is virtually independent of the quality of MOS. The correlation coefficient is not significant, and the population correlation may be assumed to be zero. Only 1% of the variation in percent improvement can be explained by the correlation.

Percent Improvement: Relatively Unbiased as Measure of Skill
Percent improvement over guidance shows definite potential as a basis of forecaster comparison. The correlation analysis indicates that percent improvement over guidance should be very nearly the same for forecasters of equal skill. However, before forecaster comparisons can be made, the effect of variability must be examined in more detail.

4. THE EFFECT OF VARIABILITY

The effect of variability can best be illustrated by examining its influence on the distributions of both mean absolute error and percent improvement over guidance.

The distributions were obtained by computing the mean and standard deviations for both verification scores for each month of 1981 and 1982. All forecasts (a total of 1460) issued by WSFO San Antonio during this two-year period were included. For each month, approximately two-thirds of the verification scores should lie between one standard deviation above the mean and one standard deviation below the mean. The resulting distributions are illustrated in figure 1.

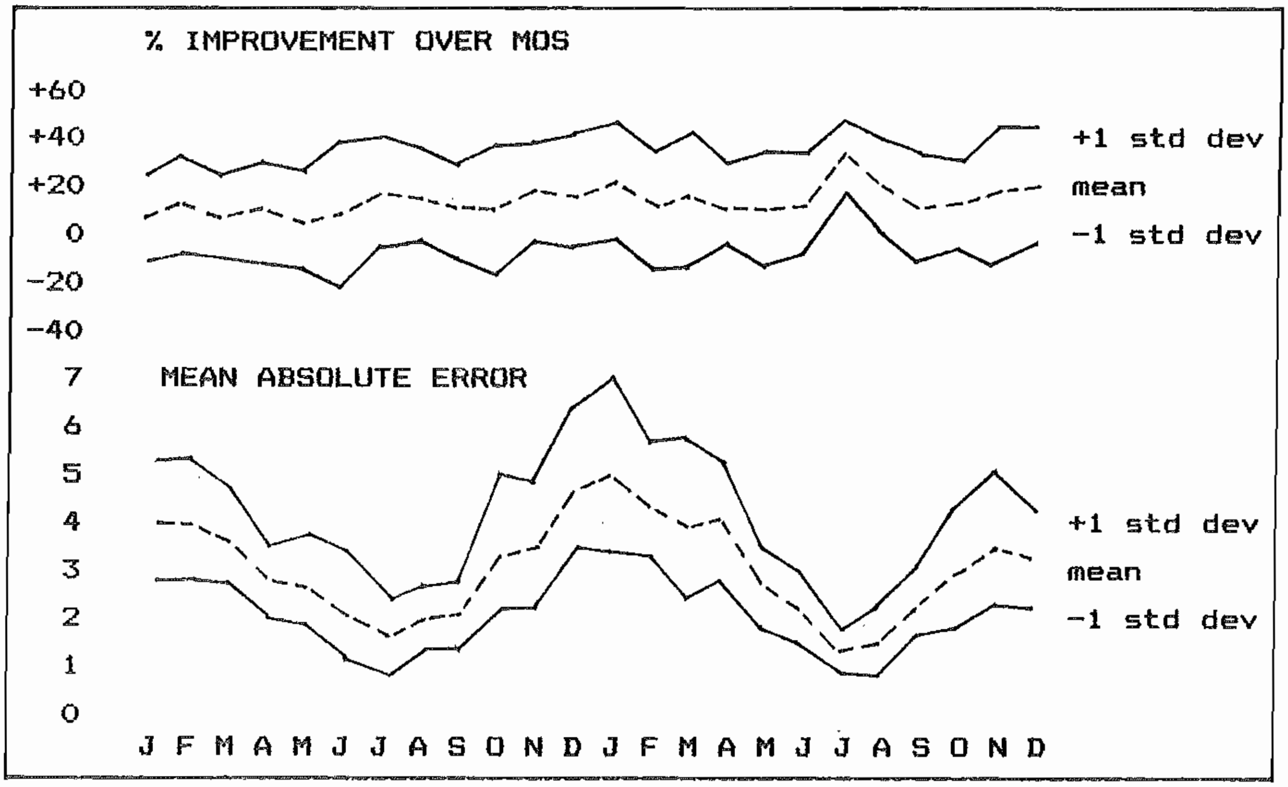

Figure 1. Distribution of Verification statistics
(1460 forecasts, 1981-1982)

From figure 1, it is immediately evident that variability has a dramatic influence on mean absolute error. Average mean absolut errors--and the range of errors--are much lower in summer than ı winter. It is clear that a forecaster who works a greater number of his shifts in winter cannot be compared with a forecaster who works more shifts in summer.

The effect of variability is much less noticeable on percent improvement over MOS. As suggested by the correlation analysis, percent improvement scores for both the summer seasons of 1981 and 1982 were slightly higher than the preceding winter seasons. But the most important feature of the distribution is that percent improvement scores--and the range of scores--tend to be *uniformly* distributed over the two-year period. The one notable exception was July 1982.

Before forecaster comparisons can be made, two very important questions concerning variability and its influence on percent improvement scores must first be answered:

1. Is the influence of variability of sufficient magnitude to adversly affect forecaster comparisons?

2. If so, how can the effect of variability be eliminated?

The first question was answered by statistically testing the following *hypothesis*:

"The effect of variability is not important. Furthermore, each of the 24 monthly samples of percent improvement scores are from the same population of scores. The variation between months, as depicted in figure 1, is due to the random effect of chance."

The hypothesis was tested by the analysis of variance (Klugh, 1974), and the results are summarized in table 2. The ratio of variation in scores between months and within months was found to be highly significant. The hypothesis that monthly variation in percent improvement scores was a result of chance was *rejected*. In order to fairly compare forecasters, the monthly variation in percent improvement scores must be eliminated.

Source of variation	Sum of Squares	df	Estimate of Variance (Mean Square)	F
Between months	61727.96	23	2683.82	5.06 **
Within months	760687.19	1436	529.73	
Totals	822415.15	1459		

Table 2. Analysis of Variance--Percent Improvement Over MOS
(** indicates F-ratio is highly significant)

5. ELIMINATION OF VARIABILITY

The monthly variation in percent improvement scores can be eliminated through the use of standard scores (Klugh, 1974). Standard scores allow comparisons to be made between distributions with widely differing means and standard distributions. Each distribution (in this case, month) is simply rescaled so that the mean becomes 0 and the standard deviation becomes 1. The rescaled measurement (in this case, percent improvement score) is given by the relation

$$Z = (X - m)/s$$

where X is the raw percent improvement score, m is the monthly mean, and s is the monthly standard deviation.

When this rescaling is accomplished for each month of the year, then samples of forecaster scores can be compared without bias. The relative importance of each score is its departure from its monthly mean--expressed in units of its monthly standard deviation. This relative importance is maintained, but the units of measurement are now common throughout the year. The mean for each month will be 0, and the standard deviation for each month will be 1.

Forecaster scores for the year are then tabulated using standard units. The resulting scores may then be transformed back into the more familiar units of percent improvment over guidance. This is done with the formula

$$X = M + S(Z)$$

where M is the annual mean percent improvement, and S is the annual standard deviation. The resulting scores then represent each forecaster's percent improvement throughout a full year in which the mean and standard deviation remained constant. The monthly variation due to variability (and other causes) has been eliminated.

6. SIGNIFICANCE TESTS

With the effect of variability eliminated, individual forecaster scores of percent improvement over guidance can be computed and then compared. The scores in table 3 represent the performance of the forecast staff at WSFO San Antonio for the year 1981. (The forecaster numbers have been changed to maintain confidentiality.)

It is important to note that each forecaster score represents the mean improvement per forecast. Percent improvement is computed for each forecast by the formula

$$IPVMT = 100\ (FP-MOS)\ /\ MOS$$

where FP is total forecaster error, and MOS is total guidance error. The mean is then computed by dividing the total IPVMT scores by the number of forecasts.

FCSTR NO.	FCSTS ISSUED	AVG % IPVMT PER FCST
39	2	19.4
42	91	19.0
49	39	17.1
37	1	16.8
35	100	16.7
51	43	15.1
43	15	14.0
48	9	14.0
46	103	12.7
44	72	11.9
40	28	10.4
38	48	9.3
41	39	7.6
45	45	6.5
32	95	5.6

Table 3. Percent Improvement Scores for 1981

But what does this all mean? More importantly, is there ar significant difference in forecaster performance? Or is the variation in forecaster scores due to chance? These questions can be answered by statistically testing the following hypothesis:

"There is no significant difference in forecaster scores. Each of the forecaster scores represents a sample mean obtained from the same population. Variation among forecasters is due to chance."

This hypothesis could be tested by the analysis of variance. This would determine whether or not the variation among forecasters was significant. However, it would not indicate which of the scores was significant.

An alternative test of the hypothesis is the comparison of means (Panofsky and Brier, 1968). From the hypothesis, each forecaster score is assumed to be a sample mean obtained from the same population. The population mean in this case is the station mean, which for 1981 was 12.5%. The hypothesis can then be tested by evaluating the statistic

$$T = (X - M)\sqrt{N} / S$$

where X is the forecaster mean improvement, M is the station mean improvement, S is the station standard deviation, and N is th number of forecasts issued.

This statistic has the Student-t distribution with N-1 degrees of freedom. If the above statistic is determined to be significant, this means that with 95% confidence, we can say the population mean for a particular forecaster is greater than (or less than) the station mean. The forecaster has demonstrated more (or less) skill than the station average.

The above hypothesis was tested for each forecaster in 1981, and the results are listed in table 4.

FCSTR NO.	FCSTS ISSUED	AVG % IPVMT PER FCST	
39	2	19.4	
42	91	19.0 *	
49	39	17.1	
37	1	16.8	
35	100	16.7	
51	43	15.1	
43	15	14.0	
48	9	14.0	
46	103	12.7	
Station	730	Mean = 12.5	Std dev = 23.1
44	72	11.9	
40	28	10.4	
38	48	9.3	
41	39	7.6	
45	45	6.5	
32	95	5.6 *	

Table 4. Percent Improvement Scores for 1981
(* indicates significant departure from station mean)

The results in table 4 are much more meaningful. Two forecasters demonstrated performance during 1981 that differed significantly from the station mean. Forecaster 42 demonstrated above average performance, while the performance of forecaster 32 was below the station average. Forecaster 42 would no doubt be pleased about his performance, while forecaster 32 might be somewhat embarassed. Forecaster 32 would most likely make a determined effort to remove the asterisk alongside his score.

The remaining forecaster scores are not significant, and no conclusions can be reached regarding relative performance. The variation in scores may quite likely be the result of chance-- rather than skill.

It might be noted that both forecasters with significant scores of percent improvement had worked relatively large numbers of forecast shifts. As a consequence of sampling theory, significant differences between means become more apparent with larger sample sizes. But how might a forecaster with a relatively small number of forecasts demonstrate significant skill?

One way to demonstrate skill in a small sample would be to have an extremely high (or low) average improvement score. But an alternative demonstration of skill would be performance that was consistently above the station average.

For example, assume that a forecaster issued only two forecasts each month over the course of a year. Furthermore, assume that his monthly average improvement scores were above the station average for 10 of the 12 months of the year. Has the forecaster demonstrated skill? This question can be answered by testing the following hypothesis:

"The forecaster has not demonstrated skill. Furthermore, his above-average performance for 10 of 12 months is due to chance."

If we assume that all forecasters have equal skill, then the probability that a forecaster's monthly score will be above the station average is .5. Essentially, what we are asking then is: what is the chance of tossing a coin 12 times and obtaining 10 (or more) heads? The probability of such an occurrence can be determined by evaluating the binomial function (Panofsky and Brier, 1968)

$$p(m) = \frac{N!}{m!\ (N-m)!}\ .5^{N}$$

where m is the number of successes and N is the number of trials. The probability of an average forecaster demonstrating above-average performance 10 or more months out of 12 months is 1.9%. Therefore, such performance is significant, and the hypothesis is rejected. The forecaster has indeed demonstrated skill.

7. SUITABILITY OF PERCENT IMPROVEMENT FOR SIGNIFICANCE TESTS

The purpose of significance tests is to reach a valid conclusion with regard to a population of data, by examining only a sample of that data. Are such tests valid when applied to forecaster scores of percent improvement over guidance?

Before most significance tests can be applied, several very important criteria must be met:

(1) The sample must be randomly obtained.

(2) Successive values of sample data must be independent of each other.

(3) The population from which the sample was drawn must b normally distributed.

Samples of percent improvement scores can be assumed randomly obtained. A forecaster might work at any time of the month or year. Furthermore, any advantage gained by working a particular month of the year has been eliminated through the use of standard scores.

Successive values of percent improvement may be considered independent of one another. Percent improvement is virtually independent of guidance and only slightly affected by variability. The effect of variability is most apparent between different months, and this influence has been eliminated.

Average scores of percent improvement may be considered normally distributed. The distribution of _individual_ percent improvement scores is skewed. However, as a consequence of the central limit theorem (Panofsky and Brier, 1968), this skewness is virtually eliminated when _sample means_--rather than individual scores--are examined. The hypothesis of normal distribution was _accepted_ after performing a Chi-square test of normality on sample mean improvement scores for two years of verification data.

Percent improvement scores, when expressed as sample means, are therefore statistically suitable for tests of significance.

8. A COMPUTER PROGRAM FOR COMPARATIVE FORECAST VERIFICATION

The ultimate objective of the investigation outlined in this paper was to develop a verification program that would provide feedback to the forecaster concerning his performance--relative to the forecast staff as a whole. An important question that each forecaster might ask is: What is the "bottom line?" How am I doing compared to the rest of the staff?

After determining that percent improvement over guidance could be used as an unbiased measure of relative forecast performance, a computer program was written to provide the feedback described above. The program uses each forecast issued by this office as data, and provides both a monthly and annual summary of individual and station improvement over guidance. Examples of this output are illustrated in tables 5 and 6.

Each month, and for each forecast, the following data is used as input:

(1) forecaster number

(2) total forecaster absolute error (all stations, all periods)

(3) total MOS absolute error (all stations, all periods)

(4) total number of forecast errors greater than 5 degrees

(5) total number of MOS errors greater than 5 degrees.

IMPROVEMENT OF MOS TEMPERATURE FORECASTS
FEBRUARY 1982

FCSTR NO.	FCSTS ISSUED	% FCSTS IMPROVED	% FCSTS WORSE	% REDUCTION BUSTS > 5 DEG	AVG % IPVMT PER FCST	NO. FCSTS ABV STA AVG
44	3	66.7	0.0	36.4	27.8	2/3
48	2	100.0	0.0	0.0	26.4	2/2
46	6	100.0	0.0	25.7	24.1	4/6
42	11	90.9	9.1	32.5	23.3	9/11 *
41	7	85.7	14.3	20.7	12.8	4/7
51	3	66.7	33.3	33.3	10.0	1/3
STATION	56	71.4	26.8	14.6	MEAN = 9.3	STD DEV = 25.6
35	7	71.4	28.6	9.8	8.2	4/7
38	4	50.0	50.0	11.5	.2	2/4
49	4	50.0	50.0	-10.0	-5.0	2/4
32	9	33.3	66.7	-19.3	-19.5 *	2/9

NOTES:

1. FCSTRS LISTED BY AVG % IPVMT.

2. VALUES FOLLOWED BY '*' ARE STATISTICALLY SIGNIFICANT. (WOULD OCCUR BY CHANCE LESS THAN 5% OF TIME)

Table 5. Example of Monthly Summary

IMPROVEMENT OF MOS TEMPERATURE FORECASTS
MARCH 1981 - FEBRUARY 1982

FCSTR	FCSTS	% FCSTS	% FCSTS	% REDUCTION	AVG % IPVMT	NO. MONTHS
NO.	ISSUED	IMPROVED	WORSE	BUSTS > 5 DEG	PER FCST	ABV STA AVG
42	103	83.5	13.6	35.4	20.2 *	11/12 *
48	11	90.9	0.0	57.9	17.5	3/4
35	96	79.2	18.8	26.8	17.5	7/12
43	12	83.3	8.3	25.3	16.9	2/5
51	35	85.7	8.6	30.9	15.7	6/10
44	75	78.7	17.3	32.2	14.2	6/12
49	40	72.5	22.5	24.3	14.2	6/10
46	107	75.7	23.4	24.3	13.7	5/11
STATION	730	75.2	21.9	25.0	MEAN = 13.3	STD DEV = 24.2
40	28	75.0	25.0	14.7	11.2	3/8
41	42	73.8	26.2	22.0	10.6	5/10
38	39	71.8	20.5	15.9	10.2	4/11
45	38	60.5	36.8	20.7	6.3	4/10
32	104	62.5	35.6	13.8	5.1 *	3/12

NOTES:

1. FCSTRS LISTED BY AVG % IPVMT.

2. MONTHLY (SEASONAL) VARIATION IN AVG % IPVMT HAS BEEN ELIMINATED.

3. VALUES FOLLOWED BY '*' ARE STATISTICALLY SIGNIFICANT. (WOULD OCCUR BY CHANCE LESS THAN 5% OF TIME)

4. MEAN STATION IPVMT OVER MOS IS HIGHLY SIGNIFICANT.

Table 6. Example of Annual Summary

The program was written in BASIC language and runs on the Apple II computer, which is now available in all Southern Region offices The program performs all calculations, including evaluation of th Student-t distribution. The data can be entered quickly, and the output is available in only several minutes.

The output consists of several statistics not yet discussed. The percentage of guidance forecasts improved and made worse by the forecaster are indicated. Also indicated is the percentage of guidance errors of more than 5 degrees that were caught by the forecaster. Errors greater than 5 degrees represent at least a two-category bust, and the elimination of such errors definitely results in an improved forecast.

The program also tests the significance of the station improvement over guidance. For each annual summary ever evaluated by this program, the station improvement over guidance has been highly significant. The forecast staff at WSFO San Antonio have clearly demonstrated skill over guidance, and this skill has resulted in improved forecasts to the public.

9. CONCLUSION

Percent improvement over guidance, unlike mean absolute error, is a relatively unbiased indicator of forecaster performance. It is virtually independent of the quality of guidance available to the forecaster, and is only slightly influenced by temperature variability. This influence can be statistically eliminated, allowin valid comparisons among forecasters.

The comparison of forecasters is not intended to embarrass forecasters. The objective of such comparisons is to provide each forecaster with an honest appraisal of his efforts relative to the forecast staff as a whole, and when deserved, a pat on the back for a job well done. As stated by Panofsky and Brier (1968), such verification procedures can provide a valuable contribution to the quality of forecasts. The existence of a checking scheme, even if imperfect, tends to keep the forecasters more alert and interested in maintaining--and improving--the accuracy of the forecasts.

ACKNOWLEDGEMENTS

I wish to thank Skip Ely, DMIC, and Bill Crouch, AM, WSFO San Antonio for their support and encouragement in preparing this paper. I also wish to thank my wife, Jerilyn, for her support and understanding--particurlarly during the many hours of my absence while at the computer. I especially wish to thank Roger Smith, WSEO, for many stimulating discussions concerning forecast verification. Roger, incidently, is forecaster number 42 in the examples--clearly in a class by himself.

Thomas M. Hicks

REFERENCES

Gregg, G.T., 1969: On Comparative Rating Of Forecasters, ESSA Tech. Memo., WBTM SR-48.

Klugh, H.E., 1974: Statistics: The Essentials for Research, John Wiley & Sons, Inc., 426 pp.

Panofsky, H.A., and Brier, G.W., 1968: Some Applications of Statistics to Meteorology, The Pennsylvania State University, 224 pp.

Roberts, C.F., Porter, J.M., and Cobb, G.F., 1967: Report on the Forecast Performance of Selected Weather Bureau Offices for 1966-67, ESSA Tech. Memo., WBTM FCST-9.

Snedecor, G.W., and Cochran, W.G., 1980: Statistical Methods, The Iowa State University Press, 507 pp.