A Comparison of Temperature and Precipitation Forecasts Issued by Telecasters and the National Weather Service

DENNIS M. DRISCOLL

Department of Meteorology, Texas A & M University, College Station, Texas

(Manuscript received 5 June 1987, in final form 19 September 1988)

ABSTRACT

Late evening weather forecasts by telecasters at major network television stations in seven United States cities, and corresponding forecasts from the National Weather Service, were monitored by meteorology students for 6 months in 1985–86. These forecasts were of temperature and precipitation for the three periods of "tonight," "tomorrow," and "tomorrow night." The accuracy of temperature forecasts was evaluated with three indexes: mean absolute error, root mean square error, and percentage of errors over 10°F. For precipitation, the indexes were Brier score (accuracy) and reliability.

The accuracy of temperature forecasts was not greatly different for the telecasters and the NWS. Three of 20 pairings show a statistically significant difference according to the sign test; this is not much more than would be expected by chance. For precipitation similar results were obtained: only 1 of 20 Brier score pairings is statistically significant. The NWS has higher reliability scores, although no test exists for determining the statistical significance of this difference.

1. Introduction

The routine weather forecasts of the National Weather Service (NWS) that the American public hears, directly or indirectly, are the result of four consecutive steps, each of which culminates in a prediction of the future state of the atmosphere. These are 1) the output of numerical models; 2) guidance, the forecast that results when the current statistical adjunct, MOS, is applied to model output; 3) the zone and city forecasts that local NWS offices prepare subjectively and release to the public; and 4) the forecast which reaches the public. This last forecast may be identical to the third, but an overwhelming majority of Americans now receive their forecasts via television (Ryan 1982), and only about 15% of the telecasters queried in a recent survey (Driscoll 1986) transmit the NWS forecast unmodified. Thus, the forecast that the public hears most often is not that released by the NWS, but one that has undergone some modification by the television weather forecaster (telecaster). In addition, an increasing number of television stations receive their forecasts from private firms, which prepare weather forecasts more or less independently of the NWS.

The question then is, how well does the telecaster—the person who presents a weather forecast on television—serve the public? Are forecasts more accurate (or more skillful) when the telecaster acts as an intermediary between levels 3 and 4? An adequate answer to this question would be possible only with a much more comprehensive study than is reported here; this is offered as a start in that direction. The intent is to record corresponding forecasts of temperature and precipitation for the same areas (those of the NWS and of television weather forecasters), prepare appropriate indexes of accuracy, compare indexes, and, if appropriate, offer suggestions as to how accuracy may be improved.

Although the NWS verifies its forecasts at levels 1, 2, and 3, and independent verifications of NWS accuracy and skill are published periodically (e.g., Murphy and Sabin 1986), nothing in the scientific literature documents the accuracy of forecasts issued by telecasters, nor is there any mention of this subject at the annual conferences on weathercasting (radio and television), at least as deduced from the conference reports in the *Bulletin of the American Meteorological Society.* Of course, telecast accuracy is a sensitive issue, and the industry is highly competitive and very visible. Comparison of one telecaster with another or with the NWS might well be harmful either to the ratings of news programs or to the NWS.

We are aware of two limited studies (Mahoney and Mass, personal communications). Tom Mahoney, Chief Meteorologist at WFRV-TV in Green Bay, Wisconsin, compared his temperature forecasts for 4 months in 1985 with those of the NWS, two private forecasting firms, and two other local television stations. For this period his forecasts ranked first, those

Corresponding author address: Prof. Dennis M. Driscoll, Dept. of Meteorology, College of Geosciences, Texas A & M University, College Station, TX 77843-3146.

of the NWS fourth. In Seattle, the students in an atmospheric sciences class at the University of Washington taught by Professor Clifford F. Mass recorded and compared the forecasts of temperature and precipitation of five television stations for 25 days in the spring of 1984.

2. Data and procedures

We engaged the services of one meteorology student at each of seven widely scattered cities in the United States. Hereafter these cities will be identified by region only; they are NW (Northwest), SW (Southwest), UM (upper Midwest), LM (lower Midwest), GL (Great Lakes), GC (Gulf Coast) and EC (East Coast). According to the "Designated Market Areas" section of the *Spot Television Rates and Data* issue of 13 December 1984 (Standard Rate and Data Service, Inc. 1984), one of these cities (as ranked by number of households) was in the top 10, two were ranked between 11 and 20, one between 31 and 40, one in the 80s, and two between 100 and 150.

For a period of any 5 or more days in a week for any 6 months (not necessarily consecutive) during the period August 1985–May 1986 the meteorology students monitored two forecasts available in the late evening, between 1000 and 1130 LST. The first was that of the telecaster(s) employed at a television station in or near the city in which the student lived. Only stations affiliated with one of the three major networks were chosen, and this choice was made by the meteorology student. The students were instructed to record only forecasts of temperature and precipitation for the periods tonight, tomorrow, and tomorrow night that appeared on the screen, except when verbal comments were clearly meant to supersede the written forecast. The second forecast was that of the NWS which was listened to at the same time, obtained from the weather wire, NOAA weather radio, or the weather channel. The verification data were obtained from the *Local Climatological Data* (LCD) publication for the local NWS station.

3. Verification methods

The intent of this study is to compare the accuracy of the telecaster with that of the NWS at each station, and to summarize these comparisons. No attempt is made to determine skill, or to compare these forecasts with those made by persistence, climatology, or some other standard. A comparison among stations is possible, but we caution that this can be misleading unless interstation differences are accounted for; this is not done here.

a. Temperature

Three objective measures were used to determine the accuracy of temperature forecasts: 1) mean absolute error (MAE), the average of the absolute differences between forecast and observed temperatures; 2) root mean square error (RMS), the square root of the mean of the squared differences; and 3) the percent occurrence of errors in temperature forecasts $\geq 10°F$. Not all forecasts were given as a single number; many, those of the telecaster and NWS alike, were given as implicit intervals (mid-50s), or as explicit intervals (35°–40°F), and it was necessary to equate these with one number. Table 1 shows these equivalents. We followed the current operational practice of expressing temperature in degrees Fahrenheit.

In verifying temperature forecasts the following was assumed about the time periods which apply to each of the three forecasts. "Tonight's low" is the lowest temperature occurring between the telecast (usually about 2200 LST but 2300 LST in the eastern time zone) and 0700 LST except LDT when applicable. "Tomorrow's high" is the highest temperature observed between 0700 and 1800 LST of the calendar day following the telecast, and "tomorrow night's low" the lowest temperature occurring between 1800 LST of the calendar day following that of the forecast and 0700 LST of the day after that.

The verification temperatures for each of these three periods were determined from the local NWS station LCD and the LCD Supplement. These give the maximum and minimum temperatures occurring during the period midnight–midnight, and the temperature every 3 h, respectively. This procedure was not definitive in every case, especially for minimum temperatures. Infrequently, the lowest temperature given in the LCD could, according to the Supplement, have occurred around 0700 LST, or in the hours just before the following midnight, as often happens after a cold front passage during the day. The errors arising in this case—when the wrong decision is made about when the low temperature shown in the LCD occurred—were probably very few and of no appreciable consequence for the accuracy scores we obtained. In addition, the inadequacies of the verification procedures affect both telecaster and NWS equally.

TABLE 1. Specific equivalents of temperature forecasts given as intervals. X can be any whole number (e.g., from −2 to 10). An arithmetic average was assumed for ranges of temperatures when (very infrequently) these were forecast.

Interval	Equivalent
High $X0$'s	$X7.5$
High to mid $X0$'s (mid to high $X0$'s)	$X6.25$
Mid $X0$'s	$X5$
$X0$'s	$X5$
Low to mid $X0$'s (mid to low $X0$'s)	$X3.75$
Low $X0$'s	$X2.5$
High $X0$'s to low $(X + 1)0$'s	$(X + 1)0$
Teens	16
Low teens	14
High teens	18
Midteens	16

b. Precipitation

The forecast the public hears, whether from the telecaster or the NWS, is of the likelihood, or probability, of precipitation (PoP) for a particular period. These statements may be qualitative or quantitative: "rain likely tonight" or "20% chance of rain tomorrow." The appropriate verification measure, then, given that our interest is only in comparing the two forecasts, and that one can assign appropriate numerical values to qualitative statements, is a modification of the score suggested by Brier (1950):

$$\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2 \qquad (1)$$

where N is number of forecasts, P_i the probability forecast, expressed in decimals, and O_i the occurrence (=1) or nonoccurrence (=0) of precipitation. This score is one-half the score Brier initially proposed, and is the measure of PoP accuracy most widely used.

When precipitation forecasts are given as PoPs, i.e., quantitatively, the verification procedure is straightforward. When this forecast is given qualitatively, however, a numerical equivalent is required for this analysis. The difficulty in assigning a number to a qualitative statement lies, first, in distinguishing among all the descriptors those which either indicate or qualify the PoP, and then finding numerical equivalents for them. We began this somewhat arbitrary process by recognizing that these descriptors apply to the following six characteristics of precipitation in addition to an explicit indication or qualification of probability (examples of each are in parentheses): 1) extent of area covered (scattered, spotty); 2) geographical variation (e.g., a higher probability when winds are upslope near mountains); 3) duration (brief, occasional); 4) time of occurrence (rain late tomorrow); 5) amount (heavy, light); and 6) character (sleet, snow, rain). Only the first of these is considered a qualifier of probability, since this fraction of the forecast area is used to arrive at the stated PoP (Curtis and Murphy 1985).

Not all quantifications of the PoP forecasts were arbitrary. We used the equivalents currently in use by the NWS shown in Table 2. This table also shows numerical equivalents for the other descriptors which were considered as indicating or qualifying probability. When one term bridged two or more numerical levels an average was used, as indicated by the braces in Table 2. This means that "few" or "isolated" equated to 10%, and "widely scattered" to 20%, but that "slight chance" meant 15%. A probability of zero was used when there was no mention of precipitation in the forecast. At the other extreme, an unqualified forecast, for example "rain tomorrow," was assigned a value of 90% in keeping with the NWS practice of assigning that value to all such categorical forecasts. Since drizzle and snow flurries do not usually lead to measurable accumulations, their mention was not considered a forecast of precipitation. However, when either of these hydrometeors was added in an "and" statement, the probability was assigned in keeping with the accompanying form of precipitation. For example, in Table 2, a forecast of "rain and drizzle tonight" was assigned a probability of 90%, "possible rain and drizzle tonight" a value of 50%, etc. When an "or" statement was used, the probability corresponding to the accompanying form was halved: "rain or drizzle tonight" becomes 45%, "possible rain or drizzle" becomes 25%. This procedure applied only to calculation of the Brier score. For reliability computations (next paragraph), probabilities ending in 5—and there were only a few of these—were divided between the adjacent categories ending in zero. When sequenced probability statements (either quantitative or qualitative, and a practice of the telecaster only) were employed, a simple average was obtained; for example 30% was assigned to the statement "20% chance of showers tomorrow morning increasing to 40% by tomorrow afternoon."

The second attribute of PoP forecasts to be evaluated is reliability. Forecasts of X percent PoP are reliable if, over the long run, precipitation occurs X percent of the time. For a collective measure of reliability we use

TABLE 2. Numerical equivalents of probability qualifiers.

Probability	Equivalents
	*Current NWS Practice**
0	none
10	slight chance, few, isolated
20	slight chance, widely scattered
30, 40, 50	chance, scattered
60, 70	likely, numerous
80, 90, 100	categorical
	Additional Equivalents Used in this Study
5	very slight chance . . . slight risk of . . . maybe a (sprinkle)
10	sprinkles . . . a few sprinkles . . . risk of sprinkles . . . risk of showers
40	some . . . spotty (showers)
50	possible rain or drizzle . . . periods of light rain or drizzle . . . areas of light rain or drizzle
80	ending
90	lingering . . . developing . . . storm warning . . . periods of

* From *NWS Operations Manual,* Chapter C-11, Section 8.3.5, p. 24. See text for additional details.

$$\frac{1}{n}\sum_{i=1}^{11} n_i(|P_i - \bar{O}_i|) \qquad (2)$$

where P_i is probability value i, $\bar{O}_i$ is the observed relative frequency of precipitation when forecast P_i is issued, n_i is the number of forecasts of P_i, and n is the total number of forecasts. The minimum value possible, zero, occurs when the forecasts are reliable ($P_i = \bar{O}_i$) in all 11 categories.

The two measures of the quality of PoP forecasts utilized here are both desirable characteristics (attributes), but they do not measure the same mathematical property and thus are incommensurate. A perfect Brier score (0) occurs only when all forecasts of zero are followed by no precipitation and those of 100% by precipitation; a prediction in any other category, regardless of the outcome, produces a value greater than zero. The worst possible score is 1.0. When the Brier score is zero the forecast reliability must also be zero (perfect). On the other hand, a perfectly reliable series of PoP forecasts which includes predictions in categories other than zero and 100% produces a Brier score higher than zero. For example, it can be shown that for an equal number of perfectly reliable forecasts in each of the 11 categories, the Brier score is 0.15. The Brier score increases for $\bar{O}_i > P_i$ in the 0%–40% categories and for $\bar{O}_i < P_i$ in the 60%–90% categories, and decreases for $\bar{O}_i < P_i$ (10%–40%) and for $\bar{O}_i > P_i$ (60%–100%). Note that $\bar{O}_i$ cannot be less than P_i at zero percent or greater than P_i at 100%. A series of forecasts of $P_i = 50\%$ results in a Brier score = 0.25 regardless of verification. It can also be shown that when a series of PoP forecasts includes all 11 categories, and each is reasonably close to reliable, the contributions to the Brier score for each category are symmetrical around the maximum contribution at $P_i = 50\%$. An application of this is shown when PoP forecasts are analyzed in section 4b.

4. Results

The data used in this analysis are summarized in Table 3. Nearly all of the forecasts of temperature and precipitation received from the meteorologists were used in the analysis. Rarely was a forecast for one period for either the telecaster or NWS missing; if it were, the other forecast type was used. Usually the number of forecasts for the three periods was the same, so the number of these for any period is the number in Table 3 divided by 3. The choice of months, and of days within the weeks, was made by the student, given the constraints of any 6 months from August to May and any 5 of 7 days weekly, as explained in section 2. Only the student at GL did not make this minimum. At NW, the telecaster did not make a forecast for tomorrow night, so this period was not analyzed.

a. Temperature

The reporting practices for the seven stations were as follows. Five of the telecasters exclusively forecast one number (e.g., 65°F), one gave explicit ranges (low tonight 63° to 67°F), and the seventh used a variety of formats, i.e., one number, implicit ranges (mid-60s), and explicit ranges. The preference of NWS forecasters was for implicit ranges, although occasionally they used one number or, less frequently, explicit ranges.

Table 4 summarizes the temperature analysis. With only one exception, GC, the three accuracy indicators invariably increase through the three periods, i.e., forecasts become less accurate. The number of months out of six for which the telecaster and NWS were superior to the other is included to give some idea of the significance of the telecaster–weather service (TC–WS) differences in mean absolute error (MAE) and root mean square errors (RMS) scores. At SW, for example, a fairly substantial difference is apparent in both. Conversely, a relatively small difference in MAD and RMS scores—UM is a good example—indicates an almost even division of months in which the telecaster and the NWS have the better score.

The percentage of forecasts in error by ≥ 10°F varies principally with climate. The lowest values occur when day-to-day weather changes are comparatively small (NW, SW), the highest in the most variable climates (UM, LM, EC). Here GC is somewhat of an exception;

TABLE 3. Data analyzed: A, S . . . A, M are the months August through May; × indicates the months with data included in the study.

	Period											
	1985					1986					Total forecasts	
City	A	S	O	N	D	J	F	M	A	M	Temp	Precip
Northwest (NW)			×	×	×	×	×	×			318*	321*
Southwest (SW)	×	×	×	×	×	×					465	468
Upper Midwest (UM)		×	×	×			×	×	×		393	396
Lower Midwest (LM)		×	×	×	×	×	×				381	387
Great Lakes (GL)		×	×				×	×	×	×	348	351
Gulf Coast (GC)		×	×	×	×	×	×	×			477	486
East Coast (EC)		×	×	×	×	×	×				393	396

* Forecast for tomorrow night not included.

TABLE 4. Results of temperature analysis. Explanation of abbreviations: TC: telecaster, WS: National Weather Service, RMS: root mean square error, MAE: mean absolute deviation, ≥10° = percent of errors greater than or equal to 10°F. Numbers in parentheses are months (out of 6) in which that forecaster had a lower score than the other; sums less than 6 indicate that in one month or more the two tied to three significant figures. Underlining indicates that one forecaster is statistically significantly better, at 5%, than the other for that period.

	Tonight's minimum			Tomorrow's maximum			Tomorrow night's minimum		
Station-Forecaster	RMS	MAE	≥10°	RMS	MAE	≥10°	RMS	MAE	≥10°
Northwest (NW)									
TC	2.87 (5)	2.24 (6)	0.6	3.51 (2)	2.75 (3)	1.9	—	—	—
WS	3.47 (1)	2.84 (0)	1.2	3.45 (4)	2.72 (3)	0.6	4.15	3.38	1.9
Southwest (SW)									
TC	3.21 (6)	2.38 (5)	1.3	3.47 (6)	2.76 (5)	0.6	3.89 (6)	2.97 (5)	2.6
WS	3.76 (0)	2.78 (1)	0.6	3.90 (0)	3.24 (1)	1.3	4.53 (0)	3.45 (1)	4.5
Upper Midwest (UM)									
TC	4.23 (3)	3.51 (4)	3.8	5.24 (3)	4.05 (3)	6.9	5.44 (3)	4.48 (4)	6.9
WS	4.26 (3)	3.45 (2)	3.8	5.18 (3)	3.99 (3)	7.6	5.58 (3)	4.65 (2)	8.4
Lower Midwest (LM)									
TC	3.65 (5)	2.85 (5)	2.4	5.56 (3)	4.35 (3)	10.2	6.10 (4)	4.98 (4)	12.6
WS	4.15 (1)	3.13 (1)	3.9	5.72 (3)	4.45 (3)	10.2	6.90 (2)	5.30 (2)	15.7
Great Lakes (GL)									
TC	4.31 (3)	3.22 (3)	5.2	4.79 (3)	3.81 (3)	5.2	5.44 (3)	4.28 (2)	7.8
WS	4.39 (3)	3.28 (3)	3.4	4.88 (3)	3.89 (3)	6.0	5.28 (3)	4.12 (4)	7.8
Gulf Coast (GC)									
TC	5.44 (0)	4.47 (0)	10.1	4.28 (2)	3.00 (3)	6.3	6.26 (2)	5.13 (2)	17.6
WS	4.42 (6)	3.44 (6)	4.4	3.95 (4)	3.06 (2)	3.1	5.75 (4)	4.45 (4)	10.7
East Coast (EC)									
TC	3.87 (3)	3.15 (3)	1.5	4.97 (3)	3.75 (2)	9.2	5.61 (2)	4.26 (3)	9.9
WS	4.07 (3)	3.32 (3)	1.5	4.31 (3)	3.47 (4)	3.8	5.15 (4)	4.07 (3)	7.6

relative to other forecasters, the telecaster did a poor job of predicting minimum temperatures. Note that his percentage of large errors averages about double that of the NWS.

To determine which of the two forecasters at each station was the better for the study period we performed a sign test (Bhattacharyya and Johnson 1977). For each forecast of temperature, for each period (total forecasts for all three periods are shown in Table 3), we determined which of the two was more accurate. Identical forecasts, and those that were equidistant from occurrence (e.g., forecasts of 75 and 77 followed by an occurrence of 76) were excluded. We rejected the hypothesis that the NWS forecaster and the telecaster are equal at 5% (two-tailed test). Table 4 shows that the telecasters at NW for the period tonight and at SW for tomorrow night, and the NWS at GC for tonight, were statistically significantly better than their counterparts according to this test. Table 4 also shows very little difference in the RMS and MAE scores at UM and GL. The possibility that forecasts are simply being relayed at these stations is considered in the discussion section.

b. Precipitation

As was the case with temperature, the reporting practices of the various telecasters and the NWS with regard to precipitation are of interest. These varied considerably among the telecasters. PoP forecasts (when there was any mention of precipitation) were never stated at two stations, were mentioned infrequently at two stations, usually at one station, and always at one. At the seventh station, PoP forecasts were rare and the verbal statements tended to be categorical, i.e., with very few qualifiers. In general however, the telecasters who relied least on numerical probabilities were most likely to use qualifiers that are different from those used by the NWS (the qualifiers equated to percentages as noted in Table 2). The NWS forecasters were much more uniform in this respect: five stations almost always used PoP forecasts, the sixth always did, and the seventh never did. However, their qualifiers were always equivalent to numbers, as required by the forecasting manual and discussed earlier (see Table 2).

Forecast verification periods for the PoP forecasts in this analysis were defined to be consistent with the commonly accepted definitions of tonight, tomorrow, and tomorrow night, i.e., 1800 to 0700 LST, 0700 to 1800 LST, and 1800 to 0700 LST of the following day, respectively. All times are local standard except for the station days to which daylight saving time applied. It was of course necessary to exclude from the "tonight" period (but not "tomorrow night") the hours 1800 to 2200 LST (2300 LST in the East), since these hours occur before the forecast is aired. A period was con-

sidered to have precipitation if a measurable amount (≥0.01 inches) was recorded during the three periods specified above.

Table 5 shows Brier scores for both telecaster and NWS at the seven cities for each of the three periods, averaged for the 6 months of study. Table 3 shows the number of forecasts for each station. As noted in section 3b the best Brier score is the minimum, zero, which occurs when all forecasts are categorical and the predicted event (precipitation or no precipitation, equivalent to 100% and 0%, respectively) always materializes. The worst and highest score possible, unity, also occurs when nothing but categorical forecasts are issued, but the unforecast event always materializes. Table 5 shows that, in practice, the values range from about 0.050 to 0.250. There is, of course, a climatological bias, with the best scores most likely to occur in the driest areas (SW) and vice versa (NW). Again, no correction is made for this aspect of the forecasts because our interest is only in comparing the telecaster and the NWS at each location.

As was the case with temperature, precipitation forecasts become less accurate as the period between release and verification (lead time) increases. Exceptions occur at LM, for both telecaster and NWS, at GL for the NWS, and at EC for the telecaster.

The statistical significance of the differences in Brier scores between telecaster and NWS at each station, for each of the three periods, was determined as shown in the Appendix. Table 5 shows that only one of the 20 pairings is significantly different; this is the number expected by chance at the 5% level. The reliability indexes (expression 2) of Table 5 show that at six of the seven stations, the NWS has the lower (better) score. We know of no test to determine the significance of the difference of these indexes.

Figure 1 shows more detail regarding the reliability aspect of PoP forecasting. Note the disparate practices of the two groups in terms of utilization of the 11 PoP categories. Excluding 100%, only two of the telecasters used all of the remaining ten categories, either explicitly or as qualifiers with numerical equivalents (Table 2), while one failed to use four categories, and the remaining four telecasters ignored three of them during the study period. On the other hand, forecasters at all but one of the NWS facilities used at least nine of these ten categories. Of course, the exclusion of some categories may be due only to the method of conversion of descriptive to quantitative forecasts (Table 2), which prevents some probability values (e.g., 30%, 50%, 60%, 100%) from ever being used. In terms of overall reliability comparisons, the NWS has better scores, i.e., a greater proportion of points which fall near the diagonal. This is most apparent at NW, GL, and EC. Underforecasting of probability values predominated at UM (both TC and WS), and at GL–WS, while the PoP

TABLE 5. Results of precipitation analysis. Underlining indicates that the difference between the Brier scores at that station-period is statistically significant at five percent. (See Table 4 for explanation of numbers in parentheses.)

	Brier score ($\times 10^3$)				
Station-Forecaster	Tonight	Tomorrow	Tomorrow night	Average	Reliability ($\times 10^2$)
Northwest					
TC	166 (2)	205 (2)	—	186	15.1
NWS	152 (4)	163 (4)	—	158	8.7
Southwest					
TC	51 (2)	63 (5)	91 (3)	68	5.5
NWS	45 (4)	70 (1)	103 (3)	71	7.0
Upper Midwest					
TC	92 (3)	131 (1)	146 (3)	123	10.0
NWS	92 (3)	120 (4)	147 (2)	120	8.0
Lower Midwest					
TC	112 (4)	97 (4)	135 (1)	115	9.7
NWS	113 (2)	91 (2)	107 (5)	104	7.8
Great Lakes					
TC	149 (1)	179 (2)	246 (0)	191	19.1
NWS	97 (5)	144 (4)	130 (6)	124	6.7
Gulf Coast					
TC	64 (3)	90 (3)	106 (3)	87	7.5
NWS	51 (3)	94 (3)	92 (3)	79	4.6
East Coast					
TC	81 (0)	138 (0)	122 (1)	114	10.4
NWS	53 (6)	94 (6)	96 (5)	81	4.7

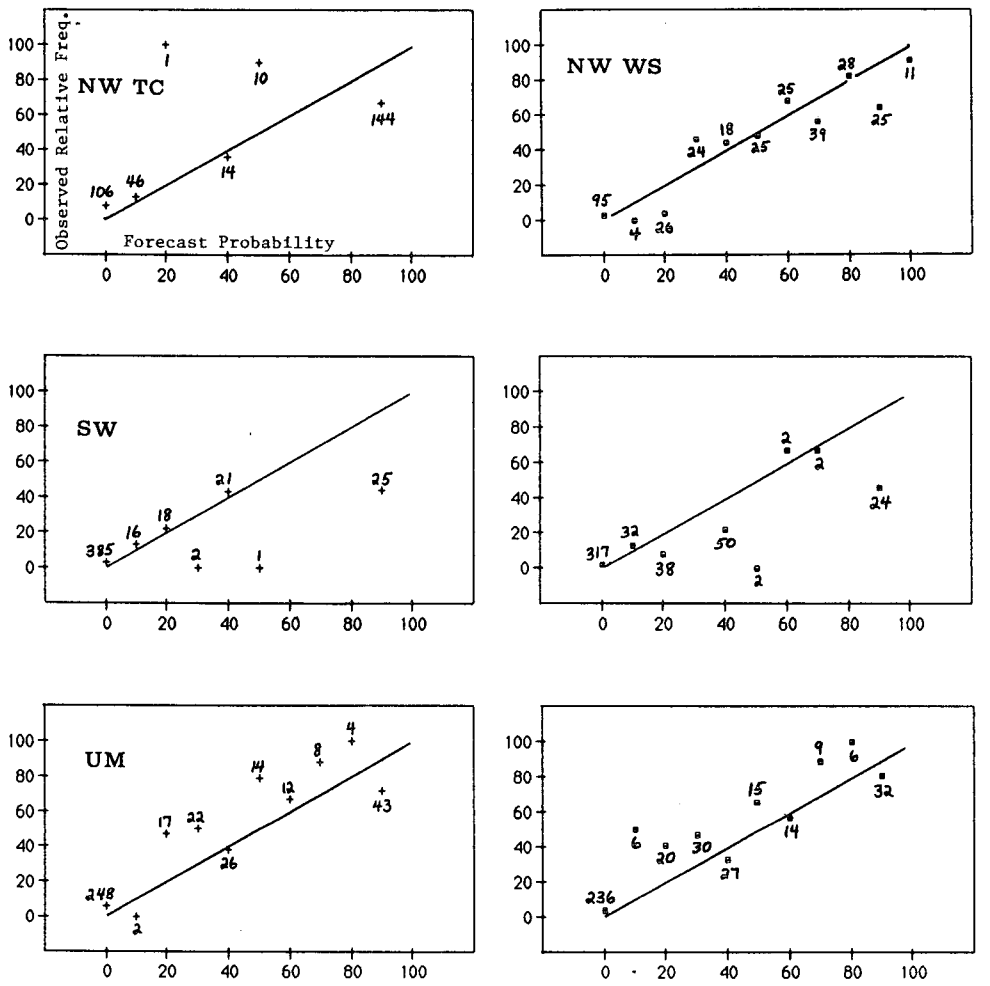

FIG. 1. Reliability diagrams. Stations are indicated by letters (see text); telecaster's diagrams are to the left, those of the National Weather Service are to the right. Numbers are number of forecasts in that particular PoP category.

was overestimated for many values at LM–WS, GC (both forecasters), and at EC–WS.

An overall index of reliability for each of the 11 PoP categories is shown near the bottom of Table 6. The NWS has lower (better) scores in seven of the ten categories in which comparisons are possible. For both groups the values among categories 10–80 are not greatly different, but the reliability for 0% is lower, and that for 90% is considerably higher, than for the intermediate categories. In other words, PoP forecasts of no precipitation (0%) are more reliable than those for the middle values, and considerably more reliable than those of precipitation (90%). This last distinction is even more marked for the telecaster (28 vs 21).

Table 6 was developed to show how the 11 PoP categories contribute to the Brier score. Recommendations for bettering this score are possible if it can be shown that some categories contribute substantially more than others. This table shows that PoP forecasts in the categories 0%, and 90%–100% contribute most. When the 11 categories are ranked according to the contribution of each to the Brier score, these two categories rank either first or second in all but two (0%) and five (90%–100%) of the 14 possible combinations of stations and forecasters.

Another perspective on the differences between telecasters and NWS forecasters, and on how the Brier score is contributed to differentially by the PoP categories, is shown by the bottom three lines of Table 6. These show the Brier score per forecast, averaged separately for all telecasters and all NWS forecasters, and the same for an "ideal" series of forecasts, one in which for all categories $O_i = P_i$ and the n_i are equal. This perspective has the advantage of correcting for the large number of forecasts in the 0% and 90%–100% categories, which comprise 72% of all forecasts. A comparison of the bottom three rows of Table 6 shows that in the middle nine categories, the deviations from this

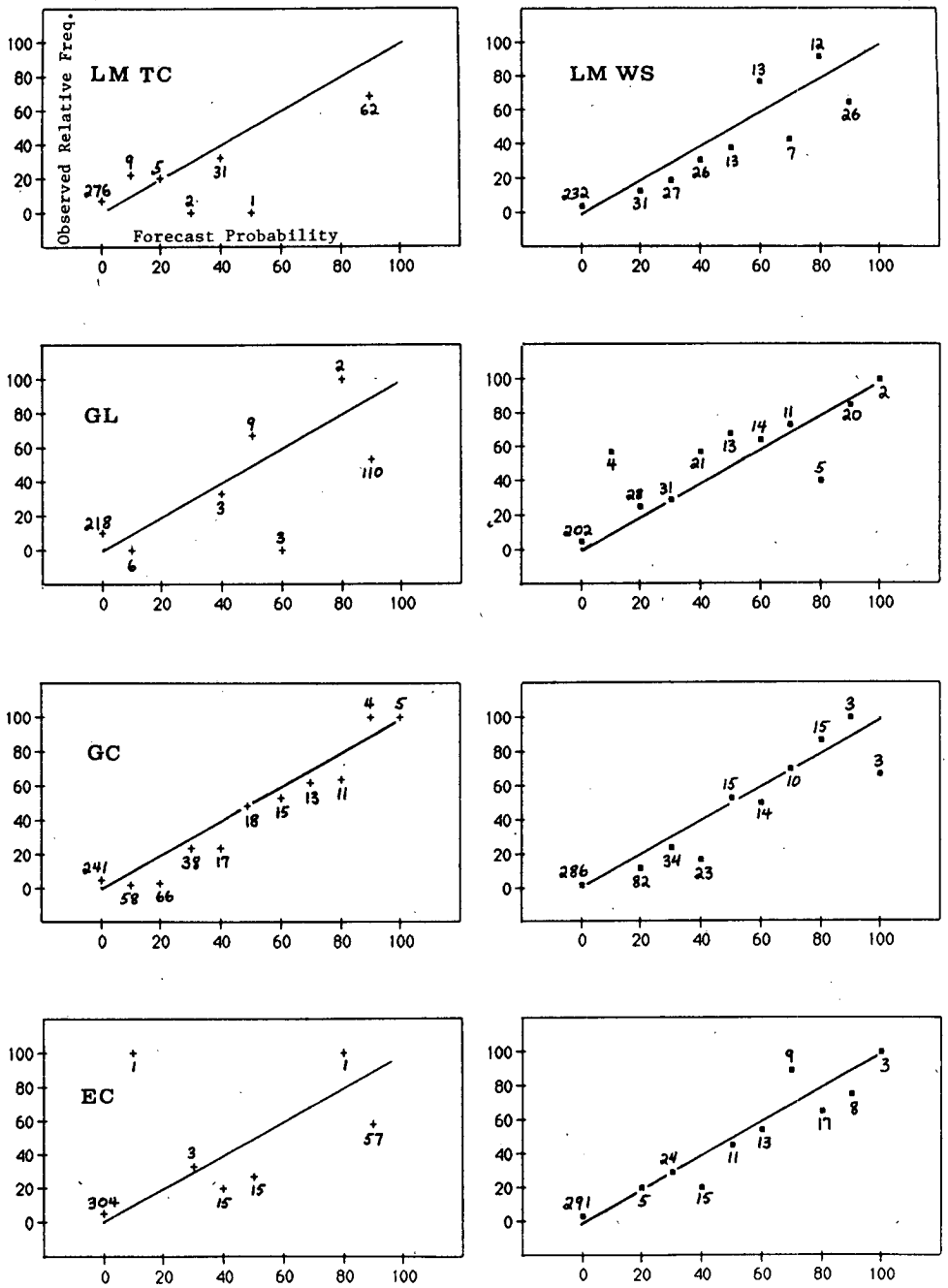

FIG. 1. (*Continued*)

"ideal" state are mostly inconsequential, but for 0%, and especially for 90%, they are quite large.

From this finding it is clear that to improve their Brier scores the forecasters for whom this applies must concentrate on improving categorical forecasts, especially those of precipitation (90% and 100%). For example, two-thirds of the telecaster's Brier score at NW was contributed to by categorical forecasts of precipitation of which two-thirds materialized (see Fig. 1 and Table 6). If we assume that only 107 categorical forecasts had been issued (instead of the 144 actual), and that the remaining 37 forecasts had been, say, 50%, then the reliability for the 90% category becomes perfect and the Brier score drops from 0.186 to 0.120. And if 90% probability had been forecast only for the 96 events in which precipitation materialized, and the 48 re-

TABLE 6. Percentage contribution of PoP categories to the total Brier score.

Station-Forecaster	PoP categories										
	0	10	20	30	40	50	60	70	80	90	100
Northwest											
TC	14	09	01	—	05	04	—	—	—	67	—
NWS	06	00	03	13	09	12	11	20	08	15	02
Southwest											
TC	31	05	10	01	16	01	—	—	—	36	—
NWS	15	10	10	—	30	01	01	01	—	31	—
Upper Midwest											
TC	28	00	11	13	12	07	05	02	00	20	—
NWS	22	05	13	18	13	08	07	03	01	11	—
Lower Midwest											
TC	40	04	02	00	16	01	—	—	—	37	—
NWS	25	—	09	11	14	08	07	06	03	18	—
Great Lakes											
TC	31	00	—	—	01	03	02	—	00	63	—
NWS	25	04	12	15	13	07	07	05	05	06	04
Gulf Coast											
TC	28	03	09	17	08	11	09	08	07	00	00
NWS	16	—	24	16	12	10	09	05	05	00	03
East Coast											
TC	36	02	—	02	07	09	—	—	01	45	—
NWS	31	—	03	16	09	09	10	04	13	05	00
Reliability ($\times 10^2$)											
TC	5.8	6.8	15.9	11.9	8.0	20.6	12.3	11.8	17.6	28.2	—
NWS	3.2	12.3	10.0	8.5	14.3	9.7	7.9	12.8	10.6	21.0	10.4
Brier score (per forecast)											
TC	56	81	129	215	226	250	253	204	173	315	0
NWS	33	163	130	217	222	250	235	226	163	254	105
"Ideal"	0	90	160	210	240	250	240	210	160	90	0

maining forecasts had again been made at 50%, the Brier score would have fallen to 0.100.

5. Discussion

This investigation was designed so that the comparison would be equitable, i.e., as much as possible no advantage would be given to either forecaster. There are some circumstances, however, which suggest that one or the other may have a slight advantage. First, we acknowledge that the matter of timing can be crucial in some situations; e.g., the effect of the time of a cold front passage on daily minimum and maximum temperatures. The telecaster's forecast is up to the minute because he or she has been able to consider all information available up to air time. Even for the relatively few telecasters who simply transmit the NWS forecast there is always the opportunity to make last minute changes as circumstances warrant, for example after observing that rain has already begun to fall. This opportunity is most likely to affect forecasts for the first of the three periods. The NWS forecast available at the time of the telecast was released during the period 2100 to 2200 LST (*NWS Operations Manual,* 1979, chapter C-11, section 6.3, p. 7), and for 2300 LST news programs (eastern time zone only) it is already an hour and a half old.

Another circumstance which may favor one forecaster over the other involves the areas (or points) for which each forecasts, and their perception of where—if at all—the forecasts will be verified. Forecasts by the NWS are verified against observations taken at an NWS or Federal Aviation Administration (FAA) facility within the forecast area. The telecaster's work is of course not routinely verified, but if he or she thinks in terms of a verification point it is likely to be the same facility, and thus there is no inequity. This is because the past weather data which the telecaster displays are likely to be from that facility, and it would be to the telecaster's disadvantage to have different forecast and verification points. Of course, if the telecaster "aims" his or her forecast for a population center (e.g., the

downtown area), and there are systematic differences in climate between that area and the NWS or FAA facility (e.g., a suburban airport), the telecaster would be at a relative disadvantage to assume that the NWS forecasts are for the verification point and not the population center.

Another possible source of inequity arises from what might be different perceptions (or interpretations) of the objective (numerical) equivalents given to the subjective descriptors. Perhaps the telecaster's equivalent of, say, "low forties," or "spotty," or "risk of" does not correspond to 42.5°, 40%, and 10%, respectively. The telecaster might then be at a disadvantage compared to the NWS, for whom these are the required equivalents (U.S. Dept. of Commerce, 1979).

Finally, the PoP forecasts of the NWS and the telecaster might be different because of various perceptions of what constitutes a precipitation event. For the NWS this is defined as the standard one-hundredth of an inch or more, but it is possible that the telecaster thinks—perhaps subconsciously if not consciously—that less than a measurable amount constitutes such an occurrence. Drizzle and flurries might then be considered events which verify as precipitation. To see what effect this redefinition has on the Brier score, we reexamined all categorical forecasts of precipitation by the telecaster at NW. Of the 144 forecasts, 96 (67%) were followed by measurable precipitation (Fig. 1). Of the remaining one-third, about half were followed by a trace sometime during the verification period. Had traces verified as precipitation, then, the percent correct in this category would have improved from 67% to 83%. The Brier score would also improve, but no estimate is possible without knowing how traces were distributed among the categories other than 90%–100%.

In summary, of all possible sources of bias we acknowledge that the several factors noted above might, in extreme circumstances, give one forecaster or the other an advantage. It seems unlikely, however, that these factors, acting singly or in combination, would make an appreciable difference in the conclusions resulting from this research.

Finally, we consider whether two truly different forecasts are being compared in this study. It seems very unlikely that the telecaster's forecast is made completely independently of that of the NWS; at the very least, he or she will check the forecast of the NWS, if only out of curiosity. At the other extreme, it has been shown that many telecasters, especially those in the smaller markets, act simply as a conduit for the NWS forecast, repeating it verbatim, or nearly so. According to Driscoll (1986) only 6% of the telecasters queried give no consideration at all to the NWS forecast.

We observed during this study that at some stations, and more often than might be expected by coincidence, the telecaster's forecast appeared to be simply a rewording of that of the NWS; for example, changing a forecast of "low tonight in the low forties" to "low tonight 42," or changing a PoP forecast from qualitative to quantitative or vice versa. It seems likely that if the telecasters we studied did this the scores would be identical, or nearly so. An examination of the actual "side-by-side" temperature forecasts, and of Table 2, shows that at the five stations NW, SW, LM, GC and EC, the differences are considerable, while at UM and GL they are quite small. For precipitation (Tables 5 and 6 and Fig. 1), the largest differences occur at NW, GL and EC, the smallest at SW, UM, and GC. From this evidence it appears likely that the telecaster at UM is simply passing on the NWS forecast or making negligible revisions. At the remaining stations the differences seem to be significant, i.e., it appears that the telecasters at these stations either prepared their own forecasts or made meaningful changes to those of the NWS.

6. Summary

The late evening forecasts of television weather forecasters at seven major network stations were compared with simultaneous forecasts of the National Weather Service. To obtain these data a meteorology student in each of the metropolitan areas logged forecasts of both for at least 5 days out of 7 for 6 months. Appropriate measures of accuracy for temperature and precipitation (PoP forecasts) were calculated from 6 months of observations for the three periods of tonight, tomorrow, and tomorrow night.

For this 6 month period, the accuracy of temperature forecasts was not greatly different for three of these pairings, for three the telecaster was more accurate, and at the remaining station the NWS was more accurate. For three station periods, the differences were statistically significant. The PoP forecasts were evaluated with two measures, Brier score (for accuracy) and reliability. At two of the stations, neither forecaster did better than the other, but at five, the NWS Brier scores were appreciably lower (better). The differences in Brier scores were significant for only one station period. The reliability rankings were even more markedly in favor of the NWS.

Acknowledgments. I wish to thank the National Aeronautics and Space Administration for the financial support of this work.

APPENDIX

Determining the Statistical Significance of Brier Score Differences between Telecaster and NWS

Are the observed differences between the scores of the telecaster and NWS at each city greater than could

have occurred by chance? To find the distribution of these differences under the null hypothesis—that the observed differences could have arisen by chance—we conducted a Monte Carlo procedure. A random number generator was utilized to obtain the values of the ten PoP categories (0.0, 0.1, 0.2, . . . , 0.9), with an equal probability of each; both 90% and 100% (0.9, 1.0) were considered a categorical forecast of precipitation (see section 3b). For each forecast situation, two of these random numbers were chosen, one for the telecaster, the other for the NWS. A partial score was then calculated by squaring the number for each. This is the same as subtracting zero from each value before squaring, as is done when precipitation does not follow the forecast. This procedure was repeated until 100 forecasts had been made for each forecaster, at which point the Brier score was calculated for each, and the difference between the two scores also determined. For this sample size, 100, the process was repeated for a total of 100 differences. The entire procedure was then repeated, except that the value 1 was subtracted from each randomly generated probability before squaring (to simulate the occurrence of precipitation following the forecast). This produced another sample of 100 differences of sample size 100.

This entire process was repeated to obtain 100 forecasts for each of the additional sample sizes 150, and 200, so that we had an interval of sample size (100 to 200) which spans the sample sizes used in this study. All of the differences in Brier scores for each sample size were then arrayed from least to greatest, and the 95% values determined. For the nonprecipitation cases these are 0.055, 0.045, and 0.044 for samples of size 100, 150, and 200, respectively. For the precipitation cases they are 0.066, 0.053, and 0.045.

It might seem that the distribution of differences between Brier scores would be the same for both eventualities, precipitation or no, and that it would not be necessary, therefore, to conduct evaluations for both. However, the mean Brier score for an equal number of forecasts in each of the ten PoP categories when precipitation is not observed is 0.285; when it is observed, the Brier score is 0.385. With a greater range possible for the latter, it follows that the difference between scores also will be greater, on average, and this must be incorporated into the procedure. As noted in the paragraph above, the Monte Carlo procedure verifies this.

To simulate the random process for an individual pairing, the 95% limits for that pair were calculated by weighting the occurrence of precipitation and nonprecipitation in that particular period. For example, for NW for the period tonight, where the values for sample size 150 apply, and noting that for that period there were 60 occurrences of precipitation of a possible 160, the 95% limit is $[100(0.045) + 60(0.053)]/160 = 0.048$. Since the observed difference for NW tonight is only 0.014, it is not statistically significant.

Repeating this entire procedure for the other 19 pairings shown in Table 5, and interpolating between 95% values according to sample size as necessary, revealed a significant difference only at GL for tomorrow night.

REFERENCES

Bhattacharyya, G. K., and R. A. Johnson, 1977: *Statistical Concepts and Methods.* Wiley and Sons.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Curtis, J. C., and A. H. Murphy, 1985: Public interpretation and understanding of forecast terminology: Some results of a newspaper survey in Seattle, Washington. *Bull. Amer. Meteor. Soc.*, **66**, 810–819.

Driscoll, D. M., 1986: A survey of the use of National Weather Service forecasts by television weather forecasters in the United States. *Wea. Forecasting*, **1**, 155–163.

Murphy, A. H., and T. E. Sabin, 1986: Trends in the quality of National Weather Service forecasts. *Wea. Forecasting*, **1**, 42–55.

Ryan, R. T., 1982: The weather is changing . . . or meteorologists and broadcasters, the twain meet. *Bull. Amer. Meteor. Soc.*, **63**, 308–310.

Standard Rate and Data Service, Inc., 1984: *Spot Television Rates and Data*, **66**(12).

U.S. Dept. of Commerce, 1979: *Operations Manual*, C-11. [Available at National Weather Service stations.]