Verification of Public Weather Forecasts Available via the Media

Harold E. Brooks, Arthur Witt, and Michael D. Eilts
NOAA/ERL/National Severe Storms Laboratory, Norman, Oklahoma

ABSTRACT

The question of who is the “best” forecaster in a particular media market is one that the public frequently asks. The authors have collected approximately one year’s forecasts from the National Weather Service and major media presentations for Oklahoma City. Diagnostic verification procedures indicate that the question of best does not have a clear answer. All of the forecast sources have strengths and weaknesses, and it is possible that a user could take information from a variety of sources to come up with a forecast that has more value than any one individual source provides. The analysis provides numerous examples of the utility of a distributions-oriented approach to verification while also providing insight into the problems the public faces in evaluating the array of forecasts presented to them.

1. Introduction

The purpose of a weather forecast should be to help people make better weather-information-dependent decisions. The public can obtain weather forecasts from numerous different sources, typically from a government weather service and from the news media. To optimize weather-information-dependent decisions, one obviously would want information that would help them get the most value out of the forecasts. Although the relationship between quality of forecasts and the value of forecasts is complex (e.g., Murphy 1993; Roebber and Bosart 1996), the quality of the forecasts represents a reasonable starting point. Unfortunately, information on the quality of public weather forecasts is difficult, if not impossible, to obtain.[1]

Driscoll (1988) discussed the relationship of televised weather forecasts to those available from the National Weather Service (NWS) from a small number of sites around the United States for forecasts of lead time up to 36 h. He found that the accuracy of temperature forecasts and probability of precipitation (PoP) forecasts was not greatly different for the telecasters and the NWS. Thornes (1996) showed results of a study of the accuracy of public forecasts in the United Kingdom, but it was based primarily on one verification parameter, the “percent correct.” As Murphy (1991) pointed out, the large dimensionality of the verification problem means that single measures of forecast quality can be misleading. Brooks and Doswell (1996) illustrated this idea with an example of the information available from what Murphy and Winkler (1987) described as a *distributions-oriented* approach to forecast verification.

To help fill (ever so slightly) this vast data void, we set out to record and verify public weather forecasts for the Oklahoma City area for a 14-month time period. Our purpose here is to illustrate some aspects of the differences in the forecast sources for a single location. In passing, we will show the utility of using information from more than one source to produce a forecast with more information (Brown and Murphy 1996). The analysis is by no means comprehensive, but it is illustrative of the power of diagnostic fore-

[1] In general, there is a complex relationship between the quality and value of forecasts (Murphy 1993), but analysis of the quality is a reasonable place to start. We plan to carry out an experiment using a model of electrical utility load forecasting to consider the value for at least one user.

Corresponding author address: Harold Brooks, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: brooks@nssl.noaa.gov
In final form 19 May 1997.

cast verification techniques to provide insight into the forecast process both for the user and the forecaster.

2. Analysis procedures

Forecast data were collected from five different sources: evening forecasts from the three network TV stations, a daily newspaper with early morning delivery, and the Oklahoma City NWS.[2] Each source issues a local forecast for Oklahoma City for at least 5 days in advance. The NWS forecast used was the local forecast issued around 1600 LT. The TV station forecasts used were those presented during the late afternoon/evening newscasts and were tape recorded. Maximum and minimum temperature forecasts (up to 5 days ahead) were evaluated for all five sources. Precipitation forecasts were evaluated only for those sources that produced numerical 24-h PoPs for day 1 through day 7 (two media sources). Because there is no way to assign numerical values to such "forecast products" as a single graphic of a cloud with a few raindrops underneath it, for example, or phrases such as "kind of crummy" or "hopefully, rain" or "maybe, even rain," we are unable to verify the other media forecast sources for precipitation. The data collection period ran from 4 January 1994 to 6 March 1995, with complete forecasts for all sources out to 5-day lead time collected on 338 days, and out to 7-day lead time for the two sources producing 7-day PoP forecasts on 321 days. Verification data came from the observations at the Oklahoma City airport (OKC). Maximum temperature forecasts were verified for the time period from 1200 to 0300 UTC (UTC = local standard time + 6 h). Minimum temperature forecasts were verified for the time period from 0000 to 1500 UTC. In the following sections, when measures are presented from more than one source (for comparison purposes), only those days when forecasts were recorded from all the sources were used.

It is important to note that it is relatively easy for "weather-interested" forecast users in the Oklahoma City market to receive forecasts from all of the sources described for any individual day. Many of the local radio stations use forecasts produced either by a television station or by the NWS. In general, the weather segments on the local news begin at slightly different times, so that simply by changing channels at the appropriate times, the forecasts can all be seen during the 1800 LT news broadcast. In addition, one station repeats the forecast from 1800 LT news during a 1830 LT broadcast, and another station repeats its most recent newscast continuously on a cable TV channel available on basic cable television stations throughout the Oklahoma City area.

3. Measures-oriented verification

a. Temperature

We have examined forecast quality using measures-oriented performance statistics [see Murphy and Winkler (1987) and Brooks and Doswell (1996) for a discussion of measures-oriented verification and Wilks (1995) for definitions] for the entire period of record (Table 1). One result consistent for all forecast sources is (as expected) that accuracy decreases as forecast lead time increases. Another result is that there are significant relative differences in accuracy among the different forecast sources.[3] For minimum temperatures, forecast source (FS) 2 has the lowest mean absolute error (MAE) for all periods, while FS 4 has the highest MAE for all periods. For maximum temperatures (Table 1b), there are significant differences for the first time period (day 1), with smaller differences for the other time periods. Once again, FS 2 had the lowest MAE for all time periods, except for day 4. All forecast sources had a consistent warm bias to their temperature forecasts, especially FS 4.

The development of a forecast that is the mean of all five sources (MEAN) leads to a forecast that has

[2] We have no way of knowing how independent the different forecast sources are. Presumbably, the media sources use the NWS forecast as an input into their forecasting process or, at the very least, look at the same numerical guidance products that are available to NWS forecasters, but we cannot know that for certain. We wrote letters to each of the media sources asking questions about their procedures and forecast descriptors, but received only one reply. Therefore, we have had to make interpretations of some aspect of the forecasts, particularly the meaning of PoP in the media forecasts. When no answer was received, we assumed they used the same definition as the NWS, although the phrasing of the forecasts implies that a different time period (24 h for the media and 12 h for the NWS) is used in the forecasts. Based on the characteristics of the forecast PoP, we do not believe this decision has a significant impact on the interpretation of the forecasts. Nevertheless, the appearance of undefined terms represents a dilemma for forecast users.

[3] For the protection of all concerned, the four media sources have been assigned numbers 1–4 randomly.

TABLE 1. Overall measures-oriented performance results for temperature forecasts for all 338 days with forecasts for all sources (FS 3) at all lead times. DA is days ahead, MAE is mean absolute error, T is total for all 5 days. Errors are in °F.

DA	FS 1 MAE	FS 1 Bias	FS 2 MAE	FS 2 Bias	FS 3 MAE	FS 3 Bias	FS 4 MAE	FS 4 Bias	NWS MAE	NWS Bias	Mean MAE	Mean Bias
(a) Minimum temperature												
1	2.9	0.2	2.7	0.3	2.8	0.4	3.2	1.0	2.9	0.6	2.7	0.5
2	3.9	0.1	3.4	0.3	3.7	0.5	4.0	1.4	3.6	0.9	3.4	0.6
3	4.9	0.0	4.2	–0.1	4.6	0.2	5.0	1.5	4.8	0.1	4.2	0.3
4	5.4	0.0	4.9	–0.1	5.1	0.2	5.7	1.4	5.5	0.0	4.9	0.3
5	5.5	0.1	5.3	0.0	5.6	0.2	6.1	1.4	5.8	0.1	5.3	0.3
T	4.5	0.0	4.1	0.1	4.3	0.3	4.8	1.4	4.5	0.3	4.1	0.4
(b) Maximum temperature												
1	3.9	0.3	3.6	0.2	3.8	0.6	4.2	0.7	3.8	0.6	3.7	0.5
2	5.2	0.1	4.8	0.1	4.8	0.8	5.1	0.9	4.8	0.8	4.7	0.6
3	6.2	–0.3	5.8	–0.1	5.9	0.5	5.9	0.7	5.9	0.3	5.6	0.2
4	6.7	–0.3	6.7	0.0	6.6	0.4	6.4	0.6	6.7	0.4	6.3	0.2
5	7.2	–0.3	7.1	0.3	7.2	0.7	7.3	1.0	7.5	0.7	6.9	0.5
T	5.8	–0.1	5.6	0.1	5.7	0.6	5.8	0.8	5.8	0.6	5.4	0.4

lower MAE than any of the individual forecast sources for maximum temperature for day 2 and beyond. The overall MAE for all maximum temperature forecasts by MEAN is 0.2°F lower than for the most accurate individual source. The mean forecast does not improve over FS 2 for minimum temperatures at any time period, except at day 5. Thus, even for a simple measure of accuracy, different strategies must be employed by users seeking the most accurate forecast possible.

We computed seasonal accuracy statistics as well. For the NWS, the MAE for maximum temperature forecasts was highest during the winter (Table 2). The seasonal difference is large enough that a day 5 forecast during the summer has a lower MAE than a day 1 forecast during the winter. Note also the difference in performance between the two winters. The forecasts, particularly at days 2–4, were much better in the second winter. The reasons for this difference are beyond the scope of this paper. The media forecasts (not shown) show similar behavior, highlighting the difficulty of cool-season temperature forecasting, at least in Oklahoma City. The minimum temperature forecasts have less extreme seasonality for all forecast sources (not shown).

b. *Precipitation*

A frequently used measure of the accuracy of probability of precipitation forecasts is the Brier score (Brier 1950). The Brier score (BS) is the mean-squared error of probability forecasts f_i, where $x_i = 0$ if it does not rain and $x_i = 1$ if it does. A perfect forecast has a Brier score of 0:

$$\mathrm{BS} = \frac{\sum_{i=1}^{N} \left(f_i - x_i \right)^2}{N}. \qquad (1)$$

TABLE 2. MAE for NWS maximum temperature forecasts by days ahead (DA) and season. Winter1 is first winter of dataset and winter2 is second winter.

DA	Winter1	Spring	Summer	Autumn	Winter2
1	5.2	3.9	2.4	3.4	4.4
2	6.9	5.6	2.9	4.2	5.0
3	8.7	6.7	3.2	5.5	5.9
4	10.2	7.2	3.3	6.0	6.6
5	10.4	7.9	4.0	6.5	8.0
T	8.3	6.2	3.1	5.1	6.0

The Brier skill score (SS) is the percentage improvement relative to a climatological baseline:

$$SS = 100 \times \frac{BS_C - BS_F}{BS_C}, \qquad (2)$$

where BS_C is the Brier score with a constant climatological forecast and BS_F is the Brier score of the forecast system being compared to it. Positive (negative) values of the skill score indicate the percentage improvement (worsening) of the forecast source compared to climatology.

The SS of both FS 1 and FS 2 get worse with increasing lead time (Fig. 1). This is to be expected as, typically, the forecasts get harder with time. By day 3 for FS 1 and day 4 for FS 2, the skill scores for the forecasts become 5% (or less) better than climatology. In other words, the forecasts would be almost as accurate if climatology was used in place of the actual forecast at those lead times.[4] By day 7, the forecasts are 15% and 7% worse than climatology for FS 1 and FS 2, respectively.

The primary reason for the poorer skill scores at long lead time is the increasingly dry bias of the forecasts as lead time increases. The mean PoP of the forecasts decreases with lead time (Fig. 2). One would expect the long-range PoP to approach the climatological frequency. Instead, with the exception of FS 2's day 7 forecast, the forecast PoPs approach zero. Indeed, the use of 0% as a forecast value generally increases with lead time and so does the frequency of occurrence of precipitation with a zero PoP, until it almost reaches the value of the sample climatological frequency of precipitation (Table 3).

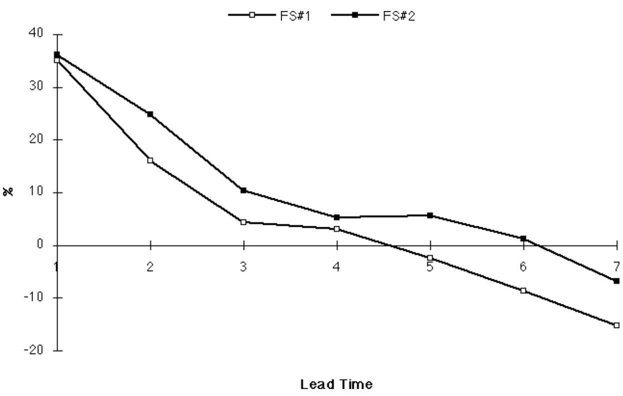


FIG. 1. Brier skill score (in percent) for PoP forecasts by lead time for FS 1 and FS 2. Positive values indicate improvement over climatology.

4. Distributions-oriented results

Distributions-oriented approaches provide a much richer picture of the characteristics of a forecast system (Murphy and Winker 1987; Brooks and Doswell 1996). In general, one wishes to describe the joint distribution of forecasts and observations. For the number of forecast sources and variables under con-

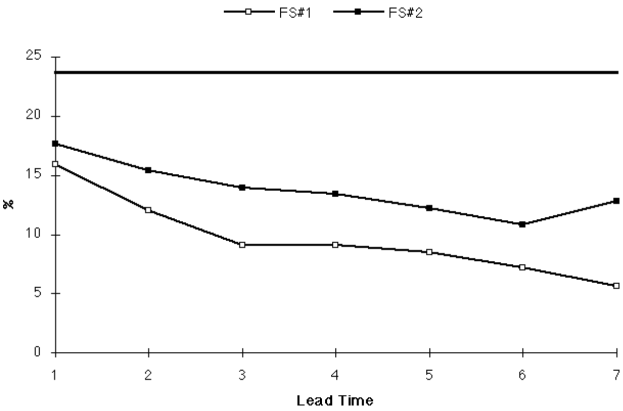


FIG. 2. Mean forecast PoP by lead time for FS 1 and FS 2. Horizontal heavy line indicates long-term climatological frequency of precipitation.

[4] This comparison uses the long-term (1961–90) climatology for each day. The sample climatology for these forecasts was 20.9%, while the long-term average across all days for Oklahoma City is 23.7%

TABLE 3. Frequency of use of 0% PoP (0% usage) and observed frequency of precipitation for 0% PoP (observed frequency) for FS 1 and FS 2 by lead time.

DA	FS 1 0% usage	FS 1 Obs frequency	FS 2 0% usage	FS 2 Obs frequency
1	53.0	4.1	50.5	3.1
2	56.7	8.8	52.6	5.9
3	64.5	12.1	54.8	11.4
4	60.1	11.4	53.6	11.6
5	62.9	14.4	52.6	11.2
6	66.0	17.9	56.4	13.3
7	72.6	20.2	43.6	20.0

sideration here (five sources with 10 temperature forecasts gives 50 arrays, even without the precipitation or intercomparisons of different sources or combinations of forecast variables), it is prohibitive to show all of the distributions. Here, we will focus on a few highlights from a distributions-oriented approach and through the use of methods to stratify forecasts (Murphy 1995).

a. Temperature

Much can be learned from looking at the joint distributions of temperature forecasts and observations [$p(f,x)$; see Murphy et al. 1989]. As an example, Table 4 shows the day 1 low-temperature forecasts from the NWS. Forecasts and observations have been grouped into 5°F bins, centered on values divisible by 5.[5] The centering was chosen so that one bin represented temperatures at or just below freezing. In general, when a particular forecast is made, the modal observation will be in the same 5°F bin. There is one exception to this and it occurs with forecasts just above freezing (33°–37°F). In that case, the mode observation is in the bin at or below freezing. Table 4 shows the high bias in the forecasts, overall, but it is particularly pronounced around freezing. In some situations (e.g., when precipitation is expected), this error seems to have the potential to cause problems for public safety.

The day 4 maximum temperature forecasts demonstrate another useful application of the distributions-oriented approach to verification. To illustrate another way of reducing the dimensionality of the verification problem, we have defined the forecast as being a forecast of the departure from climatology and then binned the forecasts and observations into 5°F bins, centered on values divisible by 5. All departures greater than or equal to 25°F are put into the ±25°F bin. Comparisons of FS 2 and FS 4 are particularly interesting (Table 5). Note that the day 4 maximum represents the only 1 of the 10 temperature forecasts (maximum and minimum) for which FS 2 does not have the lowest MAE, and in fact, FS 2 has the *highest* MAE for this forecast variable (see Table 1). In this case, FS 4 has the *lowest* MAE. However, if we consider the number of

TABLE 4. Day 1 NWS low-temperature forecasts (fcst.) and observations (obs). Data values represent 5°F bins centered on temperature at beginning of row/column; for example, there were 11 cases of observations of 28°–32°F and with forecasts of 33°–37°F. Note that first and last rows/columns (20°F, 45°F) include all temperatures below and above that value. Here, p(f) and p(x) represent the marginal probabilities of forecasts and observations for each category.

		Obs						
		20	25	30	35	40	45	p(f)
Fcst.	20	**14**	2	0	0	0	0	0.045
	25	5	**7**	3	0	0	0	0.042
	30	2	8	**13**	*3*	1	0	0.076
	35	0	3	*11*	**10**	7	0	0.087
	40	0	0	3	6	**16**	7	0.090
	45	0	0	0	0	8	**227**	0.660
	p(x)	0.059	0.056	0.084	0.053	0.090	0.067	1.000

[5] Doing this reduces the dimensionality of the verification problem, as discussed by Murphy (1991) and Brooks and Doswell (1996).

TABLE 5. Contingency table for observations and forecasts of day 4 maximum temperature anomalies for (a) FS 2 and (b) FS 4. Column headings are observed temperature changes and row headings are forecast anomalies in °F. Last row (column) is number of observations (forecasts) in each bin. Bold values indicate forecasts and observations in the same temperature bin. Number at bottom right is percentage of total forecasts in correct bin.

(a) FS 2

							Obs						
		–25	–20	–15	–10	–5	0	5	10	15	20	25	**N(f)**
	–25	**1**	1	1	0	0	0	0	0	0	0	0	3
	–20	0	**1**	1	1	1	0	0	1	0	0	0	5
	–15	2	1	**6**	1	2	6	0	0	0	0	0	18
	–10	2	1	5	**8**	6	3	6	3	0	0	0	34
Forecast	–5	1	1	4	13	**17**	8	10	4	2	1	0	61
anomalies	0	1	1	4	9	17	**24**	17	8	0	0	0	81
	5	1	0	0	5	6	17	**21**	15	7	2	2	76
	10	1	0	0	1	4	1	7	**14**	2	5	0	35
	15	0	0	0	0	0	3	3	7	**2**	3	2	20
	20	0	0	0	0	0	0	1	0	3	**1**	0	5
	25	0	0	0	0	0	0	0	0	0	0	**0**	0
	N(x)	9	6	21	38	53	62	65	52	16	12	4	**28**

(b) FS 4

							Obs						
		–25	–20	–15	–10	–5	0	5	10	15	20	25	**N(f)**
	–25	**0**	0	0	0	0	0	0	0	0	0	0	0
	–20	1	**1**	1	0	0	0	0	0	0	0	0	3
	–15	2	3	**3**	3	1	1	0	2	0	0	0	15
	–10	1	0	7	**8**	6	5	4	1	0	0	0	32
Forecast	–5	2	1	7	8	**13**	9	10	3	1	0	0	54
anomalies	0	2	1	3	13	22	**25**	14	8	2	1	0	91
	5	0	0	0	5	8	16	**17**	19	5	3	2	75
	10	0	0	0	1	3	5	17	**15**	6	2	0	49
	15	1	0	0	0	0	1	3	4	**1**	5	2	17
	20	0	0	0	0	0	0	0	0	1	**1**	0	2
	25	0	0	0	0	0	0	0	0	0	0	**0**	0
	N(x)	9	6	21	38	53	62	65	52	16	12	4	**25**

forecasts in correct 5°F bins, FS 2 has the largest number of correct forecasts of any source (95 cases or 28%) in the correct bin, and FS 4 has the fewest (84% or 25%). The apparent inconsistency between these two results occurs because FS 2 has the most forecasts more than two categories away from the observations (54% or 16%), and FS 4 has the fewest (41% or 12%). Thus, although FS 2 is more likely to be nearly correct, it is also more likely to be in serious error. *Any ranking by accuracy depends upon the definition of accuracy used.* As a result, it is not surprising that competing forecast sources can all claim to be the most accurate without "fudging" the data. As pointed out by Murphy (1993), the issue of which forecast has the most *value* to a user depends upon that user's needs and sensitivities. It is likely that different users may find the forecasts from different sources to be most valuable.

The variety of forecasts available to consumers in the area can, potentially, lead to confusion. In effect, not considering the day 6 and day 7 forecasts that are available from some of the television stations, there are 25 forecasts of a given day's maximum and minimum temperature. One way of combining all these forecasts is to consider the accuracy of the mean of the 25 forecasts as the level of agreement between the forecasts changes. To do so, we have calculated the variance of the 25 forecasts for each day (i.e., the five forecasts from each of the four media sources and the NWS) and compared it to the error of the mean of all 25 forecasts. The variance of the forecasts is correlated to the absolute error of the mean forecast at a 99% confidence level for both the minimum (correlation coefficient = 0.24) and the maximum (0.44) temperature. Thus, when the forecasts agree with each other, they are much more likely to be nearly correct than when they disagree with each other. To look at this further, we have divided the forecasts into those cases in which the variance of the forecasts is less than or greater than 10°F^2. The MAE increases with variance (Table 6). It is interesting to note that the variance in the high-temperature forecasts is quite a bit larger than in the low-temperature forecasts. Here, 182 forecasts met the low-variance criterion for the minimum temperature forecast, while only 149 did so for the maximum temperature forecast. Other things being equal, one might expect more variance in the maximum temperature forecasts, since they have a slightly longer lead time, and, more importantly, since the variance of maximum temperatures is greater than that of minimum temperatures (the standard deviation of observed minimum temperatures in this dataset is 7.9°F and that of the observed maximum temperatures is 9.3°F). Given the other results, it seems the maximum temperatures were harder to forecast during this 14-month period.

All of the television forecasters have the opportunity to use the NWS forecast as input into their forecasts,[6] so we have considered the quality of the forecasts when the private-sector forecasts disagree strongly with the NWS. To do this, we have stratified the forecasts by counting the number of times the private forecast disagreed with the NWS forecast and was closer to the observations when the forecasts disagreed by 5°F or more (Table 7). Only FS 2 increases the number of disagreements monotonically with increasing lead time of the forecast, if both maximum and minimum temperature forecasts are combined (i.e., day 1 minimum, day 1 maximum, day 2 minimum, day 2 maximum, etc.). Here, FS 2 also improves on the NWS forecast significantly (at the 99% confidence level) for 3 out of the 10 forecast periods.[7] There is a slight tendency, in general, for the media forecasts to be more accurate for disagreements at long lead times (days 4–5) compared to short lead times (days 1–2). This is particularly obvious for FS 1's maximum temperature forecast, where the source's forecasts are sig-

TABLE 6. Variance and MAE of mean temperature forecasts for cases with variance of forecasts < 10°F^2 (low variance) and > 10°F^2 (high variance).

	Min temp variance	Min temp MAE	Max temp variance	Max temp MAE
Overall	11.9	3.7	16.1	5.0
Low variance	5.3	2.9	5.4	3.4
High variance	19.5	4.6	24.6	6.4

[6] The newspaper forecast is created by a private company under contract to the paper. We do not know for certain when the forecast is made. The NWS forecast information may or may not be available to them.

[7] Significance testing was done using a Monte Carlo technique, using 100 000 trials of flipping a simulated coin the number of times that a forecast source disagreed with the NWS and counting how frequently the number of "heads" occurred by chance.

TABLE 7. Percentage of time when a given forecast source disagreed with the NWS by 5°F or more and was more accurate. Total number of disagreements in parentheses. Bold (italic) numbers indicate that the source was better (worse) than the NWS in cases of disagreement at a 95% confidence interval; underlining indicates 99% confidence interval.

DA	1	2	3	4	5	Total
(a) Minimum temperatures						
FS 1	57 (14)	*33 (41)*	46 (76)	60 (77)	57 (96)	52(304)
FS 2	69 (13)	45 (20)	**68 (63)**	**71 (80)**	55 (95)	**63 (271)**
FS 3	50 (16)	43 (21)	58 (54)	61 (51)	56 (73)	56(217)
FS 4	52 (27)	41 (38)	50 (100)	46 (107)	46 (133)	47(405)
(b) Maximum temperatures						
FS 1	*29 (21)*	*30 (47)*	49 (65)	51 (80)	**60 (127)**	50 (340)
FS 2	56 (27)	52 (42)	50 (83)	49 (100)	**62 (112)**	54 (364)
FS 3	40 (20)	39 (29)	58 (54)	**64 (61)**	60 (72)	**56 (236)**
FS 4	*35 (44)*	46 (54)	48 (103)	55 (112)	55 (130)	50 (443)

nificantly worse than the NWS for days 1 and 2 and significantly better at day 5 (Table 7b). For the most part, however, the question of whether the private forecasts are more accurate than the NWS, by this measure, represents little more than a coin flip. We offer no speculations about whether specific forecast strategies lead to the patterns, or lack thereof. At the moment, they are a curiosity, although it seems obvious that an understanding of the kinds of situations in which they improve (or do not improve) on the NWS forecast would be critical to any of the private-sector forecasters, if they are interested in improving the quality of their forecasts.

A more general aspect of this problem that forecast users have to deal with on a regular basis is what to do with conflicting forecasts from the media sources.[8] We have broken the forecasts out into those cases when any one of the forecast sources goes "out on a limb" and disagrees with all the others by 5°F or more and compared it to the mean of the other forecasts (Table 8). As might be expected, in general, a forecast that disagrees by that much from any other is likely to be wrong. Here, FS 3 is the only source that does not do significantly worse statistically than the mean of the other forecasts when it disagrees with the others. It also finds itself in that situation less often than any of the other private forecasts, perhaps indicating that the forecasters preparing FS 3's forecasts are more conservative than the other private forecasters. Table 8b also reinforces the earlier discussion about the interesting day 4 maximum temperature forecast for FS 2. Here, FS 2 takes more risks on that element than any other forecast source on any element and, in the process, does worse than the mean forecast at a 99% confidence level.

Overall, we see an increase in error of the mean forecast with increasing variance of the forecasts and significant differences in the performance of the various forecast sources when they disagree strongly with the NWS. These facts indicate the need for weather-information-sensitive users in the public to attempt to collect information from a variety of sources to get the most complete picture of the likely evolution of the weather. The relationship between variance and forecast error suggests that it is possible to quantify the uncertainty in the forecasts and, perhaps, to derive a probabilistic temperature forecast from the information. Clearly, no one source is sufficient to provide all of the useful information available within the media market.

b. Precipitation

Contingency tables for precipitation can be used to construct reliability diagrams (Wilks 1995), indicating how well the observed frequency of an event matches the forecast probability. To get larger sample sizes, we have summed the forecasts over all 7 days for FS 1 and FS 2 (Fig. 3). The general dry bias is readily apparent and is related to the tendency to overuse the 0% PoP. The observed frequency of precipitation for both sources with a PoP of zero is on the order

[8] Due to slightly different timings of the evening news presentations, the forecast portion of the television weather presentations frequently start at different times, and it is possible for someone changing channels rapidly to see all the television forecasts from newscasts nominally at the same time.

of 12%, and in fact, precipitation was observed *more* frequently when FS 2 forecast 0% PoP than 10% PoP. Overall, 86% (72%) of the forecasts of FS 1 (FS 2) were either 0% or 20%, the forecast value nearest the climatological frequency. From Fig. 3, the observed frequency of rain for both sources for both of those values was such that the points fell near the no-skill line, indicating that the forecasts contributed only marginally to skill, if at all (Wilks 1995). Thus, the most common forecasts presented to the public show little or no skill compared to climatology. Reliability diagrams for days 1, 4, and 7 (Fig. 4) show the tendency for the reliability curve to become "flatter" as lead time increases. This reflects the fact that the observed frequency of precipitation for all forecast values approaches the climatological frequency at longer lead times.

Curiously, FS 2 maintains an almost constant frequency of use of 15% at all lead times, but drops the use of 10% after day 5. It has a dramatic increase in the number of 20% forecasts at day 7, going from 71 at day 6 to 117 at day 7. Unfortunately, the number of cases where it rains on those forecasts increases only from 21 (29.6%) to 22 (18.8%). At the same time, the number of forecasts with a PoP of 0% decreases from 181 to 140, but the frequency of precipitation on those forecasts increases from 13.3% to 20.0%. Thus, the observed frequency of rain on 0% PoPs from FS 2 is actually higher than the frequency for 20% PoPs at day 7.

Typically, FS 2 produced a "wetter" forecast, although it was still dry compared to climatology. About 23% of FS 2's forecasts are PoPs exceeding the climatological frequency of precipitation, whereas only 12% of FS 1's forecasts are "wet." This difference extends up to the highest probabilities, with FS 2 using PoPs exceeding 60% 22 times (8 after day 1), including a day 3 100% PoP. In contrast, FS 1 used PoPs exceeding 60% only 13 times (never after day 1).

TABLE 8. Same as Table 7 except for when given forecast source disagreed with all other forecast sources by 5°F or more, in comparison to mean forecast of other sources. Note that there are no cases in which the source that disagreed with the others outperformed the mean at a statistically significant level.

DA	1	2	3	4	5	Total
(a) Minimum temperatures						
FS 1	0(3)	22(10)	*24(18)*	39(13)	31(13)	*27(57)*
FS 2	–(0)	50(2)	60(5)	64(14)	42(12)	55(33)
FS 3	50(4)	25(4)	43(7)	40(5)	46(13)	42(33)
FS 4	40(5)	44(9)	*11(18)*	*8(13)*	30(20)	*23(65)*
NWS	0(1)	–(0)	25(4)	22(9)	27(11)	*24(25)*
(b) Maximum temperatures						
FS 1	0(4)	*21(15)*	43(7)	33(13)	32(22)	*29(61)*
FS 2	50(6)	50(4)	29(17)	*21(25)*	53(15)	35(66)
FS 3	0(3)	50(2)	50(8)	46(11)	56(18)	48(42)
FS 4	25(12)	44(9)	*24(21)*	35(23)	35(17)	*32(82)*
NWS	100(1)	0(2)	*10(10)*	38(8)	*22(23)*	*23(44)*

5. Discussion

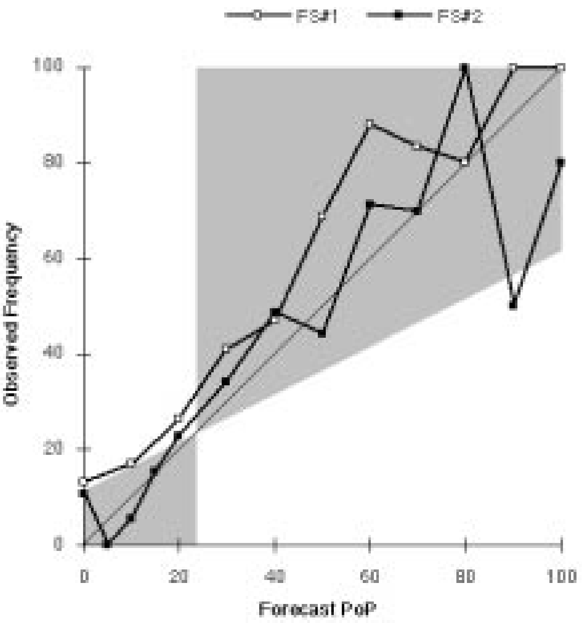
Many forecast sources are available to the public via the NWS and the media. The significant disagreements among their forecasts inevitably leads to the question of "who is the best?" Based on our analysis, we believe, as discussed by Murphy (1993), that this question is simplistic and the rich amount of information from even a cursory verification process implies that there is no universally correct answer to that question. Every one of the sources has its strengths and weaknesses, even without discussing issues such as hazardous weather preparedness. Specifically, for the media sources, they had the following strengths and weaknesses:

1) FS 1 had the least biased temperature forecasts but had the highest MAE for maximum temperature forecasts.
2) FS 2 had the lowest MAE for 9 of the 10 temperature lead times but has the largest MAE for the day 4 maximum forecasts. It was also the most likely

to be correct when it disagreed with NWS forecasts for minimum temperatures. For the longest lead time forecasts, the observed frequency of precipitation on forecasts of 0% PoPs is higher than the observed frequency for 20% PoPs.

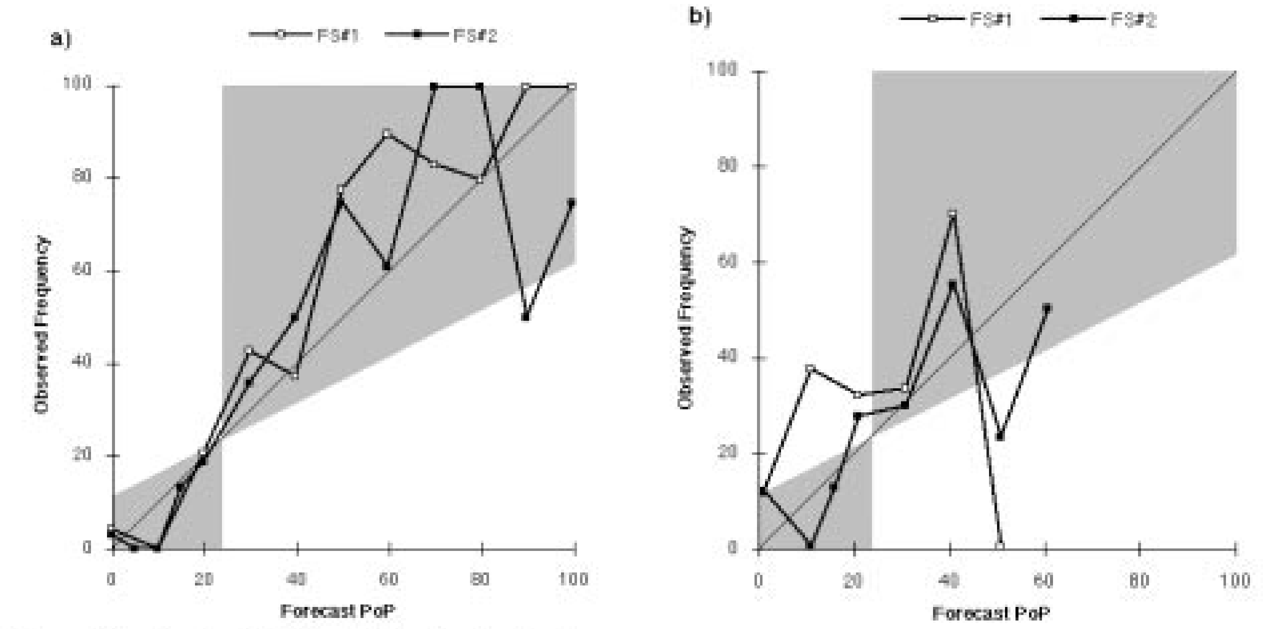
3) FS 3 was the most likely to be correct when it disagreed with NWS forecasts for maximum temperatures and was the most likely to be correct when it disagreed with all of the other sources overall for temperature forecasts. It was the most conservative forecast source, disagreeing with the NWS temperature forecast much less often than any of the other sources.
4) FS 4 was the least accurate minimum temperature forecast at every lead time but was the most accurate for the day 4 maximum.

Examination of the forecasts from sources available to the public provides fertile ground for verification specialists. More importantly, it should allow the forecasters to evaluate their own strengths and weaknesses and help in improving their products, if the quality of these forecasts is a primary concern. From what we have seen, many of the weaknesses could be improved very easily. One area of obvious improvement would be to have the long-term PoP forecasts tend toward climatology, rather than zero. Another easy improvement would be to produce unbiased tem-



FS#1	1399	41	527	149	85	16	17	6	5	1	1
FS#2	1169	18	439	269	123	61	42	10	5	2	5

FIG. 3. Attributes diagram for all 7 days lead time forecasts for FS 1 and FS 2. Diagonal line indicates perfect reliability. Shaded area is region where forecasts contribute positively to skill with respect to climatology. Numbers below figure indicate number of forecasts at each PoP from 0% to 100% by 10%. In addition, FS 2 used 5% 1 time and 15% 103 times. Total number of forecasts is 2247.



FS#1	170	9	58	28	24	9	10	6	5	1	1
FS#2	162	10	37	36	16	12	18	4	4	2	4

FS#1	193	8	88	21	10	1	
FS#2	172	1	51	37	29	13	2

FIG. 4. Same as Fig. 3 except for individual lead times: (a) day 1 (FS 2 used 5% 1 time and 15% 15 times), (b) day 4 (FS 2 used 15% 16 times), (c) (facing page) day 7 (FS 2 used 15% 16 times). Total number of forecasts is 321 for each plot.

perature forecasts for the various lead times. Given that we have been able to see this, it seems that the sources under examination 1) have not verified their own forecasts, or 2) their perception of what the public wants is different than providing the highest quality forecast, or 3) they believe that the value of their forecasts is high even if the quality is not. Given the wide range of needs of the users of publicly available forecasts and the complex relationship between quality and value, it seems unlikely that the last goal could be accomplished in any easy way. Neither of the other two options is satisfying from the public perspective. It is possible that media forecasters perceive the higher ratings as a more important goal than forecast accuracy. While this is plausible from their perspective, it may lead to forecasting strategies that do not lead to accurate forecasts. If so, we view this as a lamentable outcome and one that does not serve the public interest well. We encourage them to make forecast quality a higher priority in their process.

Further, our results highlight poor uses of probability in precipitation forecasts, as discussed by Vislocky et al. (1995). The extreme dry bias at long range is indicative of a lack of understanding (or failure to apply understanding) of climatological information. This is even without discussing the use of colloquialisms to describe the chance of precipitation in the absence of numerical probabilities or the use of verbal descriptions that are inconsistent with the numbers presented in the forecast (e.g., "Our next chance of precipitation is towards the end of the week, but for now, I'll go with just an increase in cloudiness. So, you'll need to keep watching to see how things develop."). This approach can only lead to confusion in the minds of the forecast users, the public. We find it particularly distressing given the results of Murphy et al. (1980) and Sink (1995) indicating that the public understands and prefers numerical, rather than verbal, probability of precipitation forecasts.

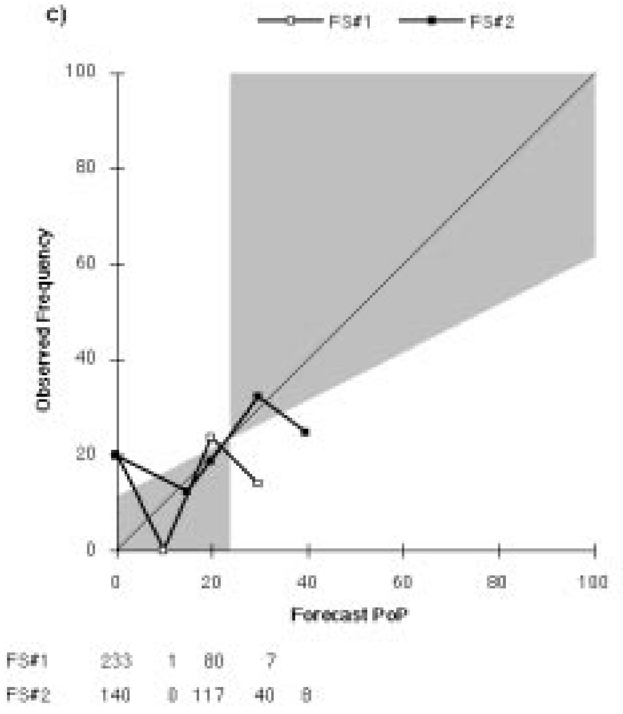


FIG. 4. (*Continued*).

Acknowledgments. J. T. Johnson helped record forecasts from the television stations. We thank Chuck Doswell, Dennis Driscoll, Allan Murphy, and anonymous reviewers for their helpful comments.

References

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

Brooks, H. E., and C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting,* **11,** 288–303.

Brown, B. G., and A. H. Murphy, 1996: Improving forecasting performance by combining forecasts: The example of road-surface temperature forecasts. *Meteor. Appl.,* **3,** 257–265.

Driscoll, D. M., 1988: A comparison of temperature and precipitation forecasts issued by telecasters and the National Weather Service. *Wea. Forecasting,* **3,** 285–295.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.,* **119,** 1590–1601.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, 1995: A coherent method of stratification within a general framework for forecast verification. *Mon. Wea. Rev.,* **123,** 1582–1588.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

——, S. Lichtenstein, B. Fischhoff, and R. L. Winkler, 1980: Misinterpretations of precipitation probability forecasts. *Bull. Amer. Meteor. Soc.,* **61,** 695–701.

——, B. G. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting,* **4,** 485–501.

Roebber, P. J., and L. F. Bosart, 1996: The complex relationship between forecast skill and forecast value: A real-world comparison. *Wea. Forecasting,* **11,** 544–559.

Sink, S. A., 1995: Determining the public's understanding of precipitation forecasts: Results of a survey. *Natl. Wea. Dig.,* **19,** 9–15.

Thornes, J. E., 1996: The quality and accuracy of a sample of public and commercial weather forecasts in the U.K. *Meteor. Appl.,* **3,** 63–74.

Vislocky, R. L., J. M. Fritsch, and S. N. DiRienzo, 1995: Operational omission and misuse of numerical precipitation probabilities. *Bull. Amer. Meteor. Soc.,* **76,** 49–52.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.