A Comparison of Measures-Oriented and Distributions-Oriented Approaches to Forecast Verification

HAROLD E. BROOKS AND CHARLES A. DOSWELL III

NOAA/Environmental Research Laboratories, National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 2 October 1995, in final form 22 January 1996)

ABSTRACT

The authors have carried out verification of 590 12–24-h high-temperature forecasts from numerical guidance products and human forecasters for Oklahoma City, Oklahoma, using both a measures-oriented verification scheme and a distributions-oriented scheme. The latter captures the richness associated with the relationship of forecasts and observations, providing insight into strengths and weaknesses of the forecasting systems, and showing areas in which improvement in accuracy can be obtained.

The analysis of this single forecast element at one lead time shows the amount of information available from a distributions-oriented verification scheme. In order to obtain a complete picture of the overall state of forecasting, it would be necessary to verify all elements at all lead times. The authors urge the development of such a national verification scheme as soon as possible, since without it, it will be impossible to monitor changes in the quality of forecasts and forecasting systems in the future.

1. Introduction

The verification of weather forecasts is an essential part of any forecasting system. Producing forecasts without verifying them systematically is an implicit admission that the quality of the forecasts is a low priority. Verification provides a method for choosing between forecasting procedures and measuring improvement. It can also identify strengths and weaknesses of forecasters, thus forming a crucial element in any systematic program of forecast improvement. As Murphy (1991) points out, however, ''failure to take account of the complexity and dimensionality of verification problems may lead to . . . erroneous conclusions regarding the absolute and relative quality and/or value of forecasting systems.'' In particular, Murphy argues that the reduction of the vast amount of information from a set of forecasts and observations into a single measure (or a limited set of measures), a *measures-oriented* approach to verification, can lead to misinterpretation of the verification results. Brier (1948) pointed out that ''the search for and insistence upon a single index'' can lead to confusion. Moreover, a measures-oriented approach fails to identify the *situations* in which forecast performance may be weak or strong.

An alternative approach to verification involves the use of the joint distribution of forecasts and observations, hence leading to the name *distributions-oriented* verification (Murphy and Winkler 1987). A major difficulty in taking this approach to verification is that the dimensionality of the problem can be very large, and hence, the datasets required for a complete verification must be very large, particularly if two forecast strategies are being compared (Murphy 1991). For a joint comparison of two forecast strategies and observations, the dimensionality, D, of the problem is given by $D = IJK - 1$, where I is the number of distinct forecasts from one strategy, J is the number from the second strategy, and K is the number of distinct observations, respectively. Thus, if each forecast strategy produces 11 distinct forecasts and 11 distinct observations (e.g., cloud cover in intervals of 0.1 from 0 to 1), the dimensionality is given by $D = (11)(11)(11) - 1 = 1330$. Clearly, the datasets needed for complete verification and the description of the joint distribution can become prohibitively large. In practice, therefore, persons making evaluations of forecasts have to make compromises between the size of the dataset and the completeness of the verification. In this paper, we show the richness of information that can be obtained from simple verification techniques using a relatively small forecast sample. We believe that the insights available from even this modest work show the importance of considering a broad range of descriptions of the forecasts and observations, in an effort to retain as much information as possible.

Murphy (1993) described three types of ''goodness'' for forecasts. We summarize those types here in

Corresponding author address: Dr. Harold E. Brooks, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069. E-mail: brooks@nssl.uoknor.edu

order to show where the present work fits. The three types are as follows.

1) Consistency: How well does a forecast correspond to the forecaster's best judgments about the weather?
2) Value: What are the benefits (or losses) to users of the forecasts?
3) Quality: How well do forecasts and observations correspond to each other?

We cannot say anything about ''consistency,'' since we have no access to forecasters' judgments. This is typically true. Consistency is the only type of goodness that is completely under the control of the forecaster, but it is difficult for others to verify. We also cannot say anything quantitative about ''value,'' since we have not done a study of the forecast's user community. We will make some general remarks about temperature forecasting, based on the premise that improvements of a few degrees in a forecast are unimportant to many users in most cases.[1]

Almost all of our attention will be focused on the ''quality'' of the forecasts. Murphy (1993) defines 10 different aspects of quality (see his Table 2 for more details). Traditional measures such as the mean absolute error (MAE) and the root-mean-square error (rmse) are related to aspects such as accuracy and skill. By using a distributions-oriented approach, the complete relationship between forecasts and observations can be examined. Forecasts can be high in quality in one aspect although being low in another. For example, forecasting the high temperatures by simply using the annually averaged high temperature every day would be an unbiased temperature forecast, but it would clearly not be very accurate over a long period. Overforecasting the high temperature by 10°F every day might be more accurate than using the annually averaged high temperature but would be biased. A perfect forecast would perform equally well at all of the various aspects of quality.

An important distributions-oriented study of temperature forecasts was done by Murphy et al. (1989), in which high-temperature forecasts and observations for Minneapolis, Minnesota, were compared. They concluded that the different measures of forecast quality gave different impressions about the quality of forecast systems. They also pointed out that the joint distribution approach highlights areas in which forecasting performance is especially weak or strong. In this paper, we will carry out a related study on a dataset for Oklahoma City, Oklahoma. Our focus will be to show the vast wealth of additional information available that can be obtained through a distributions-based verification over a ''traditional'' measures-based approach. We will point out some particularly interesting aspects of forecasting performance that, in a forecasting system that encouraged continuous verification and training, could lead to improvements in forecast quality.

2. Forecast and verification dataset

The dataset consists of 590 high-temperature forecasts from 1993 and 1994 made by the National Weather Service (NWS) Forecast Office at Norman, Oklahoma (NWSFO OUN), and verified at Oklahoma City (OKC).[2] The basic forecast systems are from the Limited Area Fine Mesh (LFM)–based Model Output Statistics (MOS), the Nested Grid Model (NGM)–based MOS, the NWSFO OUN human forecast, and persistence (PER). In addition, an average or consensus MOS forecast (CON) was created by averaging the LFM MOS and NGM MOS forecasts. Vislocky and Fritsch (1995) have shown that the simple averaging of the LFM and NGM MOS forecasts produced a significantly better forecast system over the long run than either of the individual MOS forecasts. The MOS forecasts are all based on the 0000 UTC model runs, verifying 12–24 h later, although the NWSFO forecast is taken from the area forecast made at approximately 0800–0900 UTC, verifying later that day. PER is the observed high temperature from the previous day. All days for which all four basic forecasts are available, as well as verifying observations, are included in the dataset.

3. A measures-oriented verification scheme

It is possible to develop simple measures that convey some information about the forecast performance. In particular, the bias or mean error (ME) is given by

$$\mathrm{ME} = \frac{\sum_{i=1}^{N} (f_i - x_i)}{N}, \qquad (1)$$

where f_i is the ith forecast, x_i is the ith observation, and there are a total of N forecasts. It says nothing about the accuracy of forecasts since a forecaster making 5 forecasts that are 20° too warm and 5 forecasts that are 20° too cold will get the same ME as a forecaster making 10 forecasts that match the observations exactly. In

[1] We acknowledge that many users may be sensitive to small temperature changes at some critical temperatures (e.g., a decision about travel when precipitation is forecast and the temperature is near freezing or for load forecasting for power companies where small improvements can save tens of thousands of dollars).

[2] It is important to note that none of our comments about the performance of human forecasting should be interpreted in terms of performance relative to other NWS forecast offices. We believe that, in the context of a *complete* distributions-oriented verification program, intercomparison of office performance is a desirable thing. However, since that verification program does not exist, we cannot make those comparisons.

TABLE 1. Mean error (ME), mean absolute error (MAE), and root-mean-square error (rmse) in °F for five different high-temperature forecast systems valid during the 12–24-h period for Oklahoma City. Forecast systems are persistence (PER), MOS guidance from the LFM (LFM), MOS guidance from the NGM (NGM), an average of the LFM and NGM MOS (CON), and the forecast from the NWS Forecast Office in Norman (NWSFO). Numbers in parentheses indicate percentage improvement by NWS forecast over the forecast for the given measure.

Type	PER	LFM	NGM	CON	NWSFO
ME	.15	.15	−.62	−.22	.49
MAE	6.62 (57.72)	3.35 (16.48)	3.38 (17.28)	3.04 (7.81)	2.80
rmse	9.16 (58.48)	4.48 (15.19)	4.37 (13.01)	4.04 (5.76)	3.80

order to correct that problem, the errors need to be nonnegative. There are two common ways of doing this. The MAE takes the absolute value of each forecast error and is given by

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

The rmse squares each error and is given by

$$\mathrm{rmse} = \left[\frac{\sum_{i=1}^{N} (f_i - x_i)^2}{N} \right]^{1/2}. \qquad (3)$$

Because of its formulation, the rmse is much more sensitive to large errors than MAE. For instance, suppose a forecaster makes 10 forecasts, each of which is in error by 1°, while another forecaster makes 9 forecasts with 0° error and one with 10° error. In both cases, the MAE is 1°. The rmse for the first forecaster is 1°, although it is 3.16° for the second forecaster. Thus, the rmse rewards the more consistent forecaster, even though the two have the same MAE.

For both MAE and rmse, it is possible to compare the errors to those generated by some reference forecast system (e.g., climatology, persistence, MOS) by calculating the percentage improvement (IMP). IMP is given by

$$\mathrm{IMP} = 100\left(\frac{E_R - E_F}{E_R} \right), \qquad (4)$$

where E_R is the error statistic generated by the reference forecast system and E_F is the error statistic from the other forecast system. This is often described as a skill score.

The relative performance of the various forecast systems using the simple measures described above is summarized in Table 1. NGM MOS is seen to have a cold bias (−0.62°F), although the NWSFO has a warm bias (0.49°F). Although LFM MOS has a lower MAE than NGM MOS, it has a higher rmse. The CON forecast represents a greater improvement in the MAE and rmse over either the LFM MOS or NGM MOS than the human forecasters improve over CON, according to these measures. This leaves open the question of the value (in Murphy's context) of a decrease of 0.24°F in MAE or rmse by NWSFO over the numerical guidance. By using these simple measures, we are unable to determine the distribution of the errors leading to the statistics and their dependence upon the actual forecast or observation. Therefore, it is not possible to use these measures alone to determine the nature of the forecast errors. In the hypothetical case of the two forecasters discussed above, it is likely that for most users, the forecast with ten errors of 1°F would provide more value than the forecast with 1 error of 10°F. From that view, even though any single measure is clearly inadequate, the MAE may be potentially even more misleading about forecast performance than RMSE, depending upon the assumptions about the needs of the users of the forecasts.

The MAE is one of the two temperature verification tools required by the *NWS Operations Manual* (NWS Southern Region Headquarters 1984; NOAA 1984). The other is the production of a table of forecast errors in 5°F bins.[3] We have generated this table for the various forecast systems (Table 2). As expected, PER produces the highest number of large errors. Other than PER, one of the striking aspects of the table is the frequency with which forecast temperatures have small errors. All of the forecasts are within 5°F more than 80% of the time. By using CON, the forecast was correct to within 5°F 86.1% of the time. Thus, the numerical guidance produced forecast errors exceeding 5°F approximately once per week, although the NWSFO reduced the number of such errors from 82 to 75, a decrease of 8.5%. Very large forecast errors are, of course, even less frequent. The worst forecast system by this measure (other than persistence), LFM MOS, is correct within 10°F 96.6% of the forecasts (exceeding 10°F approximately once per month); although, the best NWSFO, is within 10°F 98.6% of the time. Compared to the most accurate MOS forecast, CON, the NWSFO reduced the errors exceeding 10°F from 12 to

[3] We note that the description of those bins, as given by the *Regional Operations Manual Letter* (NWS Southern Region Headquarters 1984), is ambiguous. While we have chosen to collect the forecasts in 1–5°F, 6–10°F, 11–15°F, etc. bins, there is no guidance in the *NWS Operations Manual* as to the boundaries of the bins.

TABLE 2. Errors in forecasts (forecast−observation) by 5°F bins (except for perfect forecast) and number (percentage) of forecasts within 5° and 10°F.

Range	PER	LFM	NGM	CON	NWSFO
−36 to 40	1	0	0	0	0
−31 to −35	4	0	0	0	0
−26 to −30	4	0	0	0	0
−21 to −25	11	0	0	0	0
−16 to −20	16	0	1	0	0
−11 to −15	33	6	2	2	0
−6 to −10	52	40	53	36	27
−1 to −5	114	223	276	268	198
0	39	66	57	58	87
1 to 5	173	192	153	182	230
6 to 10	84	49	37	34	40
11 to 15	42	12	10	8	7
16 to 20	11	0	1	2	1
21 to 25	4	2	0	0	0
26 to 30	1	0	0	0	0
31 to 35	1	0	0	0	0
36 to 40	0	0	0	0	0
Errors ⩽ 5°F	326 (55.3)	481 (81.5)	486 (82.4)	508 (86.1)	515 (87.3)
Errors ⩽ 10°F	462 (78.3)	570 (96.6)	576 (97.6)	578 (98.0)	582 (98.6)

8 (33.3%). Observe that there is an important difference in the distribution of errors for NWSFO and the various MOS forecasts. In all of those cases, forecast errors greater than 10°F are much more likely to be positive (too warm) than negative (too cold). However, although the NWSFO distribution is skewed toward the overforecast (i.e., too warm) side at all bins, small MOS forecast errors are more likely to be cold than warm. This is particularly true for the NGM MOS, where 11 of the 14 (79%) errors larger than 10°F are too warm and 276 of the 429 (64%) errors of less than 6°F are too cold. Knowledge of this asymmetry could be employed by forecasters to improve their use of numerical guidance products and could be used by modelers to improve the statistically based guidance as well.

These two tables represent all the verification knowledge of temperature forecasts that is required of the forecast offices. This by no means exhausts the available information, however. The table of forecast errors (Table 2) represents one ''level'' at which a distributions-based approach to verification can be applied and is a step above the summary measures in sophistication. It gives the univariate distribution of forecast errors $p(e) = p(f - x)$. However, this approach implicitly assumes that all errors of magnitude $f - x$ are the same. A more useful approach, which we will explore in the next section, is to consider the joint (i.e., bivariate) distribution of $p(f, x)$. This latter method allows us to consider the possibility that certain values of f or x are more important than others, or that forecast performance varies with f or x.

4. A distributions-oriented verification scheme

A more complete treatment of verification demands consideration of the relationship between forecasts and observations [see Murphy (1996) for a description of the early history of this issue]. For 12–24-h temperature forecasting, an appropriate method is to consider changes from the previous day's temperature. In a qualitative sense, persistence represents an appropriate no-skill forecast for most forecast users, particularly for forecasts on this timescale. As seen in section 2, that would lead to an error of 10°F or less for almost 80% of the dataset. Thus, we have chosen to verify forecasts and observations in the context of day-to-day temperature *change.* Persistence is then reduced to a single category in the joint distribution of forecast and observed temperature changes.

The range of forecast and observed changes is 72°F (−39° to +33°F). The dimensionality of doing a complete verification comparing two forecast systems over that range of temperatures is $73^3 - 1 = 389016$. Clearly, the dataset is much too small to span that space.[4] As a result, we have chosen to count forecasts and observations in 5°F bins in order to reduce the dimensionality considerably. This also has the appeal of taking some account of the uncertainty in the observations and the variability of temperature over a standard forecast area. The bins are centered on 0°F, going in intervals of 5°F. Therefore, forecasts or observed changes of ±2°F are counted in the 0°F bin. We have

[4] We note that the use of persistence as a baseline is one way of reducing the dimensionality of the verification problem. The observed range of high temperatures over the period was 17°–103°F. The dimensionality of the evaluation of one forecast system over that range would be (87)/(87) − 1 = 7568, while for the day-to-day changes it is only (73)/(73) − 1 = 5328, a reduction of 30%. Other methods of reducing the dimensionality by stratifying the results, such as departures from climatology, also exist.

chosen to collect all changes greater than or equal to 23°F into a bin labeled ±25°F. This is due to the sparseness of the dataset even with 5°F bins. In addition, we have chosen to evaluate each forecast system individually. The dimensionality of the verification problem has been reduced significantly by these processes. Since there are now 11 forecast and observation bins for each forecast system, the dimensionality of the binned problem for each system is $11^2 - 1 = 120$.

The joint distribution of the forecasts (f) and observations (x), $p(f, x)$ contains all of the non-time-dependent information relevant to evaluating the quality of the forecasts (Murphy and Winkler 1987). These distributions for LFM MOS, NGM MOS, CON, and NWSFO are given in Tables 3a–d. Note that numbers above the bold diagonal indicate forecasts that were too cold and that numbers below the bold diagonal indicate forecasts that were too warm. Extreme temperature changes are, in general, underforecast, particularly by the numerical guidance, most especially by the LFM MOS. In the bins associated with 20°F (or more) temperature changes (of either sign), there are only 21 LFM MOS forecasts, in comparison with 34 NGM MOS, 24 CON, 30 NWSFO, and 42 observations. The extent of this becomes clear when the ratio of forecasts to observations is plotted against the forecast temperature change (Fig. 1). Ideally, this ratio should be close to unity for all forecast values. Instead, the ratio is well below unity for large temperature changes and, for the most part, slightly above one for small changes. In comparison with the numerical guidance, the NWSFO forecast is, in fact better in this respect, with large departures from unity occurring only for forecasts of cooling of 15°F and warming of 25°F, which only had one forecast.

Murphy and Winkler (1987) point out that much of the information in the joint distribution is more easily understood by factoring $p(f, x)$ into conditional and marginal distributions. In particular, we want to look at two complementary factorizations of the joint distribution following Murphy and Winkler (1987). The first is the calibration–refinement factorization, involving the conditional distribution of the observations given the forecasts, denoted by $p(x|f)$, and the marginal distribution of the forecasts, $p(f)$ (Table 4a–d). The factorization is given by

$$p(f, x) = p(x|f)p(f). \tag{5}$$

The second factorization is the likelihood–base-rate factorization, involving the conditional distribution of the forecasts given the observations, $p(f|x)$, and the marginal distribution of the observations, $p(x)$ (Table 5a–d), given by

$$p(f, x) = p(f|x)p(x). \tag{6}$$

Although we present both factorizations, we will make only brief comments about the contents.

A number of important aspects about the quality of the forecasts are apparent from the tables. The values of $p(x|f)$ and $p(f|x)$ are dominated by the diagonals in both Tables 4 and 5 on the matrix almost without exception.[5] The significant exception is related to the cold bias of the NGM MOS. Over half of the forecasts of a 5°F cooling are associated with no change in the observed temperature (Table 4b). As a result, the CON forecasts are also too cold at that range.

Reliability (also known as conditional bias or calibration) is one of the aspects of forecast quality that can be derived from the calibration–refinement factorization. It represents the correspondence between the mean of the observations associated with a particular forecast (denoted $\langle x_f \rangle$) and that forecast (f) (Murphy 1993). It can be viewed as the difference between those quantities. For perfectly reliable forecasts, the value would be zero for all forecasts, f. In the case of our four systems producing forecasts of temperature change, the differences are typically less than a degree, indicating fairly reliable forecasts (Fig. 2). However, it is worth noting that there are potentially meaningful biases of 2°–3°F at certain ranges of temperature changes. Operationally, the identification of these could be used to improve future forecasts.

Consideration of $p(f|x)$ has not received as much attention as $p(x|f)$ in forecast verification (Murphy and Winkler 1987). This is perhaps due to the standard view of verification as one of seeing what happens after a forecast has been made. Consideration of the conditional probability of forecasts given the observations requires a view of verification as an effort to understand the relationship *between* forecasts and observations, rather than just looking at what happened after a forecast was made. As an example of something that appears much clearer from the perspective of $p(f|x)$, we turn to the question of overforecasting and underforecasting the magnitude of temperature changes. It is not obvious that there is any reason to prefer one or the other, and, given that errors will occur, one would like to have overforecasts and underforecasts be equally likely. The magnitude of the asymmetry between the two appears different from an inspection of the two tables of conditional probability. Accurate forecasts are associated with the bins along the main diagonal. Underforecasting of temperature changes is associated with bins to the left (right) of the main diagonal in the upper-left (lower right) quarter of Table 4. Underforecasting of temperature changes is associated with bins below (above) the main diagonal in the upper-left (lower right) quarter of Table 5. Underforecasting of

[5] Note that in the tables of the conditional probability of observations given the forecasts (Table 4), comparisons between values must be done along a *row*, while for tables of the conditional probability of forecasts given the observations (Table 5), comparisons must be done along a *column.*

TABLE 3. Joint distribution of observed temperature changes and forecasts. Total number of forecasts or observations in that bin is *N*. The marginal probability of that forecast (observation) in percent is $p(f)$ [$p(x)$]. Number at lower right is number of forecasts in the same bin as the observed temperature change: (a) LFM MOS, (b) NGM MOS, (c) CON, (d) NWSFO.

	Observations												
	≤−25	−20	−15	−10	−5	0	5	10	15	20	≥25	*N*	$p(f)$
a)													
≤−25	**5**	1	0	0	0	0	0	0	0	0	0	6	1.0
−20	3	**6**	0	0	0	0	0	0	0	0	0	9	1.5
−15	4	4	**9**	3	0	0	0	0	0	0	0	20	3.4
−10	2	3	9	**18**	6	4	0	0	0	0	0	42	7.1
−5	0	1	2	15	**36**	21	3	1	0	0	0	79	13.4
0	0	0	1	7	29	**122**	49	8	0	0	0	216	36.6
5	0	1	1	0	3	40	**61**	26	2	1	0	135	22.9
10	0	0	0	0	0	3	18	**28**	10	1	0	60	10.2
15	0	1	0	0	0	0	1	2	**10**	2	2	17	2.9
20	0	0	0	0	0	0	0	0	0	**3**	2	5	0.8
≥25	0	0	0	0	0	0	0	0	0	1	**0**	1	0.2
N	14	16	22	43	74	190	132	65	22	8	4	298	
$p(x)$	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7		
b)													
≤−25	**6**	4	0	0	0	0	0	0	0	0	0	10	1.7
−20	5	**6**	1	0	0	0	0	0	0	0	0	12	2.0
−15	2	2	**11**	3	0	1	0	0	0	0	0	19	3.2
−10	1	1	5	**23**	16	3	0	0	0	0	0	49	8.3
−5	0	2	4	13	**35**	71	9	0	0	0	0	134	22.7
0	0	1	1	3	16	**87**	51	7	0	0	0	166	28.1
5	0	0	0	1	7	24	**52**	26	2	0	0	112	19.0
10	0	0	0	0	0	3	19	**25**	7	0	0	54	9.2
15	0	0	0	0	0	1	1	7	**9**	3	1	22	3.7
20	0	0	0	0	0	0	0	0	4	**4**	2	10	1.7
≥25	0	0	0	0	0	0	0	0	0	1	**1**	2	0.3
N	14	16	22	43	74	190	132	65	22	8	4	259	
$p(x)$	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7		
Forecasts c)													
≤−25	**4**	2	0	0	0	0	0	0	0	0	0	6	1.0
−20	6	**4**	0	0	0	0	0	0	0	0	0	10	1.7
−15	3	6	**10**	3	0	0	0	0	0	0	0	22	3.7
−10	1	2	8	**20**	5	2	0	0	0	0	0	38	6.4
−5	0	1	2	16	**42**	43	3	0	0	0	0	107	18.1
0	0	1	2	3	23	**119**	54	8	0	0	0	210	35.6
5	0	0	0	1	4	24	**59**	21	2	0	0	111	18.8
10	0	0	0	0	0	2	16	**34**	10	2	0	64	10.8
15	0	0	0	0	0	0	0	2	**9**	1	2	14	2.4
20	0	0	0	0	0	0	0	0	1	**4**	1	6	1.0
≥25	0	0	0	0	0	0	0	0	0	1	**1**	2	0.3
N	14	16	22	43	74	190	132	65	22	8	4	306	
$p(x)$	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7		
d)													
≤−25	**8**	3	0	0	0	0	0	0	0	0	0	11	1.9
−20	4	**6**	0	0	0	0	0	0	0	0	0	10	1.7
−15	1	3	**9**	1	0	0	0	0	0	0	0	14	2.4
−10	1	3	10	**20**	5	0	0	0	0	0	0	39	6.6
−5	0	1	2	14	**39**	16	2	0	0	0	0	74	12.5
0	0	0	1	7	24	**139**	43	4	0	0	0	218	36.9
5	0	0	0	1	4	30	**65**	22	2	0	0	124	21.0
10	0	0	0	0	2	5	22	**32**	6	2	0	69	11.7
15	0	0	0	0	0	0	0	7	**12**	2	1	22	3.7
20	0	0	0	0	0	0	0	0	2	**4**	2	8	1.4
≥25	0	0	0	0	0	0	0	0	0	0	**1**	1	0.2
N	14	16	22	43	74	190	132	65	22	8	4	335	
$p(x)$	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7		

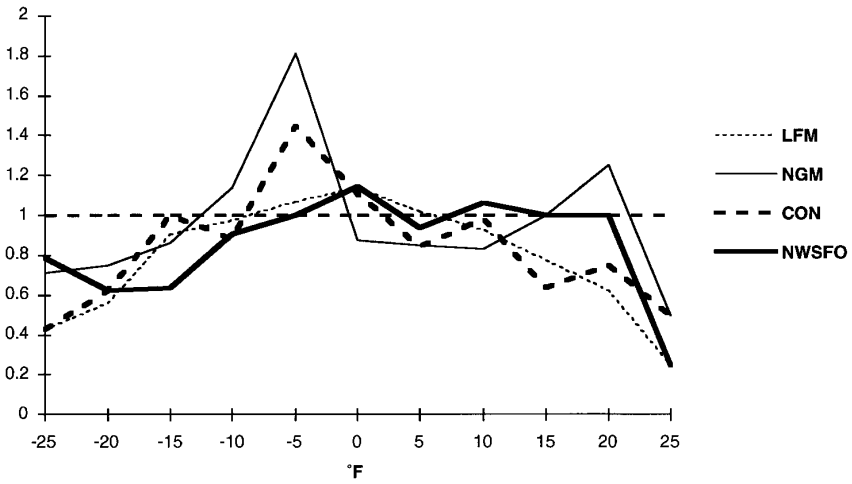

FIG. 1. Ratio of forecast to observed temperature changes by 5°F bins for four forecast systems. Abscissa is center of temperature bin. Ordinate is ratio. Unity (horizontal dashed line) indicates same number of forecasts and observations. Values greater (less) than unity indicate more (fewer) forecasts than observations in a given temperature bin.

changes in temperature appears to be a much more serious problem when viewed from the context of $p(f|x)$ instead of $p(x|f)$ (Fig. 3). This paradox can be seen upon close inspection of Table 3, where the distributions appear more skewed along columns than along rows, but it is more dramatically evident when the conditional probabilities are considered. By using $p(f|x)$, the underforecasting of extreme temperature changes becomes more apparent. In passing, we note the asymmetry in the overforecasting by the NWSFO between forecasts and observations of warming and cooling. Warming is much more likely to be associated with overforecasting than cooling is. We will return to this point in the next section.

The relationship between f and x can also be examined by creating linear regression models between the two to describe the conditional distributions, $p(x|f)$ and $p(f|x)$. The process is described in detail in appendix A of Murphy et al. (1989). To summarize, the expected value of the observations given a particular forecast, $E(x|f)$, is expressed as a linear function of the forecast,[6] by

$$E(x|f) = a + bf, \tag{7}$$

where $a = \langle x \rangle - b\langle f \rangle$ and $b = (s_x/s_f)r_{fx}$. Now, $\langle x \rangle$ and $\langle f \rangle$ are the sample means of the observations and forecasts, respectively; s_x and s_f are the sample standard deviations of the observations and forecasts, respectively; and r_{fx} is the sample correlation coefficient between the forecasts and the observations (Table 6). By plotting the departure of the expected values from the forecast [i.e., $E(x|f) - f$, rather than $E(x|f)$], the behavior of the models becomes more apparent (Fig. 4). The slope of the lines is related to the conditional bias of the forecasts. For example, the NGM MOS is high (low) for forecasts of cooling (warming). The conditional biases of the other forecasts are all of the other sign. Assuming that the bias varies linearly with the temperature forecast range, a user with that information might be able to adjust the forecasts in order to make better use of the forecasts. Over most of the forecast temperature range, the expected value of the observations associated with NWSFO forecasts departs less from the forecast than the expected value associated with the MOS products. Thus, the conditional bias of the NWSFO forecasts is less than that of the guidance products.

5. Points of interest

a. The asymmetry in forecasting warming and cooling

As mentioned earlier, there is an asymmetry in the forecasting of temperature changes by the NWSFO. Cooling is more likely to be underforecast than warming. To illustrate some facets of this asymmetry, we have considered the subset of the data related to observed moderate temperature changes of 3°–17°F (associated with the ±5°, 10°, and 15°F bins in the joint distribution tables). A cursory examination of some of the summary measures of the forecast performance reveals both the underforecasting and the asymmetry (Table 7). Positive (negative) values of ME for forecasts of cooling (warming) indicate underforecasting.

[6] As discussed in Murphy et al. (1989), a model of the expected value of the forecast given a particular observation, $E(f|x)$, can also be constructed. We have chosen to include only the model for $E(x|f)$ here.

TABLE 4. Conditional probability of observations given forecasts and marginal distribution of forecasts. Column and rows are 5°F temperature bins centered on number in heading. Total number (marginal probability) of cases in respective forecast bin is $N[p(f)]$: (a) LFM MOS, (b) NGM MOS, (c) CON, (d) NWSFO.

	Observations												
Forecasts	≤−25	−20	−15	−10	−5	0	5	10	15	20	≥25	*N*	*p*(*f*)
						a)							
≤−25	**83.3**	16.7	.0	.0	.0	.0	.0	.0	.0	.0	.0	6	1.0
−20	33.3	**66.7**	.0	.0	.0	.0	.0	.0	.0	.0	.0	9	1.5
−15	20.0	20.0	**45.0**	15.0	.0	.0	.0	.0	.0	.0	.0	20	3.4
−10	4.8	7.1	21.4	**42.9**	14.3	9.5	.0	.0	.0	.0	.0	42	7.1
−5	.0	1.3	2.5	19.0	**45.6**	26.6	3.8	1.3	.0	.0	.0	79	13.4
0	.0	.0	.5	3.2	13.4	**56.5**	22.7	3.7	.0	.0	.0	216	36.6
5	.0	.7	.7	.0	2.2	29.6	**45.2**	19.3	1.5	.7	.0	135	22.9
10	.0	.0	.0	.0	.0	5.0	30.0	**46.7**	16.7	1.7	.0	60	10.2
15	.0	.0	.0	.0	.0	.0	5.9	11.8	**58.8**	11.8	11.8	17	2.9
20	.0	.0	.0	.0	.0	.0	.0	.0	.0	**60.0**	40.0	5	0.8
≥25	.0	.0	.0	.0	.0	.0	.0	.0	.0	100.0	**.0**	1	0.2
						b)							
≤−25	**60.0**	40.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	10	1.7
−20	41.7	**50.0**	8.3	.0	.0	.0	.0	.0	.0	.0	.0	12	2.0
−15	10.5	10.5	**57.9**	15.8	.0	5.3	.0	.0	.0	.0	.0	19	3.2
−10	2.0	2.0	10.2	**46.9**	32.7	6.1	.0	.0	.0	.0	.0	49	8.3
−5	.0	1.5	3.0	9.7	**26.1**	53.0	6.7	.0	.0	.0	.0	134	22.7
0	.0	.6	.6	1.8	9.6	**52.4**	30.7	4.2	.0	.0	.0	166	28.1
5	.0	.0	.0	.9	6.3	21.4	**46.4**	23.2	1.8	.0	.0	112	19.0
10	.0	.0	.0	.0	.0	5.6	35.2	**46.3**	13.0	.0	.0	54	9.2
15	.0	.0	.0	.0	.0	4.5	4.5	31.8	**40.9**	13.6	4.5	22	3.7
20	.0	.0	.0	.0	.0	.0	.0	.0	40.0	**40.0**	20.0	10	1.7
≥25	.0	.0	.0	.0	.0	.0	.0	.0	.0	50.0	**50.0**	2	.3
						c)							
≤−25	**66.7**	33.3	.0	.0	.0	.0	.0	.0	.0	.0	.0	6	1.0
−20	60.0	**40.0**	.0	.0	.0	.0	.0	.0	.0	.0	.0	10	1.7
−15	13.6	27.3	**45.5**	13.6	.0	.0	.0	.0	.0	.0	.0	22	3.7
−10	2.6	5.3	21.1	**52.6**	13.2	5.3	.0	.0	.0	.0	.0	38	6.4
−5	.0	.9	1.9	15.0	**39.3**	40.2	2.8	.0	.0	.0	.0	107	18.1
0	.0	.5	1.0	1.4	11.0	**56.7**	25.7	3.8	.0	.0	.0	210	35.6
5	.0	.0	.0	.9	3.6	21.6	**53.2**	18.9	1.8	.0	.0	111	18.8
10	.0	.0	.0	.0	.0	3.1	25.0	**53.1**	15.6	3.1	.0	64	10.8
15	.0	.0	.0	.0	.0	.0	.0	14.3	**64.3**	7.1	14.3	14	2.4
20	.0	.0	.0	.0	.0	.0	.0	.0	16.7	**66.7**	16.7	6	1.0
≥25	.0	.0	.0	.0	.0	.0	.0	.0	.0	50.0	**50.0**	2	0.3
						d)							
≤−25	**72.7**	27.3	.0	.0	.0	.0	.0	.0	.0	.0	.0	11	1.9
−20	40.0	**60.0**	.0	.0	.0	.0	.0	.0	.0	.0	.0	10	1.7
−15	7.1	21.4	**64.3**	7.1	.0	.0	.0	.0	.0	.0	.0	14	2.4
−10	2.6	7.7	25.6	**51.3**	12.8	.0	.0	.0	.0	.0	.0	39	6.6
−5	.0	1.4	2.7	18.9	**52.7**	21.6	2.7	.0	.0	.0	.0	74	12.5
0	.0	.0	.5	3.2	11.0	**63.8**	19.7	1.8	.0	.0	.0	218	37.0
5	.0	.0	.0	.8	3.2	24.2	**52.4**	17.7	1.6	.0	.0	124	21.0
10	.0	.0	.0	.0	2.9	7.2	31.9	**46.4**	8.7	2.9	.0	69	11.7
15	.0	.0	.0	.0	.0	.0	.0	31.8	**54.5**	9.1	4.5	22	3.7
20	.0	.0	.0	.0	.0	.0	.0	.0	25.0	**50.0**	25.0	8	1.4
≥25	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	**100.0**	1	0.2

The NWSFO forecasts have the largest ME for cases of cooling and the smallest ME for warming. In terms of MAE and rmse, the NGM and CON forecasts outperform the NWSFO for cooling, although NWSFO does much better on warming. The asymmetry appears to result, for the most part, from the warm bias of the NWSFO forecasts. As seen in Table 1, NWSFO is 0.49°F warmer than the observations. If we subtract

TABLE 5. Conditional probability of forecasts given observations. Column and rows are 5°F temperature bins centered on number in heading. Total number (marginal distribution) of observations in each respective bin is $N[p(x)]$: (a) LFM MOS, (b) NGM MOS, (c) CON, (d) NWSFO.

		Observations										
		≤−25	−20	−15	−10	−5	0	5	10	15	20	≥25
						a)						
	≤−25	**35.7**	6.3	.0	.0	.0	.0	.0	.0	.0	.0	.0
	−20	21.4	**37.5**	.0	.0	.0	.0	.0	.0	.0	.0	.0
	−15	28.6	25.0	**40.9**	7.0	.0	.0	.0	.0	.0	.0	.0
	−10	14.3	18.8	40.9	**41.9**	8.1	2.1	.0	.0	.0	.0	.0
	−5	.0	6.3	9.1	34.9	**48.6**	11.1	2.3	1.5	.0	.0	.0
	0	.0	.0	4.5	16.3	39.2	**64.2**	37.1	12.3	.0	.0	.0
	5	.0	6.3	4.5	.0	4.1	21.1	**46.2**	40.0	9.1	12.5	.0
	10	.0	.0	.0	.0	.0	1.6	13.6	**43.1**	45.5	12.5	.0
	15	.0	.0	.0	.0	.0	.0	.8	3.1	**45.5**	25.0	50.0
	20	.0	.0	.0	.0	.0	.0	.0	.0	.0	**37.5**	50.0
	≥25	.0	.0	.0	.0	.0	.0	.0	.0	.0	12.5	**.0**
	N	14	16	22	43	74	190	132	65	22	8	4
	p(*x*)	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7
						b)						
	≤−25	**42.9**	25.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
	−20	35.7	**37.5**	4.5	.0	.0	.0	.0	.0	.0	.0	.0
	−15	14.3	12.5	**50.0**	7.0	.0	.5	.0	.0	.0	.0	.0
	−10	7.1	6.3	22.7	**53.5**	21.6	1.6	.0	.0	.0	.0	.0
	−5	.0	12.5	18.2	30.2	**47.3**	37.4	6.8	.0	.0	.0	.0
	0	.0	6.3	4.5	7.0	21.6	**45.8**	38.6	10.8	.0	.0	.0
	5	.0	.0	.0	2.3	9.5	12.6	**39.4**	40.0	9.1	.0	.0
	10	.0	.0	.0	.0	.0	1.6	14.4	**38.5**	31.8	.0	.0
	15	.0	.0	.0	.0	.0	.5	.8	10.8	**40.9**	37.5	25.0
	20	.0	.0	.0	.0	.0	.0	.0	.0	18.2	**50.0**	50.0
	≥25	.0	.0	.0	.0	.0	.0	.0	.0	.0	12.5	**25.0**
	N	14	16	22	43	74	190	132	65	22	8	4
	p(*x*)	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7
Forecasts						c)						
	≤−25	**28.6**	12.5	.0	.0	.0	.0	.0	.0	.0	.0	.0
	−20	42.9	**25.0**	.0	.0	.0	.0	.0	.0	.0	.0	.0
	−15	21.4	37.5	**45.5**	7.0	.0	.0	.0	.0	.0	.0	.0
	−10	7.1	12.5	36.4	**46.5**	6.8	1.1	.0	.0	.0	.0	.0
	−5	.0	6.3	9.1	37.2	**56.8**	22.6	2.3	.0	.0	.0	.0
	0	.0	6.3	9.1	7.0	31.1	**62.6**	40.9	12.3	.0	.0	.0
	5	.0	.0	.0	2.3	5.4	12.6	**44.7**	32.3	9.1	.0	.0
	10	.0	.0	.0	.0	.0	1.1	12.1	**52.3**	45.5	25.0	.0
	15	.0	.0	.0	.0	.0	.0	.0	3.1	**40.9**	12.5	50.0
	20	.0	.0	.0	.0	.0	.0	.0	.0	4.5	**50.0**	25.0
	≥25	.0	.0	.0	.0	.0	.0	.0	.0	.0	12.5	**25.0**
	N	14	16	22	43	74	190	132	65	22	8	4
	p(*x*)	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7
						d)						
	≤−25	**57.1**	18.8	.0	.0	.0	.0	.0	.0	.0	.0	.0
	−20	28.6	**37.5**	.0	.0	.0	.0	.0	.0	.0	.0	.0
	−15	7.1	18.8	**40.9**	2.3	.0	.0	.0	.0	.0	.0	.0
	−10	7.1	18.8	45.5	**46.5**	6.8	.0	.0	.0	.0	.0	.0
	−5	.0	6.3	9.1	32.6	**52.7**	8.4	1.5	.0	.0	.0	.0
	0	.0	.0	4.5	16.3	32.4	**73.2**	32.6	6.2	.0	.0	.0
	5	.0	.0	.0	2.3	5.4	15.8	**49.2**	33.8	9.1	.0	.0
	10	.0	.0	.0	.0	2.7	2.6	16.7	**49.2**	27.3	25.0	.0
	15	.0	.0	.0	.0	.0	.0	.0	10.8	**54.5**	25.0	25.0
	20	.0	.0	.0	.0	.0	.0	.0	.0	9.1	**50.0**	50.0
	≥25	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	**25.0**
	N	14	16	22	43	74	190	132	65	22	8	4
	p(*x*)	2.4	2.7	3.7	7.3	12.5	32.2	22.4	11.0	3.7	1.4	0.7

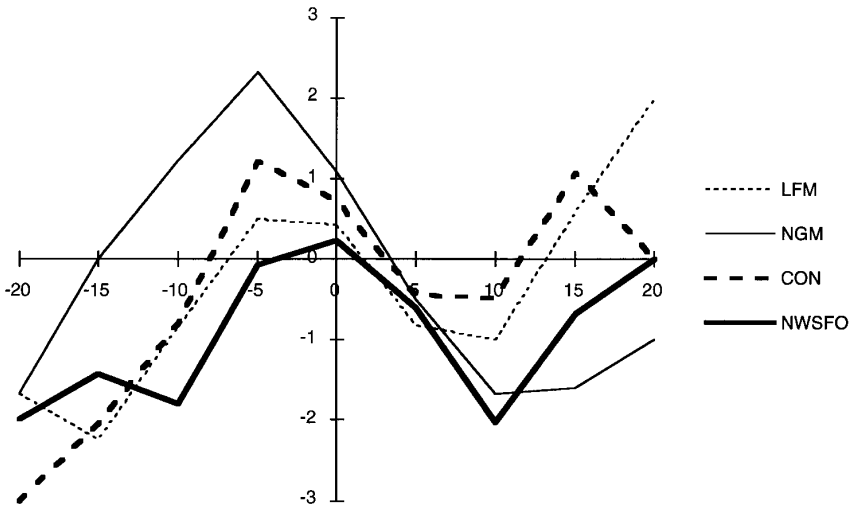

FIG. 2. Departures from perfect reliability of various temperature forecasts. Abscissa is forecast temperature change in °F. Ordinate is difference between average temperature of observations associated with forecasts and the forecasts in each bin. Positive (negative) values indicate that observations are warmer (cooler) than the forecasts.

0.49°F from each of the NWSFO forecast temperature changes in an effort to correct for the bias, we can recompute the summary measures and compare the adjusted NWSFO forecasts to the guidance (Table 8). The adjusted NWSFO performance is much less asymmetric than the unadjusted performance. Although the adjusted NWSFO still performs better in these summary measures for warming events than for cooling, the asymmetry is much less pronounced. The bias of the forecasts was a large part of the signal. This makes intuitive sense, since a warm bias will help in underforecasting of warm events, although hurting in the underforecasting of cool events.

The forecasting of extreme temperature changes gives a different picture than that of moderate temperature changes. For observed changes of more than 17°F, NWSFO improves more on guidance for cooling than for warming (Table 9). The large difference in performance of the LFM and NGM is particularly striking. It is the poor performance of the LFM in these extreme events that led to the difference seen in the overall MAE and rmse noted in section 3. It also means that, unlike for smaller temperature changes, CON is outperformed by the NGM MOS in this case. The NGM MOS is the most accurate forecast for the warming events. This is interesting in light of the overall cold bias of the NGM. Sample sizes are much smaller, of course, so that this may be an artifact. It is likely that these very large day-to-day changes in temperature have the most impact on the public and on which value can be added by providing accurate forecasts. A histogram of forecast errors highlights the difference in the various forecasting systems (Fig. 5). Despite a bias toward underforecasting changes, the NWSFO has the fewest very large errors, with only one forecast more than 12°F too low compared to five or six for the guidance. In a sense, for these very large changes, the NWSFO forecast adds a great deal of potential value for users on this small number of days by avoiding extremely large forecast errors.

b. The relationship of NWSFO to guidance

A typical question considered in verification studies involving human forecasters is that of how much ''value'' the humans add to numerically generated guidance.[7] Here we will touch briefly on this question, comparing the NWSFO to CON, which was the best of the objective guidance products discussed here. There are several possible approaches for considering the situations in which humans could add value. The first is to look at the kinds of errors associated with the spread between the LFM MOS and NGM MOS used to generate CON. In this dataset, the two MOS values never disagree by more than 12°F. Combining the ends of the distribution of the spread of MOS differences, we have calculated the improvement in RMSE over CON by NWSFO as a function of the difference between the input MOS values (Fig. 6). Although the RMSE for CON is fairly constant (between approximately 3.5° and 4.5°F), the relative performance of NWSFO varies

[7] As noted by Murphy (1993) and in the introduction here, forecasts take on value only by being used by someone. We are using the term qualitatively here, under the presumption that large (~10°F) improvements in a temperature forecast will provide value for virtually all users.

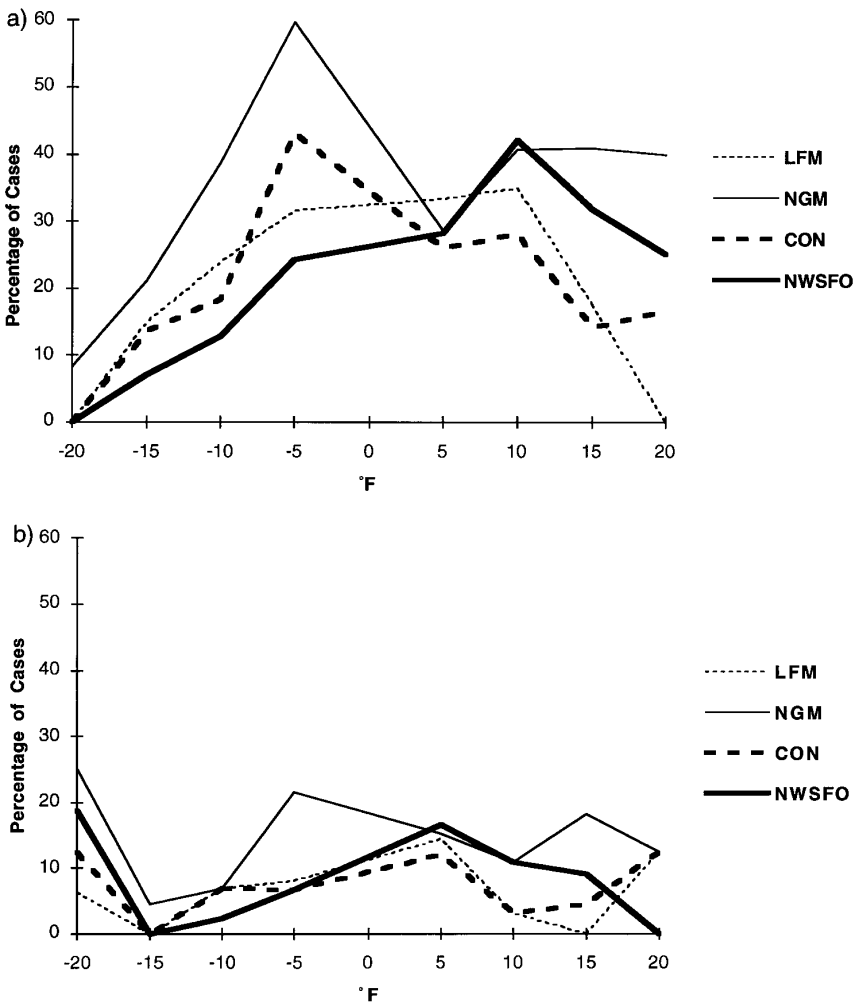

FIG. 3. Percentage of overforecasts of temperature changes by (a) forecast temperature change and (b) observed temperature change. Abscissa is temperature bin, and ordinate is percentage.

markedly. In cases where the NGM MOS is 2°–4°F cooler than the LFM MOS, the NWSFO improves over CON by approximately 20% in rmse. On the other hand, when NGM MOS is 1°–4°F warmer, the NWSFO does approximately 5%–10% *worse* than CON. This latter feature is curious, and we can offer no explanation for it, although it certainly warrants further study.

TABLE 6. Parameters associated with linear regression model for expected value of observations given forecast.

Measure	LFM	NGM	CON	NWSFO	Obs
Mean	.30	−.47	−.08	.64	.15
Standard deviation	7.56	8.41	7.60	8.17	9.17
Correlation with observations	.87	.88	.90	.91	
a	−.17	.60	.23	−.51	
b	1.06	.96	1.09	1.02	

A second approach is to look at the cases where the NWSFO disagreed with CON. In general, this did not happen very often during the period of study. There were 26 times when the NWSFO disagreed by more than 5°F with CON, 13 on each side of the CON forecast. The RMSE plotted by the difference in forecasts shows that the NWSFO, in general, slightly outperforms CON (Fig. 7). It also shows that when the two forecasts are in close agreement, they are both more accurate, in terms of the rmse. (Note that this is in contrast to the rather flat nature of the rmse of CON as a function of the difference in NGM and LFM MOS, as seen in Fig. 6.) There is approximately 2°F lower rmse when the NWSFO is 1°F warmer than the CON than the rmse when the NWSFO is either 5°F warmer or 3°F cooler than CON. An average forecast of the NWSFO and CON can be computed (''NWSCON'') and, over most of the range, it adds little value to NWSFO and CON from the standpoint of the rmse.

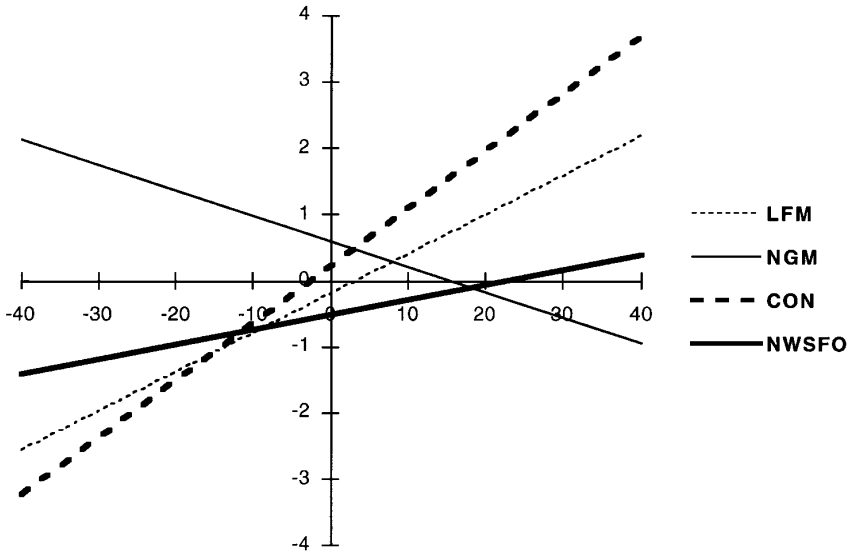

FIG. 4. Lines associated with linear regression models of the expected value of observations given forecasts. Plotted lines are $E(x|f) - f$. Abscissa is forecast temperature in °F, and ordinate is difference in °F between the expected value of the observations from the linear regression model and the actual forecast. Positive (negative) values indicate expected value of observation is warmer (cooler) than the forecast.

This implies that, at least in some statistical respects, the NWSFO and CON forecasts are not very different.

A final important step in verification is to look back at the cases that lead to some of the interesting results. As mentioned above, there were 26 times when the NWSFO and CON forecasts disagreed by more than 5°F. These cases are listed in Table 10, in order of increasing improvement by the NWSFO over CON. As would be expected, most of the cases are from the winter or transition seasons, with only one being in the summer. Seven cases have errors of opposite sign from NWSFO and CON, where the errors are large enough that the average of the two forecasts (NWSCON) beats both NWSFO and CON. In the remaining 19 cases, NWSFO is more accurate in 11 (42% of the total). Of the five disagreements of 10°F or more, the NWSFO is more accurate than CON in the two cases where the forecast errors are of the same sign.

These cases of large disagreement between NWSFO and CON provide an opportunity for improvement in temperature forecasting. Their identification means that they can be studied more closely in an effort to understand the reasons why the NWSFO disagreed with CON, and, of particular importance, it may be possible to discern when it is advantageous to disagree with the guidance products in the future. It would be hoped then that forecasters could learn (a) when they have a better

TABLE 7. Simple statistics as in Table 1, stratified by observed cooling of 3°–17°F and observed warming of 3°–17°F. Numbers in parenthesis are percentage improvement by unadjusted NWSFO forecast over guidance product. Negative values indicate guidance performed better in this parameter. Number of forecasts indicated at upper left.

	Cooling			
$N = 139$	LFM	NGM	CON	NWSFO
ME	2.45	1.63	2.24	2.78
MAE	3.63 (5.35)	3.30 (−4.14)	3.19 (−7.90)	3.44
RMSE	4.73 (4.18)	4.40 (−3.21)	4.23 (−7.28)	4.54
	Warming			
$N = 219$	LFM	NGM	CON	NWSFO
ME	−1.95	−1.86	−2.08	−1.15
MAE	3.20 (17.40)	3.31 (20.14)	3.01 (12.27)	2.64
RMSE	3.99 (15.95)	4.09 (17.86)	3.70 (9.28)	3.36

TABLE 8. As in Table 7 except for adjusted NWSFO forecast (cooled by 0.49°F).

	Cooling			
$N = 139$	LFM	NGM	CON	NWSFO
ME	2.45	1.63	2.24	2.29
MAE	3.63 (12.24)	3.30 (3.44)	3.19 (−0.05)	3.19
RMSE	4.73 (10.14)	4.40 (3.22)	4.23 (−0.60)	4.25
	Warming			
$N = 219$	LFM	NGM	CON	NWSFO
ME	−1.95	−1.86	−2.08	−1.64
MAE	3.20 (10.76)	3.31 (13.72)	3.01 (5.22)	2.86
RMSE	3.99 (11.01)	4.09 (13.03)	3.70 (3.95)	3.55

TABLE 9. As in Table 7 except for extreme temperature changes (≤ -18°F and ≥ 18°F).

		Cooling		
$N = 30$	LFM	NGM	CON	NWSFO
ME	6.83	4.73	6.07	3.47
MAE	7.43 (30.94)	6.13 (16.30)	6.80 (24.51)	5.13
RMSE	9.00 (29.10)	7.88 (19.03)	8.32 (23.28)	6.38
		Warming		
$N = 12$	LFM	NGM	CON	NWSFO
ME	−5.25	−2.00	−3.83	−3.75
MAE	5.75 (17.39)	3.83 (−23.91)	4.83 (1.72)	4.75
RMSE	7.21 (23.46)	4.30 (−28.22)	5.74 (3.99)	5.52

opportunity to improve upon MOS forecasts significantly and (b) when MOS is an adequate forecast and can be used without change.

6. Discussion

We have looked at the verification of 12–24-h high-temperature forecasts for Oklahoma City from a distribution-oriented approach. The impression one gets of the performance of the various forecast systems depends on how complete a set of descriptors one uses. If the approach to verification is limited to simple summary measures, the richness of the relationship between forecasts and observations is lost. What appear as issues of fundamental importance when considering a distributions-oriented approach to verification cannot even be asked with a measures-oriented approach, since the presentation of the data does not allow the issues to be *identified.* Simple summary measures of overall performance offer almost no information about the relationship between forecasts and errors, and, as a result, it is difficult to learn about the occasions on which human forecasters can improve significantly on numerical guidance.

If one believes that the point of human intervention in weather forecasting is to provide information that will allow users to gain value from forecasts, and that small improvements in accuracy (say 1°–2°F) have little significant impact on the large majority of users, then it is imperative to consider the *distribution* of errors. In particular, overall summary measures can confuse the potential value added in a small, but highly significant, set of cases by being swamped by information from the very large number of ''less important'' forecast situations. One interpretation of the errors in forecasting extreme temperature changes here is that the NWSFO adds significant value to the numerical guidance on about 5 days in the dataset (as measured by the reduction in very large underforecasts of large temperature changes). In comparison to the 590 days in the dataset, that number seems very small, but in comparison to the 42 days on which large changes took place, it becomes a much more significant contribution. This final point adds a cautionary note to the use of distributions-based verification systems associated with the large dimensionality of the verification problem. The use of distribution-based approaches means that the ''impressions'' of the forecast system will necessarily be based on smaller sample sizes. Thus, while the distributions-oriented verification potentially offers a more complete picture of forecast system performance, it must be used with care and adequate sample sizes collected.

We also identified two interesting features in the NWSFO forecasts. The first is a pair of asymmetries in

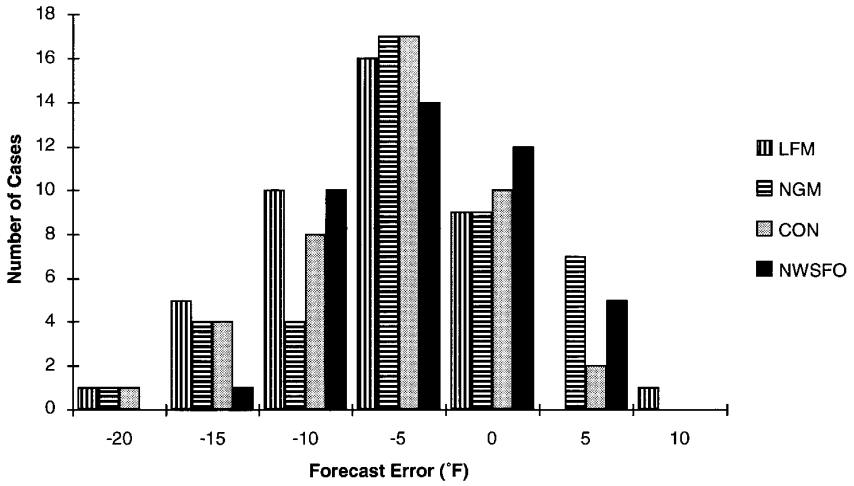

FIG. 5. Histogram of errors for forecast change for cases of observed changes more than 17°F. Errors are binned in 5°F bins centered on −20°, −15°, −10°F, etc. Negative (positive) values indicate that the temperature change was underforecast (overforecast).

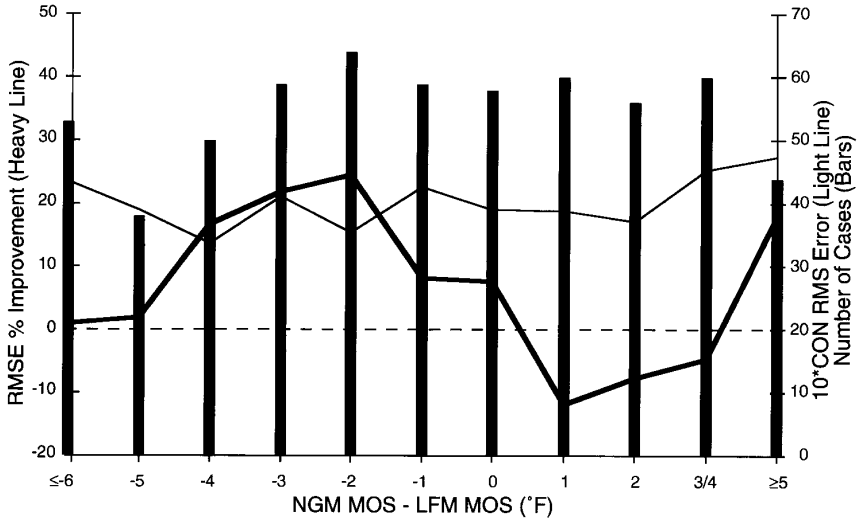

FIG. 6. rmse of CON forecast (light line) and percentage improvement by NWSFO over CON (heavy line) as a function of the disagreement between NGM MOS and LFM MOS. Light dashed line is zero improvement. Abscissa is difference between NGM MOS and LFM MOS such that positive values indicate that NGM MOS is warmer than LFM MOS. Left vertical scale indicates percentage improvement in rmse by NWSFO compared to CON. Right vertical scale indicates rmse of CON, multiplied by 10, and number of cases in each category (vertical bars).

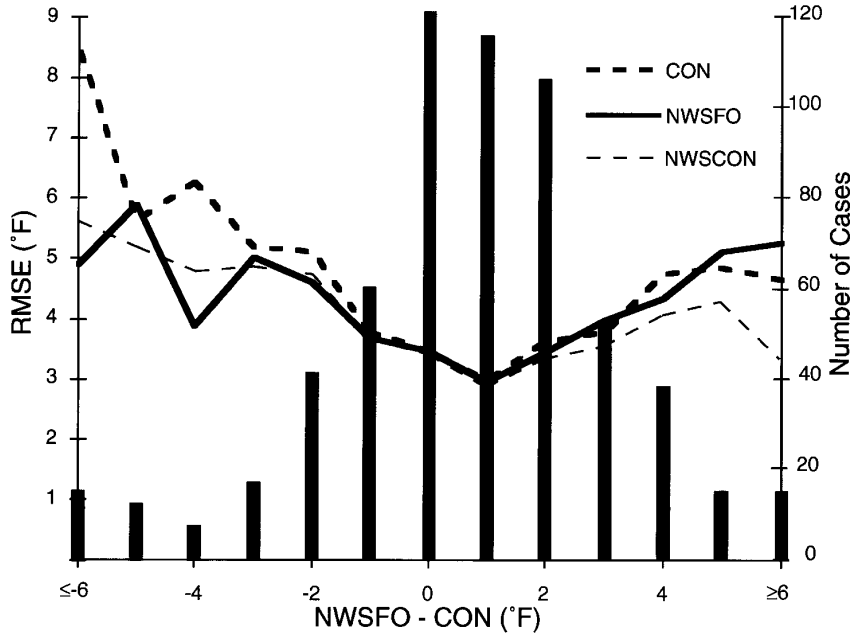

FIG. 7. rmse of CON (heavy-dashed line) and NWSFO (solid line) as function of the difference in the two forecasts. The rmse of an average of CON and NWSFO (NWSCON) is plotted as the light-dashed line. Vertical bars indicate number of cases in each category. Abscissa is the difference between NWSFO and CON in °F, with positive values indicating NWSFO forecast is warmer. Left vertical scale is rmse in °F. Right vertical scale is number of cases (vertical bars).

TABLE 10. CON and NWSFO forecasts and errors for 26 days on which NWSFO and CON disagreed by more than 5°F. ''NWSFO improvement'' is the difference in the error of the two statistics, with the sign convention such that positive values indicate lower error for NWSFO.

	Forecast			Error		
Date	CON	NWSFO	Observed	CON	NWSFO	NWSFO improvement
2 Feb 1993	−6	5	−4	−2	9	−7
22 Feb 1993	−1	5	−2	1	7	−6
4 Feb 1994	−11	−5	−11	0	6	−6
19 Feb 1994	2	8	−3	5	11	−6
27 Mar 1994	0	−6	0	0	−6	−6
10 Mar 1994	25	14	22	3	−8	−5
14 Jan 1993	4	−2	3	1	−5	−4
30 Dec 1993	0	6	1	−1	5	−4
26 Sep 1993	3	−3	1	2	−4	−2
5 Nov 1993	−28	−36	−31	3	−5	−2
9 Dec 1993	8	16	11	−3	5	−2
6 Jan 1994	−16	−27	−22	6	−5	1
31 Jan 1993	9	15	13	−4	2	2
2 Jul 1994	−9	−3	−5	−4	2	2
9 Jan 1993	−2	5	3	−5	2	3
24 Nov 1993	−31	−39	−37	6	−2	4
9 Jan 1994	2	10	8	−6	2	4
24 Feb 1994	24	33	31	−7	2	5
15 Feb 1993	−6	−12	−13	7	1	6
16 Feb 1993	−3	−9	−12	9	3	6
29 Oct 1993	−27	−21	−20	−7	−1	6
21 Feb 1994	−3	−11	−10	7	−1	6
9 Mar 1994	5	−1	−1	6	0	6
27 Dec 1993	−16	−9	−9	−7	0	7
25 Feb 1994	−15	−25	−30	15	5	10
8 Feb 1994	2	−9	−18	20	9	11

the forecasting of temperature changes. For moderate changes (3–17°F), NWSFO forecasts warming events more accurately than cooling. In fact, the NGM MOS and CON forecasts outperform NWSFO on the cooling events over this range. The asymmetry appears in large part due to a bias toward higher temperatures in the NWSFO forecasts. For extreme events ($\geq$18°F), however, the NWSFO forecasts of cooling are much more accurate than those of warming and outperform the numerical guidance. The second feature is an improvement over guidance by NWSFO for those cases where the NGM MOS is a few degrees cooler than the LFM, although doing worse when NGM MOS is slightly warmer than the LFM. These two features suggest that it should be possible to improve the accuracy of temperature forecasts by using some fairly simple strategies taking into account the performance of the various guidance forecast systems.

We have looked at only one forecast element at one forecast lead time. A complete verification would necessitate looking at all forecast elements at all lead times. In the absence of that, it is impossible to know what the current state of forecasting is. As a result, it will be impossible to monitor the impacts of future changes in forecasting techniques and in the forecasting environment, such as those associated with the modernization of the NWS. A fundamental question facing the NWS in the future is the allocation of scarce resources. An ongoing comprehensive verification system has the potential to identify needs and opportunities for improving forecasts through entry-level training, ongoing training, and improved forecast techniques. If small improvements leading to small value for users cost large sums of money, it is economically unwise to pursue them. If, on the other hand, opportunities exist for adding large potential value to forecasts, it is economically unwise to ignore them. Unfortunately, at this time, the verification system within the NWS is inadequate to provide decision makers enough information to make choices about the potential value of forecasts.

Forecast verification is, of course, of importance to more than just the NWS. Private forecasters need to show that users get increased value from their products over those freely available from the NWS. As a result, the issue of the proper approach to forecast verification goes beyond the public sector. It is of importance to anyone who makes or uses forecasts on a regular basis. It is in the interest of both parties to move toward a complete distributions-oriented approach to verification. Failing to do so will limit the value of weather forecasting in the future.

Acknowledgments. We wish to thank the staff at NWSFO OUN for their willingness to share the data we have used. Allan Murphy provided inspiration for the project through ongoing conversations over a period of several years, as well as commenting on the draft manuscript. We also thank Arthur Witt of NSSL and an anonymous reviewer for their constructive comments on the manuscript.

REFERENCES

Brier, G. W., 1948: Review of ''The verification of weather forecasts.'' *Bull. Amer. Meteor. Soc.,* **29,** 475.

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.

——, 1996: The Finley Affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

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

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

NOAA, 1984: *National Weather Service Operations Manual.* Chapter C-43. [Available from National Weather Service, Office of Meteorology, 819 Taylor Street, Rm. 10A26, Fort Worth, TX 76102.]

NWS Southern Region Headquarters, 1984: Public weather verification. 4 pp. [Available from NWS Southern Region, 1325 East-West Highway, Silver Spring, MD 20915.]

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.,* **76,** 1157–1164.