Subjective Probabilistic Tornado Forecasts: Some Experimental Results

ALLAN H. MURPHY

Department of Atmospheric Sciences, Oregon State University, Corvallis 97331

ROBERT L. WINKLER

Graduate School of Business, Indiana University, Bloomington 47405

(Manuscript received 15 December 1981, in final form 17 May 1982)

ABSTRACT

An experiment was conducted at the National Severe Storms Forecast Center during 1976 and 1977 in which National Weather Service forecasters formulated probabilistic forecasts of several tornado events in conjunction with both severe weather outlooks and severe thunderstorm and tornado watches. The results indicate that the probabilistic forecasts associated with the outlooks were quite reliable and exhibited positive skill, relative to forecasts based on sample climatological probabilities. The probabilistic forecasts associated with the watches, however, were less reliable and skillful. In view of the lack of prior experience at making probabilistic tornado forecasts, as well as the absence of feedback, comparable objective probabilistic guidance, and even appropriate past data on which to base climatological probabilities, the results of the experiment are quite encouraging. Some suggestions for further work in probabilistic tornado forecasting are provided.

1. Introduction

Severe local storm forecasts in the United States are formulated by National Weather Service (NWS) forecasters at the National Severe Storms Forecast Center (NSSFC) in Kansas City, Missouri. In this context, severe local storms are defined as the occurrence of one or more tornadoes or one or more severe thunderstorms accompanied by large hail and/or damaging winds. Two types of forecasts are issued by NSSFC forecasters, namely outlooks and watches. Outlooks are prepared routinely two or three times a day and they delineate those areas, if any, in the United States that are expected to experience severe weather during subsequent periods of fixed length (generally, 24, 21, or 16 hours). However, a watch can be issued at any time that one of the following events is expected to occur within the next six hours: a) one or more tornadoes (a tornado watch) and b) one or more severe thunderstorms (a severe thunderstorm watch). Watches are generally defined over a rectangular "box" ~93 000 km² (27 000 n mi²) in area, and they are usually valid for periods of four to six hours. For a more detailed description of the NWS severe weather forecasting and warning programs, including NSSFC activities, see Mogil and Groper (1977).

In preparing the outlooks and watches, NSSFC forecasters make use of information from a variety of conventional and nonconventional sources, including satellite and radar data. In addition, they receive guidance forecasts relating to the occurrence of thunderstorms and severe local storms from the National Meteorological Center (NMC) on an operational basis (Charba, 1979; Reap and Foster, 1979a). These objective forecasts are based on classical statistical and model output statistics (MOS) procedures, and they are expressed in probabilistic terms. On the other hand, the official outlook and watch forecasts formulated by NSSFC forecasters are expressed in categorical terms. That is, the uncertainty inherent in these subjective forecasts is not specified.

It is generally agreed that forecasts expressed in probabilistic terms offer at least two important advantages compared with traditional categorical forecasts. First, this mode of expression provides forecasters with a means of quantitatively describing the uncertainty inherent in their forecasts. Second, probabilistic forecasts provide potential users of the forecasts with valuable information for making rational decisions in uncertain situations. Moreover, recent results from the NWS operational probability of precipitation (PoP) forecasting program (e.g., Cummins, 1974; MacDonald, 1977; Charba and Klein, 1980) and from several probability forecasting experiments (e.g., Murphy and Winkler, 1977a; Winkler and Murphy, 1979; Murphy, 1981) indicate that NWS forecasters can formulate reliable and skillful probability forecasts for a variety of variables or events.

In view of the importance of severe weather outlooks and watches and their actual and potential im-
pacts on the general public, it seemed desirable to investigate, on an experimental basis, the ability of NSSFC forecasters to quantify the uncertainty in their forecasts. In addition, such a study could be expected to yield useful information concerning the degree of uncertainty associated with current outlooks and watches and the nature of any tendency on the part of NSSFC forecasters to overlook or underforecast. The purpose of this paper is to present the results of an experiment in subjective probabilistic tornado forecasting. The experiment was conducted at NSSFC during 1976 and 1977, and some preliminary results of this experiment were presented by Murphy and Winkler (1979) [see also Murphy and Winkler (1977b)]. In Section 2, we briefly describe the nature of the experiment. Section 3 contains a short discussion of the methods used to evaluate the forecasts and some problems encountered in the evaluation process. Experimental results related to the probabilistic outlook and watch forecasts are presented in Sections 4 and 5, respectively. Section 6 contains a discussion and conclusion, including some suggestions for future work in this area.

2. The experiment

The experiment involved the formulation of probabilistic forecasts of several tornado events in conjunction with both severe weather outlooks and severe thunderstorm and tornado watches. The experimental forecasts were formulated by the forecasters at NSSFC who regularly prepare these official outlooks and watches. The outlooks of concern here were issued each day at approximately 0800 GMT and delineate those areas, if any, in the United States that are expected to experience severe weather during the period from 1200 GMT that day to 1200 GMT the next day. During the period of the experiment, the forecasters who formulate the official (0800 GMT) outlooks were asked to assign probabilities to the following events: a) one or more tornadoes will occur in the severe weather area(s) delineated in the outlook during the (24-hour) period in question and b) ten or more tornadoes will occur anywhere in the United States during that period. We shall refer to the experimental probability forecasts of these two events as the O1 and O2 forecasts, respectively.

The forecasters who issue the official watches were asked to assign probabilities to the following events: a) one or more tornadoes will occur within the watch; b) three or more tornadoes will occur within the watch; and c) at least one tornado occurring within the watch will attain a rating of two (1976)/three (1977) or more on the FPP scale (Fujita and Pearson, 1973). Probabilities were assigned to these three events in conjunction with both severe thunderstorm and tornado watches. We shall refer to the experimental probability forecasts of these three events as the W1, W2, and W3 forecasts, respectively.

The experiment covered a period from February through June in both 1976 and 1977. During this ten-month period, 177 and 451 sets of experimental forecasts were made in conjunction with the official outlooks and watches, respectively. Observations related to tornado and severe thunderstorm occurrences were obtained from the NSSFC verification data base.

3. Methods of evaluation

Since the primary purpose of this experiment was to investigate the ability of NSSFC forecasters to quantify the uncertainty in their outlook and watch forecasts, we are concerned here principally with the reliability and skill of the experimental forecasts as probability forecasts. Reliability refers to the degree of correspondence between forecast probabilities and observed relative frequencies over a sample of forecasts. This attribute of the forecasts can be examined qualitatively by means of a reliability diagram in which the relative frequency of occurrence of an event is plotted against the forecast probability for specific probability values (or ranges of probability values). The broken line connecting these sample points is then compared with the diagonal 45° line representing perfect reliability (equality of probability and relative frequency). Reliability diagrams for the experimental outlook and watch forecasts are presented in Figs. 1a and 3a, respectively (see Sections 4 and 5). Reliability is an indicator of the ability of forecasters to quantify uncertainty and, as such, it is an attribute of particular importance in the early stages of a probability forecasting program when forecasters are first learning to express their uncertainty quantitatively.

In examining a reliability diagram for a sample of probability forecasts, it should be kept in mind that departures from perfect reliability may be due simply to sampling variability. Sampling variability, of course, decreases as sample size increases. Points associated with very few forecasts are subject to relatively large sampling variability (e.g., the standard error of the relative frequency can be as large as 0.10 even when as many as 25 forecasts are used to generate a point on a reliability diagram), whereas points associated with a substantial number of forecasts (e.g., 100 or more) are subject to much smaller sampling variability. Since it is desirable to provide some information concerning sampling variability in conjunction with reliability diagrams, each reliability diagram
in this paper is accompanied by a forecast relative frequency diagram in which the relative frequency of use of the various probability values is indicated (see Figs. 1b and 3b in Sections 4 and 5). These relative frequencies also provide an indication of the sharpness of the forecasts. Sharpness refers to the extent to which the sample of forecasts approaches a sample consisting only of forecasts with probabilities equal to zero and/or one (such a sample is perfectly sharp).

Skill is defined as the accuracy of the forecasts of concern, relative to the accuracy of forecasts produced by some reference procedure such as climatology or persistence. Accuracy, in turn, represents the average degree of correspondence between individual forecasts and observations. The accuracy of probability forecasts is frequently measured by the average Brier score (BS) (Brier, 1950). For a sample of $K$ forecasts in a situation involving only two events (e.g., tornado/no tornado), BS can be defined as follows:

$$\overline{BS} = \frac{1}{K} \sum_{k=1}^{K} (r_k - d_k)^2,$$

(1)

where $r_k$ is the probability assigned to the event of interest [e.g., the occurrence of one or more tornadoes in the outlook area(s) in the case of the O1 forecasts] on the $k$th occasion ($0 \leq r_k \leq 1$) and $d_k$ is the observation of the event, in which $d_k = 1$ if the event occurs on the $k$th occasion and $d_k = 0$ otherwise ($k = 1, \ldots, K$). Thus, BS is simply the mean square error of forecasts expressed in probabilistic terms. The values of BS in (1) range from zero (best possible score) to one (worst possible score). The Brier score is a strictly proper scoring rule (see Winkler and Murphy, 1968), since it encourages forecasters to make their forecast probabilities correspond exactly with their best judgments (i.e., it discourages hedging).

A skill score (SS) based on the Brier score can be defined as follows:

$$SS = [1 - (\overline{BS}/\overline{BS}_r)] \times 100\%,$$

(2)

where $\overline{BS}_r$ denotes the average Brier score for forecasts produced by some reference system (e.g., climatology). Such a skill score is positive (negative) when the forecasts of concern are more (less) accurate than the reference forecasts. Specifically, perfect forecasts ($\overline{BS} = 0$) attain a skill score of 100%, whereas forecasts equivalent in accuracy to the reference forecasts ($\overline{BS} = \overline{BS}_r$) receive a skill score of 0%. Measures of skill play an important role in forecast evaluation, since forecasts generally must pass a test of skill before they are considered to be of sufficiently high quality to be disseminated to the public or specific users.

In evaluating the results of the NSSFC experiment, some problems were encountered in obtaining reference forecasts with which to compare the experimental forecasts. These problems were due, in large measure, to the fact that the forecast area and/or valid period for the outlook and watch forecasts vary from occasion to occasion. For example, suitable data were not available from a period prior to the start of the

---

2 As defined in Eq. (1), $\overline{BS}$ is one-half of the original (average) Brier score. That is, we have neglected the score due to the forecasts $1 - r_i$ and observations $1 - d_i$ associated with the complementary event ($k = 1, \ldots, K$).

3 It should be noted that $\overline{BS}_r = 0$ only if the forecasts produced by the reference system are perfect forecasts. Thus, it is not unreasonable to assume that $\overline{BS}_r > 0$ and, as a result, that SS is finite.
experiment to determine the relevant long-term climatological probabilities. Thus, the sample relative frequencies from the experimental period were used as climatological probabilities in computing the reference score $BS$, in (2). It can be shown that the use of sample relative frequencies, instead of long-term climatological probabilities, tends to underestimate the skill of a set of forecasts (e.g., Murphy, 1974).

It also would have been desirable to compare the subjective forecasts formulated in the experiment with the corresponding objective probabilistic forecasts prepared by NMC. Unfortunately, spatial and temporal matching of the two sets of forecasts, which is required to provide the basis for a meaningful comparison, presents the evaluator with a number of difficult problems (see Charba and Burnham, 1978; Reap and Foster, 1979b). As a result, such comparisons were not attempted in this study. It should be noted that the lack of correspondence between the objective MOS forecasts and the experimental subjective forecasts implies, in effect, that the latter were formulated without the aid of directly applicable probabilistic guidance forecasts.

The average Brier score for a sample of $K$ probability forecasts can be partitioned into three terms under the assumption that the forecast probabilities (i.e., the $r_k$s in this case) possess only a finite number, $S$ say, of distinct values $r^s$ ($s = 1, \ldots, S$) (Murphy, 1973). In a two-event situation, this partition divides the sample of $K$ forecasts into $S$ subsamples, each of which contains all of the forecasts corresponding to a distinct probability value (i.e., a particular $r^s$). Under this assumption, $BS$ in (1) can be expressed as

$$BS = \bar{d}(1 - \bar{d}) + \frac{1}{K} \sum_{s=1}^{S} K^s (r^s - \bar{d}^s)^2$$

$$- \frac{1}{S} \sum_{s=1}^{S} K^s (\bar{d}^s - \bar{d})^2,$$

where $K^s$ is the number of forecasts for which $r_k = r^s$ ($\sum s K^s = K$), $\bar{d}^s = (1/K^s) \sum_k d_k$ (the summation $s$ here is taken over all values of $k$ for which $r_k = r^s$), and $\bar{d} = (1/S) \sum s K^s \bar{d}^s$. The first term on the right-hand side (RHS) of (3) is simply the variance of the observations in the sample ($\bar{d}$ is the observed relative frequency of the event of interest over the entire sample—it is sometimes referred to as the sample climatological probability). This term represents a measure of the uncertainty inherent in the observations and its values range from zero (minimum uncertainty—the event of interest occurs on all or none of the $K$ occasions) to 0.25 in two-event situations (maximum uncertainty—the event occurs on exactly one-half of the $K$ occasions). It should be noted that this term does not depend on the forecasts (i.e., on the $r_k$s) in any way.

The second term on the RHS of (3) is the reliability term, which is the weighted mean squared difference between the forecast probabilities and the corresponding observed relative frequencies when these forecast probabilities were assigned to the event of interest. This term, which is non-negative and vanishes only when the forecast probabilities and observed relative frequencies are equal for all $S$ subsamples, provides a quantitative measure of the reliability of the forecasts. The third term on the RHS of (3) is the resolution term, which is the weighted mean squared difference between the observed relative frequencies in the subsamples and the overall (i.e., sample) observed relative frequency. The sign of this term is negative and it vanishes only when all the subsample relative frequencies are equal to the sample relative frequency. Thus, the larger the reliability term the less reliable (and accurate) the forecasts, whereas the larger the resolution term the greater the resolution (and accuracy) of the forecasts. This three-term partition of the Brier score provides quantitative measures of important attributes of probability forecasts and the corresponding observations, and it is used here (in addition to other methods) to evaluate the forecasts formulated in the NSSFC experiment.

Finally, in judging the results of the NSSFC experiment, it should be kept in mind that the sample sizes associated with the experimental outlook and watch forecasts are quite small ($n = 177$ for the outlooks and $n = 451$ for the watches). Thus, attention will be focused here primarily on the reliability and skill of the unstratified samples of forecasts. However, we also have examined the results for various stratifications of the overall samples, including stratifications by year (1976–77), month (February–April/May–June), size of outlook or watch area, and forecaster. When marked differences are found for any such stratifications, they are reported in the paper.

<table>
<thead>
<tr>
<th>Type of forecast</th>
<th>Number of forecasts</th>
<th>Average number of areas</th>
<th>Average size of areas (km$^2$)</th>
<th>Average probabilities</th>
<th>Relative frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>O1</td>
<td>177</td>
<td>1.30*</td>
<td>558 400</td>
<td>0.605</td>
<td>0.571</td>
</tr>
<tr>
<td>O2</td>
<td>177</td>
<td></td>
<td></td>
<td>0.154</td>
<td>0.186</td>
</tr>
</tbody>
</table>

* Based on data for 1976 for those days on which outlook area(s) were specified.
4. Experimental results—outlooks

The average probabilities and relative frequencies of occurrence of the events for the outlook (i.e., O1 and O2) forecasts, as well as some basic information concerning the outlook areas themselves, are presented in Table 1. The average probability exceeds the relative frequency for the O1 forecasts, whereas the opposite is true for the O2 forecasts. However, these differences actually are quite small, indicating that the overall correspondence between average probabilities and relative frequencies for the experimental outlook forecasts was quite good. For the purposes of comparative evaluation, it should be noted that these relative frequencies represent the overall sample climatological probabilities of the events.

The reliability diagram for the outlook forecasts is presented in Fig. 1a. The forecasters were permitted to assign probabilities of 0.00, 0.02, 0.05, 0.10, 0.20, ..., 1.00 to the events of concern. The plotted points, then, indicate the relative frequencies of the events when these probabilities were used. For example, the O1 event [i.e., one or more tornadoes in the outlook area(s)] occurred on 7 (38.9%) of the 18 occasions on which the forecasters assigned a probability of 0.40 to this event (see solid line in Fig. 1a). The diagonal 45° line in this diagram represents perfect reliability, as noted in Section 3. In judging the reliability of the outlook forecasts, it should be kept in mind that the sample sizes for the O1 and O2 forecasts are quite small (n = 177 in both cases), and each point on the reliability diagram is based on only a portion of these observations.

Examination of the reliability diagram for the O1 forecasts in Fig. 1a reveals that these forecasts were quite reliable, although a slight tendency to over-

![Graph](image_url)

Fig. 2. Average number of tornadoes in outlook areas for O1 (solid line) and O2 (dashed line) forecasts as function of forecast probability. Straight lines are simple regression curves fitted to respective broken lines.

forecast (i.e., for probabilities to exceed relative frequencies) is evident for probability values > 0.50. However, the amount of overforecasting is very small. When the reliability of the O1 forecasts was examined as a function of the size of the outlook area(s) (reliability curves not presented to conserve space), the results indicate that a tendency to overforecast existed for the smaller areas (< 550 000 km² (160 000 n mi²) in area) and that a tendency to underforecast existed for the larger areas (> 550 000 km² (160 000 n mi²) in area). The relative frequency of use diagram in Fig. 1b indicates that a roughly uniform distribution existed for probability values > 0.20 for the O1 forecasts. Probability values < 0.20 were assigned to this event on only two days during the course of the experiment.

With regard to the reliability of the O2 forecasts, Fig. 1a reveals that these forecasts were not as reliable as the O1 forecasts. However, examination of the frequency of use distribution for the O2 forecasts in Fig. 1b indicates that the points on the reliability curve in Fig. 1a corresponding to large departures from the 45° line, which are primarily associated with higher probability values, are based on relatively few forecasts. For most probability values associated with an appreciable number of forecasts, the forecasters exhibited a tendency to underforecast. In this regard, probability values in the range from 0.02 to 0.20 were assigned to this event on almost 80% of the forecasting occasions, whereas probabilities exceeding 0.50 were used on only seven days (see Fig. 1b).

The average number of tornadoes in an area would be expected to increase as the forecast probability increases, and the relationships between forecast probability and average number of tornadoes for the O1 and O2 forecasts are depicted in Fig. 2. The straight lines in the diagram represent simple linear regressions fitted to the respective points on the broken lines describing these empirical relationships. The slopes of the regression lines in Fig. 2 confirm the fact that positive relationships exist between the average number of tornadoes in the areas of concern and the forecast probability for both the O1 and O2 forecasts. In general, positive relationships between these variables still exist when the number of severe thunderstorm reports, as well as the number of tornadoes, is considered.

The Brier scores, the terms in the partition of the Brier score, and the skill scores for the outlook forecasts are presented in Table 2. Overall, the O1 and O2 forecasts both exhibited positive skill. That is, the forecasts formulated by the forecasters were more accurate than forecasts based on the overall sample climatological probabilities for the relevant events.4

---

4 Since sample climatological probabilities are used in the computation of skill scores in this paper, it is possible to express the skill score SS (%) as follows: SS = [(Reliability - Resolution)/ Variance of observations] x 100. Thus, SS is positive when the magnitude of the resolution term exceeds the magnitude of the reliability term.
The O1 forecasts were somewhat more skillful than the O2 forecasts (19.1% versus 14.3% improvement over climatology), and this difference in performance is reflected in a smaller reliability term and a larger resolution term for the O1 forecasts than for the O2 forecasts. The difference between the magnitudes of the reliability terms for the O1 and O2 forecasts is consistent with, and provides quantitative support for, the previous discussion of the reliability of the outlook forecasts that was based on examination of Fig. 1a. Stratification of the O1 forecasts by “period” of the tornado season (February–April versus May–June) reveals that the forecasts for the February–April period were more skillful than those for the May–June period (26.3% versus 11.1%; in computing these skill scores, the appropriate sample climatological probabilities for the respective periods were used). A similar stratification of the O2 forecasts shows that the May–June period forecasts were more skillful than the February–April period forecasts for this event (26.8% versus −4.4%).

The results presented here indicate that the NSSFC forecasters who regularly formulate outlooks were reasonably successful in quantifying the uncertainty in these forecasts. Specifically, the O1 forecasts [i.e., the forecasts relating to the occurrence of one or more tornadoes in the outlook area(s)] were quite reliable and exhibited appreciable skill. Although the O2 forecasts (i.e., the forecasts concerned with the occurrence of ten or more tornadoes in the United States) were somewhat less successful than the O1 forecasts, evaluation of the former revealed a positive skill score overall and reasonably good reliability for probability values that were used frequently.

5. Experimental results—watches

Some basic characteristics of the watches of concern are summarized in Table 3 as a function of the type of watch. On the average, the watches covered an area of ≈79 800 km² (28 500 n mi²), and they had average lead times and valid times of ≈30 min and 4.75 h, respectively. The differences in these characteristics between tornado watches and severe thunderstorm watches were quite small, with the former tending to involve slightly larger areas and slightly longer lead times and valid times.

The average probabilities and observed relative frequencies for the W1, W2, and W3 forecasts, as a function of the type of watch, are indicated in Table 4. First, it should be noted that the W3 forecasts were missing on 23 of the 451 forecasting occasions. Overall, the W1 and W2 forecasts exhibited a tendency for average probabilities to exceed relative frequencies, whereas the opposite was true for the W3 forecasts. However, except for the W1 forecasts associated with tornado watches, the differences between the average probabilities and relative frequencies for the various combinations of types of forecasts and types of watches were quite small.

The reliability diagram for the watch forecasts is presented in Fig. 3a. As in the case of the outlook forecasts, the forecasters who formulated the experimental watch forecasts were permitted to assign probabilities of 0.00, 0.02, 0.05, 0.10, 0.20, . . . , 1.00 to the events of concern. Examination of the curve for the W1 forecasts (one or more tornadoes in the watch area; solid line) indicates that the forecasters tended to overforecast for this event, and this tendency was most pronounced for probability values ≈ 0.50. When the reliability of the W1 forecasts was examined as a function of the size of the watch area (reliability curves not presented), overforecasting was found to be greater for the smaller watches than for the larger watches [e.g., watches < 79 800 km² (28 000 n mi²) versus watches > 79 800 km² (28 000 n mi²)].

Fig. 3b reveals that the distribution of forecast probabilities for the W1 forecasts was bimodal with maxima at 0.20 and 0.60. The bimodal nature of this distribution is related to the differences between the probabilities assigned to tornado and severe thunderstorm watches (see Table 4). It is also of interest to note that the minimum in the distribution occurs at 0.40, a value near the sample climatological probability (0.379 for this event; see Table 4). Probability values < 0.10 and > 0.90 were seldom assigned to this event during the experiment.
The dashed line in Fig. 3a for the W2 forecasts (three or more tornadoes in the watch area) exhibits a somewhat erratic behavior, including a slight tendency toward overforecasting. On the other hand, the points associated with the largest departures from the 45° line involve relatively few forecasts. The distribution in Fig. 3b for the W2 forecasts indicates that probability values in the interval from 0.05 to 0.20 (inclusive) were used most frequently (the sample climatological probability for this event was 0.133; see Table 4). Probability values > 0.50 were assigned to this event on only seven occasions during the experimental period.

Examination of the reliability diagram in Fig. 3a for the W3 forecasts (one or more F2/F3 or greater tornadoes in the watch area; dotted line) indicates that the forecasters tended to underforecast for probability values < 0.10 and to overforecast for probability values > 0.10. Specifically, this curve possesses a positive “slope” that is somewhat smaller than that of the diagonal 45° line representing perfect reliability. Fig. 3b reveals that probability values of 0.05 and 0.10 were used most frequently (one of these two values was assigned to the event on almost 50% of the forecasting occasions), and that probability values > 0.50 were never assigned to this event.

The average number of tornadoes in the watch areas as a function of the forecast probability is depicted in Fig. 4 for the W1, W2, and W3 forecasts. The slopes of the regression lines fitted to the three empirical curves indicate the existence of positive relationships between average number of tornadoes and forecast probability for all three types of forecasts, with the slopes of the lines for the W1 and W2 forecasts exceeding that for the W3 forecasts. The relationships for the W2 and W3 forecasts are not as satisfactory as that for the W1 forecasts.

The Brier scores, the terms in the partition of the Brier score, and the skill scores for the watch forecasts are presented in Table 5. The skill scores indicate that none of the three types of experimental probability forecasts associated with the watches achieved appreciable skill. In this regard, recall that sample relative frequencies were used in computing the climatological Brier score and this procedure necessarily underestimates the skill of the forecasts (see Section 3). The difference between the Brier score for the W1 forecasts and the Brier scores for the W2 and W3 forecasts is due largely to the difference between the corresponding sample climatological probabilities (i.e., this difference in scores is not indicative of a
Severe Storms Forecast Center ( NSSFC) expressed forecasts of tornado occurrence and intensity in probabilistic terms. These forecasts were formulated during 1976 and 1977 in conjunction with the regular outlooks and watches prepared by NSSFC forecasters. In the case of the experimental outlooks, the forecasters assessed the probability of one or more tornadoes in the outlook area(s) (O1 forecasts) and the probability of ten or more tornadoes in the United States (O2 forecasts). The results presented in Section 4 indicate that the forecasters were quite successful in quantifying the uncertainty in the outlook forecasts. The outlook forecasts were reasonably reliable, with a closer correspondence between forecast probabilities and observed relative frequencies for the O1 forecasts than for the O2 forecasts. Moreover, the average number of tornadoes in the areas of concern increased, in general, as the forecast probability increased, providing further evidence of the forecasters’ ability to quantify uncertainty in the outlook forecasts in a consistent manner. With regard to skill, both types of experimental outlook forecasts exhibited positive skill relative to forecasts based on (overall) sample climatological probabilities of the relevant events. The experimental probabilistic forecasts related to watches involved three events: 1) one or more tornadoes in the watch (W1 forecasts); 2) three or more tornadoes in the watch (W2 forecasts); and 3) one or more F2 (1976)/F3 (1977) or greater tornadoes in the watch (W3 forecasts). The results presented in Section 5 indicate that the forecasters were only moderately successful in quantifying the uncertainty in these forecasts. The watch forecasts clearly were less reliable than the outlook forecasts. However, relative frequency generally was an increasing function of forecast probability, and the average number of tornadoes in the watches also tended to increase as forecast probability increased for all three types of watch forecasts. With regard to skill, none of the three types of experimental watch forecasts achieved appreciable positive skill relative to sample climatological probabilities. Of course, it is not particularly surprising that the watch forecasts were less skillful than the outlook forecasts, in view of the fact that the former involved greater specificity in space and time (i.e., smaller areas and shorter periods) than the latter. Moreover, the formulation of skillful forecasts of tornado intensity (i.e., the W3 forecasts) simply may be beyond the current state of the art of severe local storm forecasting.

6. Discussion and conclusion

In this paper we have described the results of an experiment in which NWS forecasters at the National

<table>
<thead>
<tr>
<th>Type of forecast</th>
<th>Number of forecasts</th>
<th>Brier score</th>
<th>Variance of observations</th>
<th>Reliability</th>
<th>Resolution</th>
<th>Skill score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>451</td>
<td>0.2393</td>
<td>0.2354</td>
<td>0.0262</td>
<td>0.0223</td>
<td>-1.7</td>
</tr>
<tr>
<td>W2</td>
<td>451</td>
<td>0.1130</td>
<td>0.1153</td>
<td>0.0113</td>
<td>0.0136</td>
<td>2.0</td>
</tr>
<tr>
<td>W3</td>
<td>428</td>
<td>0.1217</td>
<td>0.1189</td>
<td>0.0062</td>
<td>0.0034</td>
<td>-2.4</td>
</tr>
</tbody>
</table>
While this initial effort to quantify the uncertainty in tornado forecasts cannot be viewed as an unqualified success, we consider the results of the experiment to be quite encouraging. In this regard, it should be kept in mind that NSSFC forecasters had no prior experience in probability forecasting, that they did not receive any formal feedback concerning their performance during the experimental period, and that they did not have access to directly comparable objective probabilistic guidance information when making their forecasts. In fact, reliable estimates of the climatological probabilities for the events of interest were not even available prior to the start of the experiment. Moreover, the experimental forecasts were evaluated using sample climatological probabilities, a procedure that necessarily underestimates the skill of the forecasts (relative to a standard of reference based on long-term climatological probabilities). Obviously, these considerations must be taken into account in evaluating the performance of the NSSFC forecasters.

There are several reasons for believing that NSSFC forecasters could markedly improve upon their performance in any future experimental or operational probability forecasting program. These improvements could take the form of increases in reliability and/or skill. With regard to the former, forecast probabilities exceeded observed relative frequencies for most events and probability values considered in the experiment. It should be noted that such overforecasting also has occurred in many other probability forecasting programs, when forecasters have expressed their uncertainty in probabilistic terms for the first time (e.g., Hughes, 1965; Sanders, 1973; Gregg, 1977; Johnson et al., 1979; Daan and Murphy, 1982). This overforecasting is due in part to the constraints imposed by the forecasters' previous experience with categorical forecasting, in which they necessarily are restricted to probability values of zero or one. Some experience and feedback is required before forecasters can make full and proper use of the entire probability scale.

Moreover, it should be mentioned that overforecasting frequently is considered to be desirable when forecasts are expressed in categorical terms, especially in situations in which the forecast involves events associated with significant social and/or economic impacts. The severe weather outlooks and watches issued by NSSFC forecasters certainly satisfy such a criterion. In this regard, evaluation of these official outlooks and watches reveals a considerable amount of overforecasting (e.g., Weiss et al., 1980). In brief, the forecasters are encouraged to obtain a high probability of detection, which necessarily leads to a substantial false alarm rate. Of course, such overforecasting is not appropriate when forecasts are expressed in probabilistic terms. However, in view of the fact that the experimental forecasts were formulated by the same forecasters who issued (almost simultaneously) the official categorical outlooks and watches, it is not surprising that these forecasts exhibited some degree of overforecasting. In any case, feedback concerning the correspondence between forecast probabilities and observed relative frequencies and/or reliable probabilistic guidance forecasts could be expected to reduce the amount of overforecasting in these experimental forecasts.

Since skill depends on both reliability and resolution, the above-mentioned increases in reliability would be expected to lead to increases in skill. On the other hand, it may be more difficult to achieve appreciable improvements in resolution, because such improvements depend on advances in the state of the art of severe weather forecasting. However, the availability of comparable objective guidance forecasts could contribute to greater resolution and thereby to increases in skill. For this reason, it would be desirable to develop a means of obtaining directly relevant probabilistic guidance information from the current objective forecasts provided to NSSFC forecasters by the National Meteorological Center. Of course, the forecasters themselves might be able to increase the skill of the forecasts by improving their reliability and/or resolution. For example, as noted in Sections 4 and 5, the forecasters evidently did not take sufficient account of the size of the outlook and watch areas in formulating their experimental forecasts. This tendency to overforecast for smaller areas and to underforecast for larger areas (at least in a relative sense) is similar to the tendency exhibited by NWS forecasters in some locations to overforecast for six-hour periods in precipitation probability forecasting (see Hughes, 1980). Proper adjustment of forecast probabilities for the size of the area of concern could also lead to increases in skill.

With regard to future work, we believe that it would be desirable to conduct additional experiments at NSSFC with respect to subjective probability forecasting. The purposes of such experiments would be to investigate further the ability of NSSFC forecasters to quantify uncertainty in a reliable and skillful manner and to improve upon their performance in the initial experiment described in this paper. These experiments might also contribute to a reduction in the amount of overforecasting currently found in official NSSFC forecasts. If the experiments were successful, then probabilities could be assigned in the future to the official outlooks and watches on an operational basis. While these probabilities might be intended for use only by NSSFC (and other) meteorologists, this mode of expression of the forecasts would be more appropriate in view of the state of the art of severe tornadoes is <0.20, which restricts the range of probability values for the O1 and W1 forecasts in the experiment.
weather forecasting than the current practice of expressing outlook and watch forecasts in categorical terms. Moreover, this practice would be consistent with Sanders’ recommendation, made almost twenty years ago, that “probability be acknowledged as the proper internal language of forecasters” (Sanders, 1963, p. 201).

With regard to the possible inclusion of probabilistic information in severe weather forecasts issued to the general public and specific user groups, obviously such a practice should not be initiated without giving very careful consideration to the needs of these individuals or groups and without devoting considerable attention to educational efforts concerning the proper interpretation and use of this information. Nevertheless, it should be recognized that forecasts expressed in probabilistic terms contain more information and are of greater value—whether this value is measured in monetary terms or in terms of reductions in serious and fatal injuries—than traditional categorical forecasts. If probabilities were assigned to the official outlooks and watches on an operational basis, as suggested above, then these probabilities would be available for possible dissemination (on a request basis) to selected groups, such as agencies or organizations which must take precautionary actions when severe weather threatens. Such probabilities could appreciably enhance the information content and usefulness of current severe weather forecasts.

Acknowledgments. We would like to express our appreciation to the forecasters at the National Severe Storms Forecast Center (NSSFC) in Kansas City, Missouri, for their participation in the probabilistic tornado forecasting experiment. The assistance and cooperation of Messrs. F. P. Ostby of NSSFC and L. A. Hughes, formerly of the NWS Central Region Headquarters, in the design and conduct of the experiment are also gratefully acknowledged. This paper was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under grants ATM74-001610A01, ATM77-24060, and ATM80-04680.

REFERENCES