A Probabilistic Forecast Contest and the Difficulty in Assessing Short-Range Forecast Uncertainty

THOMAS M. HAMILL AND DANIEL S. WILKS

Department of Soil, Crop, and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 22 July 1994, in final form 5 May 1995)

ABSTRACT

Results are presented from a probability-based weather forecast contest. Rather than evaluating the absolute errors of nonprobabilistic temperature and precipitation forecasts, as is common in other contests, this contest evaluated the skill of specifying probabilities for precipitation amounts and temperature *intervals*. To forecast optimally for the contest, both accurate forecasts and accurate determination of one's uncertainty about the outcome were necessary. The contest results indicated that forecasters over a range of education levels produced skillful forecasts of temperature and precipitation relative to persistence and climatology. However, in this contest forecasters were not successful in assessing the uncertainty of their maximum or minimum temperatures from day to day, as measured by the correlation of interval width and absolute error. Though previous experiments have shown more optimistic results, the seasonal variation of forecast uncertainty can account for much of the observed correlation, suggesting that day-to-day assessment of forecast uncertainty may be more difficult than previously believed. It is argued that objective methodologies should be developed to quantify uncertainty in forecasts.

1. Introduction

This article describes a probability-based weather forecast contest conducted during the spring semester of 1994 at Cornell University and what was learned about forecast uncertainty following an analysis of the contest results. The intent of this contest was to give students experience in making probabilistic weather forecasts and to learn about probability forecasting. The contest was unusual in that probabilistic forecasts were made for both temperature and precipitation. Many nonprobabilistic contests exist or existed, such as at the University of Virginia (Colucci et al. 1992) and the National Forecast Contest (F. Gadomski 1994, personal communication). Other contests require contestants to make probability forecasts for precipitation, yet for short-range forecasts, contestants are typically required to make nonprobabilistic forecasts of temperature, such as tomorrow's maximum (max) and minimum (min) temperatures (Bosart 1983; Sanders 1986; Croft and Milutinovic 1991).

Probabilistic temperature forecasts are issued less frequently but are not inherently confusing or hard to formulate. Winkler and Murphy (1979, hereafter WM79) outline a number of ways of expressing probabilistic temperature forecasts, including "fixed probability central credible intervals." For a 50% central credible interval forecast, the two temperatures denote the upper and lower quartiles of the probability distribution for that forecast—that is, the chance the temperature is below the lower limit or above the upper limit are each 25%. Thus, a typical forecast would be "tomorrow's high temperature 58°–64°F." Typical 50% interval widths for temperature at a 1-day projection are 6°–8°F. Through the width of the interval, the forecaster can communicate daily the uncertainty in a temperature forecast, which should be useful information for many citizens and businesses. A wide range denotes greater than average uncertainty in the forecast—for example, a day when the forecaster is unsure whether a warm front will remain to the south or move north of the forecast location. Conversely, a narrow range denotes less than average uncertainty, such as might be specified during a quiescent summer pattern. In addition, if one assumes the probability distribution for the forecast temperature is Gaussian or if other credible intervals are specified (e.g., the 80% interval), the probabilities of exceeding other colder or warmer temperatures may be easily estimated.

Probabilistic temperature forecasts can be useful since they provide more information than the traditional "point" temperature forecast; the interval quantifies a forecaster's uncertainty about the temperature being forecast. Murphy and Winkler (1979) outlined a rationale for probabilistic temperature forecasts and suggested the National Weather Service (NWS) begin testing such an approach with the eventual goal of na-

Corresponding author address: Thomas M. Hamill, Department of Soil, Crop, and Atmospheric Sciences, 1126 Bradfield Hall, Cornell University, Ithaca, NY 14853.
E-mail: tmh8@cornell.edu

tionwide implementation. There are good reasons for expressing forecasts of temperatures probabilistically. First, the chaotic nature of tropospheric dynamics does not permit exact specification of future temperatures any more than for other variables. This fact constitutes the fundamental meteorological justification for probabilistic temperature forecasts. Second, many users would find such forecasts useful, since many real-world decisions are based on temperature forecasts. Specifying probabilities associated with future temperatures permit better decisions in the sense that greater economic value is realized relative to otherwise comparable forecasts that do not contain a quantitative expression of uncertainty (e.g., Murphy 1977; Krzysztofowicz 1983).

Despite the merits of probabilistic temperature forecasts, the format for public dissemination of forecast temperatures has changed little in the past three decades. NWS forecasts are nonprobabilistic, such as a city temperature forecasts for "high: 62°F," or they are at best pseudoprobabilistic, such as a zone forecast for "tomorrow's high in the mid-60s." Further, such pseudoprobabilistic forecasts are intended primarily to convey geographic variations of temperature rather than forecast uncertainty. Broadcast meteorologists have been even more reluctant to embrace probabilistic temperature forecasting; most television meteorologists forecast specific highs and lows for each of the next 5 days, although by the fifth day 10°–20°F errors are not uncommon. Either through inertia or fear of public reaction, probabilistic temperature forecasts are the exception and not the rule. However, for forecasts for a specific location, operational implementation of the simplest fixed-probability central credible interval forecasts could be achieved "transparently" and with minimal confusion; each forecast would consist of only a temperature range with the probability information expressed implicitly (through a prespecified fixed-probability central credible interval such as 50%).

Probabilistic precipitation forecasts are regularly issued by the NWS in terms of a probability of precipitation (POP). Worded forecasts are typically used to convey the type and expected intensity of precipitation. As with temperature forecasts, calibrated quantitative assessment of precipitation probabilities by category (e.g., 20% probability of 0.0 in.–trace, 50% probability of 0.01–0.10 in., 20% probability of 0.11–0.25 in., 10% probability of 0.26–0.50 in.) are economically more valuable than a fixed forecast of precipitation amount. With such a forecast, an educated decision maker can more accurately assess the likelihood of damaging weather and assess the appropriate action (Krzysztofowicz et al. 1993). Model Output Statistics (MOS) (Dallavalle et al. 1992), produce probabilistic quantitative precipitation forecasts but they are routinely distributed only in a simpler, nonprobabilistic format.

A university contest provides a natural test bed for exploring unconventional forecasting concepts. Since forecasts from the contest are not disseminated to the public, the only thing at risk is students' pride over unskillful results. During the spring semester of 1994, we conducted a forecast contest for probabilistic, 1-day forecasts of max and min temperature and precipitation by category. In addition to giving students practice, this contest was designed to explore some interesting questions: For example, can contestants set credible intervals of appropriate widths? Are variations in the widths of the intervals or the spread of precipitation probabilities across forecast categories positively correlated with forecast error, indicating skill in assessing uncertainty? This paper will review the results from this contest and will reexamine previous experiments in probabilistic forecasting in light of this contest's results. Further, this paper discusses the general difficulty in assessing forecast uncertainty day to day and the possible approaches to improve assessments of forecast uncertainty. It is hoped that the discussion of this contest and its results will encourage similar probabilistic contests and experiments elsewhere.

Section 2 will outline the contest format and scoring rules; section 3 provides an analysis of this contest's results as well as a reanalysis of similar prior experiments. A discussion follows in section 4, with emphasis on possible approaches to aid the forecaster in quantifying uncertainty. Conclusions are provided in section 5. Since NWS forecasts are typically expressed in English units of degrees Fahrenheit and inches of precipitation, these units are retained in this paper.

2. Description of the forecast contest

Twenty-nine entrants participated in this extracurricular forecast contest for the spring semester of 1994 forecasting the next day's max and min temperatures and precipitation amount probabilistically. Of these 29, 12 produced forecasts consistently throughout the semester; the increasing workload and discouragement with early unskillful forecasts led many to stop making forecasts. Contestants ranged in education level from Ph.D. to freshman, and two of the forecast entrants were automated schemes, one based on MOS and the other a blend of persistence and climatology. The contest ran each weekday over 12 weeks and 3 days, totaling 63 days of forecasts. Forecasts were required to be entered by 1900 LST each forecast day, and the forecast period of record was 0800 LST the next day to 0800 LST the following day. The contest was also run during the fall semester of 1994 with a larger (and different) set of contestants; however, analysis of the contest results will focus mainly on the first semester. Verification data were taken from Cornell University's Game Farm Road weather station. This site often received a cold drainage flow from nearby Mt. Pleasant on clear nights, thus making its min temperature climatology substantially different from nearby MOS stations at Binghamton and Syracuse.

Precipitation probabilities were forecast for each of the six MOS precipitation amount categories (0.0 in.–

trace, 0.01–0.09 in., 0.10–0.24 in., 0.25–0.49 in., 0.50–0.99 in., ≥1.00 in.). Probabilities were required to be rounded to the nearest 10% and were entered as integers summing to 10; thus, a forecast of 9–1–0–0–0–0 was acceptable, but the program would not allow 8.5–1.5–0–0–0–0. All forecasts are referenced against a baseline "persistence" forecast. For precipitation this used the previous day's rainfall amount. If no rainfall occurred (category 1), then the persistence forecast was 10–0–0–0–0–0. Similarly, if the previous day verified category 3, the persistence forecast was assigned the forecast 0–0–10–0–0–0.

Penalty points for precipitation forecasts were assessed using the ranked probability score, or RPS (Epstein 1969; Murphy 1971). Daan (1985) discusses the merits of the RPS relative to other scoring methodologies. Using a forecast distribution vector of precipitation probabilities **y**, a cumulative distribution vector **Y** is defined with components

$$\mathbf{Y}_m = \sum_{j=1}^{m} \mathbf{y}_j, \quad m = 1 \cdots 6. \tag{1}$$

Similarly, from the vector of the observations o, a cumulative distribution vector **O** is also generated:

$$\mathbf{O}_m = \sum_{j=1}^{m} \mathbf{o}_j, \quad m = 1 \cdots 6, \tag{2}$$

where $\mathbf{o}_j = 10$ if precipitation occurred in the jth category and is zero otherwise. The RPS is the squared difference between the forecaster's cumulative distribution vector and the observed cumulative distribution vector

$$\text{RPS} = \sum_{m=1}^{6} (\mathbf{Y}_m - \mathbf{O}_m)^2. \tag{3}$$

An example of the calculation of the RPS is given in Table 1. Note that because probability forecasts are entered as integers from 0 to 10, these RPS scores are 100 times higher than if probabilities were issued in the range from 0 to 1.

Temperature forecasts were entered as 50% fixed-probability central credible intervals as in WM79. Upper and lower limits were entered for both the max and min temperature forecasts. The nonprobabilistic, reference persistence forecasts used the previous day's minimum and maximum temperatures and a zero interval width. Temperature interval forecasts were penalized for wide intervals and for verifications outside the forecast intervals. Denoting the lower cutoff for the forecast temperature limit L, the upper limit U, and the verification temperature T, the penalty points P were assessed using

$$P(L, T, U) = \begin{cases} 4(L - T) + (U - L + 1) & \text{if } T < L \\ U - L + 1 & \text{if } L \leq T \leq U. \\ 4(T - U) + (U - L + 1) & \text{if } T > U. \end{cases} \tag{4}$$

Winkler (1972) showed that this penalty function is strictly proper (i.e., cannot be "hedged" or "gamed") assuming the subjective forecast distribution is normally distributed. Though forecaster's distributions were often not normally distributed, a strategy for gaming in these situations was not obvious.

Forecast skill S was calculated by comparing a contestant's accumulated penalty points P_f to reference penalty points from persistence P_p and perfect forecasts P_{perf}:

$$S = \left(\frac{\sum P_f - \sum P_p}{\sum P_{\text{perf}} - \sum P_p} \right) 100\%. \tag{5}$$

For precipitation forecasts, penalty points are just the RPS, so skill was calculated by summing the appropriate RPS values over the period of interest—for example, a week or the semester to date. Perfect precipitation forecasts have RPS = 0. Temperature skill scores were also calculated using (5). Unlike precipitation forecasts, a perfect temperature forecast with zero interval width was assessed 1 penalty point, and thus, $\sum P_{\text{perf}} = n$, where n is the number of days in the evaluation period. To rank contestants, an average of max, min, and precipitation skill was tabulated weekly for the previous week and for the semester to date.

Sample results from the contest are shown in Fig. 1. As shown, the output is divided into three sec-

TABLE 1. An example of the calculation of the ranked probability score given a forecast distribution **y** and the observed distribution **o**.

Precipitation category	0.0 in.–trace	0.01–0.09 in.	0.10–0.24 in.	0.25–0.49 in.	0.50–0.99 in.	>1.0 in.
Forecast distribution **y** (in tens)	6	3	1	0	0	0
Observed distribution **o** (in tens)	0	10	0	0	0	0
Cumulative forecast distribution **Y**	6	9	10	10	10	10
Cumulative observed distribution **O**	0	10	10	10	10	10
$(\mathbf{Y}_m - \mathbf{O}_m)^2$	36	1	0	0	0	0

$\text{RPS} = \sum_{m=1}^{J} (\mathbf{Y}_m - \mathbf{O}_m)^2 = 37$

****
*** ***
*** CORNELL FORECAST CONTEST RESULTS FOR WEEK 12 ***
*** ***
****

PLACE	NAME	CUMULATIVE SKILL	WEEKLY SKILL	MAX TMP SKILL	MIN TMP SKILL	PRECIP SKILL	TOT MAX SKILL	TOT MIN SKILL	TOT PRE SKILL
1	ART2DE2	62.3	65.5	79.1	68.9	48.6	59.3	59.7	67.9
2	DOGBERT	61.9	69.5	78.8	71.5	58.3	57.5	62.4	65.7
3	CAPT. ENSEMBLE	60.8	62.3	75.9	54.8	56.0	62.2	53.5	66.7
4	RIDGE	60.4	67.9	82.0	64.0	57.6	58.7	57.7	64.8
5	HOCKEY PUCK	58.3	70.5	76.6	78.5	56.4	59.5	51.7	63.7
6	WXBUNNY	57.7	67.4	84.8	55.3	62.1	60.8	46.8	65.5
.									
8	MOSCASTER	54.1	13.2	-23.6	18.3	44.7	60.8	40.0	61.5
.									
12	ITH STOCASTER	32.9	15.9	1.1	13.3	33.3	17.1	37.5	44.1
.									
27	PERSISTENCE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

A

THE FOLLOWING ARE THIS WEEK'S MEDAL WINNERS:

GOLD MEDAL - HOCKEY PUCK
SILVER MEDAL - DOGBERT
BRONZE MEDAL - RIDGE

BEST TEMPERATURE FORECASTER - HOCKEY PUCK
BEST PRECIPITATION FORECASTER - WXBUNNY

B

VERIFYING OBSERVATIONS

	TUE-24	WED-24	THU-24	FRI-24	SAT-24
OBS MAX TEMP	85	80	53	69	56
OBS MIN TEMP	59	38	37	41	41
PRECIP CAT	3	1	2	3	5

DAY BY DAY ANALYSIS

FORECASTER		TUE-24	WED-24	THU-24	FRI-24	SAT-24
PERSISTENCE	MX TMP PENALTY	73	21	109	65	53
	MAX TEMP FCST.	67 / 67	85/ 85	80/ 80	53/ 53	69/ 69
	MN TMP PENALTY	45	69	89	13	17
	MIN TEMP FCST.	48/ 48	55/ 55	59/ 59	38/ 38	37/ 37
	PRECIP PENALTY	0	200	100	200	300
	PRECIP. FCST.	00X000	00X000	00X000	X00000	0X0000
HOCKEY PUCK	MX TMP PENALTY	19	10	20	23	7
	MAX TEMP FCST.	82/ 76	86/ 81	63/ 56	65/ 59	62/ 56
	MN TMP PENALTY	5	24	8	7	10
	MIN TEMP FCST.	59/ 55	49/ 42	44/ 37	44/ 38	47/ 42
	PRECIP PENALTY	33	140	50	33	93
	PRECIP. FCST.	233200	044200	721000	023320	023320

C

FIG. 1. Abbreviated sample output from the forecast contest. Section A of the output summarizes the overall ordering of contestants along with summary information for the semester and the previous week. Section B indicates the best performers during the previous week. Section C gives a detailed breakdown of daily forecasts.

tions: the top section summarizing the overall contest rankings, the middle section indicating the weekly contest winners, and the last section providing a detailed description of the forecast errors day by day. In section A of Fig. 1, column 1 lists the *overall* ranking of the contestants (all three forecast variables for the semester to date), and column 2 indicates the associated name of the forecaster (pseudonyms used here). Column 3 indicates the overall skill of the forecaster, and column 4 indicates the skill during the previous week, which is used to determine the weekly winners shown in section B of Fig. 1. The remaining columns detail the weekly skill of max, min, and precipitation forecasts (col-

umns 5–7) and their skill for the semester to date (columns 8–10).

The next section (B) summarizes the best forecast performance during the previous week. The contestants with the top three overall skill scores are awarded the symbolic medals, and the best temperature forecaster (max and min combined) and best precipitation forecaster are noted.

Section C of Fig. 1 displays a day-by-day breakdown of the observed weather and the contestant's forecasts and associated penalty points. Only two forecasters are shown here for brevity. These diagnostics are useful for screening out the occasional data entry error and for providing feedback to the students on the errors they made with each forecast. The statistics are also useful for verifying the weekly skill numbers in section A of Fig. 1. For example, the persistence precipitation forecast accumulated 800 (=0 + 200 + 100 + 200 + 300) error points, and the forecaster "Hockey Puck" accumulated 349 (= 33 + 140 + 50 + 33 + 93) error points. Thus, using (5), Hockey Puck's weekly precipitation forecast skill is 100% (349 − 800)/(0 − 800) = 56.4%.

Note the formulation of the contest in skill scores tallied by week and semester to date may result in the overall skill being quite different than an equal weighting of the weekly skills. For example, a week with much precipitation that varies in amount day to day will yield many penalty points for persistence forecasts, often more points than are tallied during 2 or 3 relatively dry weeks. Hence, the overall forecast skill relative to persistence is more strongly determined during variable weather patterns than during quiescent ones.

3. Analysis of contest results

a. Temperature forecasts

The most interesting aspect of this university contest was the use of credible intervals for temperature forecasts. Little has been documented specifically on the use of these fixed probability credible intervals. Peterson et al. (1972) documented a procedure for training forecasters to set their intervals. Murphy and Winkler (1974) documented a forecast experiment in Denver, Colorado, using credible intervals. The most complete study (WM79) described an experiment conducted over a 9-month period using experienced National Weather Service forecasters in Milwaukee, Wisconson. These forecaster's 50% central credible interval widths exhibited a correlation of +0.35 with the forecast errors (the absolute difference between the interval's center and the verification temperature). Positive correlations can indicate the ability of the forecaster to assess uncertainty. However, the extent to which this correlation can be accounted for by seasonal changes in predictability rather than day-to-day changes was not clear. (Climatologically, temperatures tend to be more variable in winter than in summer.) Thus, an interesting question is whether forecasters show ability to accurately set intervals day to day, as opposed to season to season. While interval widths may vary from season to season, we presume that the day-to-day variations in interval widths are of greater interest to the forecaster and forecast user. The potential economic benefits from such daily "forecasts of forecast uncertainty" are detailed in Wilks and Hamill (1995).

Concentrating on the intervals set by the top three forecasters in the Cornell contest (as shown in Fig. 1) indicates that the interval widths were set properly in aggregate. Overall, to be consistent with a 50% credible interval, 25% of the verifying observations should fall below the lower limit, 50% between the lower and upper limits, and 25% above the upper limit. For maximum temperatures, the top three forecasters' combined percentages were exactly 25–50–25 and for minimum temperatures 35–52–13. The skewed minimum temperature forecasts reflect deep snow cover in central New York during 1994 and frequent radiational cooling that often surprised contestants.

Still, considering the intervals in aggregate does not answer the question of whether forecasters accurately set interval widths day to day, or, equivalently, whether forecasters were successful at assessing forecast uncertainty. The answer from this contest was a qualified "no." Overall, the top three contestants' interval widths for max temperature correlated with the errors at +0.03 and for min temperatures at +0.06. These low correlations may be due to a number of factors. First, though the top three finishers in the contest all had forecast experience, with a Ph.D., B.S., and M.S. in meteorology, respectively, forecasting was not their full-time, professional responsibility. Second, the contest lasted only 13 weeks, and it takes time to become familiar with any new forecasting concept. Still, even when considered together as a consensus forecast, and considered without the results for the first 7 weeks (taken as a learning period), the correlations were no higher.

Consider Fig. 2, a plot of forecast maximum temperature intervals versus verification data for the forecaster "Capt. Ensemble," the third-ranked contestant. This contestant relied heavily on the variation in temperatures in MOS forecasts to set the interval width. This forecaster used Binghamton, New York, and Syracuse, New York, MOS forecasts produced by the Limited-Area Fine Mesh and Nested Grid Model (NGM) on two successive runs, giving eight estimates of max temperatures. Differences between the eta (Black 1994) and NGM numerical output (FOUS) were also considered. The rationale behind this forecast strategy was to subjectively simulate an ensemble/lagged average forecast (Hoffman and Kalnay 1983; Tracton and Kalnay 1993). Thus, when MOS forecast temperatures diverged widely from station to station or between models or initialization times, a wider interval was used. When there was substantial agreement, a smaller interval was used. Some subjective judgment

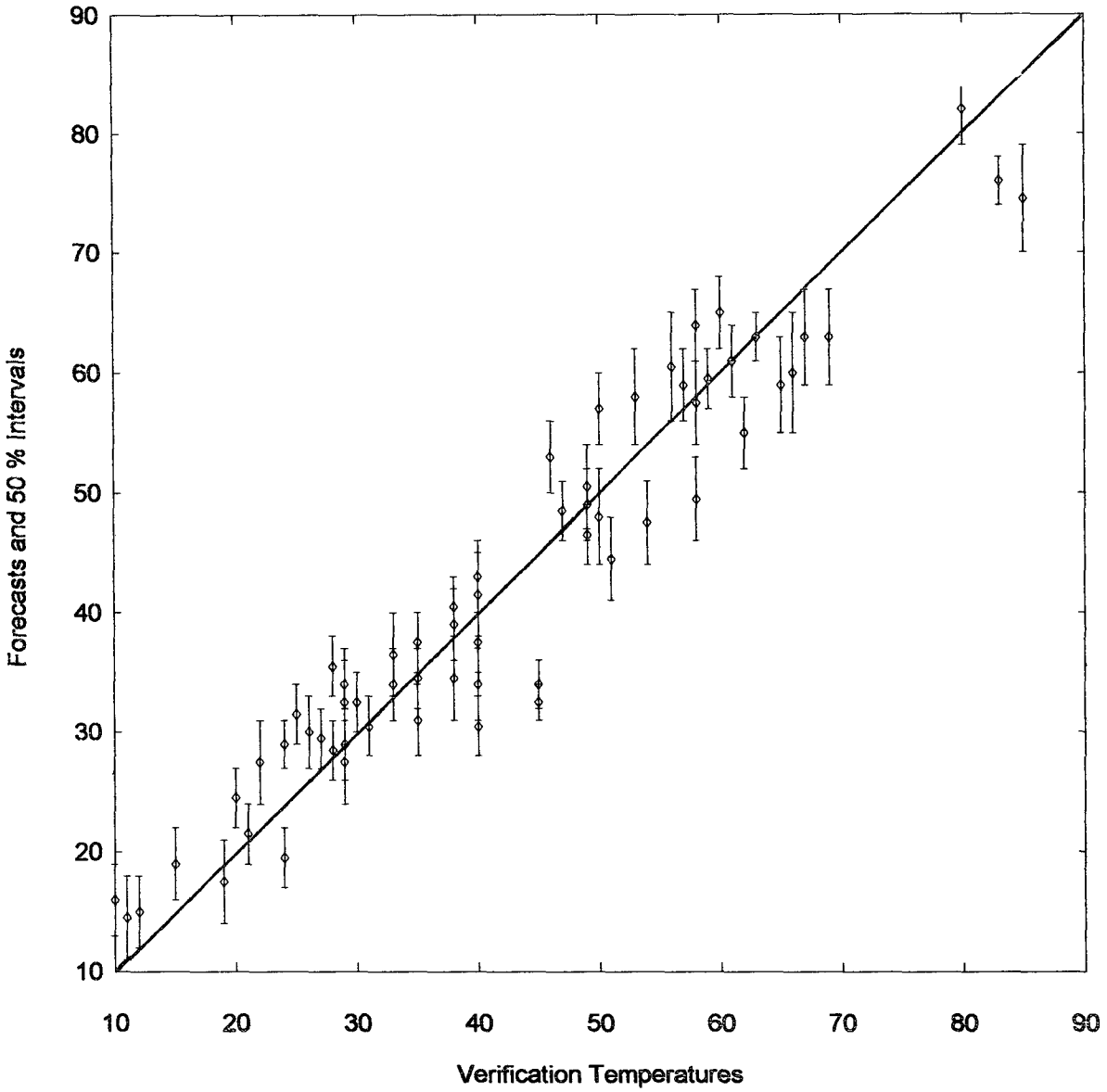

FIG. 2. Max temperature forecast intervals plotted against verification temperatures for the forecaster Capt. Ensemble. Temperatures in °F.

was also used by this forecaster. The correlation of interval widths to error for this forecaster's max temperatures was +0.03.

Of more theoretical interest than the specific contest results are what they may imply about the higher correlations observed in the Milwaukee experiment of 1979. A partial explanation might be that Milwaukee forecasters were informed of and used information on the typical temperature variability *by season.* Even if Milwaukee forecasters were not consciously aware of the seasonal effects of uncertainty, they may have used it implicitly, generating larger intervals on average in winter than in summer. Conversely, the Cornell contest, at 13 weeks in length, did not span a range of seasons, and its contestants certainly were only vaguely aware of the seasonal nature of uncertainty. Thus, determining the magnitude of seasonal effects of forecast variability is important; if automatically varying interval widths slowly through the year may achieve nearly the same positive correlation as forecasters varying them day by day, then the optimistic Milwaukee results must be interpreted with more caution. Further, it may be more appropriate in any circumstance to measure the magnitude of their observed correlation of interval width and error relative to the baseline obtained using climatological variances, as the prediction of midtro-

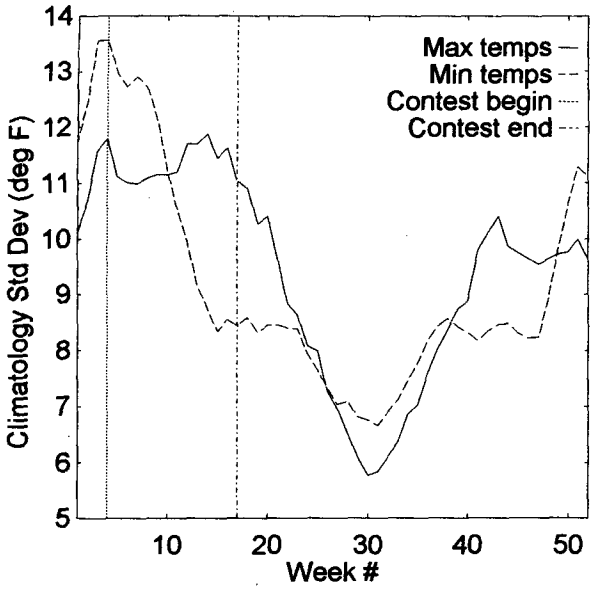

FIG. 3. Standard deviation of max and min temperatures around a running 3-week mean, plotted against the week number. Data are from Cornell University Game Farm Road, 1961–93. Vertical lines denote beginning (week 4) and end (week 7) of the forecast contest.

pospheric flow is measured relative to the climatological flow pattern using the anomaly correlation (Miyakoda et al. 1972; Murphy and Epstein 1989).

Accordingly, seasonal effects of variability were examined both for Cornell and for Milwaukee. For Cornell, using the record from 1961 to 1993, the standard deviation of daily temperatures about a running 3-week mean was calculated. The results are shown in Fig. 3. As shown, the variability of min temperatures is highest in midwinter and decreases quickly through the spring. However, the variability of max temperatures remains high through midspring. This contest took place from week 4 to week 17, denoted with vertical lines in Fig. 3; thus, the figure suggests that if forecast intervals were varied seasonally by an amount proportional to the climatological variance, the max temperature intervals would remain nearly constant throughout the contest, but the min temperature intervals would decrease sharply from the beginning to end.

To test the effect of using seasonally varying interval widths, the contest was rerun from beginning to end with climatology as a contestant. Fifty percent central credible intervals were calculated using the weekly mean temperatures and standard deviations and assuming a Gaussian distribution of errors about the climatological mean. The results indicate the usefulness of seasonal climatological variances in setting interval widths. The max temperature interval widths were correlated with the error at −0.09, the low correlation consistent with the homogeneous climatological variance of max temperatures during the 13-week contest. However, the min temperature interval widths were correlated at +0.27, better than any of the contestants and nearly to the level achieved by the Milwaukee forecasters. (Note, however, that the skill of these forecasts was low because the forecast intervals were quite wide.) This suggests a forecaster informed of the seasonal variations might achieve the high correlation observed in the Milwaukee experiment primarily through climatological information, rather than day-to-day adjustments of interval width.

After examining the possible benefits of forecasting intervals at Cornell based on climatological variances, the Milwaukee observations were examined in the same manner. Figure 4 is a plot of the standard deviations of daily temperatures about a running 3-week mean for Milwaukee using data from 1949 to 1993. Vertical lines denote the end (July) and beginning (October) of the experiment described in WM79. As shown, there is a large annual cycle in the climatological standard deviations of temperature. Noting this, climatological 50% credible interval forecasts were generated using the observations from October 1974 to July 1975 (the period for the experiment in WM79) in the same manner as done with the Cornell data described above. The interval widths were correlated with the absolute errors, yielding a correlation of −0.16 for max temperatures and +0.32 for min temperatures. The negative correlation coefficient for the max was surprising; accordingly, ranked (Spearman) correlation coefficients were also calculated since these would be more insensitive to outliers. The Spearman correlation coefficients were +0.16 for max temperatures and +0.27 for minimum temperatures, indicating that a few outliers skewed the product–moment correlation. Nonetheless, overall cli-

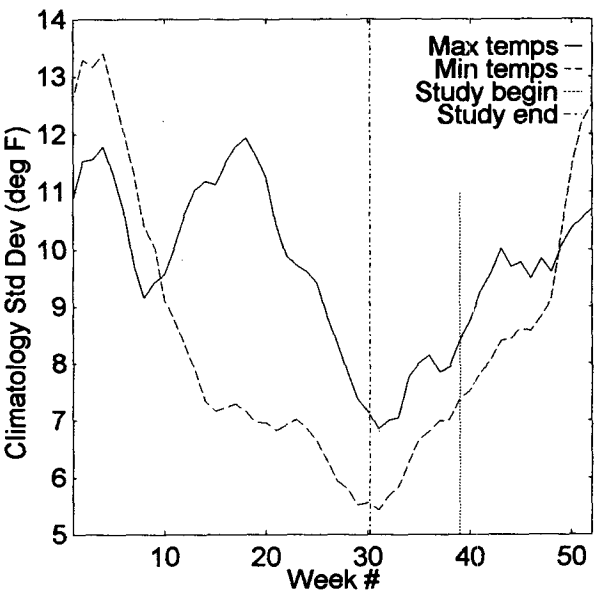

FIG. 4. Same as Fig. 3 but for Milwaukee data, using the period 1949–93. Vertical lines denote the beginning (week 39) and end (week 30) of the experiment.

matologically based correlations were all observed to be less than the Milwaukee forecasters' +0.35.

Another more appropriate test of the seasonal effects may be to use temperatures and their annual error variability from actual forecasts rather than from climatological information. Accordingly, NWS objective (Klein et al. 1959; Glahn and Lowry 1972) and subjective point max and min temperature forecasts for Milwaukee were obtained for the period January 1972–December 1976. A regression model was developed from this training data to predict interval widths, which varied seasonally. The regression model was developed using all available observations and forecasts in this dataset except those in the period of the experiment, from October 1974 to July 1975. This regression model used a one-harmonic expansion for minimum temperatures and a two-harmonic expansion for maximum temperatures. Use of a two-harmonic expansion yielded a better fit, consistent with the higher-frequency variations observed in the Milwaukee climatology (Fig. 4). The results from this experiment are summarized in Table 2. As shown, the observed correlations are stratified by lead time, by forecast cycle, and by whether the forecast was objective or subjective. Generally, the observed correlation coefficients are one-half to two-thirds of the magnitude quoted in WM79. Thus, the annual cycle of temperature variability may have contributed significantly to Milwaukee forecasters' ability to set reliable interval widths but did not account for all of the observed correlation.

The Cornell forecast contest itself was a convenient vehicle for further testing of seasonal effects. After the initially pessimistic results obtained during the first semester, a prototype scheme was developed for the next semester of the contest, conducted during the fall of 1994. This scheme automated the prediction of forecast intervals based on available local NGM MOS. The operating hypothesis in developing this scheme was that three factors could have contributed to uncertainty and could be observed in the objective MOS forecasts for nearby Binghamton (BGM) and Syracuse (SYR). The first was the seasonal nature of uncertainty, as discussed earlier. The second factor was initial condition uncertainty, which could be reflected in cycle-to-cycle differences in MOS forecasts. This is the basis of "lagged average forecasting" (Hoffman and Kalnay 1983), more commonly applied to forecasts of longer duration. A last hypothesized effect was the magnitude of atmospheric baroclinicity, as measured by differences in MOS temperatures between BGM and SYR, with systematic corrections for each to match the Cornell Game Farm Road climatology. In developing a simple regression equation to predict temperature forecast uncertainty, the second and third effects were combined. For both max and min forecasts, the most current 1200 UTC and 0000 UTC NGM MOS forecasts of max and min at BGM and SYR were used (i.e., the 36-h max forecast from the 1200 UTC run and the 48-h max from the 0000 UTC run). These forecasts were adjusted for local yearly average temperature differences between the Cornell Game Farm Road and BGM or SYR (using 1990–91 MOS data). Next, a mean Cornell Game Farm Road forecast max and min were predicted from a weighted average of the adjusted BGM and SYR data according to

$$T_{GF} = 0.15T_{00B} + 0.15T_{00S} + 0.35T_{12B} + 0.35T_{12S}, \quad (6)$$

with "GF" indicating the predicted temperature at the Cornell Game Farm Road and the subscripts for terms on the right-hand side indicating the forecast cycle and BGM (B) or SYR (S). Lower weights were given to older data, consistent with the error and the cycle-to-cycle covariance. Next, a measure of the variability from cycle to cycle and from of each of the four individual forecasts from this mean were diagnosed as a weighted sample standard deviation according to

$$\begin{aligned} s_{GF} = \{[&0.15(T_{00B} - T_{GF})^2 \\ &+ 0.15(T_{00S} - T_{GF})^2 \\ &+ 0.35(T_{12B} - T_{GF})^2 \\ &+ 0.35(T_{12S} - T_{GF})]3^{-1}\}^{1/2}. \end{aligned} \quad (7)$$

Finally, a regression equation was developed to predict the forecast rms error, and thereby the interval width, based on a one-harmonic expansion of the day of the year and s_{GF}. Using 1991–92 as training data, the regression results indicated that for min temperatures, s_{GF} was not a significant predictor of the error and was thus left out of the predictive equation, but this term was a significant predictor for the max temperatures though the assigned weight was small.

The ability of this scheme to predict uncertainty here was mixed. This scheme produced the second-highest skill in temperature forecasts of the 28 contestants participating during the fall semester of 1994. For minimum temperatures, this scheme's observed correlation during the second semester of the contest was +0.35, as high as in the Milwaukee experiment and despite the fact that the only predictor was the first harmonic of the day of the year. However, for max temperatures, the observed correlation was −0.14, worse than if uniform intervals were used. Upon analysis, it appeared that intervals for maximum temperatures were set too widely late in the semester. The predicted intervals for both max and min increased throughout the fall, from an average of 6° at the beginning of the contest to 8° at the end. For minimum temperatures, overall, the percentage of forecasts below/within/above the credible intervals were 22–53–25, consistent with the definition of a 50% credible interval. However, for the maximum temperature forecasts, the allocation was 18–65–17.

Because of the small sample size of the contest and of the dataset for training the regression model, these

TABLE 2. Correlations of climatologically derived interval widths and forecast absolute errors for objective and subjective temperature forecasts at Milwaukee for the period October 1974–July 1975. Forecasts are stratified by lead time, by objective/subjective, and by forecast cycle.

	12–24 h	24–36 h	36–48 h
0000 UTC (Objective)	0.28	0.21	0.15
0000 UTC (Subjective)	0.29	0.16	0.22
1200 UTC (Objective)	0.20	0.24	0.21
1200 UTC (Subjective)	0.10	0.22	0.16

results alone are not conclusive. However, along with analysis of the Milwaukee forecasts and the Milwaukee and Cornell climatologies, seasonal forecast uncertainty is clearly an important factor. Though forecasters in the Milwaukee experiment evaluated intervals day by day, results suggest the observed correlation might nearly have been matched simply by judicious use of the error climatology. The implications of this will be discussed in section 4.

b. Precipitation forecasts

Participants from semester 1 of the Cornell contest uniformly showed skill in allocating probability across precipitation categories relative to persistence forecasts. For example (as shown in Fig. 1, output from week 12 of the forecast contest), all of the top contestants' precipitation skill for the semester to date exceeded 50%.

The skill scores alone do not present a complete picture, however. Another informative diagnostic is a comparison between the average forecast distribution and the sample climatological distribution of precipitation during the contest. The correspondence between these relates to the bias, or "reliability in the large," of the forecasts. These results are presented in Table 3 for the top three forecasters and two automated schemes, one based on the Binghamton and Syracuse MOS and the other a blend of persistence and climatology. The difference between actual relative frequency and average forecast probability is rather small for most categories, though all forecasters underforecast the relative frequency of the category 6 events ($\geq 1.0''$). To some degree, this underforecasting results from the requirement that probabilities be rounded to the nearest 10%. However, the underforecast of category 6 was primarily due to one storm (2–3 March 1994) in which forecasters uniformly followed model guidance, which severely underforecast snow amounts (over 20 in. fell when 7 in. were predicted).

Another interesting question is whether the spread of precipitation forecasts is positively correlated with the probability-weighted absolute error of the precipitation forecast. For example, is the error between the average forecast category and the verification category higher on a day when the forecaster issues a widely distributed forecast (e.g., 2–2–3–2–1–0) versus a more specific forecast (e.g., 0–8–2–0–0–0)? If so, this would indicate proficiency in determining the uncertainty of the forecast situation. Accordingly, the forecasts of the top three contestants were examined in this way, correlating the standard deviation of the precipitation forecasts to their absolute errors of the verification category minus the mean forecast. The results indicate better success than with temperatures. The top three forecasters' correlations were +0.49, +0.52, and +0.59, respectively. The rank (Spearman) correlations were +0.75, +0.62, and +0.73, respectively. A scatterplot of the absolute error versus the standard deviation of the forecast is shown for the third-ranked forecaster, Capt. Ensemble, in Fig. 5.

The higher correlations for precipitation might have been expected. Though a notable failure was indicated earlier for the 2–3 March 1994 snowstorm, these days composed a small part of the full sample. Overall, there were many days when forecasters were quite confident in their forecasts for no precipitation and issued a 10–0–0–0–0–0 forecast, which usually verified accordingly. Analysis of an experiment in probabilistic quantitative precipitation forecasting by NWS personnel (Murphy et al. 1985) and MOS (Murphy et al. 1985; Wilks 1990) indicated that much of the forecast skill resided in the probability for the 0.0–trace category, that is, in the POP forecast. Also, the observed high correlation may also be accounted for by accurate assessments of uncertainty by season or by regime. Forecasters understand that summertime forecasts are often more variable due to the convective, hit-or-miss nature of precipitation and typically issue less concentrated forecasts.

TABLE 3. A comparison of the distributions of precipitation forecast probabilities over the duration of the contest for the top three forecasters and the two automated schemes compared to the verifying relative frequencies.

Precipitation category (j)	0.0 in.–trace	0.01–0.09 in.	0.10–0.24 in.	0.25–0.49 in.	0.50–0.99 in.	>1.0 in.
Verification	55.5%	17.5%	11.1%	6.3%	6.3%	3.2%
ART2D2	50.3%	23.6%	14.1%	7.3%	4.0%	0.6%
Dogbert	50.3%	20.9%	14.3%	7.7%	5.7%	2.5%
Capt. Ensemble	52.4%	21.5%	12.4%	7.6%	4.6%	1.4%
MOScaster	54.0%	24.1%	10.7%	10.0%	7.9%	0.0%
Stocaster	50.6%	24.4%	13.4%	7.7%	3.6%	0.0%

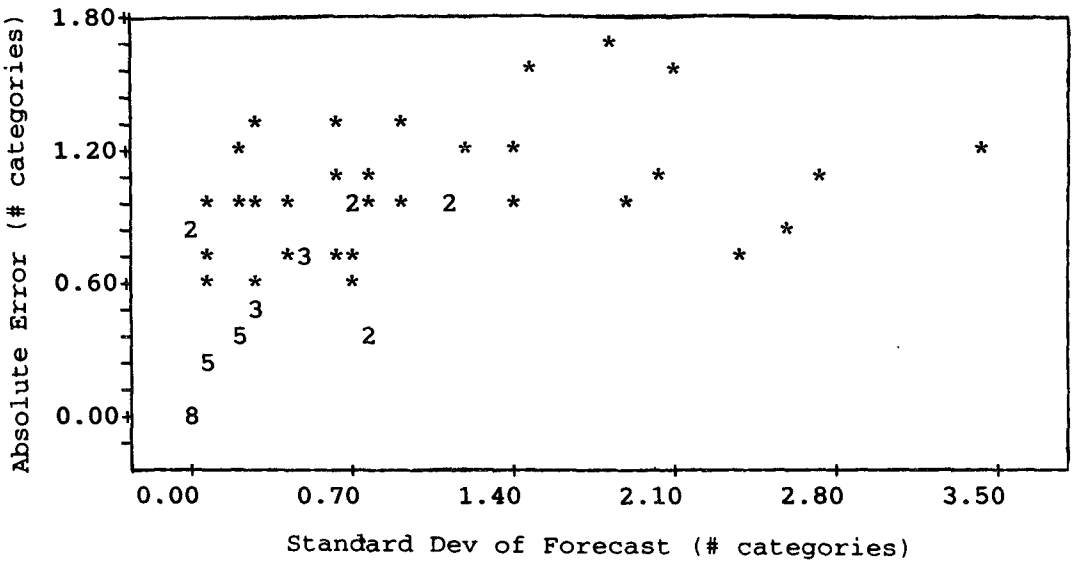

FIG. 5. Scatterplot of the absolute error of precipitation forecasts plotted against the standard deviation of the forecasts. Data are from forecasts generated by the contestant Capt. Ensemble.

An inaccurate assessment of the uncertainty of an ordinary event, however, is certainly much less important to the forecast user than assessments for large precipitation events. Thus, the two badly missed forecasts for 2–3 March did not dramatically affect the correlation coefficient, but they would have been prominent failures in the mind of a forecast user.

4. Discussion

The analysis in section 3 focused primarily on comparisons between the Milwaukee experiment, conducted with NWS forecasters, and the Cornell University experiment, conducted with student, faculty, and staff contestants. The higher correlation between interval width and forecast error observed with Milwaukee forecasters is likely due to the different level of experience but may also be due to the shorter duration of the Cornell contest, which masks seasonal effects. Two implications emerge from this analysis. First, though it is easy to learn the rudiments of correctly forecasting the most likely weather outcome, quantifying uncertainty in possible outcomes is much more difficult. Second, since the seasonal effects are large, the ability to assess forecast skill day to day may be less pronounced than previously believed. The Milwaukee experiment exhibited a correlation of interval width and absolute error of +0.35. If compared against a climatological baseline, with correlations of approximately +0.20, the success appears less impressive. This certainly does not invalidate the concept of nor the need for probabilistic forecasts. Rather, it highlights the necessity of training and research into ways to assess forecast uncertainty.

a. Subjectively evaluating uncertainty

Peterson et al. (1972) detailed one method for training forecasters to set central credible intervals; more recent literature on the subject is available in Murphy (1985, pp. 353–356) and Clemen (1991, Chap. 8). To set the credible intervals, the forecaster first decides which temperature $T_{0.5}$ is equally likely to have a verifying temperature above or below it. This is the median of the forecaster's subjective probability distribution. Using this temperature, the forecaster next selects a temperature $T_{0.25}$, which is equally likely to have the verifying temperature below it and between it and $T_{0.5}$. This temperature is the lower bound of the 50% credible interval. In the same way, the forecaster next evaluates $T_{0.75}$, the upper bound of the 50% central credible interval.

Developing a realistic subjective distribution of weather event probabilities (and thereby credible intervals) is probably learned over time through day-in, day-out exposure to different forecast situations. Over the long term, a forecaster can develop a mental database of the reliability or unreliability of a model's performance under varying weather patterns. Further, an experienced forecaster may be better able to understand the implications of discrepancies between data sources, such as satellite or radar data indicating a deviation from the short-term model forecast. These considerations may partly explain the lower correlation of interval widths to error for the contestants in this forecast compared with the NWS forecasters studied by WM79. Additionally, experienced forecasters may have an understanding of the synoptic climatology of uncertainty. The presence of a strong baroclinic zone over the area

or upstream may trigger a lack of confidence in a specific forecast. Further, perhaps an experienced forecaster looks for variability in humidity, boundary layer temperatures, or precipitation amount from model to model or from cycle to cycle as useful diagnostics of uncertainty.

All of the above may be tenable hypotheses, yet there has been little or no objective testing to validate most of these rules; thus, "experience" may reflect the use of both valid rules and invalid rules the forecaster thinks are valid. For example, some of the more experienced Cornell contestants used the differences in predicted precipitation between the eta and NGM models to assess the spread of probabilities to assign to their precipitation forecast. The rule of thumb was that consistency between model runs implied high forecast confidence, and vice versa. However, for the 2–3 March 1994 snowstorm, snowfall was underpredicted in both the eta model and NGM during all cycles. Thus, this particular rule did not work when needed most, and the extent of its validity is unknown.

b. Objectively evaluating uncertainty

If there are indeed forecast users who would benefit from accurate quantitative assessments of uncertainty day by day, then the objective evaluation of uncertainty is a legitimate area for research. This research could take at least two paths. First, one could test the validity of the rules of thumb developed by experienced forecasters, evaluating, for example, the validity of assessing uncertainty from differences between model forecasts. To the extent that these rules are synoptically general and not specific to one locale, such research would benefit new forecasters, helping them quickly build the missing experience base. There is an extensive literature on the error characteristics of particular forecast models, which may help forecasters evaluate their confidence in a particular model realization. However, as forecast models improve, systematic errors are decreasing, and forecast model errors are increasingly attributable to random errors, that is, sensitive dependence on initial conditions (Reynolds et al. 1994). Thus, long-term experience with a new model such as the eta model may not be as profitable as with previous models.

The second area of research likely to be of use is stochastic-dynamic prediction, or more specifically here, ensemble forecasting applied to the short range. With ensemble forecasts, multiple forecasts are run from the same model with slightly different initial conditions. Alternative initial conditions are generated to produce divergent forecasts, yet, which are dynamically consistent and physically plausible considering the observations. Ensemble forecasts are useful, first, because the mean of an ensemble of forecasts is typically more skillful than the majority of its members (Toth and Kalnay 1993). Further, the variance among ensemble members has been shown to be correlated to forecast uncertainty (Palmer 1994). Consider a particular day when the local evolution of the forecast is very sensitive to specific small perturbations in the initial conditions. The ensemble will consist of a collection of very dissimilar weather forecasts, and such a day will present a more uncertain forecasting situation. Conversely, if another day's forecast evolution is less sensitively dependent on the initial condition, then its ensemble spread will likely be smaller and the forecast uncertainty lower. Perhaps a short-range ensemble forecast (Brooks and Doswell 1993) will aid in defining the probability of damaging weather. For example, should one-third of the forecasts in an ensemble predict a heavy snow event, this may indicate a greater likelihood for that event than if one-tenth of the members make this prediction. Recently, Mullen and Baumhefner (1994) demonstrated the utility of the ensemble methodology in making short-range forecasts of explosive cyclogenesis, though that study concentrated on errors in the position and intensity of the cyclone rather than the sensible surface weather.

5. Conclusions

This paper has described a probability-based forecast contest conducted at Cornell University during the spring semester of 1994. The interesting aspect of this contest relative to other forecast contests around the United States was the inclusion of probabilistic temperature forecasts using central credible intervals. Forecasters showed little success in assessing the day-to-day uncertainty in temperature forecasts, as measured by the correlation of credible interval width to absolute error. This result is in contrast to WM79, where professional NWS forecasters' temperature intervals were positively correlated with forecast. Several factors, including the shorter length of the Cornell contest and inexperience of the forecasters, could account for the lower correlation. However, it was demonstrated that more than half of the observed correlation in the Milwaukee experiment of WM79 could be accounted for by the seasonal nature of forecast variability, even if forecasters were not explicitly accounting for this. Thus, results of this forecast contest suggest that short-term forecast uncertainty may be more difficult to evaluate day to day than previously believed, even for experienced forecasters.

Explicit guidance on assessing the uncertainty inherent in different forecast situations will certainly be useful. Ideally, rather than relying only on forecast experience, it would be desirable to also develop and test objective methodologies for estimating uncertainty. There are at least two fruitful areas of research. First, rules of thumb of experienced forecasters could be validated, determining the true usefulness of currently uncodified subjective algorithms for assessing uncertainty. A second method is the application of short-

range ensemble forecast technology. For probabilistic ensemble forecasting to be successful, days when the ensemble exhibits widely divergent forecasts locally should have greater forecast error on average than for days with more homogeneous ensemble forecasts. Short-range ensemble forecasting is still in the research stage; however, the technology appears very promising given the dramatic improvements made in medium-range forecasting with ensembles.

Despite the limited success of our contestants' probabilistic temperature forecasts, such forecasts would be valuable to and are desired by many weather forecast users. Fixed-probability temperature interval forecasts can be implemented by operational forecasters transparently, since the publicly issued product is simply a temperature range. Thus, we encourage others to experiment in making such forecasts. With time, documentation, and training, forecasters will undoubtedly develop the expertise to set intervals more accurately.

Acknowledgments: The authors thank Allan Murphy and John Jannuzzi for their careful reviews of this manuscript. The authors would like to thank other participants in this forecast contest, including Matt Annello, Jeff Baskin, Jeff Berardelli, Justin Berk, Wayne Bresky, Matt Briggs, Jason Cali, Richard Caniglia, Mike Carestia, Adam Cohen, Jeffrey Cuming, Art DeGaetano, Gregg Dehner, Chris Erbig, Peter Hall, Chris Leonardi, Michael Mischna, Jim O'Sullivan, Jeff Schultz, David Whitehead, Mary Wicks, and Mark Wysocki.

REFERENCES

Black, T. L., 1994: The new NMC mesoscale Eta model: Description and forecast examples. *Wea. Forecasting,* **9,** 265–278.

Bosart, L. F., 1983: An update on trends of skill of daily forecasts of temperature and precipitation at the State University of New York at Albany. *Bull. Amer. Meteor. Soc.,* **64,** 346–354.

Brooks, H. E., and C. A. Doswell, 1993: New technology and numerical weather prediction—A wasted opportunity? *Weather,* **48,** 173–177.

Clemen, R. T., 1991: *Making Hard Decisions: An Introduction to Decision Analysis.* PSW Kent, 557 pp.

Colucci, S. J., P. C. Knappenberger, and T. K. Ceppa, 1992: Evaluation of a nonprobabilistic weather forecast experiment. *Wea. Forecasting,* **7,** 507–514.

Croft, P. J., and J. D. Milutinovic, 1991: The Rutgers University forecasting contest: Forecaster performance versus model guidance. *Natl. Wea. Dig.,* **16,** 2–12.

Daan, H., 1985: Sensitivity of the verification scores to the classification of the predictand. *Mon. Wea. Rev.,* **113,** 1384–1392.

Dallavalle, J. P., J. S. Jensenius Jr., and S. A. Gilbert, 1992: NGM-based MOS guidance—The FOUS14/FWC message. Technical Procedures Bulletin 408, NOAA/National Weather Service, 9 pp.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte-Carlo forecasting. *Tellus,* **35A,** 100–118.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.,* **16,** 672–682.

Krzysztofowicz, R., 1983: Why should a forecaster and a decision maker use Bayes' theorem? *Water Resour. Res.,* **19,** 327–336.

——, W. J. Drzal, T. R. Drake, J. C. Weyman, and L. A. Giordano, 1993: Probabilistic quantitative precipitation forecasts for river basins. *Wea. Forecasting,* **8,** 424–439.

Miyakoda, K., G. D. Hembree, R. F. Strickler, and I. Schulman, 1972: Cumulative results of extended forecast experiments I: Model performance for winter cases. *Mon. Wea. Rev.,* **100,** 836–855.

Mullen, S. L., and D. P. Baumhefner, 1994: Monte Carlo simulations of explosive cyclogenesis. *Mon. Wea. Rev.,* **122,** 1548–1567.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

——, 1977: The value of climatological, categorical, and probabilistic forecasts in the cost-loss ratio situation. *Mon. Wea. Rev.,* **105,** 803–816.

——, 1985: Probabilistic weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 337–377.

——, and R. L. Winkler, 1974: Credible interval temperature forecasts: Some experimental results. *Mon. Wea. Rev.,* **102,** 784–794.

——, and ——, 1979: Probabilistic temperature forecasts: The case for an operational program. *Bull. Amer. Meteor. Soc.,* **60,** 12–19.

——, and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.

——, W.-R. Hsu, R. L. Winkler, and D. S. Wilks, 1985: The use of probabilities in subjective quantitative precipitation forecasts: Some experimental results. *Mon. Wea. Rev.,* **113,** 2075–2089.

Palmer, T., 1994: The ensemble prediction system (EPS): Status and plans. *ECMWF Newsletter,* **5** (spring issue), 3–15.

Peterson, C. R., K. J. Snapper, and A. H. Murphy, 1972: Credible interval temperature forecasts. *Bull. Amer. Meteor. Soc.,* **53,** 966–970.

Reynolds, C. A., P. J. Webster, and E. Kalnay, 1994: Random error growth in NMC's global forecasts. *Mon. Wea. Rev.,* **122,** 1281–1305.

Sanders, F., 1986: Trends in skill of Boston forecasts made at MIT, 1966–1984. *Bull. Amer. Meteor. Soc.,* **67,** 170–176.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.

Wilks, D. S., 1990: Probabilistic quantitative precipitation forecasts derived from PoPs and conditional precipitation amount climatologies. *Mon. Wea. Rev.,* **118,** 874–882.

——, and T. M. Hamill, 1995: On the economic value of forecasting forecast skill. Preprints, *14th Annual Conf. on Weather Analysis and Forecasting,* Dallas, TX, Amer. Meteor. Soc., J5(7)–J5(12).

Winkler, R. L., 1972: A decision-theoretic approach to interval estimation. *J. Amer. Stat. Assoc.,* **67,** 187–191.

——, and A. H. Murphy, 1979: The use of probabilities in forecasts of maximum and minimum temperatures. *Meteor. Mag.,* **108,** 317–329.