THE RUTGERS UNIVERSITY FORECASTING CONTEST: FORECASTER PERFORMANCE VERSUS MODEL GUIDANCE

Paul J. Croft and James D. Milutinovic

Department of Meteorology and Physical Oceanography
Cook College, Rutgers University
New Brunswick, New Jersey

Abstract

Forecast contest results for two years were examined to compare student and faculty (s/f) participants to model guidance. Forecasts of maximum and minimum temperatures (TEMP), extended maximum and minimum temperatures (ETEMP), probability of precipitation (POP), and probability of precipitation amount (POPA) were made for New Brunswick, New Jersey. LFM FOUS and NGM FOUE guidance for LGA, IPT, and PHL (for POPA); and LFM MOS predictions for EWR, ABE, and PHL (for TEMP and POP); and New Brunswick Climatology (for TEMP and ETEMP), routinely available to forecasters, were also entered into the contest for comparison to s/f participants. During the first contest year model guidance error scores were not disclosed to s/f forecasters. Results indicated that although most model guidance (MOS PHL, ABE, and EWR) did place among the top ten of more than thirty forecasters, several participants ranked higher in each forecast category. Not unexpectedly, FOUS POPA forecasts and CLIMATOLOGY TEMP and ETEMP forecasts did poorly. Biases in MOS TEMP predictions indicated varying tendencies of over- and under-predicting New Brunswick temperatures which were somewhat mimicked by the s/f participants. POPA forecasts showed virtually no skill. A comparison between the first and second contest years, when model guidance error scores were posted regularly for s/f forecasters to see, indicated that twice as many s/f forecasters were able to "beat" MOS TEMP and POP guidance the second year. Although MOS TEMP guidance did not lose appreciable ground in the standings the second year, error scores averaged nearly sixteen points higher. MOS POP guidance dropped as many as eight places in the standings the second year as error scores increased nearly three times over the previous year. Results indicate that when s/f forecasters know the ranking of model guidance relative to their own, they may improve their forecasting abilities by studying the magnitude and direction of model guidance error versus their own. This is of significant educational value in terms of participants' understanding of how to use MOS and FOUS guidance effectively and how to recognize the strengths and weaknesses of model guidance.

1. Introduction

National Weather Service and private forecasting firms alike rely on numerical guidance in the preparation and issuance of forecasts to public and private users. As a result, the accuracy and integrity of forecasts can be assessed by comparing human forecasters to numerical guidance, particularly when that guidance has been shown to improve with time (see Carter, et al., 1989). Many authors have put forth forecast verification and scoring algorithms since Brier (1950) introduced the "p-score" to measure the accuracy of probability of precipitation forecasts. Sanders (1963 and 1967), Murphy and Epstein (1967), Glahn and Jorgensen (1970), Gulezian (1981), Gordon (1982), Daan (1985), Murphy (1986), and Murphy and Winkler (1987) have all tested and/or decomposed the Brier score and found it to be a suitable measure of forecast accuracy.

Various scoring techniques have been used to evaluate the performance of both numerical guidance and human forecasters to detect changes in forecast accuracy or forecaster ability. Sanders (1973, 1979, and 1986) and Bosart (1975) both found that the improvement of student forecasters over numerical guidance showed little change with time. This led Snellman (1977) to introduce the concept of "meteorological cancer" to explain why the improvement of forecaster skill over guidance did not change with time. Snellman considered meteorological cancer to be the result of human forecasters' realization (or belief) that model guidance possessed the same, or better, skill than they. As a result, the forecaster would begin to use model guidance directly, and blindly, in forecast preparation. After an initial increase in skill, the forecaster's own skill would show little change.

Profiles of forecaster traits have also been examined and related to skill. Gedzelman (1978) studied the skill of beginner forecasters and concluded that most forecasting skill was acquired by the time they had made thirty forecasts. A profile of forecasters indicated that there was apparently little advantage to being meteorologically educated (although at least some basic coursework was essential) and that experience was more important in the forecast of unusual weather situations. Also, the amount of time spent on forecast preparation was found to be proportional to forecaster skill. Bosart (1983) found that new forecasters could routinely make skillful forecasts using numerical guidance without any appreciable meteorological understanding. Vislocky and Young (1988) pointed out that forecasters could enhance their skill scores given knowledge of the scoring method. By knowing how the scoring technique minimizes error, a forecaster may adjust his or her forecast accordingly to obtain the greatest benefit. However, as Murphy (1989) points out, this also inevitably leads to some hedging by forecasters.

There are two issues to be examined here. The first is that if forecasters rely heavily on numerical guidance in making general public forecasts, then it is important to assess their ability relative to guidance. This would provide an opportunity to determine what forecast skills could be improved by the forecaster, particularly if the forecaster's skill is not appreciably different from that of numerical guidance. The second issue evolves from the first in that most numerical guidance is site specific (e.g., numerical guidance and model output statistics) and is used in a "broad-brush approach" to predict local area conditions. It is of interest to assess how well (or poorly) site specific numerical guidance performs

when applied in this way by examining its skill in prediction for an alternate location. Further, since the forecaster is using site specific guidance as surrogate guidance for the local area, it is of interest to determine whether the guidance is doing a better job than the human forecaster for that area.

In order to evaluate forecaster skill versus model guidance, accumulated error scores may be compared. Although not a direct measure of skill, error scores offer an objective technique by which a forecaster's ability can be compared to that of the model guidance, *provided* the skill or error of the guidance is unknown to the forecaster. In this situation, the forecaster cannot be directly biased by knowledge of the skill of the guidance (except through a priori or acquired knowledge). Conversely, the error scores can also be compared when forecasters know how the guidance is doing. Although this eliminates true statistical independence between forecasters and numerical guidance, the knowledge gained by the forecaster may be useful in improving forecasts and at the same time provide an incentive to "beat" the guidance. Although error scores are not a direct measure of skill (since they fail to give the direction of the errors or make any comparisons to a control forecast) they do serve as a surrogate measure for the comparison of skill between forecasters and model guidance, independent of climatology, and between forecasters. Still, error scores are not truly independent since even in the first instance the forecaster usually has, or quickly develops, a knowledge of the model guidance's performance simply by seeing or using it. This is the case whether the forecaster knows model errors explicitly or not. Further, model and forecaster biases may be identified by the examination of individual or seasonal forecasts. These issues were examined using 1988–89 and 1989–90 data from the New Brunswick Forecast Contest. During the first year guidance error scores were withheld from forecasters while during the second year they were not.

Name			Date	
	Min	Max	Min	Max
TEMP				
ETEMP				
	12 HR	24 HR	36 HR	48 HR
POP				
POPA	TONIGHT		TOMORROW	
0-T				
0.01-0.10				
0.11-0.25				
0.26-0.50				
0.51-1.00				
>1.00				

	0-T	0.1-0.9	1.0-3.9	4.0-7.9	8.0+
SNOW					

Fig. 1. New Brunswick Forecast Contest Form (as described in text).

2. The Forecast Contest

The New Brunswick Forecast Contest is held twice weekly (every Monday and Wednesday) in the Department of Meteorology and Physical Oceanography, Cook College Rutgers University during the fall and spring semesters. Twenty-seven forecasts are made each semester for a total of fifty-four each contest year. Predictions are made for categories of temperature, extended temperature, probability of precipitation, probability of precipitation amount, and snowfall amount for the New Brunswick cooperative weather station for a four day (ninety-six hour) period. The forecasting form used is shown in Figure 1. The contest is open to all undergraduate and graduate students and faculty in the department. All forecasts must be completed by 6:00 p.m. of the forecast day to be accepted, and the participants are free to use any guidance or other information available in the Department (e.g., climatological records, the difax circuit, McIDAS, etc.). Contest rules are listed in Appendix A.

Forecasts of maximum and minimum temperature (TEMP) are made for four consecutive 12-hour periods, beginning 0000 UTC of the forecast day, to express overnight lows (0000–1200 UTC) and daytime highs (1200–0000 UTC) for the subsequent 48-hour period. Maximum and minimum temperature forecasts for the 48- to 96-hour period are also made (ETEMP) and are treated as a separate forecast category. All values are expressed in whole degrees Fahrenheit. Probability of precipitation (POP) forecasts are made for four consecutive 12-hour periods, the first beginning at 0000 UTC ("tonight") of the forecast day, and are expressed in tens of percent (as are public forecasts). The probabilities may range from zero to 100 and indicate the likelihood that 0.01 inches (0.254 millimeters) or more of precipitation will occur. Probability of precipitation amount (POPA) forecasts are made for two consecutive 12-hour periods, the first from 0000 to 1200 UTC ("tonight") and the second from 1200 to 0000 UTC ("tomorrow"), beginning at 0000 UTC of the forecast day. Probabilities may be expressed in multiples of five percent in each of the six class intervals shown in Figure 1 and must add to 100 percent. The difference between POP and POPA forecasts (5% versus 10% increments) was merely to allow forecasters more leeway in "spreading out" their probability estimates of precipitation amounts. Snowfall forecasts were similar to POPA but covered a 48-hour period. However, due to the limited number of occurrences of snowfall during any forecast period, further study of snow probabilities was omitted from this study.

During the 1988–89 and 1989–90 academic year, selected model guidance (and climatology the first year) forecasts were entered into the contest in order to compare their performance to student and faculty (s/f) forecasters. POPA forecasts were obtained from the FOUS output of the Limited-Area Fine Mesh (LFM) Model and the FOUE output of the Nested Grid Model (NGM) for IPT (Williamsport, PA), LGA (La Guardia Field, NY), and PHL (Philadelphia, PA) during the first year. POPA forecasts were obtained directly from numerical guidance such that a 100 percent probability was assigned to that class which matched the accumulated precipitation for the period. This was done so that "blind use" of model precipitation amount forecasts could be compared to

forecasters'. MOS quantitative precipitation forecasts were not used because the categorical forecasts did not match those of the POPA classes in the contest. Further, TEMP and POP forecasts were obtained from Model Output Statistics (MOS) based on the LFM for ABE (Allentown, PA), EWR (Newark, NJ), and PHL both years. The climatological mean daily maximum and minimum temperatures for New Brunswick were entered as TEMP and ETEMP CLIMATOLOGY forecasts during the first year. The model guidances were selected to serve as surrogate FOUS, FOUE, and MOS predictions for New Brunswick in spite of climatic non-homogeneities. The FOUS and FOUE guidance sites were selected to cover the forecast region about New Brunswick. The ABE MOS was chosen because of its climatological similarity to the New Brunswick site in terms of mean temperatures and precipitation. The EWR and PHL MOS were chosen in spite of their urban-heat island bias because of their proximity to the New Brunswick site. These selections were made to simulate the "real-world" forecasting problem encountered by National Weather Service and private forecasters alike in that forecasts are made for a city or area which is not a FOUS or MOS site. It is because of these non-homogeneities that simple linear interpolation of numerical guidance does not necessarily provide for an accurate forecast. The sites are shown in Figure 2 and provide a reasonable estimate of the anticipated weather conditions in New Brunswick.

Although all of the aforementioned products were routinely available to forecasters, the error scores of the objective guidance were not during the first contest year (1988–89). This was done for two reasons, first, in order to prevent s/f forecasters from being biased towards the use of any particular product during the contest (although, inevitably, some forecasters would be biased based upon their past or acquired experience); and second, to provide for further analysis. It was felt that this would allow for a somewhat objective, although not purely independent, comparison between model guidance and forecaster performance. During the 1989–90 contest year the error scores were posted with forecaster error scores so that s/f forecasters could compare their latest error scores with those of numerical guidance for each forecast day.

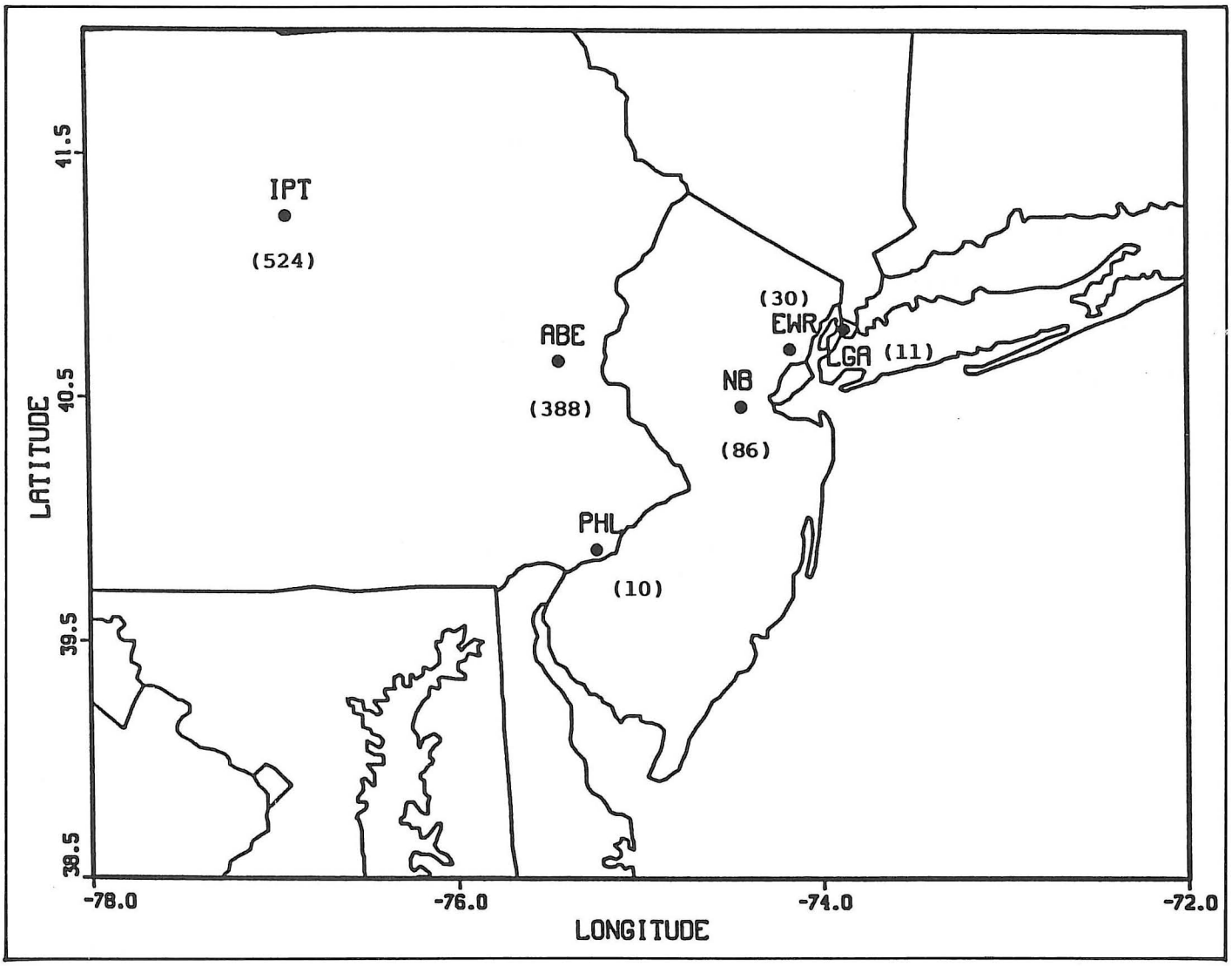

Fig. 2. Location of FOUS/FOUE and MOS model guidance used for New Brunswick (NB). LFM and NGM FOUS for Williamsport, Pennsylvania (IPT), LaGuardia Field, New York (LGA), and Philadelphia, Pennsylvania (PHL) were used for POPA forecasts. LFM MOS for Allentown, Pennsylvania (ABE), Newark, New Jersey (EWR), and PHL were used for TEMP and POP forecasts. Values in parentheses are elevations above mean sea level (in feet).

3. Verification and Scoring

All forecast categories were verified based on observations recorded at the New Brunswick, New Jersey cooperative weather station. Temperatures were verified using a digital maximum and minimum thermometer in combination with a hygrothermograph. Precipitation probabilities and amounts were verified by the standard eight-inch National Weather Service rain gauge and a recording weighing rain gauge.

For scoring purposes, error points for TEMP and ETEMP were calculated by taking the absolute value of the difference between predicted and observed temperatures. These differences were summed across TEMP and ETEMP for each forecast and accumulated over all forecast days. Error points for POP and POPA forecasts were determined using a variation of the Brier score where the error score (E_{pop}) was determined by:

$$E_{pop} = 100\ (F - O)^2 \qquad (1)$$

Where **F** is the forecast probability (in tenths) ranging from zero to one and **O** the observed probability (zero or one). The observed probability was set equal to zero when no precipitation was observed and to one when precipitation was observed. Further explanation of error score calculations for POP and POPA forecasts are given in Appendix B.

Error points for each forecast category were accumulated once a week for all participants for their Monday and Wednesday forecasts. These were posted as soon as verification was made and prior to the next contest day. The participants were then ranked by error points for each forecast category and standardized against the participant with the least error points in that category (i.e., the scores of the remaining participants were given relative to the forecaster with the fewest error points). For presentation purposes, the lead forecaster was then listed as having "0" error points and all other forecasters were then listed relative to the lead score. It was felt that this method of presentation was more meaningful to the participants and would encourage competition. In TEMP and ETEMP for example, the presentation allowed forecasters to determine how many "degrees they needed to make up" in order to improve their standing. Forecast standings were also subdivided to separate those participants who had participated in at least sixty percent of all forecasts from those who had not ("others"). This was done to discourage participants from not forecasting in order to maintain their standing in the contest. In order to appear in the final contest year standings, a forecaster was required to have participated in at least sixty percent of all forecasts as well.

Whenever a forecast (or a forecast category) was missed or filled out incorrectly, or a participant simply chose not to forecast, the fortieth percentile error score (based on the pool of active forecasters for that day) of each category was added to that participant's previous error scores. The fortieth percentile is that value below which forty percent of the observations (in this case, forecasters) lie. This was done in order to eliminate the problem of calculating a participant's daily or weekly error scores when forecasts were missed. These error scores were assigned each day for each category and were dependent on the number of forecasters participating each contest day and how well (or poorly) they forecasted as a group. The fortieth percentile was used to ensure that a forecaster did not "gain ground" in the standings when not forecasting. When a forecaster's POPA probabilities did not add to 100 percent, the fortieth percentile error score was assigned to that participant.

4. Results

Since some changes have been made in the New Brunswick Forecasting Contest during the last several years it was not possible to examine trends in skill. Instead the final standings of each contest year were tabulated according to accumulated error scores and normalized against the first place participant, as shown in Tables 1 and 2. All student and faculty (s/f) forecasters were letter-coded for confidentiality. Those s/f forecasters participating in sixty percent or less of the forecasts appear under the "others" heading. Because of changes in student enrollment, only sixteen s/f forecasters participated both years. There were thirty-seven participants the first year and thirty-nine the second although only a maximum of thirty forecasters were listed in the standings at any one time. This was done to favor forecasters who participated most often. The MOS for ABE, EWR, and PHL; CLIMATOLOGY (New Brunswick mean temperatures); and the LFM FOUS and NGM FOUE for IPT, LGA, and PHL; which were hidden during the duration the first contest year, are also shown. Each sample consisted of fifty-four contest days. Consensus forecast scores (representing the "state of the art") were computed for TEMP and POP during the 1988–89 contest. The consensus forecast was based on all s/f and MOS forecasts. The consensus was not computed during the second contest year as comparison between s/f forecasters and model guidance was of more concern.

4.1. Temperature

Results for 1988–89 indicated that eight forecasters scored better than the MOS TEMP forecasts and that all easily did better than CLIMATOLOGY (see Table 1). This was not surprising since, in the former instance, each of the MOS TEMP forecasts are not specifically for New Brunswick, and in the latter, CLIMATOLOGY is generally a poor forecast. In ETEMP, although forecasters again beat CLIMATOLOGY, the accumulated error points for CLIMATOLOGY (as compared to the first place participant) were approximately half those accumulated by CLIMATOLOGY for the TEMP category. This probably reflects the tendency of forecasters to trend toward climatological normals when forecasting ETEMP. CLIMATOLOGY was not included during the 1989–90 contest based on its performance during the 1988–89 contest. When calculated, the consensus forecast beat all but one forecaster in the contest.

Results for 1989–90 (Table 2) indicated that fifteen forecasters (nearly twice as many as the first year) scored better than the MOS TEMP forecasts. It was noted during this contest year that ABE MOS placed among the top five forecasters for seven consecutive weeks (and as high as first and second) from late September to mid-November before falling in the standings. The high performance of ABE MOS early in the contest led to an early advantage for model guidance and was difficult for forecasters to overcome. MOS TEMP forecasts generally placed lower the second year. One possible explanation is that knowledge of MOS performance allowed forecasters to improve their standing. It is also possible that forecasters learned to "play the verification game" better relative to the MOS and/or that model guidance simply had a "bad year" during the 1989–90 contest. MOS accumulated error point totals (not shown) averaged 15.9 degrees higher, although ABE MOS "improved" by a similar amount.

MOS TEMP forecasts were further examined each contest year to determine whether a systematic bias existed when

Table 1

TEMP	PTS BACK	ETMP	PTS BACK	POP	PTS BACK	POPA	PTS BACK
B	.0	W	.0	W	.0	R	.0
W	30.8	G	13.0	R	40.6	V	33.1
V	66.8	F	46.0	D	81.4	W	57.5
D	68.8	B	68.0	G	127.4	B	89.7
G	91.0	V	78.2	MOS PHL	143.0	E	93.8
F	91.2	E	103.6	MOS EWR	149.0	H	176.4
E	108.6	R	158.8	H	150.2	F	193.2
MOS PHL	112.2	H	189.8	B	163.4	D	195.8
R	112.6	D	277.0	MOS ABE	187.0	G	218.8
MOS EWR	124.2	CLIMATOLOGY	435.0	J	266.2	J	413.3
J	147.6			F	341.4	FOUS LFM PHL	929.5
MOS ABE	153.2			E	347.2	FOUS NGM LGA	938.3
H	180.4			V	350.2	FOUS NGM PHL	1201.5
CLIMATOLOGY	953.2					FOUS LFM LGA	1516.1
OTHERS							
K	111.0	K	174.2	AA	120.2	AA	253.7
A	120.8	A	174.6	K	297.0	N	363.9
N	138.6	BB	188.0	CC	338.4	U	424.7
M	138.6	N	192.2	BB	343.4	CC	484.0
S	147.6	M	204.0	N	456.8	T	486.4
T	148.4	T	222.6	T	466.4	K	496.0
CC	157.8	Q	224.4	M	490.6	Q	541.2
BB	158.6	Z	241.6	Z	579.8	M	557.0
Z	168.2	CC	242.8	X	643.2	X	566.9
Q	176.4	L	291.2	L	648.6	BB	632.2
U	190.4	I	303.6	Q	742.0	L	644.9
I	227.2	X	323.6	U	812.0	A	664.7
X	234.4	U	324.0	A	845.8	I	672.5
L	249.2	S	329.6	S	1035.2	S	773.7

Table 1. Final standings of New Brunswick Forecast Contest for 1988–89 academic year. Student and faculty forecasters were letter coded for confidentiality while FOUS/FOUE and MOS guidance, and CLIMATOLOGY were included following contest completion. Points back (PTS BACK) represents the difference between each category's first place participant's error score and all other participants' error scores. Those listed under "OTHERS" participated in less than sixty percent of all forecasts.

they were used to predict New Brunswick temperatures. The average deviation of the observed temperature from the predicted temperature was determined for each forecast period (1 = "tonight," 2 = "tomorrow," etc.) by each MOS "season" TEMP forecast. The MOS seasons considered were fall (Sep 1–Nov 30), winter (Dec 1–Feb 28), and spring (Mar 1–May 31). Table 3 lists the average deviations by season for ABE, EWR, and PHL; and for an average MOS (based on the mean of three stations combined) for the 1988–89 contest year and Table 4 for the 1989–90 contest year. The percent frequency with which New Brunswick lows and highs were over (positive values) or under-predicted (negative values) are also shown. An obvious signal was apparent in the fall and winter of the 1988–89 contest during which ABE MOS consistently predicted maximum and minimum temperatures to be one to four degrees colder than those actually observed in New Brunswick. During the 1989–90 contest (see Table 4), this signal was less clear in the fall but more emphatic in the winter for forecast periods three and four. During the spring season, ABE MOS exhibited a much weaker and somewhat diurnal signal in both years such that predicted minimum temperatures were higher than observed in New Brunswick while predicted maxima were generally lower. The fall and winter differences were felt to be related to the more continental climate of ABE and its slightly higher elevation (388 feet versus 86 feet) while those of the spring were not as easily explained.

An apparent diurnal signal was evidenced in EWR MOS (both years) and PHL MOS (the first year) predictions during the fall as they over-predicted minimum temperatures and underpredicted maximum temperatures in New Brunswick by one to three degrees. This minimum temperature difference is believed related to differences between urban and suburban climates of New Jersey and is well documented by DeGaetano and Shulman (1984) and O'Reilly et al., (1988). The tendencies of EWR and PHL MOS to over-predict minimum and maximum temperatures in the third and fourth periods in 1988–89 and under-predict them in 1989–90 may be related to differences in weather regimes each year. The spring season signals for EWR and PHL MOS and all MOS combined were quite strong and consistent both years (particularly EWR). In both the EWR and PHL MOS, the first

Table 2

TEMP	PTS BACK	ETMP	PTS BACK	POP	PTS BACK	POPA	PTS BACK
DD	.0	EE	.0	G	.0	W	.0
W	6.0	F	40.8	F	.0	F	23.5
G	20.2	GG	58.6	W	66.8	G	45.1
F	21.0	D	66.8	GG	158.6	DD	102.8
EE	49.2	Z	67.4	D	209.8	Z	119.1
Z	54.8	W	78.2	Z	236.6	GG	121.9
FF	63.0	DD	115.2	LL	240.0	KK	135.1
DD	75.8	G	121.4	MOS ABE	282.6	LL	191.5
GG	91.6	JJ	176.6	DD	282.8	J	216.1
J	91.6	FF	196.4	HH	308.6	JJ	258.6
HH	112.8	LL	358.8	JJ	318.2	HH	310.3
MOS PHL	121.8	S	415.4	MOS EWR	323.6	D	320.8
JJ	125.2			MOS PHL	383.6	EE	464.5
MOS ABE	137.8			EE	432.8	FF	480.0
S	153.8			KK	462.4	S	857.2
MOS EWR	177.8			S	616.6		
KK	193.2			FF	716.6		
LL	238.8			J	998.0		
OTHERS							
B	33.8	B	94.4	B	255.2	AA	145.4
N	59.4	MM	128.6	MM	283.4	B	171.6
MM	78.2	N	135.2	QQ	285.8	MM	232.0
NN	114.6	U	152.0	AA	299.2	PP	298.2
U	125.0	HH	155.4	NN	378.4	U	309.6
PP	159.0	QQ	168.8	U	514.2	N	324.5
QQ	247.6	KK	194.0	N	529.4	QQ	430.7
RR	257.6	PP	289.8	RR	708.8	NN	605.7

Table 2. Same as in Table 1, but for 1989–90 contest year.

Table 3

	ABE			EWR			PHL			ALL MOS		
Fcst Per	Ave Dev	%+	%−	Ave Dev	%+	%−	Ave Dev	%+	%−	Ave Dev	%+	%−
FALL												
1	−1.43	34.8	56.5	2.39	65.2	30.4	0.83	52.2	34.8	0.59	50.7	40.6
2	−2.65	13.0	78.3	−1.13	17.4	60.9	−0.83	26.1	56.5	−1.54	18.8	65.2
3	−0.57	39.1	56.5	3.48	73.9	13.0	2.22	73.9	26.1	1.71	62.3	31.9
4	−3.74	17.4	78.3	−1.83	26.1	60.9	−1.65	30.4	65.2	−2.41	24.6	68.1
WINTER												
1	−4.00	7.1	85.7	1.21	64.3	35.7	0.36	57.1	35.7	−0.81	42.9	52.4
2	−4.86	0	78.6	−1.14	28.6	64.3	−1.00	35.7	50.0	−2.33	21.4	64.3
3	−2.21	35.7	64.3	1.71	57.1	35.7	0.93	50.0	28.6	0.14	47.6	42.9
4	−1.29	28.6	71.4	1.86	71.4	28.6	1.57	57.1	28.6	0.71	52.4	42.9
SPRING												
1	0.82	70.6	23.5	4.19	81.3	18.8	4.00	76.5	17.6	2.98	76.0	20.0
2	−1.18	41.2	52.9	−1.38	31.3	62.5	1.41	64.7	29.4	−0.36	46.0	48.0
3	0.59	52.9	29.4	3.56	81.3	18.8	3.06	82.4	17.6	2.38	72.0	22.0
4	−0.41	47.1	41.2	0.75	56.3	43.8	3.12	64.7	29.4	1.16	56.0	38.0

Table 3. Average deviations (Ave Dev) and relative frequencies of over (%+) and under (%−) prediction of New Brunswick temperatures by ABE, EWR, and PHL MOS, and all MOS combined, for each MOS season and by forecast period (Fcst Per) during the 1988–89 contest. Frequencies which do not sum to 100 indicate the relative frequency of predictions which exactly matched observations.

and third period minimum temperature predictions were two to four degrees higher than those observed in New Brunswick. In both years, second period maximum temperatures were under-predicted by EWR MOS and over-predicted by PHL MOS with both over-predicting fourth period maximum temperatures.

Although only two years of data were used, it was felt that the magnitude, direction, and consistency of some of the

Table 4

	ABE			EWR			PHL			ALL MOS		
Fcst Per	Ave Dev	%+	%−	Ave Dev	%+	%−	Ave Dev	%+	%−	Ave Dev	%+	%−
FALL												
1	0.87	56.5	34.8	4.57	82.6	8.7	3.30	65.2	21.7	3.00	69.6	21.7
2	−1.70	26.1	65.2	−0.52	39.1	43.5	0.26	43.5	30.4	−0.65	26.1	52.2
3	−0.09	43.5	39.1	2.22	65.2	30.4	2.04	69.6	30.4	1.35	65.2	30.4
4	−1.78	26.1	65.2	−0.22	39.1	52.2	0.13	47.8	47.8	−0.65	39.1	47.8
WINTER												
1	−2.13	26.7	66.7	3.07	80.0	20.0	1.60	73.3	20.0	0.73	66.7	33.3
2	−3.07	20.0	66.7	0.20	46.7	53.3	0.47	46.7	46.7	−0.67	46.7	53.3
3	−4.93	6.7	93.3	−0.53	33.3	46.7	−2.13	40.0	60.0	−2.47	33.3	66.7
4	−5.67	20.0	80.0	−2.33	20.0	66.7	−2.60	20.0	73.3	−3.60	20.0	80.0
SPRING												
1	0.94	43.8	56.2	4.38	81.3	18.7	4.13	81.3	12.5	3.19	62.5	18.8
2	−0.44	43.8	56.2	−0.94	31.3	56.3	1.56	43.8	50.0	0.00	43.8	56.2
3	−0.38	37.5	50.0	3.13	56.3	31.3	2.38	56.3	43.7	1.75	50.0	43.8
4	−0.44	50.0	43.8	1.00	50.0	37.5	1.56	56.3	37.5	0.88	62.5	37.5

Table 4. Same as in Table 3, but for 1989–90 contest year.

bias signals were reasonable since the MOS equations are relatively stable. The biases were felt to be related to factors such as urbanization (EWR and PHL), elevation (ABE), and possibly latitude (all). However, differences in atmospheric circulation from year to year confound these apparent biases. The individual deviations were also plotted against time for each period for each MOS to determine whether a trend existed, during any season, but no trend was evident.

A similar procedure was followed for s/f forecasters for the TEMP category and results are shown in Table 5 for both contest years. Forecaster biases paralleled those for all MOS combined in nearly every period each season of the first contest year. This was not true during the second year when forecaster biases matched those of the MOS only during the winter season for the third and fourth forecast periods. In general, s/f forecaster biases were of lesser magnitude the second year and may reflect improved forecasting ability based on knowledge of model guidance performance.

4.2. Precipitation

The final standings for POP forecasts indicated that only six forecasters were able to beat PHL MOS during 1988–89 (Table 1). Four of these forecasters had more than four years of meteorological education and/or experience, while one was a freshman undergraduate student participating in the contest for the first time. Three of these four were able to beat the POP consensus forecast during the 1988–89 contest. In the 1989–90 contest eight forecasters (nearly twice as many) were able to beat ABE MOS. Four had at least four years of experience, while the remaining three had at least two years of experience. MOS POP guidance for PHL and EWR fell six to eight places in the standings during the second contest year while ABE moved up one position. MOS POP the second year had between two-and-one-half and three times greater accumulated (normalized) error points when compared to the leading s/f forecasters' error scores and may reflect either an advantage to the s/f forecasters of "seeing" how the MOS guidance was performing during the contest, or as a significant improvement in POP forecasting ability by s/f forecasters.

The frequency of "blown" POP forecasts, that is, when the precipitation probability assigned by a forecaster was

Table 5

	Average Deviation	Relative Frequency (%+)	Relative Frequency (%−)
FALL			
1	−0.51/0.42	41.2/50.0	50.8/41.0
2	−1.40/1.07	30.3/39.0	61.7/48.6
3	0.71/0.10	51.7/44.8	41.0/48.3
4	−1.74/0.68	30.5/34.8	61.3/54.0
WINTER			
1	−0.97/−0.02	35.7/45.4	55.3/47.6
2	−1.93/−0.16	23.0/62.6	64.7/31.1
3	1.08/−2.44	56.6/29.3	35.7/64.5
4	0.84/−2.81	47.7/17.9	42.1/80.2
SPRING			
1	0.98/0.43	58.5/49.6	32.2/40.5
2	−1.07/−0.04	43.7/44.1	50.8/47.4
3	1.03/1.16	57.4/51.8	35.0/40.9
4	−1.09/1.30	45.4/62.8	50.3/30.0

Table 5. Average deviations and relative frequencies of over (%+) and under (%−) prediction of New Brunswick temperatures by s/f forecasters for 1988–89/1989–90 contest years.

zero and precipitation occurred, or conversely when a probability of 100 was assigned and no precipitation occurred, was also examined. Although many forecasters had blown a POP forecast at least once, in either of the manners described above (fifteen and eighteen, respectively for the 1988–89 forecast year; and twenty-five and seven during 1989–90), the MOS forecasts for ABE, EWR, and PHL never did the first year; and only six times (a maximum of three times for PHL) the second year. Virtually all "blown" MOS POP forecasts occurred in the fourth forecast period (only once for the third period) whereas s/f forecasters "blew" POP in all forecast periods. The s/f forecasters were up to five times more likely to "blow" a POP forecast in the third and fourth periods than the MOS guidance. It was apparent from these results that the MOS POP forecasts were hard to beat in terms of "blown" forecasts. Discussion with forecasters who had "blown" a forecast indicated that they were either "just going for it," particularly in the third and fourth forecast

periods, or being overconfident in the first and second periods.

POPA forecasts made by LFM FOUS and NGM FOUE guidance did poorly in the 1988–89 contest and were easily beaten by all forecasters. This was not unexpected as FOUS POPA forecasts were effectively binary ("all or nothing") while s/f forecasters were able to "spread out" their probability forecasts. FOUS POPA were used this way so that it would not be necessary to develop decision making tools as to what type of distribution to use (e.g., lognormal) in "spreading out" the probability. The objective was to determine whether forecasters could beat "blind" guidance, and in this case, they could. Further, quantitative precipitation forecasts based on LFM guidance have been shown to exhibit relatively little skill (see Gyakum and Samuels, 1987). It was most likely the ability to spread out probabilities that gave s/f forecasters the edge in POPA over the binary guidance. The most "accurate" FOUS POPA forecasts (when applied to New Brunswick) were PHL (LFM) and LGA (NGM) model output.

5. Conclusions

The analysis of forecaster performance and comparison between forecasters and model guidance (as suggested by Murphy, 1989), particularly for locations for which site specific guidance is not available, is important in providing insight to forecaster development and improvement. When compared to model guidance, s/f forecasting ability in each forecast category was found to be generally higher. It was noted that TEMP forecasts made by s/f participants sometimes mimicked the prediction biases of the LFM MOS. In POP forecasting, s/f participants with more than four years of experience and/or education (with one exception) were able to "beat" the MOS predictions. This may indicate experience to be a more important factor in POP forecasting. The impact of participants' knowledge of model guidance performance was assessed in 1989–90 when model guidance error scores were shown in the weekly standings. Nearly twice as many s/f forecasters were able to score better than model guidance in both TEMP and POP during the second contest year. This may be related to their knowledge of the performance of model guidance during the contest and from knowledge gained from viewing the previous year's results.

The results indicate that although the use of model guidance as a surrogate model forecast for an alternate location (evidently a common practice) does appear to be reasonable, steps should be taken by the forecaster to account for climatological differences between MOS and non-MOS stations when forecasting. Although linear interpolation may provide a "first-guess" forecast, a more methodical approach, such as that offered by Walts and Pochop (1977), would produce better forecasts. Further, the results suggest that continuous comparisons between forecasters and model guidance are of educational value and allow for a better appreciation of MOS biases, strengths, and weaknesses by the user. The results also present a challenge in that NGM MOS guidance, presently available, will eventually replace LFM MOS guidance (see Carter, et al., 1989). Therefore it is important that comparisons between s/f forecasters' performance and that of the NGM MOS begin as soon as possible to allow forecasters to get a "feel" for the new guidance. The 1990-91 contest incorporated this information to assess NGM MOS performance for comparison with s/f forecasters and is being continued in the 1991–92 contest.

Acknowledgments

This is a paper of the Journal Series (D-13001-2-90), New Jersey Agricultural Experiment Station, Cook College, Rutgers University, New Brunswick, New Jersey. Thanks are extended to Nathan M. Reiss and Mark D. Shulman for their review and commentary of the manuscript. Credit is also given to Anthony J. Broccoli and John R. Lanzante who created the original contest and source code. The authors also thank the NWD reviewers for their helpful comments and revisions toward publication of this paper.

Authors

Paul J. Croft received his B.S. and M.S. degrees in Meteorology at Cook College, Rutgers University in 1985 and 1987. He has recently completed an Interdisciplinary Ph.D. in Agricultural Meteorology at Rutgers University. He has served as Assistant State Climatologist of New Jersey and is presently an Instructor with the Department of Meteorology and Physical Oceanography at Cook College. His research interests are in Applied Meteorology and Applied Climatology.

James D. Miluntinovic received his B.S. in Meteorology and Mathematics at Cook College, Rutgers University in 1983. He received his M.S. in Meteorology and Interdisciplinary Ph.D. in Meteorology, Oceanography, and Remote Sensing from Rutgers University in 1985 and 1990 respectively. He worked on the FAA Central Weather Processor Project and is currently employed at Accu-Weather, Inc. His research interests are Remote Sensing, Image Processing, and Graphical Presentation of Meteorological data.

APPENDIX A

NEW BRUNSWICK FORECASTING GAME RULES

Forecasts will be made twice each week, on Monday and Wednesday. Each forecast will include the following parameters: Temperature, Extended Temperature, Precipitation Probability, Probability of Precipitation Amount, and Probability of Snow Amount. Scores will be computed for each of these five parameters, and separate standings will be maintained for each parameter. Scoring rules for each variable will be discussed below.

1. Temperature

A temperature forecast will be issued for each of four consecutive 12-hour periods, running from 00Z–12Z, 12Z–00Z, 00Z–12Z, 12Z–00Z. The temperature forecast should be the minimum, maximum, minimum and maximum, respectively for each of the four periods. The digital maximum/minimum recording thermometer in conjunction with

the hygrothermograph will be used to verify these forecasts. Error points will be calculated by taking the absolute value of the forecast error (that is, the forecast minus observed).

2. Extended Temperature

An extended temperature forecast will be issued for each of four consecutive 12-hour periods beginning immediately following the last period of the standard temperature forecast. The extended temperature forecast should be the minimum, maximum, minimum, and maximum, respectively, for each of the four periods. On Monday these will cover Wednesday nighttime, Thursday daytime, Thursday nighttime, and Friday daytime. On Wednesdays, these will cover Friday nighttime, Saturday daytime, Saturday nighttime, and Sunday daytime. Verification and scoring will be the same as used in the temperature section.

3. Precipitation Probability

A precipitation probability forecast will be issued for each of four consecutive 12-hour periods beginning at 7:00 P.M. EST on the day the forecast is issued. These are the same periods for which the temperatures are forecast. The probability of significant precipitation, *DEFINED AS THE PROBABILITY OF RECEIVING 0.01 INCHES,* for each 12 hour period should be forecast. Only percentages which are even multiples of ten (e.g., 0, 10, 20, etc.) are allowed. These forecasts will be scored using a variation of the half-Brier score, a standard scoring method for probability forecasts. Error points are assigned according to the following tables.

For No Rain Observed

Forecast	0	10	20	30	40	50	60	70	80	90	100
Score	0	1	4	9	16	25	36	49	64	81	100

For Rain Observed

Forecast	0	10	20	30	40	50	60	70	80	90	100
Score	100	81	64	49	36	25	16	9	4	1	0

The standard eight (8) inch rain gauge measurement in conjunction with the weighing rain gauge will be used to determine the occurrence or non-occurrence of measurable precipitation.

4. Probability of Precipitation Amount

A forecast of probability of precipitation amount is made for each of the first two 12 hour periods for which precipitation probability forecasts are made. This is *not* a conditional forecast—it will be scored *every day* even if rain does not occur. Estimate the probability of the precipitation amount falling in each of the six categories for each 12 hour period.

Error points will be calculated using a variation of the Brier score (Panofsky and Brier, 1968). Error points will be weighted so that the closer a forecast is to the verification amount, the fewer the error points that will be penalized. Error points for the VERIFICATION category are calculated according to the tables in Section 3 (POP). Error points for the NON-VERIFICATION category are calculated using the following formula:

$$\mathrm{ERROR}_{\mathrm{POPA}} = \frac{\mathrm{ERROR}_3}{(\mathrm{N} - \mathrm{DIST}_{\mathrm{CAT}})}$$

Where $\mathrm{ERROR}_{\mathrm{POPA}}$ is POPA error, ERROR_3 is the error for this category as calculated in the tables of section 3, N is the number of categories for POPA (N = 6), and $\mathrm{DIST}_{\mathrm{CAT}}$ is the number of categories away the forecast is from the verification category.

The probabilities must add up to 100 percent; any forecaster not adhering to this requirement will be given the fortieth percentile score for that category. Again, the standard 8 inch rain gauge measurement along with the weighing rain gauge trace will be used to verify this forecast.

5. Probability of Snow Amount

A forecast of probability of snow amount (including sleet, freezing rain, etc.) is made for the entire 48-hour period beginning at 00Z (7 P.M. EST) on the day the forecast is issued. Estimate the probability of the total snowfall falling in each of the five categories. *The probabilities must add up to 100 percent;* any forecaster not adhering to this requirement will be given the fortieth percentile score for that category. This is *not* a conditional forecast—it will be scored each day even if snow does not occur. Snowfall will be scored in the same manner as probability of precipitation amount, as described in section 4. The official New Brunswick observations will be used to verify this parameter. *Probabilities should be estimated in increments of 5 percent.*

Miscellaneous

All forecasts should be written on the forecast forms and placed in the box by 6 P.M. Once you have put your forecast in the box, you may not alter it. If two forecasts with the same name are found in the box, neither forecast will count. If a forecaster does not make a forecast on a given day, he/she will receive a score equal to the fortieth percentile score of the forecasts made on that day. Similarly, if any portion of a forecast is missing or invalid, the forecaster will receive the fortieth percentile score for that portion of the forecast. In order to appear in the standings for a particular variable a forecaster must have made *at least 60 percent of the possible forecasts* for that particular variable.

References

Panofsky, H. A., and G. W. Brier, 1968: *Some Applications of Statistics to Meteorology,* The Pennsylvania State University, pp. 201–203.

APPENDIX B

POP Scoring

The error points associated with the POP forecast are scored based on equation (1) as presented in section 3. Error scores for various predicted probabilities are given for the occurrence and non-occurrence of precipitation.

For no precipitation observed:

Forecast (%)	0	10	20	30	40	50	60	70	80	90	100
Error Score	0	1	4	9	16	25	36	49	64	81	100

For precipitation observed:

Forecast (%)	0	10	20	30	40	50	60	70	80	90	100
Error Score	100	81	64	49	36	25	16	9	4	1	0

POPA Scoring

The POPA forecast was treated as an unconditional forecast and therefore was scored whether or not measureable precipitation was observed. Scoring of each POPA class in the POPA category (see Figure 1) involved a two-part scoring process. The verifying forecast class ($POP_{(v)}$, one of the six) error score was derived in the same manner as the POP forecast and then added to the error points accumulated in each of the non-verification classes ($POP_{(nv)}$, the five remaining). The error score for each non-verification class was computed by first calculating the error points for that class according to equation (1) and then dividing by a weighting factor (WF). WF was defined as the number of classes (six) minus the number of classes between the non-verification class and the verification class as expressed by:

$$WF = (N - D) \tag{2}$$

Where N is the number of POPA forecast classes (six) and D is the "distance" (or difference) between the POPA non-verifying forecast class and the verifying POPA forecast class (recall Figure 1). The expression for scoring the non-verification classes for POPA is then:

$$E_{popa} = E_{pop(v)} + \sum_{n=1}^{5} \{E_{pop(nv)}/WF\} \tag{3}$$

$$E_{popa} = E_{pop(v)} + \sum_{n=1}^{5} \{E_{pop(nv)}/WF\} \tag{3}$$

where $E_{pop(v)}$ is defined by equation (1), and WF by equation (2). This scoring procedure was designed to reward those whose forecasts came nearest to the verification class.

For example, the scoring of POPA is broken down into two parts below. The error of the verification class and the error associated with the non-verification classes are scored separately and summed. The verification class is scored according to equation (1). The non-verification classes are scored in the same manner and then adjusted by the weighting factor defined in equation (2).

Three forecasters (A, B, C) make the following POPA predictions with verification in the "0 – T" POPA Class:

POPA Class	Forecaster A	Forecaster B	Forecaster C
0 – T	0	0	80
0.01 – 0.10	80	0	20
0.11 – 0.25	20	0	0
0.26 – 0.50	0	0	0
0.51 – 1.00	0	80	0
> 1.00	0	20	0
Error points based on verification category "0 – T"	100	100	4
Error points accumulated in non-verification categories	13.8	36	0.8
Total Error Points	113.8	136	4.8

When the first POPA class verifies, forecaster error scores of 100, 100, and 4 points, respectively, are obtained using equation (1). Error points for each non-verification class are weighted with respect to the distance between predicted and observed conditions and give 13.8, 36, and 0.8 additional error points, respectively, to the forecasters. Therefore, total POPA error for forecasters A, B, and C are 113.8, 136, and 4.8 respectively. It is clear that without the scoring of the non-verification classes or the weighting factor no differentiation would be made between forecasters A and B despite A's superior forecast (as compared to B).

References

Bosart, L. F., 1975: SUNYA experimental results in forecasting daily temperature and precipitation. *Mon. Wea. Rev.*, 103, 1013–1020.

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

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

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, 4, 401–412.

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

DeGaetano, A. T., and M. D. Shulman, 1984: An evaluation of the New York City northern New Jersey urban heat island effect. *Nat. Wea. Dig.*, 3, 27–30.

Gedzelman, D., 1978: Forecasting skill of beginners. *Bull. Amer. Meteor. Soc.*, 59, 1305–1309.

Glahn, H. R., and D. L. Jorgensen, 1970: Climatological aspects of the Brier p-score. *Mon. Wea. Rev.*, 98, 136–141.

Gordon, N. D., 1982: Evaluating the skill of a categorical forecast. *Mon. Wea. Rev.*, 110, 657–661.

Gulezian, D. P., 1981: A new verification score for public forecasts. *Mon. Wea. Rev.*, 109, 313–323.

Gyakum, J. R., and K. J. Samuels, 1987: An evaluation of quantitative and probability-of-precipitation forecasts during the 1984–85 warm and cold seasons. *Wea. Forecasting*, 2, 158–168.

Murphy, A. H., 1986: A new decomposition of the Brier score: formulation and interpretation. *Mon. Wea. Rev.*, 114, 2671–2673.

Murphy, A. H., and E. S. Epstein, 1967: A note on probability forecasts and "hedging." *J. Appl. Meteor.*, 6, 1002–1004.

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

Murphy, A. H., 1989: Comments on some aspects and impacts of forecast verification. *Nat. Wea. Dig.*, 14, 41–42.

O'Reilly, L. A., R. E. Sharples, and M. D. Shulman, 1988: A statistical evaluation of the New York City-northern New Jersey urban heat island effect during spring and autumn. *Nat. Wea. Dig.*, 13, 29–33.

Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, 2, 191–201.

Sanders, F., 1967: The verification of probability forecasts. *J. Appl. Meteor.*, 6, 756–761.

Sanders, F., 1973: Skill in forecasting daily temperature and precipitation: some experimental results. *Bull. Amer. Meteor. Soc.*, 54, 1171–1179.

Sanders, F., 1979: Trends in skill of daily forecasts of temperature and precipitation, 1966–78. *Bull. Amer. Meteor. Soc.*, 60, 763–769.

Sanders, F., 1986: Trends in skill of Boston forecasts made at MIT, 1966–84. *Bull. Amer. Meteor. Soc.*, 67, 170–176.

Snellman, L. W., 1977: Operational forecasting using automated guidance. *Bull. Amer. Meteor. Soc.*, 58, 1036–1044.

Vislocky, R. L., and G. S. Young, 1988: Improving your weather forecasts through a better knowledge of skill scores. *Nat. Wea. Dig.*, 13, 15–17.

Walts, D. S., and L. O. Pochop, 1977: Operational objective temperature forecasts at non-MOS stations. *Mon. Wea. Rev.*, 105, 3–8.