skill in forecasting daily temperature and precipitation: some experimental results

Frederick Sanders
Massachusetts Institute of Technology
Cambridge, Mass. 02139

Abstract

Estimates of skill in prediction of daily temperature and precipitation are obtained from a six-year record of real-time forecasts for Boston made in the Department of Meteorology at Massachusetts Institute of Technology. No secular increase of skill is found during the period 1966–1972, despite continuous improvement of predicted synoptic-scale flow patterns at the surface and at 500 mb, which were used for guidance. Skill, defined as incremental accuracy of the forecasts over forecasts of the climatological mean, is near 50% for the first day. The decrease of skill with increasing range is more rapid for precipitation than for temperature, the 10% level being reached in two and one-half days and four days, respectively. A pronounced seasonal variation of skill, with a summer minimum, is attributed to a similar variation in the importance of synoptic-scale, as opposed to mesoscale, sources of weather variability. The latter sources are seen as primarily responsible for the present limit of first-day skill. Bias in the probability forecasts of precipitation is relatively small. For the first day, unbiased forecasts representing near certainty are often made, while for the fourth day probability statements representing departures of more than 10 or 15% from the climatological probability are not reliable. Comparison of our results with trends in skill of temperature and precipitation forecasts made at the National Meteorological Center confirms that an increase in the skill of synoptic circulation prognoses is no guarantee of increased skill in predicting the weather.

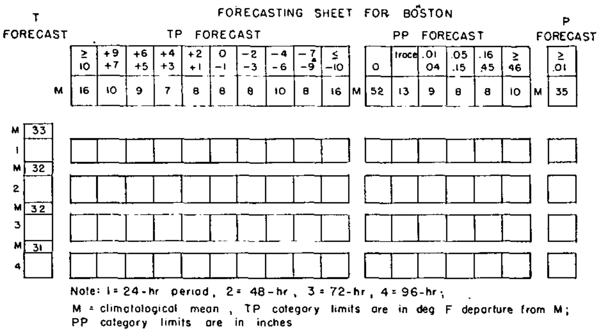
1. Introduction

For many years current forecasts have been made on weekdays in the Department of Meteorology, Massachusetts Institute of Technology, with the primary purpose of providing instruction in the ways of the real atmosphere. During the past six years these forecasts have been made in exactly the same format and have been systematically verified and evaluated in exactly the same way. It seems appropriate, therefore, to examine these forecasts to assess secular and seasonal trends in skill, differences in skill in temperature and precipitation forecasting, the rate of loss of skill with increasing forecast range, and the level of skill itself.

2. Description of the forecasts

The forecasts refer to the observation site at Logan International Airport, Boston, and to each of the next four consecutive 24-hr periods, beginning at 1800 GMT of the current day. The first such period is called the 24-hr forecast and the last the 96-hr forecast. For each period there are four forecasts: the minimum temperature in °F (the T forecast), its probability distribution about the climatic mean for the calendar date (the TP forecast), the probability that there will be at least 0.01 in of precipitation (the P forecast), and the probability distribution over six categories of precipitation amount (the PP forecast). Over the year one of these categories (namely, no precipitation), occurs about half the time. The limits for the other five are chosen so that each occurs about 10% of the time. In the TP forecast there are 10 categories of departure from the climatic mean, with limits chosen so that each category occurs about 10% of the time over the year as a whole. A sample forecast form is shown in Fig. 1.

Climatological mean values are used for guidance and also serve as a control forecast. A smoothed annual march of minimum temperatures was derived from a sample of observations from 1949 to 1966 yielding a value, to the nearest whole °F, for each calendar date. The same 18-year record was used to derive, for each month, percentage frequencies of occurrence of the various categories used in the TP, P, and PP forecasts. These



FORECASTING SHEET FOR BOSTON

	T FORECAST		TP FORECAST											PP FORECAST							P FORECAST
			≥ 10	+9 +7	+6 +5	+4 +3	+2 +1	0 −1	−2 −3	−4 −6	−7 −9	≤ −10		0	trace	.01 04	.05 .15	.16 .45	≥ 46		≥ .01
		M	16	10	9	7	8	8	8	10	8	16	M	52	13	9	8	8	10	M	35
M	33																				
1																					
M	32																				
2																					
M	32																				
3																					
M	31																				
4																					

Note: 1 = 24-hr period, 2 = 48-hr, 3 = 72-hr, 4 = 96-hr;
M = climatological mean, TP category limits are in deg F departure from M;
PP category limits are in inches

FIG. 1. A sample forecast form, with climatological mean values for a prediction made on 2 December.

frequencies are given to the nearest 1% and are regarded as climatological mean probabilities.

Synoptic guidance is obtained from the National Facsimile Circuit and from Service A and C teletype circuits. Further, we derived an approximate annual march of 500-mb height, thickness from 1000 mb to 500 mb, and 850-mb temperature from climatological mean maps, so that we might evaluate forecast values in terms of departure from the mean.

The forecasts each day must be completed no later than 1330 LST. When Boston is on Eastern Standard Time, this completion time is 30 minutes after the beginning of the first forecast period, but we feel that "gifts" during this half hour are quite rare and that the overlap has little statistical effect on the measured forecast skill. Forecasts in which, because of careless preparation, the sum of the probabilities in the TP or PP forecasts is not 100% are discarded. There is no requirement, however, that the P forecast be consistent with the PP forecast or that the forecast minimum temperature lie at the middle of the TP probability distribution, but the forecasters are aware of the advantages of such consistency and almost always respond reasonably. Most forecasters tailor their TP distribution to their expected error in forecasting the minimum temperature, using past experience as a guide.

The temperature forecasts are in °F. The probability forecasts may be given to the nearest 1%, but many forecasters choose to work in increments of 5%, feeling that their ability to make fine discriminations is counterbalanced by their tendency to make arithmetic mistakes.

3. The forecasters

The forecast activity is open to anyone, though formal academic credit is given only in special circumstances; and any individual may make any or all of the 16 forecasts (T, TP, P, and PP forecasts for each of the 4 periods) on any day. The forecasters are mainly graduate and undergraduate students of this and other MIT departments but include also departmental faculty and professional and technical staff. Experience level ranges from 47 years for one of my esteemed colleagues to zero for a number of freshmen. Given the available guidance and a short period of intensive instruction this latter group is able to participate quite successfully. The number of forecasters on a given day has been as low as two and as high as about 30, but is typically about a dozen. Many decline to make the 72-hr and 96-hr forecasts, probably because of the diminishing amounts of explicit guidance and demonstrable skill. Most forecast deliberation occurs between noon and 1330 LST, but the facsimile and teletype data flow is available for study at any time.

4. Verification and scoring of the forecasts

Verification is based upon the 1800 GMT synoptic observation at Logan Airport. Scoring is cumulative over each of three contiguous periods during the year, starting on the academic registration day of each term. Thus the "fall" forecast term begins about the second week of September, the "spring" term about the first week in February, and the "summer" term about the second week in June. The average number of forecast days for these terms was 92, 89, and 60, respectively, over the six-year period.

For the T forecasts the score for a single forecast is simply the absolute error. For the P forecasts the Brier score (1950) is used. The Ranked Probability Score (RPS) of Murphy (1971) recently replaced the Brier score for the TP and PP forecasts, since it alleviates short-term inequities that can arise when the latter score is applied to these multiple-category forecasts. The RPS was applied after the fact to all forecasts made before the change, and forms the basis of the results to be shown here.

Each day, a consensus forecast is made for each forecast in which at least two individuals participate. The consensus is simply the average of the forecasts of the participating individuals, to the nearest 1F and 1% of probability. For each individual, the difference between the forecaster's score and the consensus score is accumulated during a forecast term for only those days in which the individual contributes to the consensus score and is the basis of the ranking of the forecasters which appears weekly, along with other data concerning the forecasts. Thus a forecaster cannot appear more skillful than his peers simply by electing to forecast only on days when the forecast is relatively certain or only on days when the climatological mean forecast looks poor. Over the six-year period there was little secular trend in the average number of forecasters comprising consensus on an individual day.

The climatological mean forecast is also similarly compared with the consensus forecast and is ranked along with the individual forecasters. This control forecast, however, does not enter into the calculation of the consensus forecast. Our results will refer mainly to the percentage improvement of the consensus forecast over the climatological mean forecast.

The skill of a set of forecasts can be defined as the incremental accuracy of the forecasts over and above what can be obtained from the best of available simple control forecasts. (In the context of probability forecasting we will consider "accuracy" to be equivalent to "lowness" of Brier Score or RPS.) One might then ask whether it is fair to use the climatological mean forecast as a control. In view of the persistence of temperature and precipitation regimes, some combination of the initial value and the climatological mean (or a persistence expectancy in the case of the TP, P, and PP forecasts) might be a better control forecast than the climatological mean itself. From a limited amount of experimentation we found that persistence expectancies yielded forecasts which were about 5% better than climatological mean forecasts the first day ahead but which gained little on the fourth day ahead. We shall nevertheless refer to improvement over the climatological mean forecast as "skill," and the reader is invited to make appropriate corrections if he wishes.

5. Results

The main emphasis will be on the performance of the consensus forecasts, since no individual had a complete record of forecasts. It should be borne in mind, however, that few if any individuals who made a substantial number of forecasts outperformed consensus on the average, thus confirming earlier experience reported by this author (Sanders, 1963).

Time series of consensus improvement over the climatological mean forecasts are given in Fig. 2. A striking result is lack of a secular trend. This lack, which is apparent to the eye, is further detailed in Table 1, in which the forecast performance in the earlier three years is compared with the gains in the later three years. In most instances, particularly in the precipitation forecasts, performance deteriorates in the later period, although the term-to-term variability is so large that limited significance can be attached to the downward trend. The lack of trend is especially surprising in view of the improvement and proliferation of the guidance material provided by the National Meteorological Center (NMC). The experience level of the group probably declined during the six-year period; in the 1967 fall term, for example, the forecaster group comprised 14 graduate students, 2 undergraduates, 3 staff and 1 faculty member, while in the 1971 fall term the group was composed of 17 graduates, 15 undergraduates, 2 staff and 2 faculty members. It might be argued that the consensus performance, while remaining more or less steady in an absolute sense, actually deteriorated relative to a rising state of the art, due to inexperience. If this were true, one might expect that the performance of the author would rise relative to consensus during the six-year period. Such, however, appears not to be the case, as can be seen from Fig. 3. There is considerable term-to-term variability in the author's performance relative to consensus, but no distinct trend.

TABLE 1. Percentage improvement of consensus over climatology. First half *vs* second half of period.

Forecast type	Mean Fall '67–Summer '69	Mean Fall '69–Summer '72	Change
24-hr T	54.80	50.09	−4.71
48-hr T	31.18	35.44	+4.26
72-hr T	19.37	20.27	+0.90
96-hr T	10.28	9.50	−0.78
24-hr TP	56.00	53.03	−2.97
48-hr TP	32.75	35.89	+3.14
72-hr TP	20.95	20.50	−0.45
96-hr TP	10.43	9.31	−1.12
24-hr P	49.59	48.59	−1.00
48-hr P	23.80	19.66	−4.14
72-hr P	7.33	6.09	−1.24
96-hr P	2.37	0.64	−1.73
24-hr PP	46.00	47.60	+1.60
48-hr PP	21.40	16.43	−4.97
72-hr PP	6.21	3.75	−2.46
96-hr PP	0.91	0.17	−0.74

The time series in Fig. 2 and the data in Table 2 show

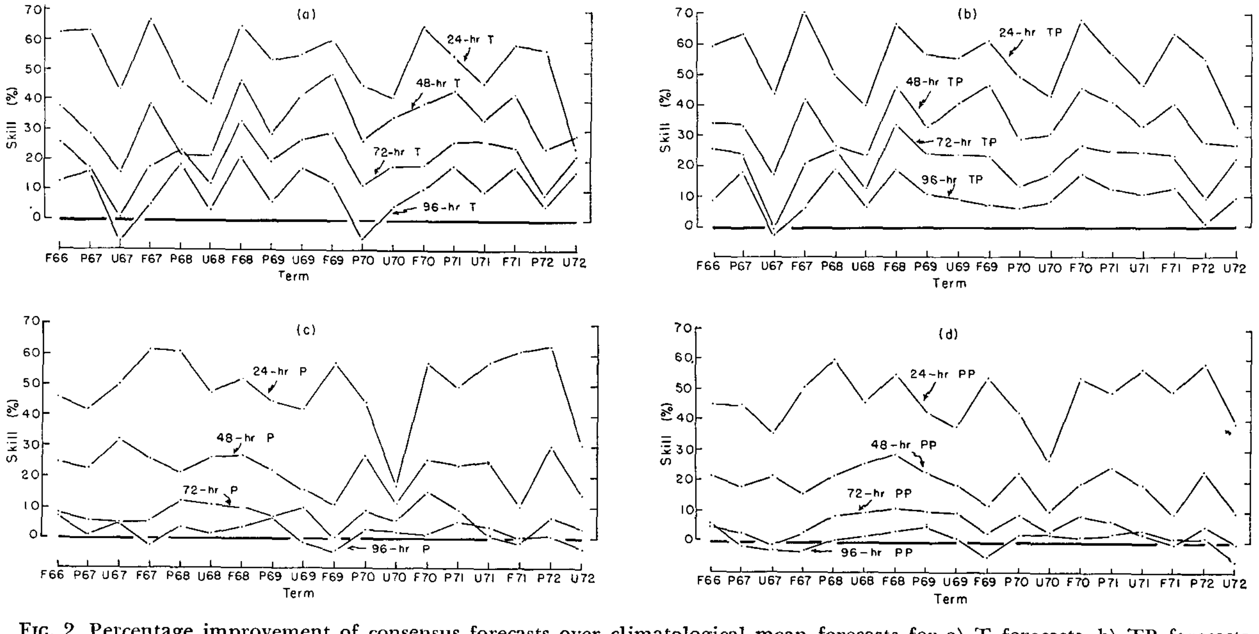


FIG. 2. Percentage improvement of consensus forecasts over climatological mean forecasts for a) T forecasts, b) TP forecasts, c) P forecasts, and d) PP forecasts. The two digits denote the year, F indicates the fall term, P the spring term, and U the summer term.

a distinct seasonal trend. Skill in forecasts of both temperature and precipitation is least in the summer term, while skill in temperature forecasting reaches a pronounced peak in the fall term.

So far as temperature forecasting is concerned, it is reasonable to suppose that variations in skill might be associated with the variability of temperature about the climatological mean and with its average algebraic departure from the mean during the forecast term. From Fig. 4 it is apparent that high skill in the fall term and low skill in summer are indeed associated, respectively, with generally high and low root-mean-square deviations of daily temperature values from their means during these terms. In Fig. 4, there is a similar association of skill with magnitude of deviation of term-mean temperature from the climatological mean. Within a given term, however, there appears to be no systematic relationship between skill and temperature variability. It is my impression that temperature variability on the predictable synoptic scale is large and dominant in the fall term, while variability in summer is small but strongly influenced by subtle differences in maritime versus continental trajectory, which are difficult to predict.

In the case of precipitation forecasting, the low level of skill in the summer term (cf. Table 2) is probably attributable to the convective character of warm-season rainfall. The number of individuals comprising consensus was smaller in summer than in the other terms, but also more experienced. Thus it does not seem that the lower skill was due to a smaller capability of the summer forecast group.

To determine whether amount or frequency of precipitation was related to skill in the P and PP forecasts, we tabulated the deviation of total precipitation at Boston from the long-term average, and the number of days with measurable precipitation, for each month during the six-year forecast period. The months October through January were associated with the fall forecast term, February through June with the spring term, and July through September with the summer term. Little relationship appeared. For example, the average four-period P-forecast skill for the 9 terms of most frequent precipitation was 19.97%, while the average for the 9 terms of least frequent precipitation was 19.54%. For the 9 wettest and 9 driest terms the average skills were 19.96% and 19.56%, respectively.

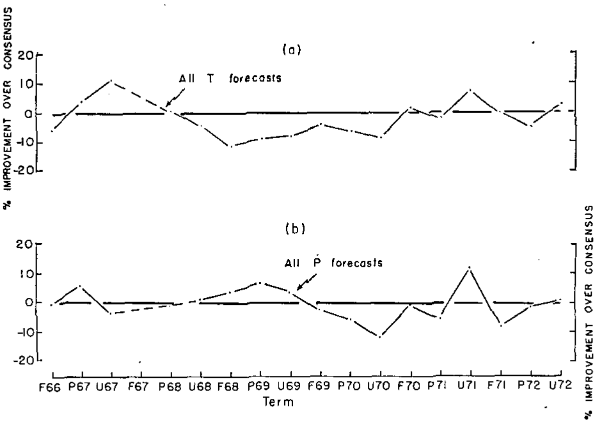


FIG. 3. Percentage improvement of author's forecast over consensus; a) T forecasts, averaged over four periods, b) P forecasts averaged over four periods. Term designation as in Fig. 2.

Note in Table 2 that the skill in the PP forecasts is slightly but systematically smaller than the skill in the P game. When the RPS is used to verify the former, the scoring process is equivalent to applying the Brier score to five dichotomous forecasts (one across each internal category boundary), and then averaging the results, as pointed out by Muench.[1] The threshold for the P forecasts, 0.01 inch, is one of the internal category boundaries. Therefore, it is clear that the skill in forecasting this boundary is larger than in forecasting the average of the others. This result is at least partly attributable to the emphasis given in the guidance information to the 0.01 in threshold, and is consistent with the author's observation (Sanders, 1963) that greater skill was shown in forecasting events which happen about half the time than in forecasting rare or very frequent events.

To examine what relationship might exist between our forecast skill and the skill of the NMC 500-mb prognostic charts (on which we rely heavily), we examined monthly mean NMC verification statistics for the North American area. The average of the correlation coefficients between numerically predicted and observed height changes for the 24-hr, 48-hr, and 72-hr forecasts

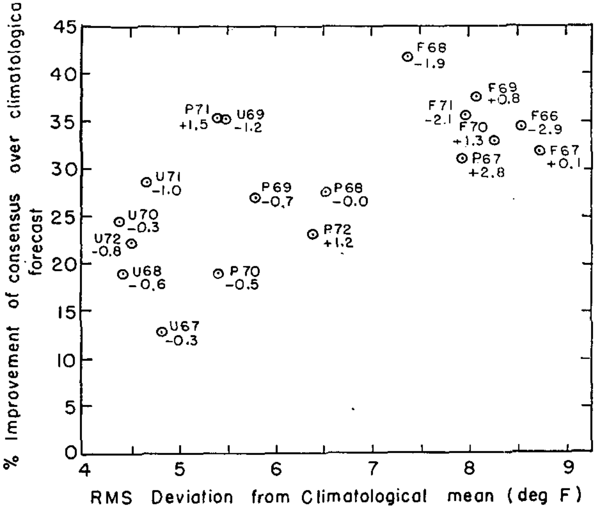


FIG. 4. Percentage improvement of consensus T forecasts over climatological mean forecasts versus root-mean-square deviation of temperature from the climatological mean, for individual terms. Term designation as in Fig. 2. The algebraic departure of term-mean temperature from the climatological mean is indicated next to data point for each term.

[1] H. S. Muench. Personal communication.

was obtained for each forecast term (with the association between month and forecast term as above). The result is shown as a time series in Fig. 5, along with the corresponding times series of our forecast skill, averaged for each term over the four forecast periods. In the case of temperature prediction there is no relationship between our skill in the T and TP forecasts and the NMC 500-mb prognostic skill. On the other hand, a weak correlation is suggested between NMC skill and our skill in the precipitation forecasts. There is a seasonal variation in both of these latter skills but there is a slow upward secular trend in NMC skill, opposite to that noted earlier in precipitation forecasting skill. Despite this opposition, close examination of the data shows a direct relation between the two skills within a given term, except in summer. The interpretation of these results is by no means clear.

The time series in Fig. 2 suggest a roughly exponential loss of skill with increasing forecast range, which is confirmed by Fig. 6. For all types of forecast, the percentage improvement of consensus over the climatological forecast is near 50 for the first period. The rate of loss of skill as range increases, however, is decidedly smaller in temperature predictions than in precipitation forecasting. Thus in the temperature forecasts about 57% of each day's skill remains a day later, while in the precipitation forecasts only about 37% is retained. Or, put another way, skill drops to 10% four days ahead in temperature forecasting and only two to three days ahead in precipitation forecasting.

This difference in behavior is doubtless attributable to the difference in characteristic spatial and temporal scales of precipitation and temperature anomaly. The latter, being more extensive and persistent, are predictable at greater range.

In Fig. 6 all the data sets deviate slightly from the exponential line in the same way. That is, the percentage loss of skill is larger in the later pair of days than in the earlier pair. Exotic interpretations of this result could be made, but none seems compelling. The effect may be due simply to the nearly total lack of NMC guidance for the third and fourth days.

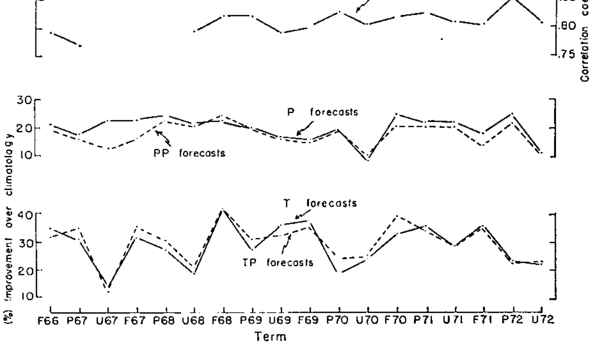


Fig. 5. Consensus skill versus correlation coefficient between predicted and observed 500-mb height changes for NMC numerical predictions. See text. Term designation as in Fig. 2.

Table 2. Percentage improvement of consensus over climatology. By forecast terms.

Forecast type	Mean Fall term	Mean Spring term	Mean Summer term
24-hr T	62.7	53.3	41.1
48-hr T	42.0	28.7	29.2
72-hr T	24.4	17.5	17.5
96-hr T	13.0	9.3	7.0
24-hr TP	64.8	55.2	43.4
48-hr TP	42.4	31.8	28.8
72-hr TP	25.7	19.9	13.8
96-hr TP	11.5	11.2	6.9
24-hr P	55.8	50.7	40.7
48-hr P	20.4	24.2	20.6
72-hr P	6.7	7.6	5.8
96-hr P	0.5	3.0	1.0
24-hr PP	51.2	49.4	39.9
48-hr PP	17.4	21.9	17.4
72-hr PP	4.5	6.8	3.7
96-hr PP	0.4	1.6	−0.4

The mean algebraic errors (bias) and the root-mean-square errors of the T forecasts are set out in Table 3, since they are interesting in themselves. The biases are almost exclusively negative. The T forecasts averaged about a quarter of a degree colder than the climatological mean while the average verifying temperature was warmer than the mean by about the same amount. The rms values were known to the forecasters as they evolved, and have been used by the forecasters as guidance in preparing their TP forecasts.

6. Evaluation of the probability forecasts

The consensus probability forecasts, and the author's, were evaluated according to the scheme proposed by the author (1963). Results for the P forecasts are shown in Fig. 7. In the evaluation process the forecasts were stratified according to the departure from climatological probability that each represented. Five percent departure categories were selected, centered on multiples of 5% departure from the climatological probability. The sorting gain and the bias penalty, expressed as percentages of the Brier score for the climatological forecasts,

Table 3. Bias and root-mean-square error of the temperature forecast (°F).

Fall		Spring		Summer	
Mean error	RMS error	Mean error	RMS error	Mean error	RMS error
24-hr −0.52	3.29	−0.52	3.03	−0.50	2.70
48-hr −0.73	4.87	−0.79	4.45	−0.61	3.37
72-hr −0.42	6.13	−0.14	5.28	−0.38	4.06
96-hr −0.54	7.00	+0.06	5.50	−0.32	4.61

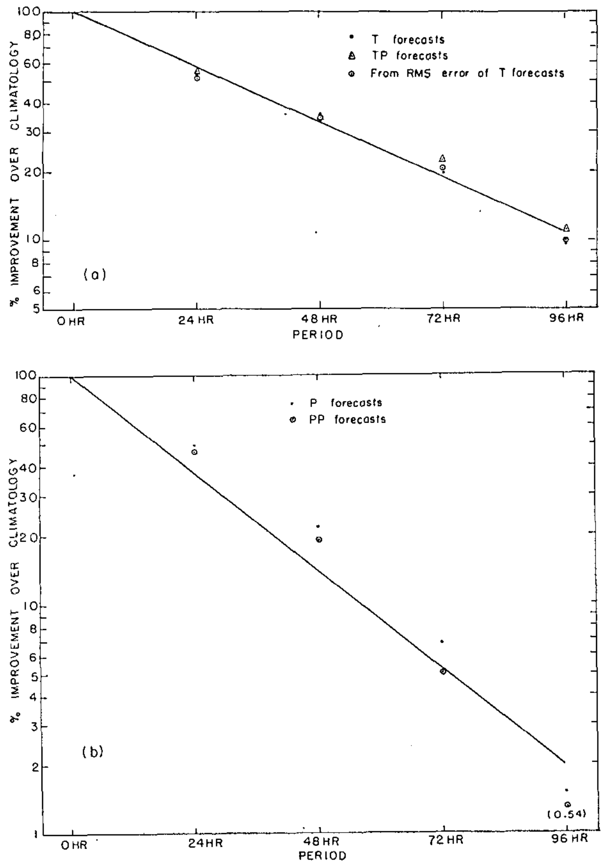


FIG. 6. Skill versus range of forecast: a) for temperature forecasts, and b) for precipitation forecasts.

were then calculated from the author's (1963) Eq. (9). That is, for each departure category the sorting gain was computed from the departure of the observed relative frequency of occurrence of precipitation from the climatological frequency. The overall sorting gain was taken as an average weighted by the number of times each departure category was used. The overall bias penalty was obtained from a comparison of the observed and predicted relative frequencies for each departure category, as a similarly weighted average.

The results of the evaluation show that the bias penalty is much smaller than the sorting gain, confirming earlier experience (Sanders, 1963). The loss of skill with increasing range is due to the decreasing ability to distinguish classes of days on which occurrence or nonoccurrence of precipitation can be predicted with near certainty, and the skill would not be significantly advanced if bias were entirely absent. Note that in the 24-hr forecasts a full range of deviations from 70% (i.e., 100% probability of precipitation in summer months) to —40% (i.e., zero percent probability in late winter and early spring) is used. At 48 hr the range of deviations is somewhat reduced and forecasts near the climatological probability occur more often. At 72 hr those few forecasts that stray more than about 25% from the climatological probability prove to be generally ill-advised while at 96 hr a departure of 10 or 15% seems to be a practical limit.

Small as the bias penalty is, it does show some systematic characteristics. The weighted bias penalty for each forecast-departure category can be attributed to overprediction or underprediction of precipitation, depending on whether the forecast probability is larger or smaller than the relative frequency of occurrence of precipitation. If this difference has the same sign as the forecast departure from the climatological probability, the weighted bias penalty can be attributed to overconfidence; if the sign is opposite, the weighted bias penalty can be attributed to underconfidence. Bias penalty incurred for the forecast departure category centered on zero is discarded. The excess of overprediction penalty over underprediction penalty can be taken as a measure of systematic overprediction in the forecasts as a whole. Similarly, the excess of overconfidence penalty over underconfidence penalty is regarded as a measure of the systematic overconfidence in the forecasts. If the bias is corrected for these systematic errors, a random bias remains.

These measures, expressed as percentages of the Brier score for the climatological forecasts (and therefore in the terms by which we measure skill in our forecasts), are given in Table 4. The 24-hr forecasts, for both consensus and the author, suffer from tendencies to overpredict precipitation and to be underconfident. The tendency to overpredict extends to the 48-hr period. Underconfidence changes to slight overconfidence, on the whole, beyond the first 24 hr. I suspect that the immediacy of verification exerts a psychological pressure in the preparation of the 24-hr forecasts, and that forecasters tend to avoid being caught with their probabilities down, so to speak. This pressure is absent for forecasts beyond the first day, for who will remember?

It is interesting that the systematic biases are nearly the same for consensus and for the author, except for the latter's 72-hr forecasts in which the large values stem mainly from 4 forecasts in the —35% departure category and 23 in the +25 category which fared particu-

TABLE 4. Systematic sources of bias penalty, expressed as percentages of Brier score for climatological forecasts. Positive values indicate overprediction or overconfidence.

		(Over/under) prediction	(Over/under) confidence
Consensus	24-hr P	(+) 1.3	(—) 0.8
	48-hr P	(+) 1.1	(+) 0.5
	72-hr P	(—) 0.0	(—) 0.2
	96-hr P	(—) 0.2	(+) 0.3
Sanders	24-hr P	(+) 1.5	(—) 0.6
	48-hr P	(+) 0.9	(+) 0.4
	72-hr P	(—) 1.3	(+) 1.2
	96-hr P	(—) 0.2	(+) 0.4

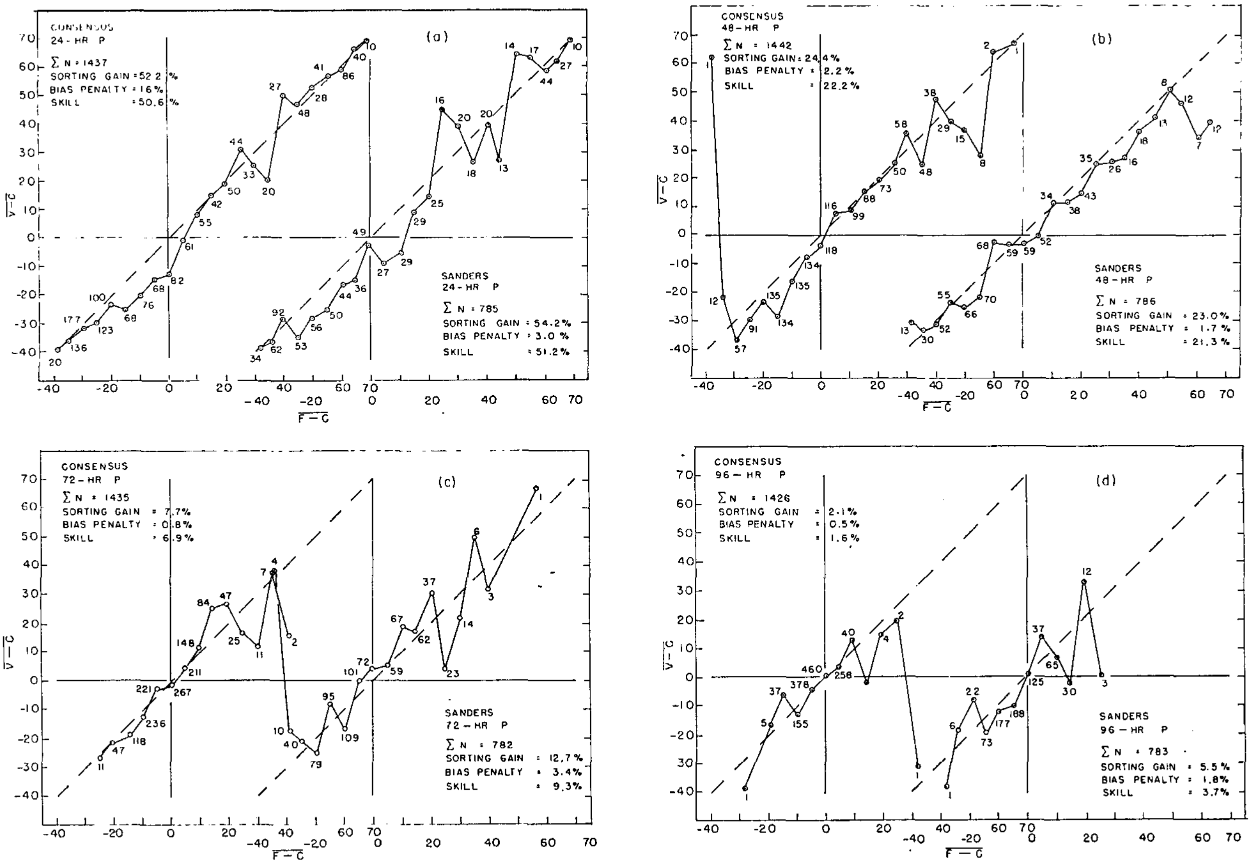


FIG. 7. Evaluation of consensus and author's P forecasts: a) 24-hr period, b) 48-hr period, c) 72-hr period, and d) 96-hr period. The abscissa, $\overline{F–C}$, is the forecast departure from the climatological percentage frequency of occurrence. The ordinate, $\overline{V–C}$, is the corresponding departure of the observed percentage frequency from the climatological frequency. Perfectly unbiased forecases would lie along the dashed diagonal. The number adjacent to each data point is N, the number of times the departure category was forecast.

larly poorly. Since the bias penalty is larger in the author's forecasts, the random bias, (which is presumably a function of sample size) is larger. Note that the author's sample contains only slightly more than half the number of forecasts in the consensus sample. As the random bias penalty is reduced in ever larger samples, the sorting gain must also be reduced; which is to say that some of it is "undeserved," since the net gain over the climatological forecast would not be expected to increase as the sample size becomes larger. This characteristic is a weakness of the present method of evaluating probability forecasts, but perhaps not a serious one since the biases at worst are slight.

7. Concluding discussion

We have set out a record of our skill in predicting daily minimum temperature and precipitation amount, two quantities that are of continuing interest to the public. We believe that this record is representative of the state of the art for a geographical location which is strongly influenced by the synoptic scales of variability. Our verification methods reward unbiased forecasts; and our measures of skill, for both probabilistic and other forecasts, are straightforward and capable of at least some analysis. We urge others to present comparable information, for meteorology is sorely in need of a quantitative basis for appraising present forecast skill, to say nothing of its variation with element and with location, its trend over the years, and its rate of loss with increasing forecast range.

Perhaps the most striking, and sobering, result is the lack of a systematic increase in forecast skill over the last six years. In fact, our skill in precipitation forecasting has shown a slight downward trend, an experience which seems to have been shared by forecasters at NMC. Fig. 8 is constructed from information given by Cooley and Derouin (1972) and from other unpublished NMC material. We see a downward trend in the verification scores for the one-inch isohyet throughout the period 1966–1971, whether for the raw computer guidance or for the human forecast. Scores for the forecasting of measureable precipitation peaked in 1968 and have declined since. Throughout the period the skill of the sea-level pressure prognostic charts has risen

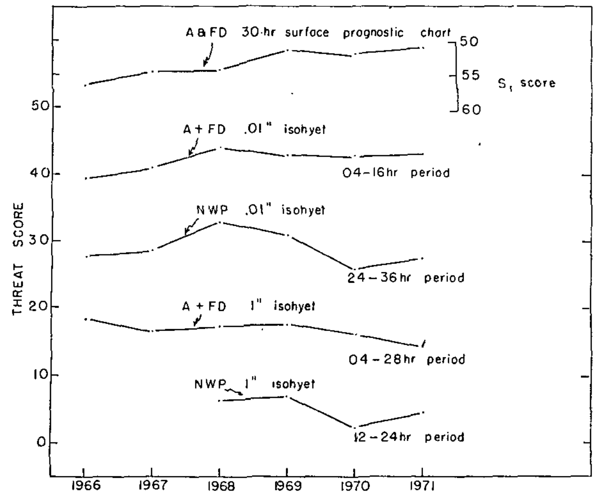


FIG. 8. NMC skills in surface prognosis and in precipitation forecasting, 1966–71. A&FD refers to manually prepared predictions, based on computer guidance. NWP refers to direct computer output. The forecast period is the number of hours beyond the time of the latest available data.

slowly, paralleling the slow improvement in the numerical 500-mb prognoses (cf. Fig. 5). Brown and Fawcett (1972) have recently commented on the lack of secular trend since 1964 in the skill of numerical precipitation forecasts at NMC. Whatever other explanations might be adduced for this state of affairs, we must concede that atmospheric perversity (which we cannot distinguish from chance variability) probably plays a major role, and that the downward trend in skill would not be expected to continue.

In temperature prediction we found increases in skill for some periods and losses in others, with a net trend very close to zero over the six-year period. We have no knowledge of comparable trends of skill in other subjective temperature forecasts, but we have kept a nearly continuous record, for our forecast days, of the behavior of the statistical forecasts of minimum temperature described by Klein and Lewis (1970). These forecasts were extracted from teletype messages available at the time our forecasts were made or shortly thereafter. Though these forecasts refer to the minimum temperature for the calendar day, they were verified as ours were: that is, on the basis of the minimum temperature from 1800 GMT to 1800 GMT. The verifying value in the two periods is usually the same and is rarely much different. The time history of the consensus improvement over this statistical forecast is shown in Fig. 9. It is apparent that the guidance product improved, relative to consensus, until 1970. However, it seems that it never has been accurate enough to induce an improvement in our forecasts, which rely on it in part.

To summarize, it is clear that an increase in the skill of synoptic-scale circulation prognoses is no guarantee of increased skill in forecasting the weather itself.

My interpretation of this paradox is that we are now squeezing virtually all the blood we can from the short-range synoptic turnip. That is, our large errors in the first period do not arise from large errors in the synoptic-scale prediction of motion, and in deep-layer predictions of thickness and relative humidity. Rather, they arise from the occurrence of shallow fronts, from the influence of the urban heat island and the land-sea contrast, from the occurrence of convective showers, from inaccuracy of predicting the time of beginning and end of none-convective rainstorms, and from other members of a virtually nondenumerable set of mesoscale influences. Thus improvement in first-period skill must depend on an improved ability to cope with these influences.

The benefit to be expected from improvement in synoptic-scale prediction beyond the first day would be a reduction in the rate of loss of skill with increasing range, but the data in Table 1 offer little support for the notion that the rate of loss has been decreasing in the past six years. Thus something more than the recently observed rate of increase in synoptic skill, probably something approaching a breakthrough, would be required to have a significant effect on this aspect of prediction. The limiting skill, moreover, would be something smaller than the present 24-hr skill, because the latter reflects some ability to take mesoscale influences into account, and this ability will surely be extremely small or absent altogether after the first day ahead.

Our finding that skill in prediction drops to 10% in two and one-half days for daily precipitation and in four days for daily temperature leaves the impression of a large gap between present practice and theoretical limits, which are asserted to be as long as two weeks. But perhaps the gap is not so large, for the theoretical limits refer to the synoptic-scale predictions of the mass field. Present skill in prediction of, say, 500-mb heights, doubtless deteriorates more slowly than the skills we have measured, but it is difficult to translate this skill into predictions which are meaningful to anyone but meteorologists.

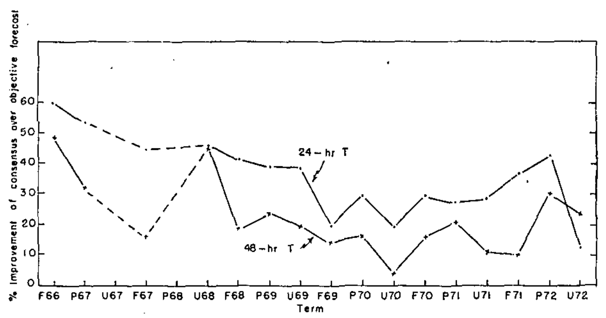


FIG. 9. Improvement of consensus temperature forecasts over objective guidance forecasts. Term designation as in Fig. 2.