Trends in Skill of Daily Forecasts of Temperature and Precipitation, 1966–78

Frederick Sanders
Department of Meteorology
Massachusetts Institute of Technology
Cambridge, Mass. 02139

Abstract

Forecasts of minimum temperature and precipitation amount at Boston have been made and evaluated in the Department of Meteorology, MIT, in essentially the same format since 1966. These forecasts refer to the first through fourth 24 h periods in advance and are partly categorical and partly probabilistic in form. The skill level in the consensus forecasts, relative to forecasts of the long-term mean, is slightly more than 50% for the first day and around 10% on the fourth, except for conditional quantitative precipitation forecasting, which is decidedly less skillful.

Regression analysis shows, except for the first day, slight increases in the skill of these predictions, at a rate of about six-tenths of a percent per year.

The skill of the guidance temperature forecasts of the National Meteorological Center has lagged the skill of the consensus forecasts by a decreasing amount from 1966 to 1972. The lag from 1972 to date has not changed significantly and varies between 10 and 30% of consensus skill, depending on range and season. The guidance for the conditional quantitative precipitation forecast is inferior to both consensus and the long-term median control forecasts.

1. Introduction

The MIT departmental weather forecasting experience from fall 1966 through summer 1972 was described and analyzed by Sanders (1973). Perhaps the most striking finding was the lack of perceptible improvement over the six-year period in skill in the prediction of minimum temperature and total precipitation amount at Boston for one, two, three, and four days ahead.

Now another six years have passed, and we wish to make a further report. Only one major change in the forecast format has occurred: starting in the fall 1974 contest the forecast probability of measurable precipitation—the P forecast as described by Sanders (1973)—was replaced by a conditional forecast of precipitation amount, given that measureable precipitation occurs. If not, this forecast is ignored in the verification and evaluation procedure. The forecast is in terms of category of equivalent water depth, one being 0.01–0.05 in, two being 0.06–0.10 in, ... 20 being 0.95–1.00 in, and so forth. The scoring rule, analogous to that for the minimum temperature, is the absolute magnitude of the difference between forecast and observed categories; and the appropriate control forecast is the median category on a day of measureable precipitation. This value for each month (category three or four) was taken from the most recent available decade of Boston observations.

The change was made to determine skill in prediction of amounts, aside from the considerable skill known to exist in forecasting whether or not measureable precipitation would occur. This threshold value of 0.01 in remained as one of the interior thresholds of the PP forecast, which continued in its original form (the probability distribution over six categories of precipitation amount) (Sanders, 1973).

All forecasts, as before, refer to the minimum temperature and the total precipitation reported by the WSFO, Logan Airport, Boston, during the four sequential 24 h periods following 1800 GMT, the approximate time of submission of the forecasts. The scoring rules, as before, are absolute error in the T (minimum temperature) and P forecasts and the ranked probability score (Epstein, 1969; Murphy, 1971) in the TP and PP probability distributions. The control forecasts are the long-term daily mean for the T forecasts, monthly median for the P forecasts, and monthly mean for the TP and PP forecasts. Error points are accumulated over discrete periods beginning on academic registration days in September (fall), February (spring), and June (summer).

TABLE 1. Regression trends in skill: T forecasts.

$\widehat{\text{Skill}}\ (\%) = a_0 + a_1\ (\text{year} - 1900)$.

r^2 is the fractional variance explained by the regression estimate.

	a_1	r^2	Year ($\widehat{\text{skill}} = 85\%$)	$\overline{\text{Skill}}$
Day 1				
Fall	−0.336	0.25	1903	62.0
Spring	−0.310	0.03	1871	53.6
Summer	+0.355	0.01	2086	44.5
Day 2				
Fall	+0.448	0.19	2071	43.2
Spring	+0.872	0.10	2033	32.2
Summer	+0.972	0.15	2027	31.9
Day 3				
Fall	+0.821	0.24	2042	27.5
Spring	+0.484	0.03	2105	20.8
Summer	+1.605	0.27	2013	20.3
Day 4				
Fall	+0.344	0.06	2179	13.7
Spring	+0.524	0.05	2111	12.3
Summer	−0.004	0.000003	16042BC	9.4

0003-0007/79/070763-07$05.75

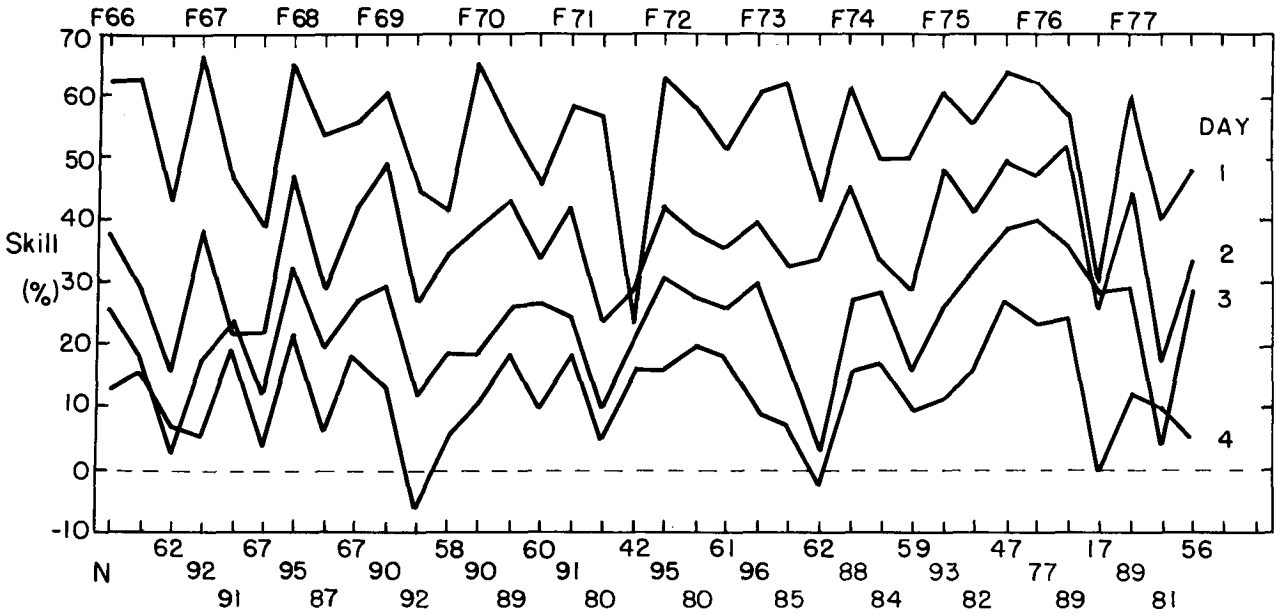

FIG. 1. Skill MIT consensus forecasts of minimum temperature as a function of time, for the first through fourth 24 h periods in advance. Along the bascissa, F66 represents Fall 1966 and the last point is for Summer 1978. N is number of forecasts for the first 24 h period.

2. Trends in skill

Skill is defined as the percentual increment of accuracy of the actual forecasts over that of the control forecasts. A record of this skill in the group-consensus minimum-temperature forecasts appears in Fig. 1. Large seasonal and irregular fluctuations are apparent, but no strong trends appear. Careful examination, however, suggests a slow improvement on days 2, 3, and 4 and a slight deterioration on the first day. The data in Table 1 confirm and elucidate these trends, by individual season. Except for the first day ahead, linear regression indicates improvement at the rate of several tenths of one percent per year. These estimates are precarious, to judge from the smallness of the explained variance, which lacks statistical significance in a number of instances. The consistency of sign of the trends, however, lends credibility to the results. The modest size of the rate of increase is underscored by the extrapolation

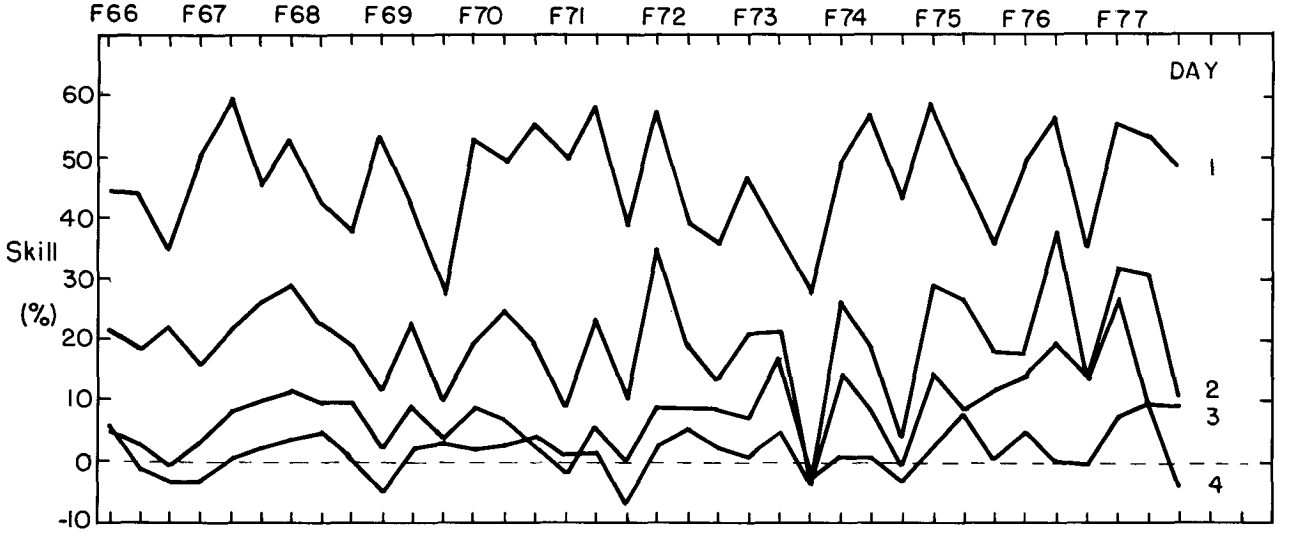

FIG. 2. Same as Fig. 1, for forecast probability of six classes of precipitation amount: none, trace, 0.01–0.04 in, 0.05–0.15 in, 0.16–0.45 in, >0.45 in.

TABLE 2. Regression trends in skill: PP forecasts.
$\widehat{\text{Skill}}\ (\%) = a_0 + a_1\ (\text{year}-1900)$.

	a_1	r^2	Year ($\widehat{\text{skill}} = 85\%$)	$\overline{\text{Skill}}$
Day 1				
Fall	+0.368	0.10	2061	52.0
Spring	+0.399	0.03	2063	48.9
Summer	+0.088	0.001	2496	38.6
Day 2				
Fall	+0.785	0.13	2052	21.9
Spring	+0.932	0.36	2038	23.7
Summer	−1.177	0.28	1912	13.2
Day 3				
Fall	+1.479	0.53	2023	9.3
Spring	+0.690	0.30	2082	9.2
Summer	+0.486	0.09	2137	5.2
Day 4				
Fall	+0.374	0.14	2194	1.6
Spring	+0.483	0.28	2142	3.0
Summer	−0.259	0.09	1641	−0.9

estimate of the year in which a nominal skill of 85%[1] would be achieved; this year falls in the 21st or 22nd century. The loss of skill on the first day may be a chance fluctuation except during the fall period, when the trend explains one-quarter of the variance and seems well established.

The mean skill for the 12-year period also appears in Table 1, for each season and forecast range. Seasonally, it tends to follow the variability of temperature, being largest in the months September through January and smallest during summer. The erosion with range leaves each day with about 60% of the skill of its predecessor.

Skill in the forecast probability distribution of minimum temperature (TP forecast) follows the results shown above closely and is not further illustrated here. The values are nearly the same, the time series are highly correlated, and the trends are closely parallel.

The record of skill in consensus probability forecasts of six classes of precipitation amount is illustrated in Fig. 2. It will be recalled that the classes are, respectively, none, a trace, 0.01–0.04 in, 0.05–0.15 in, 0.16–0.45 in, and >0.45 in. The first occurs about 50% of the time and the others about 10% each; thus the skill in these forecasts depends substantially on the ability to distinguish those days in which no precipitation occurs.

Again, there are no robust trends. Again, close inspection indicates slight improvement, except perhaps on the first day. The data in Table 2 document the improvement beyond the first day except during the summer season, when the predominantly convective precipitation continues to elude us. Regression shows the trend on the first day to be upward, but also to be only weakly established. The extrapolated year of achievement of 85% skill is again in the 21st or 22nd century. The 12-year mean skill is least during the summer and its erosion rate is about 60% per day, except on the fourth day when it disappears almost completely.

Only a limited sample of conditional quantitative precipitation forecasts is available for analysis because, as explained above, this forecast commenced in the fall of 1974 and because it is effective on only approximately one-third of all days, when measureable precipitation occurs. Nevertheless, the time series of skill in these forecasts is shown in Fig. 3 and the regression data are given in Table 3. The results for summer 1977 were not used because measureable precipitation occurred on only three of the days on which forecasts were made. The trends are mainly upward, but the extreme irregularity precludes any confident conclusion save that the improvement has not been rapid, being about the same as in the other forecasts. The general level of skill, moreover, is low, being greater than 25% only on the first day in the fall and spring seasons, and being less than 10% at all ranges during the summer and in all seasons on the fourth day ahead. There is clearly extra room for improvement in this important forecast.

3. Comparison with guidance forecasts

In the preparation of our forecasts we refer to the most recent guidance information available on the National Facsimile Circuit and on Service C teletypewriter. This information, aside from prognostic charts issuing from the National Meteorological Center (NMC), includes the statistical minimum-temperature forecast based on

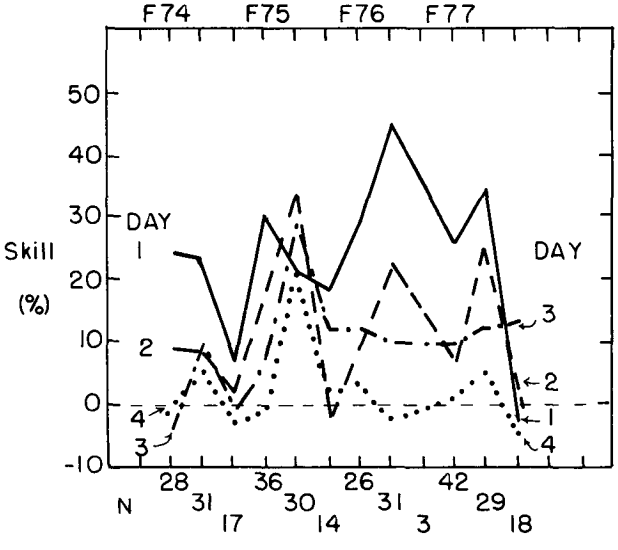

FIG. 3. Same as Fig. 1, but for conditional quantitative precipitation forecasts. Forecasts start in Fall 1974. N is the number of forecasts for the first 24 h period.

[1] Unquestionably an advance on the pervasive value of 83% correct, discussed by Neuberger (1976).

TABLE 3. Regression trends in skill: P forecasts since fall 1974.

$$\widehat{(\text{Skill} - \text{Seasonal mean skill})}\ (\%) = a_0 + a_1\ (\text{year} - 1900).$$

	a_1	r^2	Year $\widehat{(\text{skill} = 85\%)}$	$\widehat{\text{Skill}}$	Seasonal Mean Skill Fall	Spring	Summer
Day 1	+0.783	0.02	2055	22.8	26.8	30.6	6.9
Day 2	+0.425	0.01	2149	11.4	10.3	21.8	−1.1
Day 3	+2.502	0.18	2006	9.6	5.9	14.8	7.7
Day 4	−0.526	0.01	1818	1.8	−0.2	6.9	−2.4

output from dynamical prediction models, the minimum-temperature forecast prepared as part of the five-day forecast package, and the quantitative precipitation forecast produced by one or another of the dynamical prediction models. Of these, only the five-day forecast contains a subjective element; the others are entirely automated. In light of the widespread concern that the automated prediction may replace much of the judgemental part of the forecaster's job, and because of the recent concern expressed, for example, by Snellman (1977) that the forecaster is relying too heavily on the guidance (thus reinforcing the replacement in a kind of circular process), it seems useful to compare our performance with that of the guidance forecasts.

To avoid the potential objection that the guidance forecast might be based on 0000 GMT data while we have access to the 1200 GMT material, we use in this comparison the statistical temperature forecast received on facsimile and teletypewriter around 2000 GMT, the five-day forecast received around 2200 GMT, and the quantitative precipitation forecast produced from the 1200 GMT initial data. Of these we regularly see only the last prior to submission of our own forecasts, and this only in the recent past since the advance of its time of appearance on teletypewriter. In this respect, then, the guidance forecasts have an advantage. On the other hand, we put the guidance temperature forecasts at a disadvantage by regarding them as a 24 h minimum for the period beginning and ending at 1800 GMT, whereas they refer in fact to the 12 h period, 0000 GMT to 1200 GMT, in which the minimum usually falls. Thus the comparison is not as clear as we might wish, but it is the best we can do.

The record of performance of the guidance minimum-temperature forecast relative to our consensus forecast appears in Fig. 4. The guidance forecasts are statistical ones for the first two days and ones taken from the five-day forecast for the last two days, except recently when a statistical forecast for the third day was used when available. It is apparent that the accuracy of the guidance forecast rises above that of our consensus only occasionally. It is also clear that a rapid improvement in the guidance forecast occurred from the middle 1960s to the early 1970s. There was no parallel response in the skill of the consensus forecasts; the guidance forecast became worth looking at in the early 1970s but nothing especially good happened to our temperature forecasts as a consequence of this attention. An interesting question is whether the guidance forecast has leveled off in its pursuit of our skill, so to speak, or whether it is still gaining on us. The answer depends on the point at which one starts counting; a start in fall 1971 would show continuing advance while a start in summer 1974 would show a retreat. We chose to examine the last half of the record, starting in fall 1972, having little to go on except neatness.

The results of the regression analysis appear in Table 4 along with the mean consensus performance in the first and last halves of this most recent six-year period. It appears that the guidance is no longer overtaking the consensus prediction. The regression rates indicate that the guidance forecast is catching up at

TABLE 4. Regression trends in performance of guidance forecast relative to consensus forecast: T forecasts. Fall 1972–Summer 1978

$$\left(100 \times \frac{\widehat{\text{consensus} - \text{guidance}}}{\text{consensus}}\right)(\%) = a_0 + a_1\ (\text{year} - 1900).$$

	a_1	r^2	Year (consensus − guidance = 0)	$\left(100 \times \frac{\text{consensus} - \text{guidance}}{\text{consensus}}\right)$ Mean F 72–Su 78	Mean F 72–Su 75	Mean F 75–Su 78
Day 1	+0.408	0.002	2048	−29.9	−29.5	−30.3
Day 2	+1.557	0.04	1983	−12.3	−13.1	−11.4
Day 3	−2.133	0.08	1965	−20.9	−15.6	−26.3
Day 4	+0.342	0.01	2004	−10.0	−11.1	−8.8

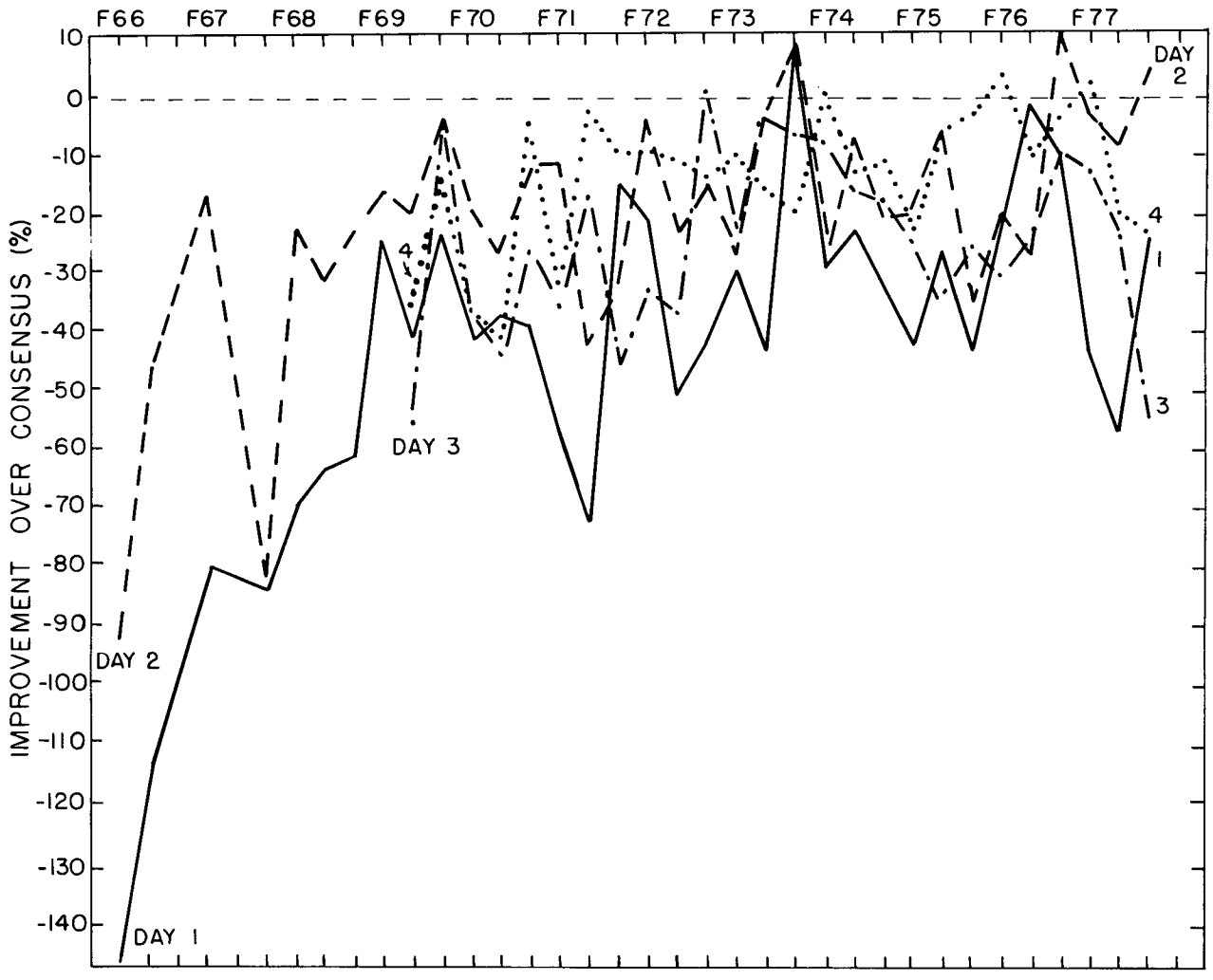

FIG. 4. Performance of guidance minimum-temperature forecasts relative to the consensus forecasts, for the first through fourth period in advance. Format as in Fig. 1. A negative value means that the guidance forecast was less accurate than the consensus forecast.

three of the four ranges, but the sum of the four rates is close to zero and the regression-explained percentage of variance is small. The mean shortfall of guidance below consensus has increased in the most recent three years compared to the preceding three at two of the four ranges, the sum of the changes again being close to zero. Similar results have been reported by Cook and Smith (1977).

The performance of the quantitative precipitation forecast from the NMC dynamical model is shown in Fig. 5 and Table 5. A comparison can be made only for the first day. Here we see great irregularity in the time series, a positive trend to which almost no credence can be given because of the miniscule explained variance, and a level of performance that is far behind the modest skill of the consensus forecast and which in fact loses to the climatological control forecast, on the average. It is disappointing not to see marked improvement, since NMC model changes during the comparison period have been aimed in part at improvement of this important prediction.

4. The state of the art

The sample of forecasts we have discussed is small, and the question naturally arises whether our results represent the state of the art, by which we mean a plateau of high performance reached by fully trained and skilled professionals that can be exceeded at any given time only by the most extraordinary individual, if by any one at all. The question is whether our consensus, comprised in part of the forecasts of students with little or no experience, is as skillful a forecast as present knowledge and resources afford. Exact comparisons are difficult to make, but we believe that our consensus forecasts do not suffer by comparison with those of the National Weather Service. Despite the provocation we hoped our earlier report (Sanders, 1973) would provide, little verification information by others appeared. Bosart (1977) presented results of forecasting, also in the university environment, which are entirely consistent with ours. Pierce (1976) has reported improvement in the NWS forecasts for Boston but the rate of increase is

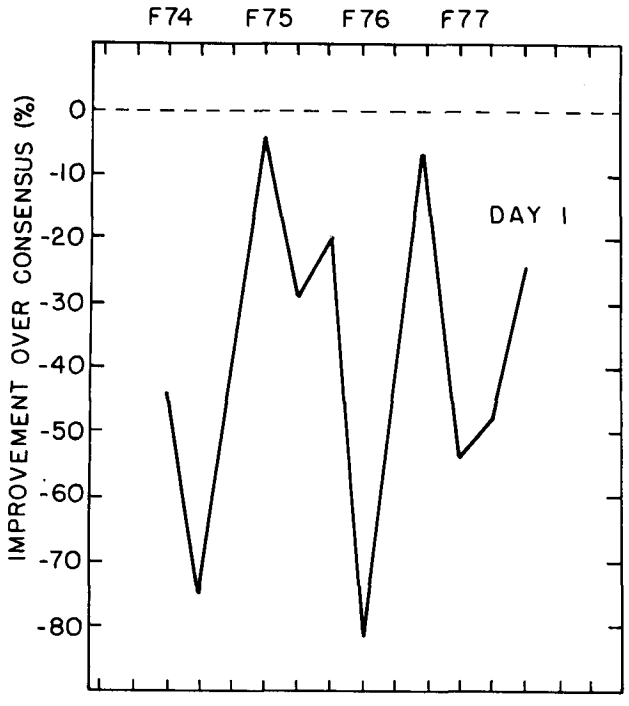

FIG. 5. Same as Fig. 4, but for conditional quantitative precipitation forecasts.

about as small as ours, and comparative levels of skill are difficult to assess. A greater increase in skill in a less accurate forecast would represent a closing of the gap, as in the case of the minimum-temperature guidance forecasts, rather than an advance in the state of the art.

It might be argued that the state of the art has advanced, but that our results fail to show it because the everchanging mix of students who constitute a major part of the consensus are less experienced and skillful. If this were the case the performance of this investigator relative to that of the group should rise with time. The data in Table 6, however, show only tiny trends in which we can have little confidence. The data show further that even this experienced forecaster is outperformed, on the average, by a consensus dominated by his students. The possibility that the state of the art is advancing, that our consensus is deteriorating, and that the present investigator is degenerating, cannot, of course, be ruled out but appears remote.

In short, lacking evidence to the contrary, we believe our results are representative of the state of the art.

5. Concluding discussion

We have found some evidence of improvement in forecast skill over the past 12 years, except in predictions for the first day. This improvement has been occurring at a rate generally less than one percent per year. It is reasonable to attribute this improvement to enhancement of the quality and quantity of guidance information provided directly or indirectly by dynamical prediction models. We say this because it appears that other approaches to forecasting research have been largely stifled during the past 25 years or so. In fact, the improvement in forecasting can be said to have been dearly bought, because the great increase of resources devoted to numerical simulation models has apparently yielded only a slight increase in the skill of forecasting at the level of the consumer. While we cannot be certain, it is likely that much of this skill, at least for the first two days, was present in methodology developed before the advent of numerical prediction (e.g., Riehl *et al.*, 1952). Our failure to find a general increase of skill in the first day probably reflects a relatively small reliance on the numerical models at short range. Perhaps it is time to encourage a kind of neo-naturalism, in which atmospheric behavior is looked at directly rather than through the filter imposed by regarding it solely as a problem of numerical simulation.

We do not suggest abandonment of present numerical capabilities as Ramage (1976) recommends, if for no other reason than that we have become accustomed to it and, more importantly, it is reasonable to conclude that it is responsible for whatever modest skill we may have in daily prediction beyond 48 hours. We simply urge that something else also be encouraged at levels where encouragement can make a real difference.

Our sample is admittedly puny, however, and we may be drawing unwarranted conclusions. Further, taking the pulse of that part of meteorology deemed most im-

TABLE 5. Regression trend in performance of guidance forecast relative to consensus forecast: P forecasts.

Fall 1974–Summer 1978, Day 1 only

$$\left(100 \times \frac{\widehat{\text{consensus}} - \text{guidance}}{\text{consensus}}\right)(\%) = a_0 + a_1\,(\text{year} - 1900).$$

a_1	r^2	Year (consensus − guidance = 0)	100 × (consensus − guidance) ÷ consensus
+0.450	0.0006	2069	−42.0

TABLE 6. Regression trends in performance of FS forecast relative to consensus forecast: T forecasts.

$$\left(100 \times \frac{\widehat{\text{consensus}} - \text{FS}}{\text{consensus}}\right)(\%) = a_0 + a_1\,(\text{year} - 1900).$$

	a_1	r^2	100 × (consensus − FS) ÷ consensus
Day 1	+0.276	0.01	+0.476
Day 2	+0.246	0.01	−4.153
Day 3	−0.025	0.001	−2.300
Day 4	+0.008	0.0002	−2.332

portant to the larger community that sustains us is hardly the job of a single academic department. Surely the American Meteorological Society has an obligation to establish and monitor benchmarks of forecast skill, quantitative enough for the elementary kinds of analysis we have brought to bear. This obligation should not be a captive of the potential political embarrassment that the results might bring. Do chips ever fall exactly where we would like them to; and are we not better off ultimately for letting them fall where they may?

Acknoweledgment. The author is grateful to the AMS Council and Executive Committee of 1967 for encouraging a number of us to consider how the state of the art might be measured and monitored, and for subsequently declining to provide resources for carrying out such a project, thus stiffening the author's resolve to show how cheaply it could be accomplished. He is also indebted to Glenn Brier and Harlan Saylor of NWS for helping him consider how forecast skill might be defined and measured. Finally, this work would have been impossible without the acquiescence over the years of students and others in the Department of Meteorology who have been willing to abjure Arago's Admonition.[2]

[2] "Never, no matter what may be the progress of Science, will honest scientific men who have regard for their reputations venture to predict the weather."

References

Bosart, L. F., 1977: SUNYA experimental results in forecasting daily temperature and precipitation. *Mon. Wea. Rev.*, **103**, 1013–1020.

Cook, D., and D. R. Smith, 1977: Trends in skill of public forecasts at Louisville, Ky. *Bull. Am. Meteorol. Soc.*, **58**, 1045–1049.

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

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

Neuberger, H., 1976: A historical note to "Prognosis for Weather Forecasting." *Bull. Am. Meteorol. Soc.*, **57**, p. 805.

Pierce, C. H., 1976: Are weather forecasts improving? *Weatherwise*, 29, 136–137.

Ramage, C. S., 1976: Prognosis for weather forecasting. *Bull. Am. Meteorol. Soc.*, **57**, 4–10.

Riehl, H., J. Badner, J. G. Hovde, N. E. LaSeur, L. L. Means, W. C. Palmer, M. J. Schroeder, L. W. Snellman, and others, 1952: Forecasting in middle latitudes. *Meteorological Monographs*, Vol. 1, No. 5, AMS, Boston, 80 pp.

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

Snellman, L. W., 1977: Operational forecasting using automated guidance. *Bull. Am. Meteorol. Soc.*, **10**, 1036–1044.

●

---

announcements *continued from page 762*

PSMSL mean sea level data publications

Monthly and Annual Mean Heights of Sea Level, Vol. 3, is the final part of a revised publication that gives world coverage of mean sea level data. This third volume contains data for Japan, the Philippines, Australasia, and the Pacific islands.

The Permanent Service for Mean Sea Level (PSMSL) is the publisher. The PSMSL, housed within the Institute of Oceanographic Sciences at Bidston on Merseyside, U.K., is charged with the task of accessing and collating mean sea level data from all known tide gage stations on a worldwide scale. This information has been published in a number of volumes, the first of which appeared in 1940. By the late 1960s, several of the earlier publications had become unobtainable and the opportunity was taken to review the entire method of data publication. The outcome was the introduction of a Revised Local Reference (RLR) as data for publication of sea level series. By definition, RLR—referred to a specific year—is an integral number of decimeters below the primary Tide Gage Bench Mark. The method of data presentation has also been reviewed and future publications will be of a loose-leaf nature. The information for each station commences on a new page, which will facilitate subsequent procedures of updating and amendment.

The preparation of data held by PSMSL for publication in the new RLR format has involved considerable time and, in total, data from 629 stations have been included in the revised publication. The coverage of the three volumes is: *Volume 1:* Europe, Africa, India, and the Far East; *Volume 2:* North, Central, and South America; and *Volume 3:* Japan, Philippines, Australasia, and the Pacific Islands. All three volumes and a catalog are available on request to PSMSL at the address below. In accordance with the loose-leaf nature of the publication, updated and amended pages for the three volumes will be issued to users regularly. Contact: The Permanent Service for Mean Sea Level, Institute of Oceanographic Sciences, Bidston Observatory, Birkenhead, Merseyside L43 7RA, United Kingdom.

Continued on page 787