

SUNYA Experimental Results in Forecasting Daily Temperature and Precipitation

LANCE F. BOSART

Department of Atmospheric Science, State University of New York at Albany, Albany 12222

(Manuscript received 1 April 1975; in revised form 18 July 1975)

ABSTRACT

An analysis of skill in predicting daily temperature and precipitation is presented for six years (1969–1975) of forecasts made for the Albany County Airport by students and faculty in the Department of Atmospheric Science of the State University of New York at Albany. The daily consensus forecast (made up by averaging the forecasts of all forecasters) shows no significant secular increase in skill for temperature. An apparent increase in the consensus skill in precipitation forecasting is noted with most of the increase occurring in the spring 1972 semester. Possible reasons for this increase are discussed. The skill (defined as the percentage improvement over a persistence climatological forecast) of the ensemble of forecasters over a persistence climatological control is near 50% for the first day decaying to 10% and near zero by the 3rd and 4th day for precipitation and to just under 10% for temperature by the 4th day. These results are consistent with the results presented by Sanders (1973).

Some relationship is found for skill to be a function of the variability of the daily temperature about the climatological mean. Skill, however, appears to be insensitive to the frequency of days with radiational cooling, a major local forecast problem. Likewise skill appears to be independent of daily rainfall amount or frequency. These findings are consistent with those found for Boston by Sanders (1973).

Finally, the trend towards a plateau in skill noted by Sanders (1973) is confirmed for a different location.

1. Introduction

Sanders (1973) has recently presented some experimental evidence that seems to indicate that forecasting skill may be on a relative plateau. The results were based on six years (1966–1972) of daily forecasts of temperature and precipitation out to 96 hours at the Massachusetts Institute of Technology (MIT). Sanders strongly urged the presentation of comparable information for other locations in order to better appraise current forecast skill. The purpose of this paper is to evaluate similar forecasts made at the State University of New York at Albany (SUNYA) over the period 1969–1975 and compare results, where possible, for the two different locations.

2. Forecast description

The SUNYA forecasts are addressed to the following questions for Albany, N. Y., for each of the next four consecutive 24 h periods, beginning at 1800 GMT of the present day: 1) will the minimum temperature be below the climatological minimum anytime during the period, and 2) will there be one hundredth of an inch or more of melted precipitation during the period? Verification is based upon the 1800 GMT surface observation from the National Weather Service Forecast Office at the Albany County Airport.

The presumed goal of each forecaster is to maximize his or her gain over the climatological forecast. Clima-

tological forecasts are prepared according to the format in Table 1 where the numbers refer to number of chances in 10 of the event verifying. This persistence climatology was derived from Albany Airport data for the 1946–1968 period. As an example, note that if the previous period verified wet (defined to be ≥ 0.01 inch melted precipitation) during the winter then there are 5 chances out of 10 that the next 24 h period will be wet; 3 chances out of 10 that the 24–48 h period will be wet, and so on. Overall, the climatological results suggest that warm spells tend to last longer than cold spells but that the latter are more intense, a result in keeping with synoptic experience in the northeastern United States. Likewise, precipitation is more frequent during the colder half of the year. Such a persistence climatology would appear to be especially suitable for a simple threshold probability forecast game in establishing a relative basis for forecast skill. The current 30-year normals (1941–1970) are used to establish the smoothed daily normal minimum temperature threshold.

3. Forecast mechanics

All forecasts are expressed as numbers from 0 to 10 representing percentages rounded to the nearest 10%. Each day a forecaster then writes 8 numbers, 4 for temperature and 4 for precipitation. Forecasts must be made no later than 1830 LST. The forecasters rely on a potpourri of synoptic guidance obtained from the National Weather Service National Facsimile Circuit

TABLE 1. Persistence climatology (number of chances in 10) used in SUNYA daily probability forecasts. Wet refers to ≥ 0.01 inch; cold, to the probability of the temperature falling below the daily climatology minimum.

	Winter (Nov.-Apr.)				Summer (May-Oct.)			
	Temperature		Precipitation		Temperature		Precipitation	
	Cold	Warm	Wet	Dry	Cold	Warm	Wet	Dry
1st Period	7	3	5	3	7	3	4	2
2nd Period	6	4	3	4	6	4	3	3
3rd Period	5	4	4	4	5	4	3	3
4th Period	5	4	4	4	5	4	3	3

and Service "A" and "C" teletype circuits as well as selected local climatological statistics.

The forecast contest at SUNYA was initiated in October 1969 as a supplement to the more traditional classroom aspects of education. There was no previous forecasting experience at SUNYA with the exception of the author who benefited from five years of probability forecasting while he was a student at MIT from 1964-1969. We started from scratch. Over the years the SUNYA forecast contest has averaged 10-15 people daily, one faculty member and a couple of undergraduates, the remainder being graduate students. The fall semester begins around Labor Day and ends in mid-December, while the spring semester begins in mid-January and ends in early May. A typical semester consists of approximately 75 forecast days. There is no forecasting during the summer. Scoring is cumulative over each individual semester.

4. Forecast verification

All forecasts are verified by a score similar to the Brier (1950) score. A consensus forecast similar to that of Sanders (1973) is computed for each forecast period except that the SUNYA consensus is rounded to the nearest 10%, consistent with the allowable forecast categories. Likewise, a persistence climatological forecast that depends upon the previous days' verification is similarly computed. Individuals, including the climatological forecast, are then ranked against the consensus forecast in the weekly forecasters' ladder. Sanders (1967, 1973) discusses the advantages of this procedure in more detail.

5. Forecast evaluation

All probability forecasts and results presented in this paper were evaluated according to the procedure outlined by Sanders (1963, 1967) and summarized as follows:

Introduce a climatological control probability r and, for each forecast probability, define the departure of the observed relative frequency of occurrence in that category from the control probability by $E \equiv 0 - r$ and the departure of the forecast probability from the control value $d \equiv f - r$. The average score for the forecasts in

the k th of these categories is

$$F_k = \frac{1}{M_k} \sum_{i=1}^{M_k} (d_k - E_{ki})^2, \quad (1)$$

where M is the number of forecasts in that particular K category. Similarly the climatological forecast is scored according to

$$C_k = \frac{1}{M_k} \sum_{i=1}^{M_k} E_{ki}^2 \quad (2)$$

so that the ultimate improvement over the climatological forecast for that forecast category is then

$$C_k - F_k = \bar{E}_k^2 - (d_k - \bar{E}_k)^2. \quad (3)$$

The first term, \bar{E}_k^2 , on the right-hand side of Eq. (3) rewards the forecaster for recognizing those synoptic situations which depart strongly from the climatological frequency (sorting gain) while the second term, $(d_k - \bar{E}_k)^2$, penalizes the forecaster who writes biased forecasts (bias penalty). Eq. (3) is then summed over all forecast categories, weighted by the number of cases in that category to yield the improvement over climatology for the entire forecast sample. Skill can

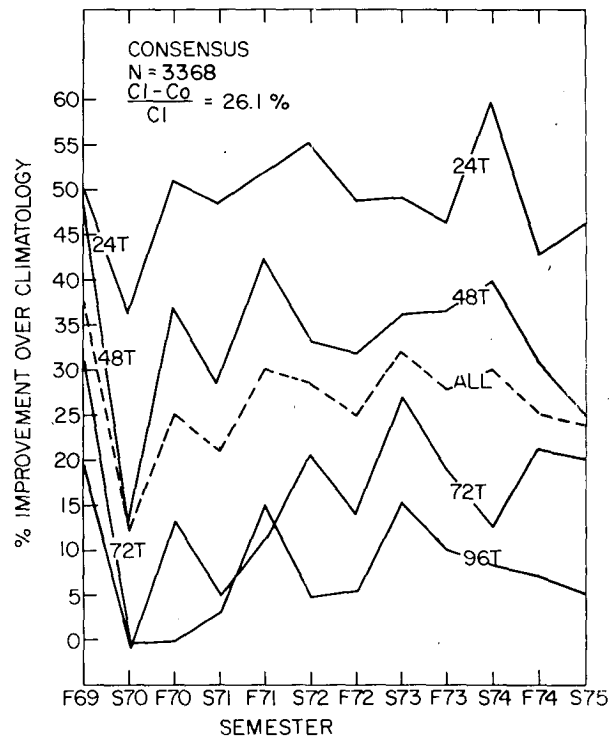


FIG. 1. Percent improvement of consensus (Co) over climatology (Cl) for 24, 48, 72 and 96 h temperature forecasts as a function of semester. N refers to the total number of forecasts, "ALL" is the average for all four forecast periods. $(Cl - Co)/Cl$ refers to the 12-semester percentage improvement of consensus (Co) over climatology (Cl) for all forecast periods. F and S refer to the fall and spring semesters respectively.

then be measured by dividing Eq. (3) by Eq. (2) and converting to percent.

6. Results

Figures 1 and 2 show the consensus results for temperature and precipitation respectively. No real trend is evident in the consensus temperature forecast. The regression curve fitted to "ALL" forecasts in Fig. 1 is $y = 24.95 + 0.23x$ ($x = 1$ corresponds to F69; y is the percentage improvement of consensus over climatology). This corresponds to a reduction of variance of 2% with the standard error of the estimate as 5.66. The regression curve applied to "ALL" forecasts in Fig. 2 is $y = 10.75 + 1.25x$ (again $x = 1$ corresponds to F69; y is the percentage improvement of consensus over climatology) with a reduction of variance of 39%, which borders on significance at the 5% level according to the standard F -test. The interpretation of this result is extremely difficult, however. The experience level of the student forecasters had increased significantly by the spring 1972 semester. The gains on the 3rd and 4th day probably reflect a gradual realization on the part of forecasters of the futility of frequent large departures from the climatological forecast. The gains on the first two forecast periods might conceivably reflect the introduction of the Limited Fine Mesh model (LFM) (National Oceanic and Atmospheric Administration, National Weather Service, National Meteorological Center, 1971) at the National Meteorological Center or the Model Output Statistics (MOS) (Klein and

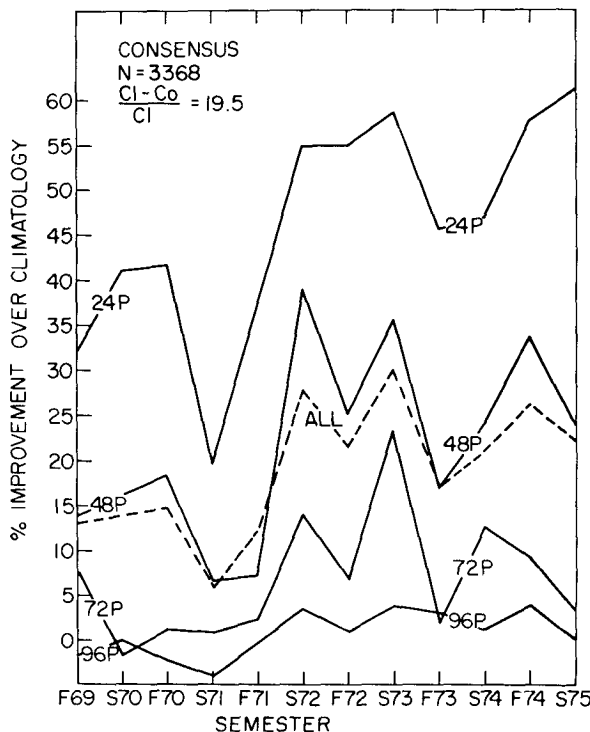


Fig. 2. Same as Fig. 1 except for precipitation.

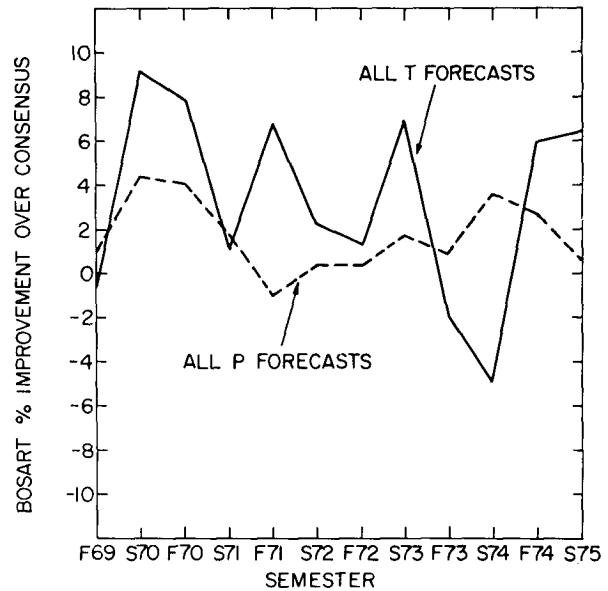


Fig. 3. Percentage improvement of the author over consensus for all temperature (T) and precipitation (P) forecasts.

Glahn, 1974) of the Techniques Development Laboratory. It is my subjective opinion that the LFM product is more significant in that it is available from the 1200 GMT data at forecast time whereas the MOS product is only available from 0000 GMT. A dramatic start is evident during the fall 1969 semester followed by a sophomore slump and then a slow rise, especially on the 3rd and 4th day. Large and persistent negative temperature anomalies during November and December of 1969 and January 1970 contributed to the favorable scores. Scores plummeted during the following spring semester when the inevitable zonal flow patterns returned. Some of the consensus decline on the 2nd through 4th day can probably be attributed to overconfident, and hence biased, forecasts on the part of novice forecasters, especially following the introductory semester when bold forecasting reaped large dividends. Overall, the experience level of forecasters making up consensus probably peaked during the spring 1972 semester. Since that time some veterans have graduated to be replaced by newer forecasters.

The consensus precipitation results in Fig. 2 are rather chaotic, with a disastrous spring 1971 semester followed by a sharp upturn to the spring 1972 semester and then a relative plateau since that time.

Similar verification results for the author (not shown) show no obvious secular trend for temperature. In fact, a slight downward trend is to be noted on the 24 h temperature forecast skill with peaks every spring and troughs every fall semester. The author's corresponding precipitation results show an upsurge from the spring 1971 to spring 1972 semesters but nowhere near as strongly as previously noted for consensus.

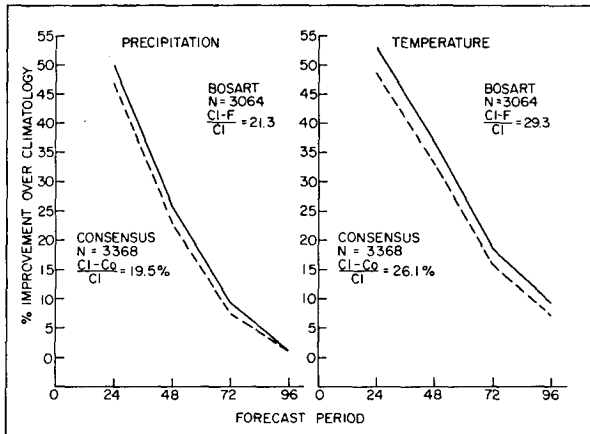


FIG. 4. Decrease of skill with time for the author (solid) and consensus (dashed) for temperature and precipitation. All other symbols as in Fig. 1.

Figure 3 shows the author's performance relative to consensus over the 12 semesters. Considerable term-to-term variability but with no obvious trend is to be noted. Evidently if the state of the art has been rising relative to a consensus that is undergoing considerable turnover it does not show up in the author's forecasts above. The overall results are summarized in Fig. 4, which shows the loss of skill as a function of forecast verification for consensus and the author. A nearly exponential loss of skill with time is noted for the temperature forecasts with this skill falling to the 10% level four days in the future. The corresponding loss of skill in the precipitation forecasts was even steeper, reaching the 10% level by the third forecast period.

The greater deterioration with time of precipitation versus temperature forecasts is hardly news to the meteorological community. What is perhaps remarkable is the great quantitative similarity between Fig. 4 of the present paper and the comparable results presented by Sanders (1973, Fig. 6). This is true despite the greater continentality of Albany versus Boston and the different time periods involved (1966-1972 versus 1969-1975). The nature of the climatological control also differs slightly. Sanders uses a climatology that changes monthly but is not a function of the previous days' verification. The present paper employs a persistence climatology as outlined earlier in Table 1. Additional results compiled by Sanders (1975¹) continue to support the conclusions of his earlier paper.

Figure 5 shows the percentage improvement of consensus over the climatological forecast versus the standard deviation of the daily mean temperatures from their climatological means during the various semesters. The corresponding regression curve is $y = 4.49 + 2.55x$ with a reduction of variance of 25% and a standard error of estimate of 5.21. While this is not significant at the 5% level according to the

¹ Private communication.

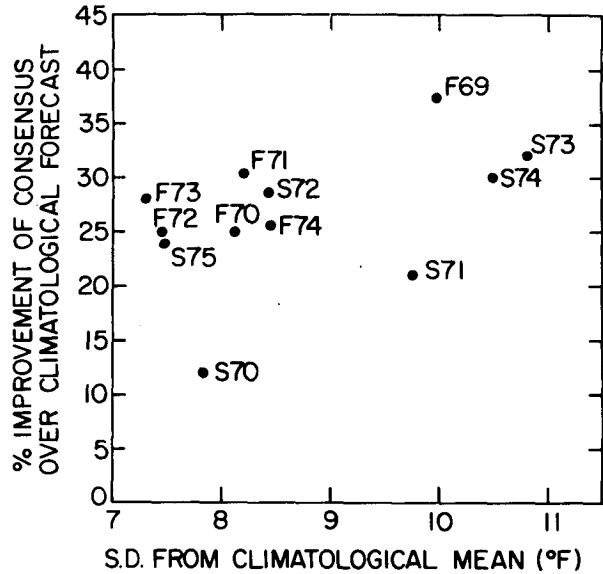


FIG. 5. Consensus temperature skill versus standard deviation (SD) of daily temperature from the climatological mean. F and S refer to the fall and spring semesters respectively.

F-test there is the suggestion of a relationship although not as pronounced as noted by Sanders (1973) for Boston. Albany is more continental than Boston, while the addition of May and part of June to the MIT results undoubtedly reduces the temperature variability during the spring semester.

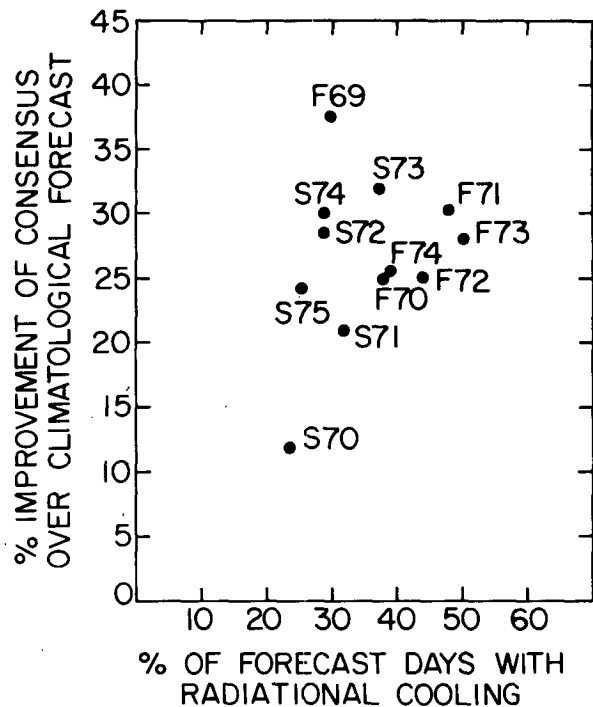


FIG. 6. Consensus temperature skill versus percentage of forecast days with radiational cooling. F and S refer to the fall and spring semesters respectively.

Figure 6 shows the relation of the consensus skill to the percentage of forecast days with radiational cooling, a distinct local forecast problem. The local climatological data together with WBAN 10 surface records for Albany, published by the Environmental Data Service of the United States Department of Commerce (1969-1975), were combed for occurrences of radiational cooling during these years. The criteria for radiational cooling was the occurrence of clear or thin high overcast skies with winds under 1 m s^{-1} for at least one consecutive three-hour period. Another clue was the suppression of the minimum temperature below the previous evening's dew point temperature. It is rather common at Albany for the minimum temperature to plunge $10\text{--}15^\circ\text{F}$ lower than the ambient dew-point temperature on nights of ideal radiational cooling. Ambiguous cases in the data sample were resolved by appealing to the hourly data tabulated on the WBAN 10 forms. The results presented in Fig. 6 indicate that radiational cooling influences 40-50% of the fall nights at Albany during this particular period. The regression curve is $y = 18.63 + 0.22x$ with a reduction of variance of 9%, so any relationship is rather weak. It is the author's observation at SUNYA that sizeable gains over climatology in radiational cooling situations can be substantially reduced by the tendency of the forecaster to predict too early an end to such episodes. Ideal radiational cooling conditions often persist at Albany in light southerly wind regimes to the west of a slow moving anticyclone, much to the chagrin of a forecaster expecting a surge of warm advection and middle cloudiness to hold up minimum temperatures.

Sanders (1973) also found no evidence for any differences in the consensus skill as a function of precipita-

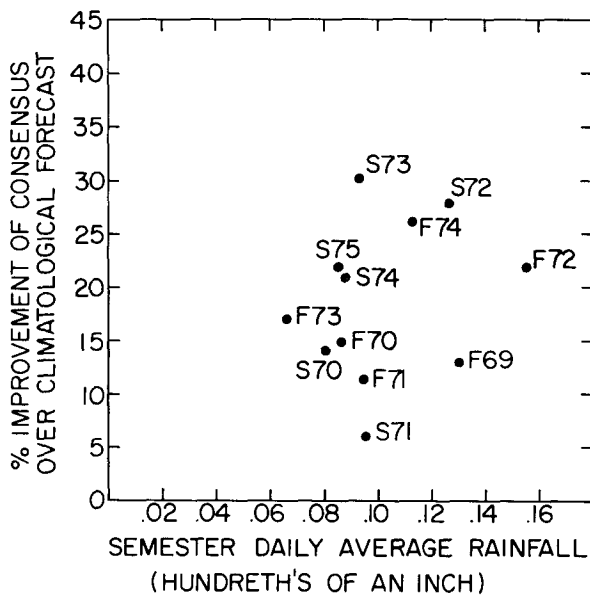


Fig. 7. Consensus precipitation skill versus mean daily rainfall. F and S refer to the fall and spring semesters respectively.

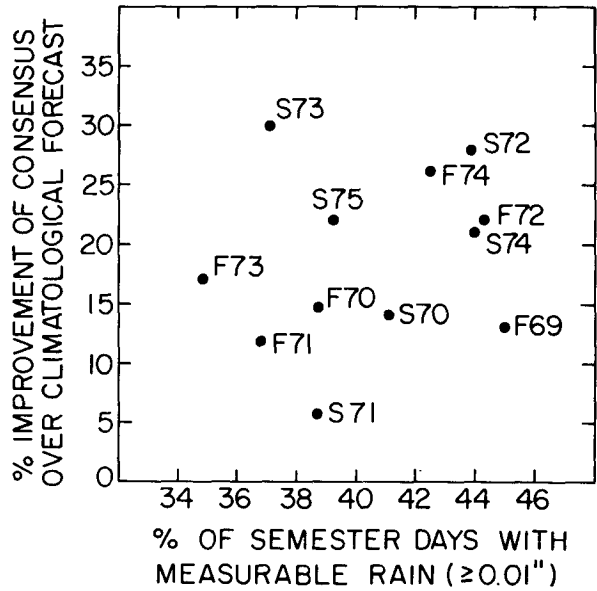


Fig. 8. Consensus precipitation skill versus frequency of measurable rain. F and S refer to the fall and spring semesters respectively.

tion amount and frequency. Our results are shown in Figs. 7 and 8. The corresponding regression curves are $y = 11.80 + 68.71x$ and $y = -0.71 + 0.48x$, which explain 6% and 5% of the variance respectively. There is clearly no significant relationship here.

Finally, the author's forecasts and consensus were analyzed by means of Eq. (3) and the results plotted in Fig. 9. Only the precipitation results can be compared directly with those of Sanders (1973, Fig. 7). Throughout the first three forecast periods (0-24, 24-48, 48-72 h) the consensus precipitation forecast is underconfident, as would perhaps be expected. The author's improvement over the consensus forecasts for these corresponding periods can be attributed both to a better sorting gain and to less bias penalty despite fewer forecasts. By the 96 h forecast period the author's reluctance to depart from the climatological forecast resulted in a reduced sorting gain. The credibility of the SUNYA and MIT findings, it is hoped, is enhanced by the significant quantitative similarity of the precipitation results. Overall, the MIT results would appear to be slightly superior for precipitation not allowing for station differences, forecast procedures, different time periods, etc.

The SUNYA temperature results shown in Fig. 10 suggest that consensus is overforecasting (overconfident) in cold situations and underforecasting (underconfident) in warm situations in the 24 h period with the underforecasting persisting throughout all time periods. The author is somewhat better calibrated despite fewer forecasts and his sorting gain is 5-10% superior to that of consensus. A few peculiarities nevertheless stand out. Consider the author's 24 h precipita-

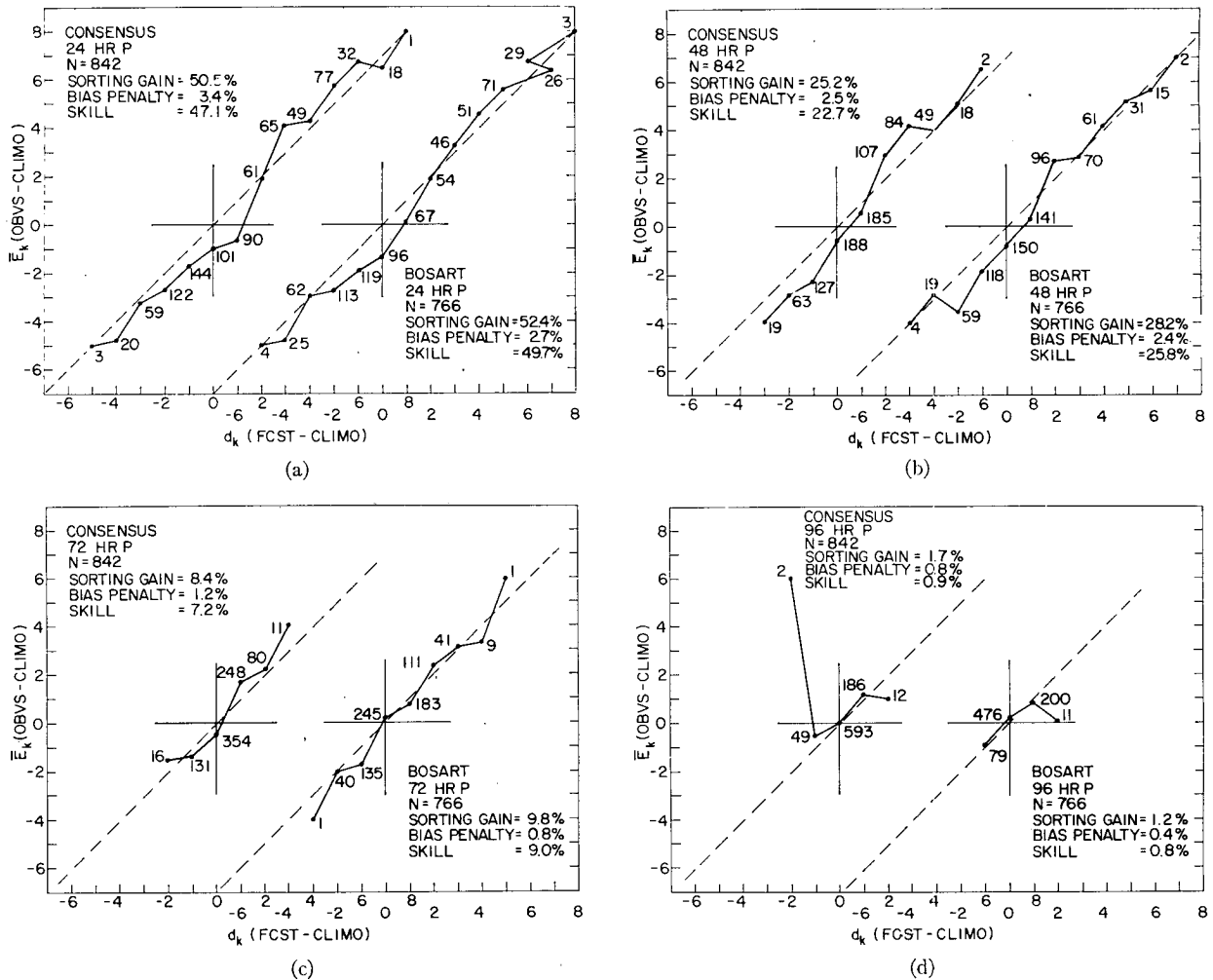


FIG. 9. Evaluation of consensus and author's precipitation forecasts: a) first (0-24 h), b) second (24-48 h), c) third (48-72 h), and d) fourth (72-96 h) periods. The abscissa, d_k , is the forecast departure from the climatological percentage frequency of occurrence (nearest 10%). The ordinate, E_k , is the corresponding departure of the observed percentage frequency from the climatological frequency (nearest 10%). Perfectly unbiased forecasts would lie along the dashed diagonal. The number adjacent to each data point represents the number of times the departure category was forecast.

tion forecast. Sizeable sorting gains were lost, inexplicably, in the $d_k = +6$ and $+7$ categories, where the forecasts appeared to be reversed. See Sanders (1967, 1973) for a good discussion of the psychological factors affecting the interpretation of these forecast statistics.

7. Conclusions

The verifications of SUNYA temperature and precipitation probability forecasts over the period 1969-1975 suggests that the gain of consensus has shown little change for temperature and a slight improvement for precipitation. The slight improvement on the 3rd and 4th day for both temperature and precipitation is undoubtedly a partial reflection of the overall increase in experience of consensus. Consensus appeared to show a modest gain in skill between 1971 and 1972 for precipi-

tion forecasts on the 1st and 2nd day, but this trend has not continued. The introduction of the Limited Fine Mesh model (LFM) and Model Output Statistics (MOS) on an operational basis at the National Meteorological Center during this period may possibly account for some of this gain. Overall, however, the consensus results are somewhat difficult to interpret because of the continued experience change of the individual forecasters.

No increase in skill in temperature forecasting has been found for the author's forecasts. If anything, the author is losing skill for the first forecast period! An ever so slight upward trend is noted for the author's precipitation forecasts but the period of record is too short to make any definitive conclusion.

Overall, the SUNYA results appear to have great quantitative similarity to the MIT results presented by

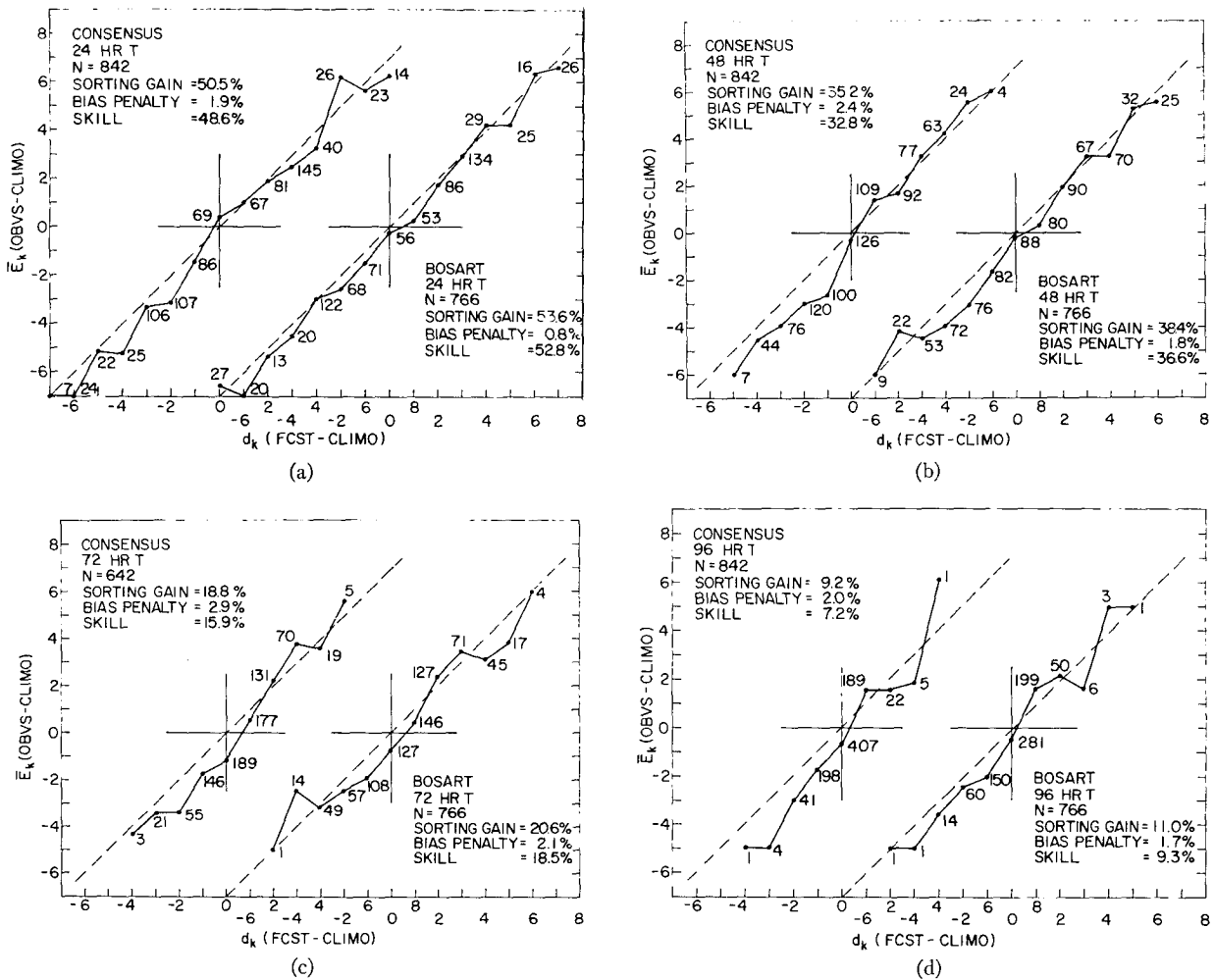


FIG. 10. Same as Fig. 9 except for temperature.

Sanders (1973, 1975²) despite considerable local differences influencing the climate of Albany and Boston. Skill in forecasting precipitation falls to 10% by the 3rd day and to near zero by the 4th day ahead. The loss of skill in temperature forecasting is not as rapid, reaching just under 10% by the 4th day ahead.

Skill in temperature forecasting appears to be somewhat dependent upon the standard deviation of the daily temperature from the climatological mean, although not as pronounced as found by Sanders (1973). Additionally, there appears to be little, if any, relationship between skill and the percentage of forecast days with radiational cooling at Albany. Likewise, for precipitation the consensus skill appears to be independent of the mean daily rainfall amount or rainfall frequency.

The suggestion of an apparent relative plateau in precipitation forecasting made by Sanders (1973) is further reinforced by the SUNYA results with the evidence extended to temperature forecasts as well. The long term verification statistics (36 h 500 mb

forecasts and 30 h surface forecasts) on National Weather Service and National Meteorological Center forecasts presented by Cooley and Derouin (1972), Cooley (1975³), and Sadowski and Cobb (1973, 1974a, 1974b) likewise suggest a pronounced decrease in the rate of increase of skill since the late 1960's with a very slight deterioration from 1972 to 1974. Perhaps the most significant part of this paper is the inability to find any truly major significant secular trend in skill. Contrary evidence for other geographic locations would indeed be welcomed.

Sanders (1973) has suggested that the relative plateau in forecasting skill for the first period results from "squeezing virtually all the blood we can from the short-range synoptic turnip" while errors in predicting an assortment of mesoscale phenomena remain undiminished. The author shares this view and he can think of numerous occasions where, given a perfect 24 h 500 mb forecast, he would be hard pressed to draw a valid surface pressure map, let alone delineate areas

² Private communication.

³ Private communication.

of significant weather. Ongoing research efforts in mesoscale meteorology must consider the scale interaction problem. An ultimate breakthrough in this dilemma probably awaits an increased understanding of how the physics and dynamics of mesoscale phenomena interact with synoptic scale features. Finally, more research is required to increase our understanding of the significance of the *vertical* fluxes of heat, momentum, and moisture. Forecasters are always concerned with the motion of existing short waves, but what of the creation of new short waves? Ninomiya (1971) and Starr (1973) have touched upon this problem, but existing operational prediction models are lacking in this area. Perhaps this is where a breakthrough can be made on the 3rd and 4th day in better coupling synoptic scale motions in the free atmosphere to the "weather" as actually observed at the surface.

Acknowledgments. The author expresses his deep appreciation to all those students who have been willing to participate in daily probability forecasting at SUNYA and especially to those who have suffered more than their share of atmospheric perversity. Miss Sally Young typed the manuscript. Partial support was provided by NSF Grant # A023897000.

REFERENCES

- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.
- Cooley, D. S., and R. G. Derouin, 1972: Long term verification trends of forecasts by the National Weather Service. NOAA Technical Memorandum NWS FCST-18, 11 pp. [Available from National Technical Information Service, Springfield, Va.]
- Klein, W. H., and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55**, 1217-1227.
- National Oceanic and Atmospheric Administration, National Weather Service, National Meteorological Center, 1971: The Limited-area Fine Mesh (LFM) model. NOAA, NWS Technical Procedures Bulletin No. 67. [Available from Weather Analysis and Prediction Division, National Weather Service, NOAA.]
- Ninomiya, K., 1971: Mesoscale modification of synoptic situations from thunderstorm development as revealed by ATS III and aerological data. *J. Appl. Meteor.*, **10**, 1103-1121.
- Sadowski, A. S., and G. F. Cobb, 1973: National Weather Service May 1971-April 1972 Public Forecast Verification Summary. NOAA Technical Memorandum NWS FCST-19, 26 pp. [Available from National Technical Information Service, Springfield, Va.]
- and —, 1974a: National Weather Service Heavy Snow Forecast Verification 1962-1972. NOAA Technical Memorandum NWS FCST-20, 29 pp. [Available from National Technical Information Service, Springfield, Va.]
- and —, 1974: National Weather Service April 1972 to March 1973 Public Forecast Verification Summary. NOAA Technical Memorandum NWS FCST-21, 64 pp. [Available from National Technical Information System, Springfield, Va.]
- Sanders, F., 1963: On subjective probability forecasting. *J. Appl. Meteor.*, **2**, 191-201.
- , 1967: The verification of probability forecasts. *J. Appl. Meteor.*, **6**, 756-761.
- , 1973: Skill in forecasting daily temperature and precipitation: some experimental results. *Bull. Amer. Meteor. Soc.*, **54**, 1171-1179.
- Starr, V. P., 1973: Remarks on the progress of general circulation studies. *Tellus*, **25**, 1-11.