The Contributions of Education and Experience to Forecast Skill

PAUL J. ROEBBER

Department of Geosciences, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin

LANCE F. BOSART

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

(Manuscript received 31 January 1995, in final form 26 June 1995)

ABSTRACT

An analysis of nine semesters of temperature and precipitation forecasts at the State University of New York at Albany has been conducted with the goal of investigating the dependence of forecasting skill on education and experience. The results show that forecast skill is largely determined by experience. The relative advantage of highly experienced forecasters is secured by virtue of the larger set of cases from which they may draw upon: given a set of forecast information (e.g., moisture, winds and cloud cover), such a forecaster is in a better position to maximize linear consistency between that information and the expected evolution of surface temperature and precipitation (given similar conditions, make a similar forecast) than someone with less forecasting experience. However, the experienced forecaster also gains substantially by recognizing those instances in which these linear relationships no longer apply and by forecasting accordingly. Such instances can often be recognized using simple rules. Consequently, there is a rapid growth of skill with experience for initially inexperienced forecasters; this progression continues through several clearly defined stages and reflects the forecaster's increased ability to implement these simple forecasting strategies. The skill advantage of human forecasters over numerical guidance continues to diminish and now largely reflects the human ability to recognize occasional departures from the linear relationship between forecast information and future observations.

1. Introduction

a. Background

How does forecast skill, including the use of statistically adjusted numerical model output such as Model Output Statistics (MOS), change with education and forecast experience? The conventional wisdom is that experience is an important aspect of overall skill. Gedzelman (1978, 1979), Firestone (1979), and Wernly (1979) have raised a number of pertinent issues in this area but lacked sufficient data to fully examine them. Ramage (1978) discussed the possible erosion of forecast skill in the face of improving numerical weather prediction and suggested that skill can be maintained only through unremitting practice (i.e., experience).

Given the diminishment in the gap between automated and human forecasts in the past two decades (Glahn 1985; Sanders 1986), the basic human strategy now is to determine those occasional situations in which the MOS predictions are likely to fail and to forecast accordingly. Perhaps of more fundamental importance is for forecasters to focus upon those occasions in which extreme events are likely; such events may or may not be represented in the MOS output (for example, an arctic cold wave versus severe convection, respectively). This represents a practical demonstration of Sanders's (1986) proposed restructuring of the forecast enterprise into routine-forecast specialists and warning-forecast specialists. The focus of this paper is primarily upon the problem of routine forecasts of temperature and precipitation, although we have not specifically excluded extreme events that may be defined by a marked deviation from the climatological normals of those variables (such as with an outbreak of arctic air). What are the keys that allow human forecasters to skillfully identify instances in which MOS will fail? Does education and/or experience play a role in this recognition?

Psychological research suggests that what Murphy (1988) called conditional bias in the skill score (which reveals how well the variance of the forecasts reflects the lack of a perfect correlation between forecast and observation; see section 2a) is introduced because people tend to make extreme forecasts when a predictor takes on an extreme value, even when the relation between the predictor and the forecast variable is weak (Stewart 1990). Decomposition of the skill score (Murphy 1988; Stewart 1990; see section 2b) as a function of forecaster experience should reveal whether ex-

Corresponding author address: Dr. Paul J. Roebber, Department of Geosciences, University of Wisconsin—Milwaukee, Lapham Hall 352, P.O. Box 413, Milwaukee, WI 53201.

perience reduces conditional bias; in other words, the experienced forecaster is not as easily misled by extremes as the neophyte. In addition, an examination of the growth of skill with experience, as was done on a more limited basis by Gedzelman (1978), will be performed.

By partitioning forecasts into an MOS component and a deviation from MOS ($F = M + d$, where F is the forecast value, M is the MOS value and d is $F - M$) and performing a correlation analysis subject to a stratification of the forecasters on the basis of education and forecast experience, it should also be possible to examine the relationship between education, experience, and forecast skill, as defined by the proportion of the variance accounted for by the deviation from MOS. These techniques will also allow an analysis of the supposed erosion of human skill with the recent advances in numerical guidance (which has continued to improve since the implementation of the Nested Grid Model; see section 3c).

b. Forecast data

To address these questions, we have analyzed forecast data obtained from the State University of New York at Albany (SUNYA) forecasting contest. Forecasts of temperature and precipitation have been made routinely by members of the Department of Atmospheric Science at SUNYA during the fall and spring academic semesters since September 1969. In this paper, we shall be dealing with results from the game for the period September 1988 through December 1992, totaling nine forecast semesters. The verification dates for this period are listed in Table 1.

The game underwent some evolution to arrive at its present form, which has been in place since September 1986. For details of the game prior to 1986, interested readers should consult Bosart (1975, 1983). Forecasts of maximum and minimum temperature, probability of measurable precipitation (that is, probability of precipitation greater than 0.01 in., hereafter referred to as POP), and precipitation amount are made for the four sequential 12-h periods beginning at 0000 UTC of the following day. Thus, a forecast submitted by 2045 UTC of day 0 (the approximate forecast submission time) would include a minimum temperature forecast for the period 0000–1200 UTC of day 1, a maximum temperature forecast for 1200–0000 UTC of day 1, a minimum temperature forecast for 0000–1200 UTC of day 2, and a maximum temperature forecast for 1200–0000 UTC of day 2. In addition, POP and precipitation amount forecasts would also be submitted for each of these four periods. Figure 1 provides a diagram of the forecasting sequence for the first 24 h.

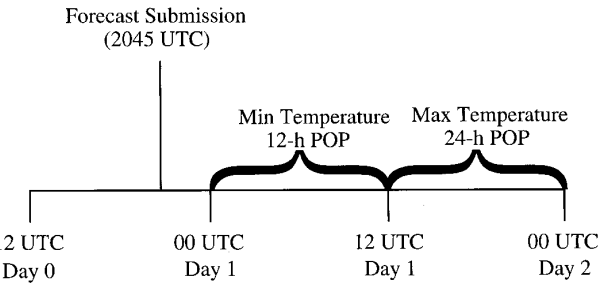

FIG. 1. Time sequence for the SUNYA forecasting game.

Daily forecasts (Monday through Friday) are made by a collective group of 10–20 individuals of varying experience and background, with approximately 66 forecast days per semester. The scoring is based on absolute error in temperature–precipitation amount forecasts and the Brier score (Brier 1950) for POP forecasts. Errors are tabulated and skill scores are presented with respect to the relatively weak control of climatology and the strong control of consensus, defined as simply the group average forecast for any given day in which at least three forecasters have played the game. Interested readers should refer to Bosart (1975, 1983) for further details concerning the scoring methods of the game.

It is natural to ask to what extent this dataset is representative of what forecasters do in operational settings with regard to information available, deadline pressures, and related issues. One of us (PJR) has had considerable experience in both operational forecast settings as well as university games, while another of us (LFB) is quite familiar with National Weather Ser-

TABLE 1. Forecast verification periods in the Albany game, fall 1988 through December 1992.

Semester	Start date	End date	Forecasts	Notes
Fall 1988	8 September	15 December	66	
Spring 1989	19 January	6 May	69	Spring break (20–24 March)
Fall 1989	7 September	14 December	64	
Spring 1990	18 January	5 May	71	Spring break (09–13 April)
Fall 1990	6 September	13 December	63	
Spring 1991	25 January	8 May	67	Spring break (25–29 March)
Fall 1991	5 September	13 December	64	
Spring 1992	24 January	2 May	63	Spring break (23–27 March)
Fall 1992	3 September	11 December	64	

vice (NWS) conditions. Information available to forecasters is virtually identical in both settings. In our view, the most important distinction is that in the university setting the deadline is softer, so that the forecaster can, if motivated, examine the data to a greater extent than can the operational forecaster, who has many duties and a much stricter timetable. However, we also suspect (although we cannot examine this question with the present dataset) that someone in continuous contact with the evolving weather situation (such as an operational forecaster) is in a better position by the end of a work session to have a ''feel'' for what pattern is emerging than a person with intermittent contact throughout the day (such as a university researcher or student). Existing data, such as that presented by Bosart (1983) and Sanders (1986) suggest that consensus forecasts from university games compete quite favorably with that of the NWS public forecasts and provide a good measure of the state of the art of temperature and precipitation forecasting. These findings are confirmed by the data under consideration here (see Tables 4–7 in section 3).

Approaches to the analysis of forecast data in real-world settings must take into account some of the limitations imposed by practical considerations. The reality is that forecast contests must be simple and fun to attract student participation; a ''statistically pure'' study (using the ideas of experimental design under controlled conditions) would probably have few participants. Our concern needs to be whether the data collected from studies such as ours can satisfy certain assumptions inherent in our analysis procedures.

Frequently, it is assumed that individual data must represent independent events sampled from a random distribution of events. Clearly, such a condition cannot be met under our study, since we do not randomly sample forecast days, and the forecast events are not completely independent. The major threat in violating this condition is autocorrelation due to persistence. In this study, only about 60% of the days (depending upon the forecast event) actually also have the previous day included in the sample, with about 30% of the events separated by at least five days. The Durbin–Watson statistic indicates little autocorrelation in our sample.

A second concern sometimes raised regards the independence of forecast errors, both within and across forecasters. In the first case, the concern is that the forecast errors of an individual who makes a series of forecasts will not represent independent events. Clearly, these errors are not independent, nor should they be, since these errors and their source are the object of study. (What might a forecaster do to improve his or her performance?) In the second case, the concern is whether forecast errors on a given day made by a group of people who talk to each other about the weather conditions before submitting their forecasts can be considered independent events. To the extent that such interactions between highly experienced forecasters and neophytes occur (with the latter presumably being more easily swayed), we have confidence that for a sufficiently large number of forecasts, the effects of these interactions on consensus are minimal. In any case, such interactions would likely bias the forecasts of the less experienced individuals toward the ideas of their more experienced colleagues, lessening the likelihood of detecting significant differences between the two groups. These interactions aside, it is also important to realize that strict independence of forecast errors across the set of forecasters should not be expected, since one wishes to understand what causes the collective (consensus) to deviate from the observations. The conditions that cause forecaster A to deviate from the eventual observations are likely also operative in the forecast of forecaster B.

2. Analysis techniques

a. Murphy's decomposition

Murphy (1988) showed that a skill score (SS), defined with respect to the mean square error (MSE) of a reference forecast that is the mean of the variable being forecast (climatology given a sufficiently long time series and a stationary climate, denoted here as MSE_c) can be written as

$$\mathrm{SS} = 1 - \frac{\mathrm{MSE}_f}{\mathrm{MSE}_c} = (r_{fo})^2 - \left(r_{fo} - \frac{s_f}{s_o}\right)^2 - \left(\frac{\bar{f} - \bar{O}}{s_o}\right)^2, \quad (1)$$

where r_{fo} is the correlation between the forecast and the observed event, and s_f, s_o, $\bar{f}$, and $\bar{O}$ are the standard deviations and the means of the forecast and observed events, respectively. The MSE_f is defined by

$$\mathrm{MSE}_f = \frac{1}{N}\sum_{i=1}^{N}(f_i - O_i)^2, \quad (2)$$

where f_i and O_i are the ith forecast and observation, respectively. Correspondingly, the MSE_c is defined by the substitution of climatological values for the forecasts f_i in (2). Equation (1) corresponds to (12) in Murphy (1988); readers interested in the details of its derivation should consult section 3a of that reference. Stewart (1990) has also discussed this decomposition in some detail, and here we provide merely a brief summary.

The skill score defined by (1) is 1.0 for perfect forecasts, 0.0 for forecasts that are only as accurate as the reference (climatology), and negative for forecasts less accurate than the reference. In the case of dichotomous events (e.g., rain/no rain), the correlation term r_{fo} measures the ability of the forecast to discriminate between occurrence and nonoccurrence of the event (attaining the value of 1.0 for a perfect linear relationship between the forecasts and observations). More generally, the

correlation term represents a measure of the strength of the linear relationship between the forecasts and observations. Because the skill score attains the level of $(r_{fo})^2$ only if all bias has been removed (the second and third terms of the equation), the correlation is a measure of the ''potential'' skill of the forecasts.

The second term, which is denoted conditional or regression bias, is introduced when extreme forecasts are made on the basis of extreme values of the predictors, even when the relation between the predictor and the forecast variable is weak. Since the slope of the regression line relating the forecasts to observations can be expressed as $(s_o/s_f)r_{fo}$, it is apparent that this bias is eliminated only when the slope is equal to 1; in this context, the bias is eliminated when there exists a one-to-one relationship between the forecasts and the observations. The third term, which is denoted unconditional bias, is eliminated only when the mean of the forecasts matches that of the observations. In the case of precipitation, it measures how well the forecasts match the actual base rate of the observed event. In the limit of a long time series and a stationary climate, this base rate takes on the climatological value, and the bias would represent the deviation of the average forecast from climatology.

This decomposition allows for an examination of the different properties of the joint distribution of forecasts and events (Murphy and Winkler 1987). However, considerable additional insight can be gained into the forecast process and relative forecast skill by considering how different forecasters use basic information to generate their forecasts. This decomposition is summarized in the next section.

b. Lens model equation

Stewart (1990) has provided a detailed discussion of the lens model equation in the context of meteorological forecasts. Essentially, the correlation coefficient r_{fo}, which appears in the skill score formula (1), can be further decomposed into terms that give insight into the forecasting context. First, the forecast (observed) event is partitioned into a multiple linear regression (MLR) equation expressing the relationship between the forecast (observation) and the available information:

$$f = M_{f|x}(x_1, x_2, \cdots, x_N) + E_{f|x} \quad (3)$$

$$O = M_{o|x}(x_1, x_2, \cdots, x_N) + E_{o|x}, \quad (4)$$

where the x_N are the predictors (which we shall refer to hereafter as cues), $M_{f|x}$ ($M_{o|x}$) represent MLR models that describe the relationship between the forecast (observation) and the cues, and $E_{f|x}$ ($E_{o|x}$) are the model residuals or errors. It should be understood that a ubiquitous example of an MLR model in the meteorological context is represented by MOS, which correlates a range of outputs from the numerical model forecasts with observations. For such a partitioning, the lens model equation is then written as

$$r_{fo} = GR_{f|x}R_{o|x} + C\sqrt{1 - (R_{f|x})^2}\sqrt{1 - (R_{o|x})^2}, \quad (5)$$

where G is the correlation between the model of the forecast ($M_{f|x}$) and the model of the observations ($M_{o|x}$), $R_{f|x}$ is the correlation between the forecast and $M_{f|x}$, $R_{o|x}$ is the correlation between the observations and $M_{o|x}$, and C is the correlation between the model residuals.

It should be understood that the models $M_{f|x}$ and $M_{o|x}$ need not necessarily be linear, as is specified here. For example, one could use neural networks to simulate the presumably nonlinear human thought process. However, experience has shown that linear approaches work exceedingly well across a variety of disciplines, including meteorology, even when the practitioners of those disciplines report that their judgment processes involve nonadditive, synergistic aggregation of information (Stewart et al. 1989). Finally, it should be noted that nonlinear relationships between cues and observations can sometimes be accounted for within the context of an MLR model through transformation of the cue. Radiative considerations suggest a logarithmic relationship between precipitable water and minimum temperature, which can be accounted for by using the logarithm of precipitable water as a forecast cue in an MLR model. Since for other types of models, (5) may not hold exactly (Stewart 1990) and the use of nonlinear models presents no obvious advantage, we have opted for the relative simplicity of MLR in this paper.

A measure of the systematic match between the models of the forecast and observations is G; high values of G give an indication that forecasters are making good use of the linear information in the cues. A measure of the consistency of the forecasts, in the sense that identical information leads to identical forecasts, is $R_{f|x}$. For example, a perfectly consistent forecaster would always produce the same forecast, given the same weather conditions as expressed by the cues. A measure of the optimal linear predictability of the observations, for the given set of cues, is $R_{o|x}$; it provides an upper bound on the correlation coefficient r_{fo} achievable by an individual forecaster. The accuracy of this measure depends upon the adequacy of the fit between the model and the observed data. If a better model were used, it would provide a better fit to the data and would indicate a higher optimal predictability of the observations. The variable C measures the residual correlation, and so provides an estimate of the nonlinear relations between the forecasts and the observations. For example, experienced forecasters know that cold-air outbreaks do not infiltrate quickly into Albany (see section 3a), and will adjust minimum temperatures upward from that that might have been expected from an examination of 850-hPa temperatures and/or thickness patterns alone. In the analysis that follows in section 3 we shall refer to the first and second

TABLE 2. Forecast cues for temperature and POP forecasts.

Temperature cue	Physical basis
Climatology	Seasonal cycles
Persistence	Synoptic-scale weather pattern
NGM MOS	MLR of forecast model output
Surface dewpoint (0000 and 1200 UTC)	Saturation temperature limit/radiation
0000 UTC surface temperature (minimum only)	Stability (with 850-hPa temperature)
850-hPa temperature (0000 and 1200 UTC)	Airmass temperature/mixing
Wind speed (0000 and 1200 UTC)	Mechanical mixing
Logarithm of precipitable water (0000 and 1200 UTC)	Tropospheric moisture–radiation
Snow on ground	Radiation
POP cue	**Physical basis**
Climatology	Seasonal cycles
Persistence	Synoptic-scale weather pattern
NGM MOS	MLR of forecast model output
Trenberth (1978) ascent forcing (0000, 1200 UTC and 12-h maximum)	Quasigeostrophic theory
250-hPa vorticity advection (0000, 1200 UTC and 12-h maximum)	Quasigeostrophic theory
850-hPa wind components (0000 and 1200 UTC)	Orographic influences
K index (0000 and 1200 UTC)	Convective instability
Lifted index (0000 and 1200 UTC)	Convective instability
Surface dewpoint (0000 and 1200 UTC)	Boundary layer moisture
Midlevel mean relative humidity (0000 and 1200 UTC)	Midlevel moisture
Cloud thickness	Low to midlevel moisture
Precipitable water (0000 and 1200 UTC)	Tropospheric moisture

terms of (5) as the linear and residual terms, respectively.

Finally, as mentioned in the introduction, the forecasts and observations can be partitioned into an MOS component and a deviation from MOS, such that

$$F = M + d_F \tag{6}$$

$$O = M + d_O, \tag{7}$$

where F and O are the forecast and observed values, M is the MOS value, and d_F and d_O represent the residual or deviations from MOS. Then, a correlation analysis can be performed on the deviations of the forecasts and observations from MOS in order to isolate that portion of the forecasts that requires skillful interpretation of the information beyond that that can be provided by MOS. Thus, a skilled forecaster would know when substantial deviations from MOS might be justified, and in that sense, skill can be defined according to the square of the deviation correlation (R_d^2), which represents the proportion of the observed variance accounted for by the forecast deviations from MOS.

c. *Description, selection, and representativeness of the cues*

To analyze forecast skill using the technique discussed above, it is necessary to know the information (cues) used by the forecasters. Although these data are not routinely recorded in the Albany forecast game, it is possible to rebuild them through a variation on what is known as the ''perfect prog'' approach (Klein et al. 1959; see below). The forecasters have at their disposal a wide array of observational and numerical modeling data up to the time of forecast submission. To estimate these data, we have assumed that the numerical models can provide a perfect picture of the atmosphere (within the limits of observational error) for short-range forecasts out to 24 h. Consequently, we use actual observations valid during the time period of the forecast verification to simulate the numerical modeling data that we cannot easily reconstruct. This assumption represents an upper limit on the accuracy of the data actually available to the forecasters; however, as discussed below, this assumption is tenable provided that we restrict our analyses to periods up to roughly 24 h (DiMego et al. 1992).

It is possible to define a priori which cues should correlate well with observed temperatures and precipitation, based upon valid physical arguments. These cues will vary, depending upon the specific parameter that we wish to forecast. For example, the cues used for a forecast of overnight minimum temperature are substantially different than those for POP. Table 2 lists the cues identified for the temperature and precipitation forecasts, along with the physical basis for the proposed correlation. The cues for temperature and precipitation were derived from a combination of surface and sounding data at Albany, New York. In addition, the Trenberth (1978) ascent forcing and the 250-hPa vorticity advection were computed using gridded data from the National Meteorological Center (NMC) dataset (Mass

et al. 1987). The gridded values were then linearly interpolated to the location of Albany.

Examination of correlations between specific cue pairs indicates fairly substantial cross correlations for some of the variables, as was noted earlier in the development of the MOS equations (Glahn and Lowry 1972). This means that it will not in general be possible to make meaningful statements concerning the relative weighting of various cues by different forecasters, since the equivalent information could be obtained through the use of a separate cue that is highly correlated with the original variable. However, it also suggests that departures from ''optimal'' cue weights may not result in substantial error, provided the information contained in the incorrectly weighted cue is accounted for in another (redundant) cue.

This point can be examined by comparing the adjusted R^2 of the MLR model ($M_{o|x}$) using the full set of cues with another MLR model that uses a subset of the original cues. We find that for each of the forecasting tasks (minimum temperature, maximum temperature, 12- and 24-h POP), a reduced set of cues can achieve approximately the same fit as with the full set of cues. For example, minimum temperature can be fitted with an adjusted R^2 of 0.95 using the 13 cues of Table 2. Approximately the same fit (adjusted R^2 of 0.94) can be achieved for this dataset using only 5 of the original 13 cues. Nonetheless, we shall employ the full set of cues in the study, since they have each been identified on the basis of physical principles, and the mean square error for independent data is usually not very sensitive to the number of predictors within rather broad limits (see Murphy and Katz 1985, pp. 305–308).

As noted above, it is necessary to establish the reliability and validity of the forecast cues that form the basis of the statistical comparisons. The typical temperature and precipitation forecast, prepared between 1800 and 2000 UTC every day, covers 12-h periods beginning 0000 UTC the next day (that evening). We have chosen to select as forecast cues parameters that are unavailable to forecasters in real time. Our approach is to use archived sounding observations for Albany and operationally derived gridded meteorological datasets, valid 0000 or 1200 UTC, to reconstruct the forecast cues. In the perfect prog approach a long time series of isobaric and sounding observations are regressed to derive temperature and precipitation prediction equations for 12-h forecast periods for selected locations. The resulting regression equations are next applied to an independent dataset where the observations, assumed to be ''perfect,'' are derived from existing operational numerical ''weather'' (circulation) simulations.

As an illustrative example, imagine that the time is 1800 UTC and a forecaster is trying to decide what to forecast for a maximum temperature and the POP for the 12–24-h forecast period (the next day). The forecaster would likely consider such parameters as the expected cloud cover and tropospheric mean relative humidity, the 850 hPa temperature and wind direction, and the configuration of the upper-tropospheric jet stream in preparing the forecast. To reconstruct forecast cues we might use such parameters as the precipitable water integrated from the surface to 500 hPa as a surrogate moisture variable, the 1000–700-hPa thickness and its departure from the long-term climatological normal as a surrogate for the temperature structure of the lower part of the troposphere, and the 850-hPa temperature in conjunction with the model forecast temperature lapse rate in the planetary boundary layer as an indicator of the potential maximum temperature assuming a well-mixed atmosphere.

At issue then is the legitimacy of using sounding and operational forecast model-derived parameters as forecast cues when the forecaster does not have direct access to this information. DiMego et al. (1992) present statistical evaluations of biases and standard deviation errors of geopotential heights, temperatures, and winds for the 850-, 500-, and 250-hPa levels obtained from the gridded initialized (0 h) and 12-, 24-, 36-, and 48-h forecast projections from the NMC Regional Analysis and Forecast System (RAFS). The gridded initialized and forecast fields are objectively interpolated to the locations of the observed radiosonde stations over North America at the previously mentioned pressure levels in order to calculate the biases and standard deviation errors. According to DiMego et al. (1992) the typical biases for the initialized (0 h) fields range from 0 to 2 m, 0.0° to 0.9°C, and 2.2 to 3.5 m s^{-1} for the 850-, 500-, and 250-hPa levels, respectively. (The larger values are found at 250 hPa.) The corresponding standard deviation errors range from 6.6 to 12.8 m, 0.9° to1.3°C, and 1.4 to 2.2 m s^{-1}, respectively. Note that these errors are quite comparable to typical radiosonde height, temperature, and wind speed errors. Accordingly, the uncertainty in the RAFS initialized gridded mass and wind fields is comparable to the observations on which the RAFS analysis and initialization procedure is based. The corresponding numbers for the 24-h RAFS forecasts are (for the biases) −3.4 to 0.4 m, −0.2° to 1.0°C, and 4.6 to 7.2 m s^{-1}, respectively, and (for the standard deviation error) 19.0 to 28.9 m, 1.4°to 2.3°C, and 3.0 to 4.7 m s^{-1}, respectively. These 24-h forecast standard deviation error bands remain within a factor of 2 of the typical radiosonde errors and are considerably less (by up to an order of magnitude) than the observed 24-h height, temperature, and wind speed changes at 850, 500, and 250 hPa associated with typical synoptic-scale transient weather regimes. Given these findings, we feel comfortable with our outlined procedure for the preparation of the forecast cues.

The cues identified in Table 2 were thus obtained from a combination of surface, sounding, and gridded observational data, and these variables were then forced into the regression equations, defined by $M_{f|x}$

and $M_{o|x}$ (where the equations $M_{f|x}$ are different for each of the consensus forecast groups and $M_{o|x}$, representing the perfect prog equation, is identical in all cases). To maximize the available cases, these equations were generated without regard to seasonal stratification. However, since the forecasts are tied to the academic semesters, these data can be loosely categorized as cold season (September to early May).

3. Results

a. Skill as a function of education and experience

The techniques described in section 2 were applied to the Albany forecast data for the full set of nine semesters. The forecasts were screened to ensure that a consensus forecast existed for each of the following forecaster categories for each day included in the test sample: faculty/staff, undergraduate, high experience, and low experience. The experience level of the individual forecasters was subjectively assessed for each semester by LFB. Essentially, this process involved an evaluation of the forecaster's relative familiarity with the map room products, general forecasting experience (at Albany and elsewhere), and intellectual curiosity about the weather. In this regard, it should be noted that the distinction between education and experience is not precise. For example, students can gain experience through the study of particular cases in a synoptic lab. This caveat aside, the stratification and screening resulted in approximately 200 forecasts for each of the four forecast categories (minimum and maximum temperature, 12- and 24-h POP).

As an initial test on the significance of education versus experience in determining forecast skill, a one-way analysis of variance (ANOVA) was performed. The results, which are summarized in Table 3, indicate that differences in mean absolute error between high- and low-experience forecasters are statistically significant at the 95% level for both types of temperature forecasts, but the differences are not significant when stratified according to educational level (faculty/staff versus undergraduate). The differences in the forecast 12-h POP mean error squared are also significant at the 95% level, when stratified according to the level of forecaster experience. Again, the differences are not significant when stratified according to educational level. For 24-h POP forecasts, the differences are not significant at the 95% level for either group; however, the significance of the experience stratification falls just below the 90% level, suggesting that there may be real differences between the experience groups for this variable. The ANOVA results provide strong evidence (subject to the caveats discussed at the end of section 1b) that it is forecaster experience rather than education (beyond some meteorological minimum that we have not attempted to quantify here) that is the decisive factor in forecast skill. To gain further insight into these differences, the skill score decompositions described in section 2 were applied to each forecast variable and are presented in Tables 4–7.

TABLE 3. One-way ANOVA results (experience versus education).

Group pair	*F* test	*P* value	Conclusion
Minimum temperature (217 cases)			
High–low experience	16.127	0.0001	Significant at 95% level
Faculty/staff–undergraduates	2.185	0.1408	Not significant at 95% level
Maximum temperature (208 cases)			
High–low experience	6.230	0.0133	Significant at 95% level
Faculty/staff–undergraduates	3.775	0.0534	Not significant at 95% level
12-h probability of precipitation (196 cases)			
High–low experience	6.219	0.0135	Significant at 95% level
Faculty/staff–undergraduates	1.173	0.2801	Not significant at 95% level
24-h probability of precipitation (199 cases)			
High–low experience	2.646	0.1054	Not significant at 95% level
Faculty/staff–undergraduates	0.422	0.5167	Not significant at 95% level

For minimum temperature (Table 4), the results show that skill levels are uniformly high and that forecast bias is minimal. While distinctions are relatively small, the skill advantage attained by the highly experienced forecasters (with respect to the less experienced forecasters) appears to be primarily due to their ability to make better linear use of the available forecast information (G); in other words, their ''mental model'' of the systematic relationship between the forecast cues and the observed event better matches that of the observations. Overall, the SUNYA consensus, representing the simple average of all individual forecasts on a given day, fares about the same as both the high-experience and faculty/staff forecast classes, and somewhat better than the National Weather Service Forecast Office (WSFO). The advantage over the WSFO appears to be largely the result of the residual term of (5), representing a nonlinear usage of available information as discussed above. Thus, it would appear that these nonlinearities are shared to some degree by all of the forecasters that make up the SUNYA consensus, since it is preserved and enhanced by the group. Finally, we note that the skill distinction between the faculty/staff and undergraduate subgroups is relatively small, as intimated by the ANOVA results discussed above.

In principle, examination of the normalized regression weights of the high- and low-experience MLR models would yield insight into how emphasis on particular cues lead to higher skill. However, because of

TABLE 4. Decomposition of skill score–minimum temperature (217 cases).

Forecaster	Skill score	r_{fo}^2	Conditional bias	Unconditional bias
Faculty/staff	0.909	0.914	0.003	0.002
Undergraduates	0.900	0.902	0.001	0.001
High experience	0.908	0.912	0.002	0.002
Low experience	0.890	0.892	0.001	0.001
SUNYA consensus	0.905	0.907	0.001	0.001
Albany WSFO	0.890	0.895	0.002	0.003
NGM MOS	0.860	0.871	0.006	0.005

Forecaster	r_{fo}	G	$R_{f\|x}$	$R_{o\|x}$	C	Linear term	Residual term	R_d^*
Faculty/staff	0.956	0.982	0.982	0.974	0.389	0.939	0.017	0.593
Undergraduates	0.950	0.974	0.984	0.974	0.389	0.943	0.007	0.534
High experience	0.955	0.980	0.983	0.974	0.412	0.938	0.017	0.585
Low experience	0.944	0.971	0.983	0.974	0.350	0.930	0.014	0.480
SUNYA consensus	0.953	0.978	0.969	0.974	0.290	0.923	0.030	0.562
Albany WSFO	0.946	0.977	0.979	0.974	0.309	0.932	0.014	0.497

* Note: R_d is the correlation between forecast and observed deviations from NGM MOS.

the high correlations between the forecast cues (e.g., temperature and dewpoint temperature) noted in section 2c, it is difficult in practice to isolate this information. Given high intercorrelation, one cue may easily be substituted for a different one with little loss of forecast information, complicating interpretation.

Table 4 also shows that there are substantial differences in the proportion of the variance of the observed deviations from MOS that are accounted for by the high- and low-experience forecast groups (34% and 23%, respectively). These results suggest that skill in predicting observed deviations from MOS is closely tied to overall skill ranking, a conclusion that holds true for all of the subcategories listed in Table 4. This is not surprising, in view of the fact that the MOS approach seeks to minimize the MSE by correcting for the systematic bias in the numerical model forecasts taken as a whole; in doing so, it does not necessarily account for bias that may be specific to certain synoptic situations. In other words, the optimal ''weighting'' of certain cues may vary, depending upon the pattern. Consider the case of overnight warm advection (850-hPa temperatures increasing). The presence of snow on the ground may result in a surface-based inversion, and the stability of the layer may prevent the warmer air aloft from mixing down to the surface. However, in the case of bare ground, the inversion layer is less likely to form, and the warm advection may result in warming surface temperatures as well. Thus, in these two cases, the cue weight for 850-hPa temperatures may be dependent upon snow on the ground.

Accordingly, we have screened the data to examine in some detail the flow patterns associated with successful forecasts of deviations from MOS. This was accomplished by searching for those dates for which consensus forecasts and the corresponding observations both differed in the same sense from MOS by at least 2.8°C (5°F). These dates were further subdivided according to whether the observed temperatures were colder or warmer than the MOS forecast. Composite surface and 500-hPa analyses were constructed from the NMC compact disc dataset fields (Mass et al. 1987) corresponding to the identified dates at 1200 UTC (the end of the verification period).

The composite fields for those dates in which the observed minimum temperatures were at least 2.8°C (5°F) warmer than MOS (and for which consensus forecasts a correspondingly large deviation from MOS), representing 11 cases, are presented in Fig. 2. The pattern shows that upper-level flow was primarily zonal (Fig. 2a), while the surface conditions in the northeastern United States on those dates were dominated by strong anticyclones (Fig. 2b); a strong thickness gradient with the coldest thicknesses poleward of Albany is also apparent. Clearly, the composite cold anticyclone moved across the region during the overnight period, a conclusion that is consistent with the observed composite 1.4°C fall in the 850-hPa temperature during the 12 h of the verification period (0000 UTC to 1200 UTC). Composite temperatures at the surface fell from 4.7°C at 0000 UTC to a minimum of −3.3°C during this period, compared to the −7.3°C minimum forecast by MOS. In contrast, consensus forecasts a minimum of −3.8°C; consensus ''knows'' that cold air does not infiltrate very quickly into Albany during cold-air outbreaks, due both to the presence of the Hudson River valley and the westerly fetch (at 850 hPa) across the warmer waters of the lower Great Lakes upstream.

The composite fields for those dates in which the observed minimum temperatures were at least 2.8°C (5°F) colder than MOS (and for which consensus forecast a correspondingly large deviation from MOS), representing eight cases, are presented in Fig. 3. The

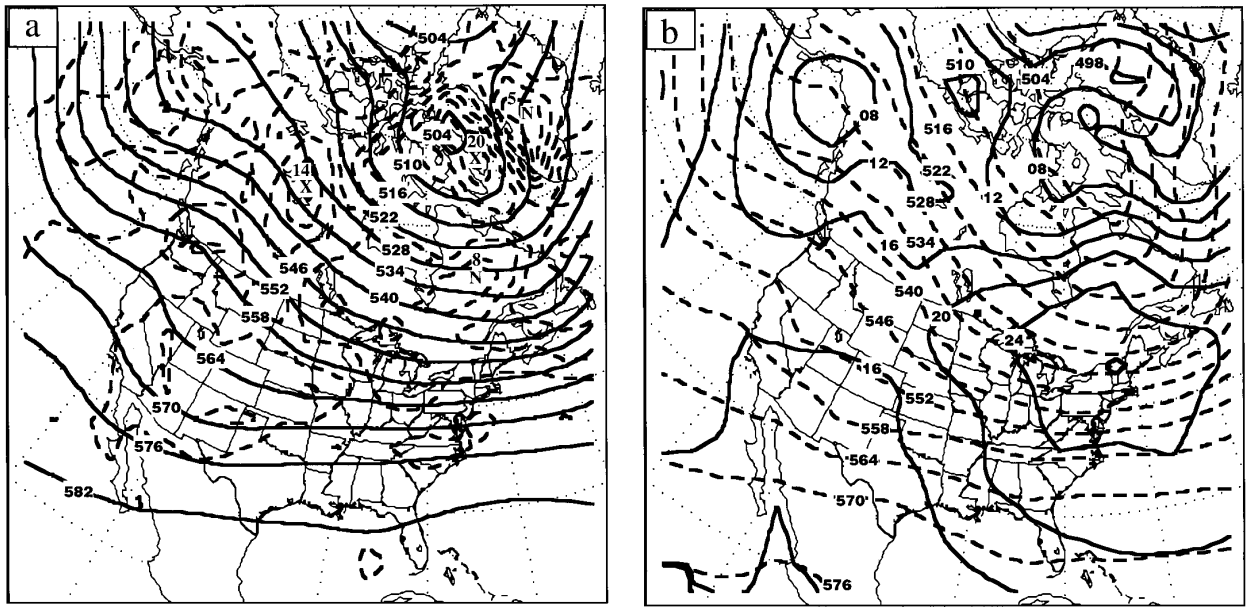

FIG. 2. Warm deviation from MOS composite for 12-h minimum temperature. Shown are (a) the 500-mb height (solid) and geostrophic absolute vorticity (dashed) fields and (b) mean sea level pressure (solid) and 1000–500-mb thickness (dashed) fields. Contour interval of absolute vorticity is 2×10^{-5} s^{-1}; only selected values are labeled for clarity.

pattern shows a composite 500-hPa short wave upstream of Albany by period end (Fig. 3a). The weather in the vicinity of Albany for these dates was dominated by receding surface anticyclones, with zonally oriented 1000–500-hPa thicknesses (Fig. 3b). This pattern is suggestive of strong radiational cooling, a statement that is supported by the composite observations at 0000 and 1200 UTC. The composite surface temperature (dewpoint) was 5.2°C (1.2°C) at 0000 UTC, falling to −0.5°C (−0.8°C) by 1200 UTC, even as 850-hPa temperatures rose from −0.3°C to +3.1°C (in association with warm advection around the backside of the anti-

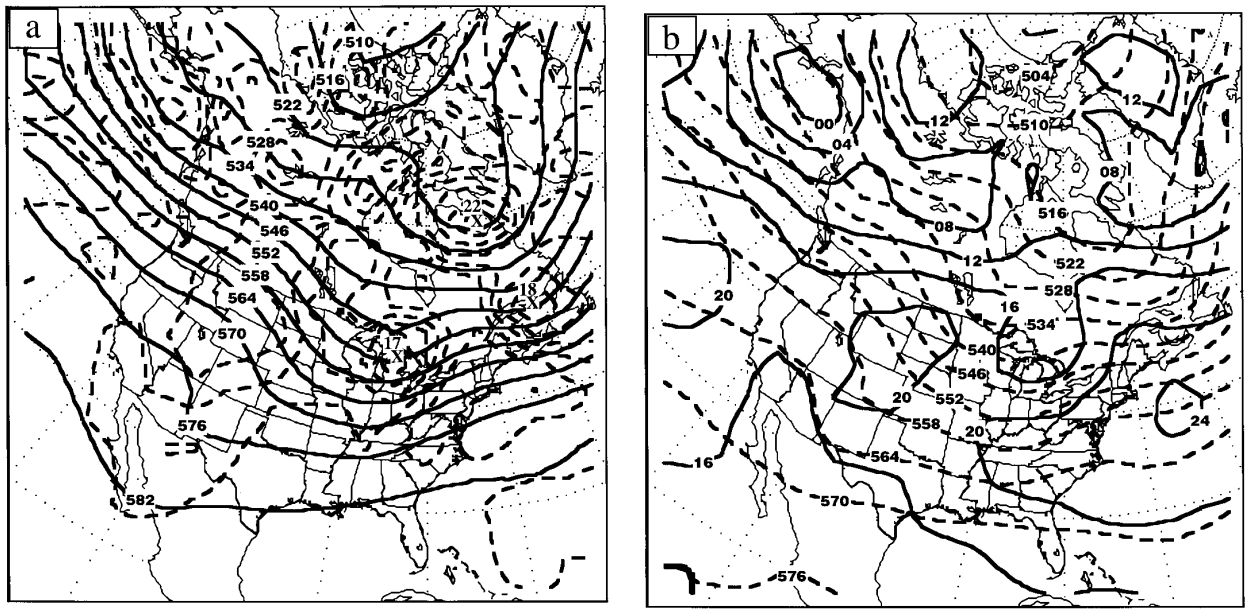

FIG. 3. Cold deviation from MOS composite for 12-h minimum temperature. Plotting convention corresponds to that of Fig. 2.

cyclone). We conclude that consensus ''knows'' that surface radiational cooling will be significant in these cases (even without snow cover, as was the case in seven instances), to the extent that forecasters predict a minimum temperature below the 0000 UTC dewpoint temperature.

These two composites represent a total of 19 of the 217 cases that compose the minimum temperature dataset. Consensus accounts for 32% of the variance of the observed deviations from MOS over all 217 cases. However, when these 19 cases are excluded, consensus accounts for only 18% of the variance. Thus, a few simple forecast rules that can be learned through experience (i.e., do not rush in the cold air, but once it has arrived the minimum temperature can fall below the evening dewpoint on a radiational cooling night) can contribute substantially to the overall skill of the minimum temperature forecasts.

Skill levels are also quite high for maximum temperature forecasts, and bias is small (Table 5). Here, the advantage gained by the highly experienced forecasters with respect to their less-experienced colleagues appears to be the result of nonlinear use of the cue information, represented by the residual term in (5). We note, however, that this nonlinear usage is retained by consensus, which again suggests that the nonlinearity is at least shared in some sense by all of the forecasters. Consensus demonstrates the highest skill for maximum temperature; that advantage appears to be largely due to enhanced consistency in the linear use of cue information ($R_{f|x}$). This consistency implies that given the same set of cues, the forecaster would essentially make the same prediction. Deviations from such perfect consistency are typical of human forecasters and reflect all of the day-to-day vagaries of operational experience. Such deviations, if largely random, could be expected to cancel among a group of forecasters, as appears to be the case here. At the same time, the preservation of the nonlinear contribution to forecast skill suggests that forecasters generally know when such a nonlinear deviation is warranted and forecast accordingly. The skill distinction between the SUNYA consensus and the Albany WSFO is largely due to this linear inconsistency in the latter group of forecasters. Again, we note that the skill differences between the faculty/staff and undergraduate subgroups is small.

In an effort to further investigate these issues, the compositing procedure discussed above was repeated for the maximum temperature forecasts. The composite fields for those dates in which the observed maximum temperatures were at least 2.8°C (5°F) warmer than MOS (and for which consensus forecast a correspondingly large deviation from MOS), representing 11 cases, are presented in Fig. 4 (valid for 1200 UTC, the beginning of the verification period). The composite fields again show zonal 500-hPa flow (Fig. 4a), with surface conditions dominated by high pressure in the wake of a frontal trough (Fig. 4b). Observations indicate that the composite temperature rose quickly from a minimum of 1.6°C to a high of 12.9°C, while dewpoint temperatures remained steady at −1.5°C. This radiative warming may also have been enhanced by warm advection on the backside of the anticyclone, with composite 850-hPa temperatures rising 3.5°C during the verification period.

The composite fields for those dates in which the observed maximum temperatures were at least 2.8°C (5°F) colder than MOS (and for which consensus forecasts a correspondingly large deviation from MOS), representing 10 cases, are presented in Fig. 5 (valid time 1200 UTC). For these dates, cyclonic flow with associated warm advection was affecting Albany (Fig. 5b). However, the upper-level flow (Fig. 5a) is consistent with the notion that the composite cyclone

TABLE 5. Decomposition of skill score–maximum temperature (208 cases).

Forecaster	Skill score	r_{fo}^2	Conditional bias	Unconditional bias
Faculty/staff	0.880	0.883	0.002	0.002
Undergraduates	0.873	0.876	0.000	0.003
High experience	0.883	0.885	0.001	0.001
Low experience	0.867	0.870	0.000	0.003
SUNYA consensus	0.907	0.909	0.000	0.002
Albany WSFO	0.863	0.865	0.001	0.001
NGM MOS	0.859	0.862	0.003	0.000

Forecaster	r_{fo}	G	$R_{f\|x}$	$R_{o\|x}$	C	Linear term	Residual term	R_d^*
Faculty/staff	0.940	0.994	0.962	0.948	0.379	0.907	0.033	0.505
Undergraduates	0.936	0.992	0.966	0.948	0.328	0.908	0.028	0.474
High experience	0.940	0.993	0.964	0.948	0.384	0.907	0.033	0.510
Low experience	0.933	0.992	0.965	0.948	0.304	0.908	0.025	0.453
SUNYA consensus	0.953	0.995	0.984	0.948	0.432	0.928	0.025	0.592
Albany WSFO	0.930	0.992	0.955	0.948	0.338	0.898	0.032	0.471

* Note: R_d is the correlation between forecast and observed deviations from NGM MOS.

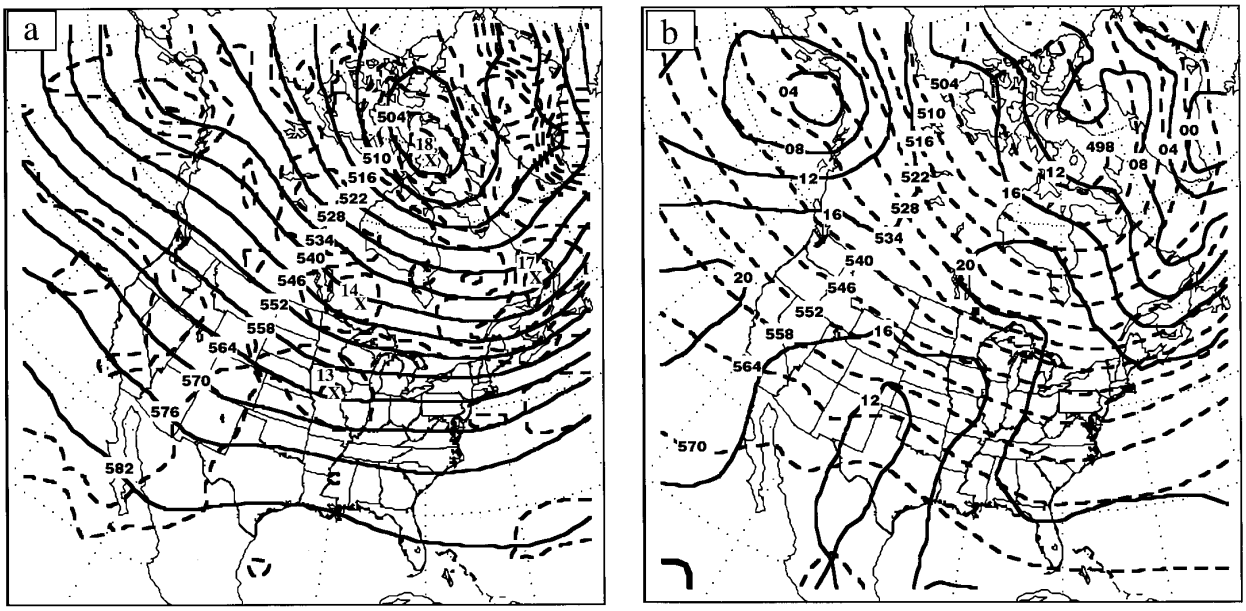

FIG. 4. Warm deviation from MOS composite for 24-h maximum temperature. Plotting convention corresponds to that of Fig. 2.

would propagate quickly through the region, and the observations show that composite values of precipitable water and 850 hPa temperatures had fallen by 0000 UTC. Nevertheless, measurable precipitation occurred at Albany in 80% of these cases, indicating that the forecast period was marked by cyclonic vorticity advection (Fig. 5a), warm advection (Fig. 5b) and cloudiness early in the period, and clearing skies and colder temperatures by the end of the period. The observed composite maximum was 9.2°C, falling to 3.8°C by 0000 UTC, compared to the MOS forecast high of 14.9°C; composite consensus maximum temperatures were 10.9°C, indicating that the aggregate forecast was able to properly account for the combined effects of reduced solar insolation (clouds) and cold advection late in the period. These two composites represent a

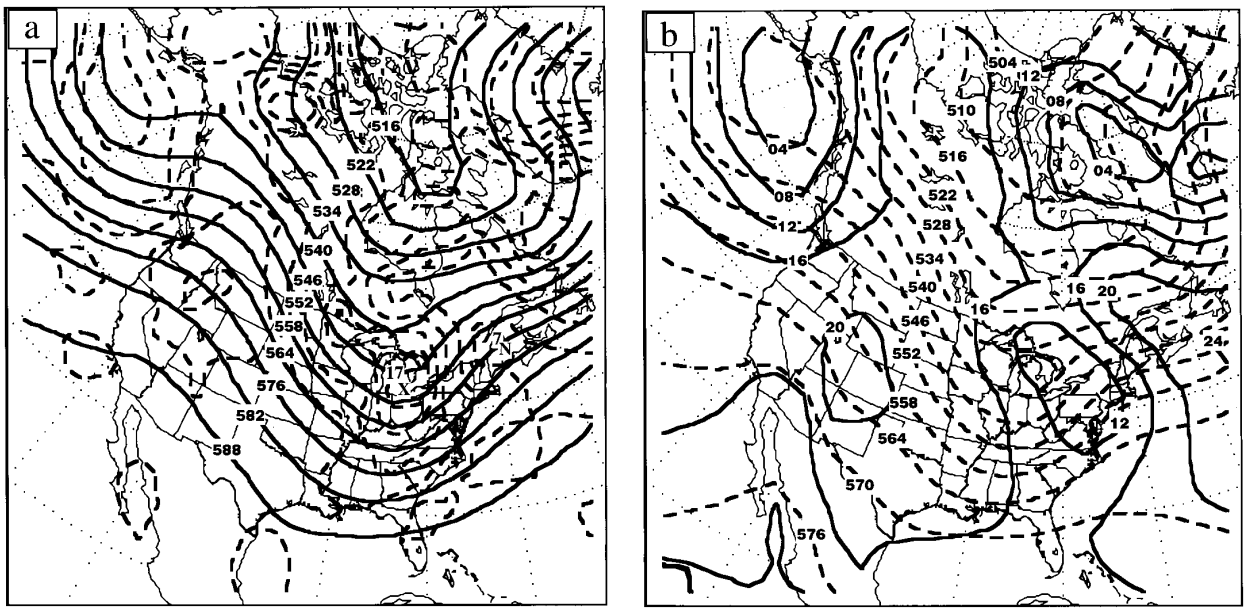

FIG. 5. Cold deviation from MOS composite for 24-h maximum temperature. Plotting convention corresponds to that of Fig. 2.

TABLE 6. Decomposition of skill score–12-h POP (196 cases).

Forecaster	Skill score	r_{fo}^2	Conditional bias	Unconditional bias
Faculty/staff	0.473	0.508	0.000	0.035
Undergraduates	0.445	0.468	0.002	0.020
High experience	0.489	0.512	0.000	0.023
Low experience	0.429	0.454	0.004	0.021
SUNYA consensus	0.547	0.561	0.001	0.013
Albany WSFO	0.545	0.561	0.001	0.016
NGM MOS	0.482	0.496	0.002	0.012

Forecaster	r_{fo}	G	$R_{f\|x}$	$R_{o\|x}$	C	Linear term	Residual term	R_d^*
Faculty/staff	0.713	0.908	0.901	0.796	0.232	0.651	0.062	0.299
Undergraduates	0.684	0.911	0.887	0.796	0.143	0.643	0.041	0.235
High experience	0.716	0.913	0.900	0.796	0.231	0.654	0.062	0.304
Low experience	0.673	0.906	0.884	0.796	0.125	0.638	0.035	0.211
SUNYA consensus	0.749	0.920	0.960	0.796	0.266	0.703	0.046	0.360
Albany WSFO	0.749	0.921	0.946	0.796	0.284	0.694	0.055	0.365

* Note: R_d is the correlation between forecast and observed deviations from NGM MOS.

total of 21 of the 208 cases that compose the maximum temperature dataset. Consensus accounts for 35% (R_d^2) of the variance of the observed deviations from MOS over all 208 cases. When these 21 cases are excluded, consensus accounts for only 13% of the variance. Thus, for maximum temperature (as with minimum temperature), some simple forecast rules (radiative effects are important in forecasting maximum temperature: dry air warms rapidly, while cloud cover drastically reduces the warming that might otherwise be expected) can contribute substantially to overall forecast skill.

The 12-h POP forecast results (Table 6) show considerably reduced skill relative to the temperature forecasts (Tables 4 and 5). In contrasting the high- and low-experience forecasters, it is noteworthy that although conditional bias is somewhat larger for the latter group, the overall bias remains quite small for both sets of forecasters. The apparent skill advantage of the highly experienced forecasters appears to be due to

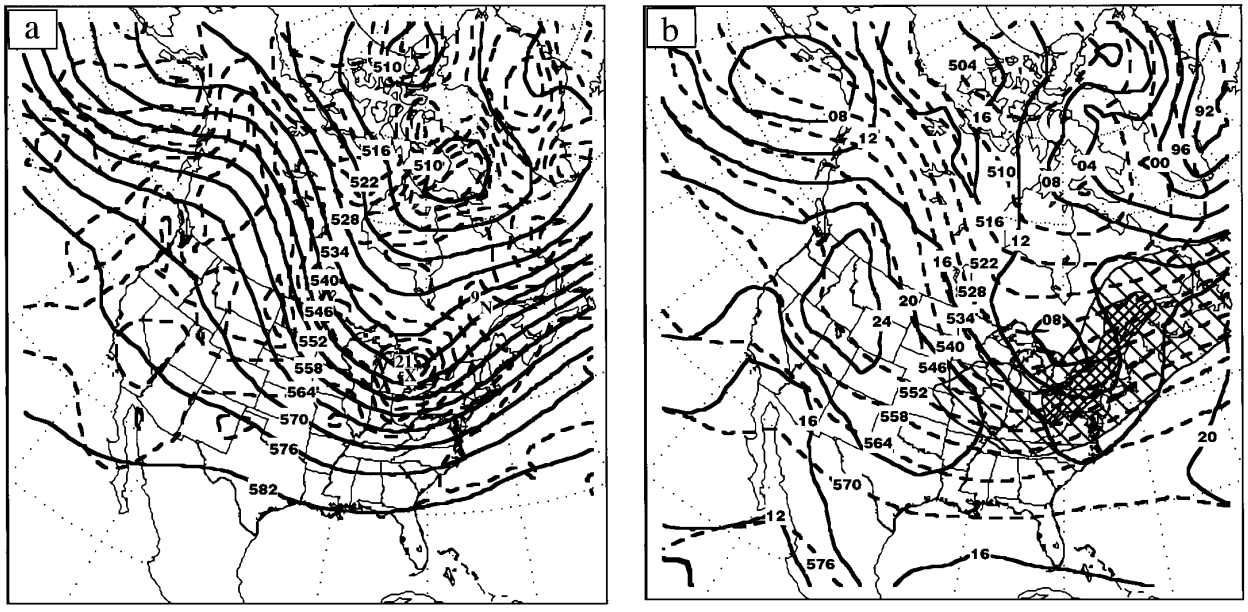

FIG. 6. Wet deviation from MOS composite for 12-h POP. Plotting convention corresponds to that of Fig. 2. In (b) the light (heavy) hatching corresponds to 250-hPa winds above 35 (45) m s^{-1}.

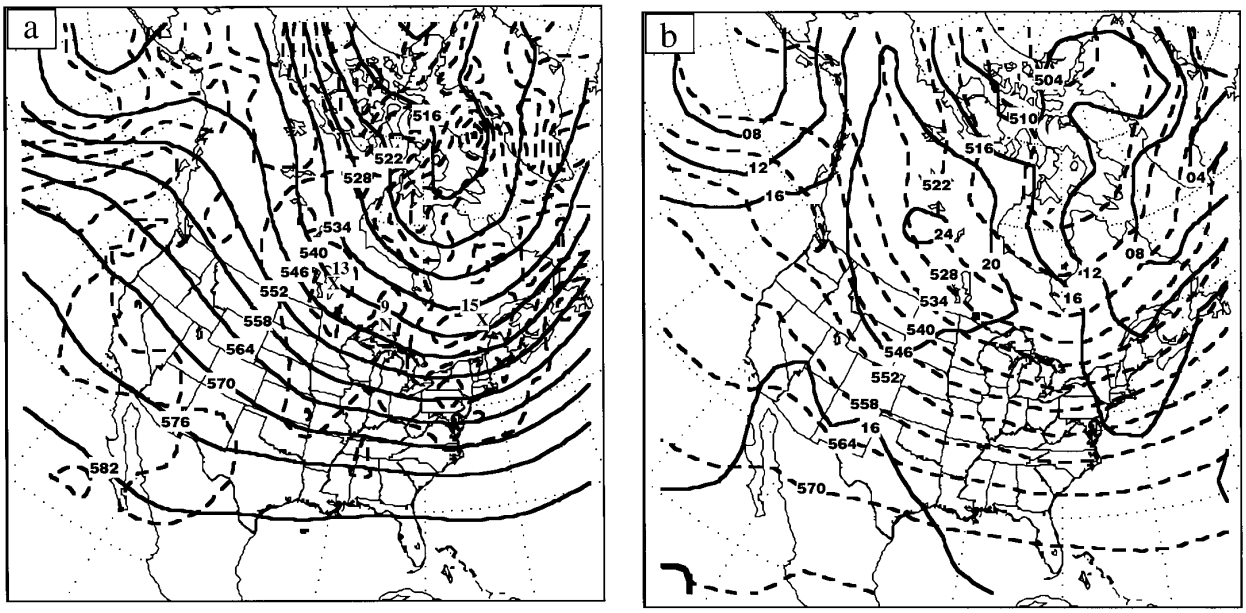

FIG. 7. Dry deviation from MOS composite for 12-h POP. Plotting convention corresponds to that of Fig. 2.

both better linear use of the available information (through enhanced forecast consistency, $R_{f|x}$) and the recognition of important nonlinear information in the cues. We note that SUNYA consensus again displays the highest forecast skill, which is the result of better linear forecast consistency. However, most of the skill advantage that was displayed by the SUNYA forecasters with respect to the WSFO for temperature is not evident for this variable. Although the skill differences between the faculty/staff and undergraduate subgroups have increased (largely due to better nonlinear use of information), these differences are still not statistically significant (Table 3).

Composites of surface and 500-hPa fields associated with skillful consensus deviations from MOS were again constructed. Figure 6 shows the composite surface and 500-hPa analyses (valid for 1200 UTC, the end of the verification period) for those cases in which both consensus and the observed POP were at least 20% higher than MOS (and thus represent measurable precipitation events, with a sample size of 5). The composite flow is characterized by large-scale geostrophic southwesterlies at 500 hPa, with the main axis of the jet lying just poleward of Albany. A short wave is embedded in the jet upstream, such that 500-hPa cyclonic geostrophic vorticity advection, providing midtropospheric forcing of ascent, is maximized in southern Quebec. At 1200 UTC, Albany was positioned just downstream of a frontal trough; with baggy isobars south of New England, there is the suggestion of composite cold-air damming, resulting in augmented warm advection, further contributing to the midtropospheric forcing of ascent. More importantly, Albany's position near the right-front quadrant of the upper-level jet (Fig. 6b) yields substantial upper-level forcing of ascent, a situation that persists throughout the 12-h verification period. Composite surface temperatures cooled only slowly, from 6.1°C at 0000 UTC to an overnight minimum of 3.8°C; cooling was restricted by the combination of warm advection, rising dewpoints, and cloud cover. Thus, the cues taken as a whole are suggestive of the potential for measurable precipitation. As a consequence, the consensus forecast POP was 63%, compared to 41% for MOS. The forecast rule here is somewhat more subtle, but the essential implication is clear: one must consider all the elements that contribute to atmospheric vertical motions and not focus only on 500-hPa cyclonic vorticity advection.

Similar composites were constructed (Fig. 7) for those cases in which both consensus and the observed POP were at least 20% lower than MOS (and thus represent nonprecipitating events, with a sample size of 12). Here, the flow was characterized by westerly geostrophic flow at 500 hPa, with no clearly defined jet-level forcing of ascent in the vicinity of Albany. Furthermore, midtropospheric forcing of ascent was short lived. At the surface, a composite cold-frontal trough was passing Albany at 1200 UTC. However, moisture was less abundant for these cases, with peak precipitable water values in the verification period of 13.5 mm (30% less than for the composites of Fig. 6). The consequence of these combined effects was to reduce the consensus POP forecast to 23% from the MOS value of 45%. Our suspicion is that these cases likely represent a timing problem, wherein consensus has determined that any precipitation that might occur will end

by the beginning of the forecast period; our dataset is insufficient to examine 6-h MOS values, but we speculate that the bulk of the high MOS POP was derived from the first 6 h of the 0000–1200 UTC verification period. When the 17 cases that compose these two 12-h POP composites are excluded, consensus accounts for less than 1% of the variance of the observed deviations from MOS (compared to 13% over all 196 cases). Thus, careful consideration of some of the subtle interrelationships inherent in forecasting precipitation can contribute to overall forecast skill. However, it is also clear that these gains are relatively smaller than those obtainable for temperatures, a result of the generally lower skill of precipitation forecasts.

The decomposition results for 24-h POP are quite similar to those for the shorter-range POP forecasts (Table 7). Overall, the Albany WSFO finally surpasses the SUNYA consensus, a distinction largely due to a very small advantage in the residual term of (5). We note that both the WSFO and consensus show a sizable skill advantage with respect to the rest of the groups, gains that are realized through higher linear consistency ($R_{f|x}$). In comparing the high- and low-experience forecaster groups, we see that the relative skill advantage of the former group is fairly small (and perhaps not statistically significant as noted earlier). Likewise, distinctions between the faculty/staff and undergraduate groups are virtually nil. The results of Table 7 show that only the SUNYA consensus and WSFO forecasts showed higher skill than MOS. Because of this general loss of skill, it is more difficult to isolate cases of appropriate consensus deviations from MOS (which account for only about 5% of the variance of the observed deviations); in general, the sample sizes are smaller and the results are less robust.

The composite for those cases in which precipitation occurred and consensus was at least 20% higher than MOS composes four events. The composite is not surprisingly quite similar to that of Fig. 6 (for 12-h POP), with a strong southwesterly flow (in the base of a trough) at 500 hPa and the main short wave positioned in such a way that it will pass to the north of Albany. However, with a frontal trough to the west, there is the suggestion that measurable precipitation at Albany is still possible in association with layer warm advection and weak vorticity advection, given the available moisture (approximately 14-mm precipitable water). The consensus forecast (50%) reflects the level of uncertainty that would be inherent 24 h previously, given the relative position of the main forcing of ascent. However, this forecast is well above MOS for the same set of events (28%), suggesting a greater reliance on the part of the human forecasters on frontal precipitation in southwest flow.

The composites for the nonverifying cases for which consensus was at least 20% less than MOS (eight events, valid at 1200 UTC) are shown in Fig. 8. The 500-hPa flow exhibits a fairly strong geostrophic westerly jet over Albany, within which weak short waves are embedded; the closest of these is near the upper Great Lakes at 1200 UTC. With a surface low pressure center to the west, a pattern of geostrophic warm advection is present; this is the primary source of midtropospheric ascent forcing for this set of cases. Measures of precipitable water indicate that sufficient moisture is present for precipitation (16.6 mm at 1200 UTC). However, the weather on these days appears to have been dominated by warm sector conditions, with temperatures rising almost 17°C from the morning minimum under relatively cloud-free skies. Consensus demonstrated greater skepticism concerning the possibility of measurable precipitation in these cases, perhaps recognizing that the warm advection precipitation band is narrow and mostly poleward and westward of Albany,

TABLE 7. Decomposition of skill score–24-h POP (199 cases).

Forecaster	Skill score	r_{fo}^2	Conditional bias	Unconditional bias
Faculty/staff	0.373	0.375	0.001	0.001
Undergraduates	0.357	0.363	0.006	0.000
High experience	0.372	0.373	0.001	0.000
Low experience	0.334	0.343	0.009	0.000
SUNYA consensus	0.488	0.490	0.002	0.000
Albany WSFO	0.503	0.503	0.000	0.000
NGM MOS	0.469	0.470	0.001	0.000

Forecaster	r_{fo}	G	$R_{f\|x}$	$R_{o\|x}$	C	Linear term	Residual term	R_d^*
Faculty/staff	0.612	0.943	0.880	0.740	0.006	0.610	0.002	0.082
Undergraduates	0.603	0.937	0.879	0.740	0.020	0.597	0.006	0.065
High experience	0.611	0.938	0.871	0.740	0.018	0.605	0.006	0.096
Low experience	0.585	0.933	0.882	0.740	−0.074	0.608	−0.023	0.016
SUNYA consensus	0.700	0.948	0.953	0.740	0.156	0.668	0.032	0.222
Albany WSFO	0.709	0.962	0.936	0.740	0.183	0.666	0.043	0.278

* Note: R_d is the correlation between forecast and observed deviations from NGM MOS.

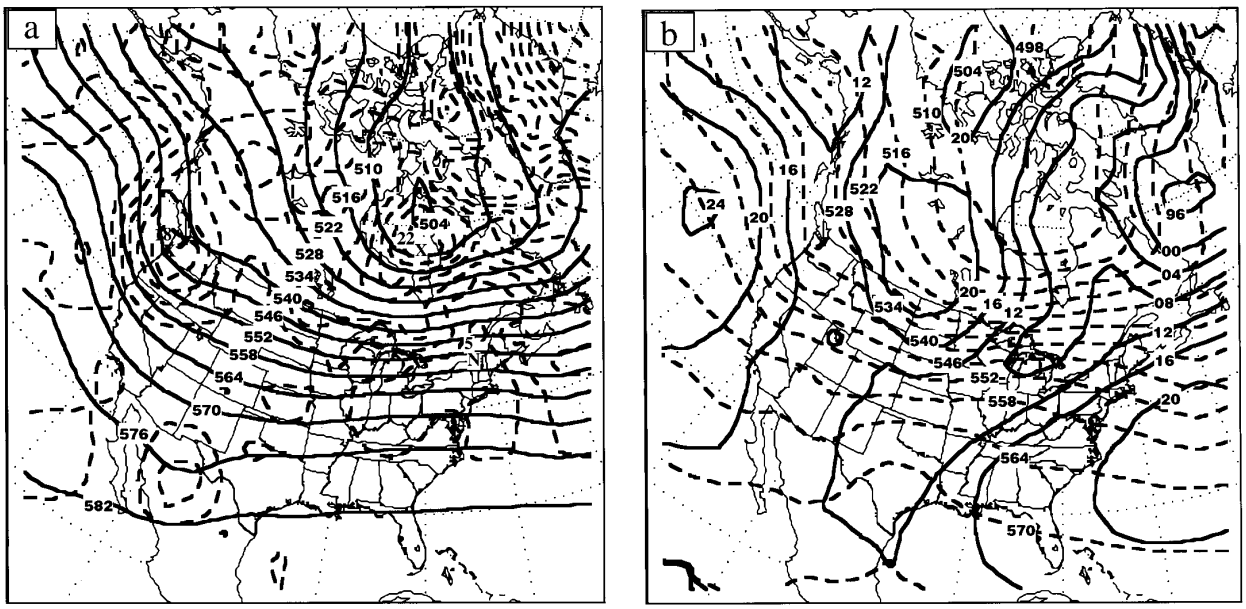

FIG. 8. Dry deviation from MOS composite for 24-h POP. Plotting convention corresponds to that of Fig. 2.

with a composite POP forecast of 21% (compared to 43% for MOS). Overall, the set of 12 24-h POP events accounts for virtually all of the consensus skill with respect to MOS.

In summary, the data suggest that it is forecast experience rather than educational level, as stated previously, that correlates with weather forecasting skill. Nevertheless, it would be erroneous to conclude that meteorological theory is irrelevant to the forecast problem. Rather, the theory provides a background framework within which the forecaster operates (for example, in the development, use and interpretation of numerical models). The advantage that is gained through forecast experience is the result of a combination of better linear and nonlinear use of forecast information. In short, the experienced forecaster remains largely consistent from day to day in a linear sense but also better recognizes when significant nonlinear deviations in the forecast are warranted by the available information (for example, in the recognition of specific synoptic patterns that may enhance radiational effects on temperature beyond that that is quantified by MOS). The greater skill of consensus forecasts relative to most of the subgroups of the study (experience, education, WSFO) is due to the greater consistency that is maintained by the aggregate forecast, eliminating the vagaries that plague any individual forecast on a given day.

b. Growth of skill with experience

The SUNYA forecast results were further analyzed to examine the growth of skill with forecasting experience (the learning curve). All forecasts made for the nine total forecast semesters were tabulated according to the number of forecasts made by a particular individual since the first forecast date (6 September 1988). For example, if forecaster A had forecast on 6, 7, and 9 September 1988 (but not on 8 September), then his forecast on 6 September would correspond to $N = 1$, 7 September to $N = 2$ and 9 September to $N = 3$, where N is the number of forecasts made by that forecaster. A second forecaster, who did not begin making forecasts until 12 September would have the forecast for that day tabulated as $N = 1$. All forecast errors (mean absolute error for temperature, squared error for precipitation probability) were computed relative to the complete consensus forecast composed of all forecasters on a given day. [see Sanders (1963) for a discussion of the fundamental ideas.] The forecast data were then stratified according to the subjective experience-level ranking of each forecaster on the date of his or her first forecast. The forecast errors were then compared as a function of the number of forecasts N in order to determine the point at which substantial differences in forecast performance were detectable. Specific skill stages were then identified and are summarized in Table 8.

The results of Table 8 indicate some important distinctions in the growth of skill between the low- and high-experience forecasters. There appears to be an initial ''break-in'' period for low experience forecasters, amounting to the first 5–10 forecasts, during which time forecast errors relative to consensus are high. This suggests that these forecasters, with no prior experience, are effectively learning the logistics of making

TABLE 8. Experience stages (determined as the number of forecasts within each group at which substantial differences in forecast performance are detectable) for minimum and maximum temperature and 12- and 24-h POP. Shown are the subjectively defined experience group, the number of forecasts for that group N, and the mean forecast error defined relative to consensus. This error is mean absolute error for temperature and squared error for POP.

Group	Forecasts	Mean error (consensus forecast)
	Minimum temperature	
Low experience	$N \leqslant 10$	−0.79
Low experience	$10 < N \leqslant 70$	−0.48
Low experience	$N > 70$	−0.38
High experience	$N \leqslant 50$	−0.28
High experience	$N > 50$	−0.09
	Maximum temperature	
Low experience	$N \leqslant 5$	−0.71
Low experience	$5 < N \leqslant 70$	−0.51
Low experience	$N > 70$	−0.42
High experience	$N \leqslant 30$	−0.44
High experience	$N > 30$	−0.06
	12-hour POP	
Low experience	$N \leqslant 8$	−1.77
Low experience	$8 < N \leqslant 100$	−0.88
Low experience	$N > 100$	−0.29
High experience	$N \leqslant 65$	−0.95
High experience	$N > 65$	−0.33
	24-Hour POP	
Low experience	$N \leqslant 10$	−2.35
Low experience	$10 < N \leqslant 70$	−0.96
Low experience	$N > 70$	−0.35
High experience	ALL	−0.49

temperature and precipitation forecasts. Following the ''break-in'' period, the low-experience forecasters enter a second stage, lasting from 70 to 100 forecasts, during which time the forecasters may be gaining true knowledge of the local meteorology; following this second stage, forecast errors relative to consensus are significantly smaller. In contrast, the high-experience forecasters do not show any ''break-in'' period but rather exhibit an abbreviated training period (lasting from 30 from 65 forecasts).

We note that for both the low- and high-experience forecasters, the training period is longer for POP than temperature (except for 24-h POP for the high experience forecasters, for which the data support no evidence of any training period at all). Related to this issue is the result that POP skill levels of both the low- and high-experience groups are roughly equivalent after the training stage, even though the high experience forecasters retain a sizable advantage for the temperature variables. We speculate that these results reflect the relative complexity of forecasting precipitation. At 12 h, there is still room for human ''mesoskill'' in differentiating important factors governing the occurrence of precipitation, but by 24 h, model limits for POP are being reached, and these factors cannot be used. With fewer degrees of freedom to the temperature forecast problem, perhaps the high experience forecasters are able to better isolate the forecast ''signal'' for this variable.

The results of Table 8 also may have some bearing on the movement of personnel within entities such as the NWS. By comparing the forecast errors of highly experienced forecasters before and after the ''training period'' defined above, these data can be used to simulate the temporary loss of skill that such a forecaster could expect to encounter upon moving to a new forecast site. These data indicate that skill with respect to consensus (which we use here as a measure of the state of the art) might fall by amounts on the order of 5% for minimum temperature, 13% for maximum temperature, and 12% for 12-h POP. These skill losses would then be erased (over periods ranging from 30 to 65 forecasts, representing timescales of 1–3 months) as the forecaster becomes ''trained'' for the particular conditions of the new site.

c. Erosion of human skill

To investigate trends in human skill, all forecast days for the period 1988–92 for which consensus forecasts were available were examined. This leads to a total of 575 forecast days. Skill scores for the Albany group consensus with respect to climatology and MOS were then computed, based on a moving window of 30 forecasts for the available set of 575 forecasts. Linear trends were also computed, based upon the skill scores for every set of 30 forecasts. The results of these analyses are presented in Figs. 9–12.

These data suggest that the overall forecast skill remained relatively flat with respect to climatology (Figs. 9a, 10a, 11a, 12a), a basic reference point for human skill. Indeed, at 24 h, the skill trend is slightly negative for both POP and temperature. We note that a similar trend is also apparent in the WSFO forecasts for this period (not shown). A continued loss of skill with respect to the numerical guidance (as represented by MOS) is clearly evident (Figs. 9b, 10b, 11b, 12b), extending the trend noted by Sanders (1986) for the period 1966–84 at Massachusetts Institute of Technology. The numerical guidance continues to improve: the Nested Grid Model (NGM) MOS equations were rederived several times during these nine semesters to reflect a larger data sample and important changes in the model physics after the initial implementation. (The NGM went operational in early 1986 with a crude boundary layer that was later modified.) It would seem that human forecasters are approaching or have already reached an absolute skill plateau that MOS is quickly overtaking (presumably through a thorough diagnosis of the synoptic-scale environment). Whether one

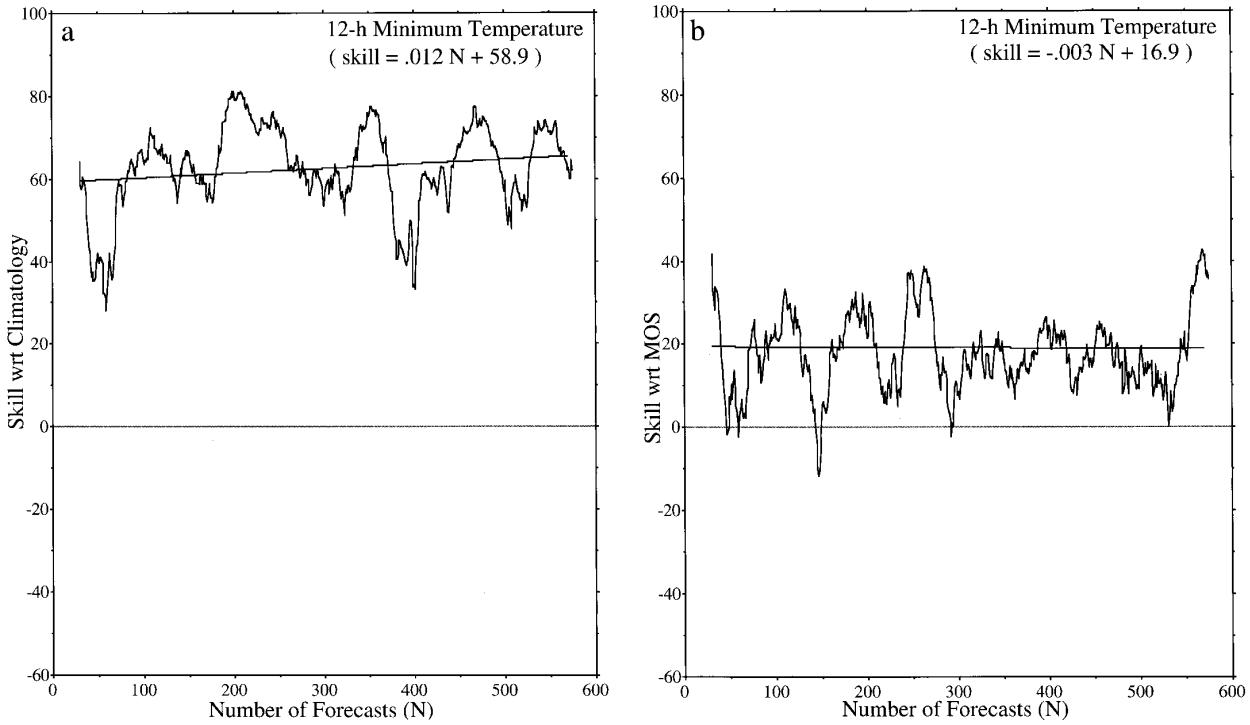

FIG. 9. Overall consensus skill for 12-h minimum temperature for the period 1988–92. Results are for (a) with respect to climatology and (b) with respect to MOS. The regression equations are also shown under the label.

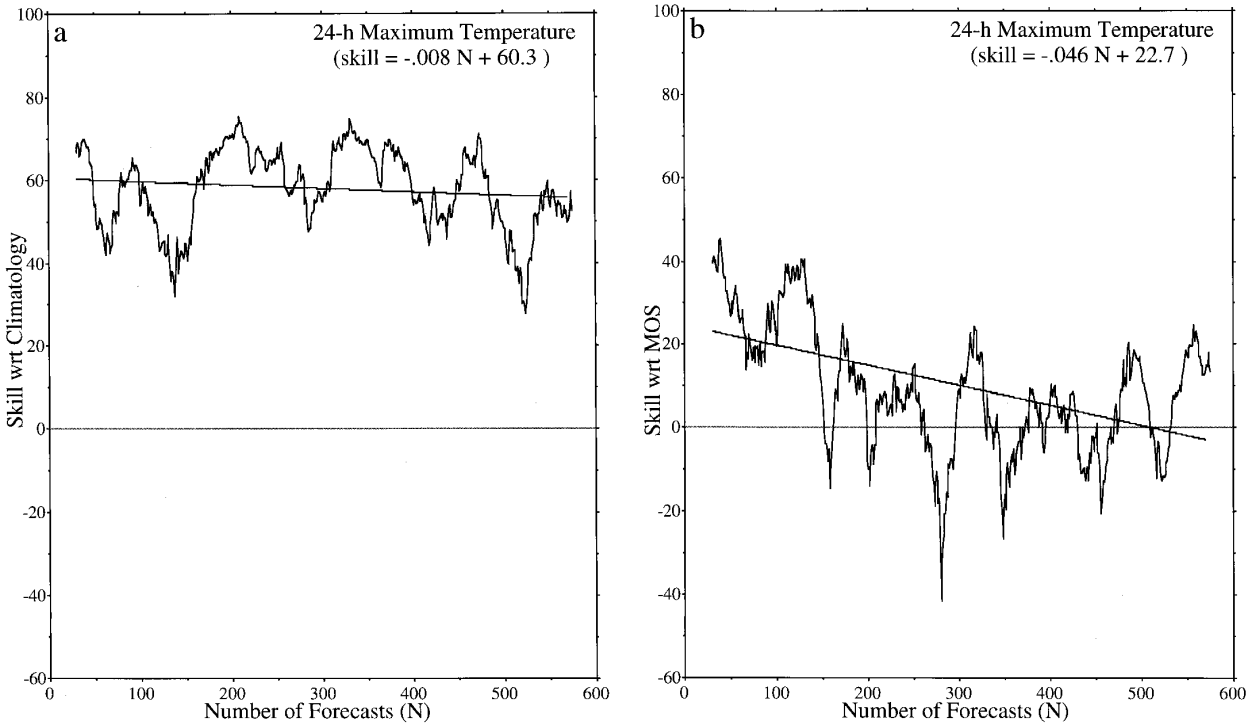

FIG. 10. As in Fig. 9 but for 24-h maximum temperature.

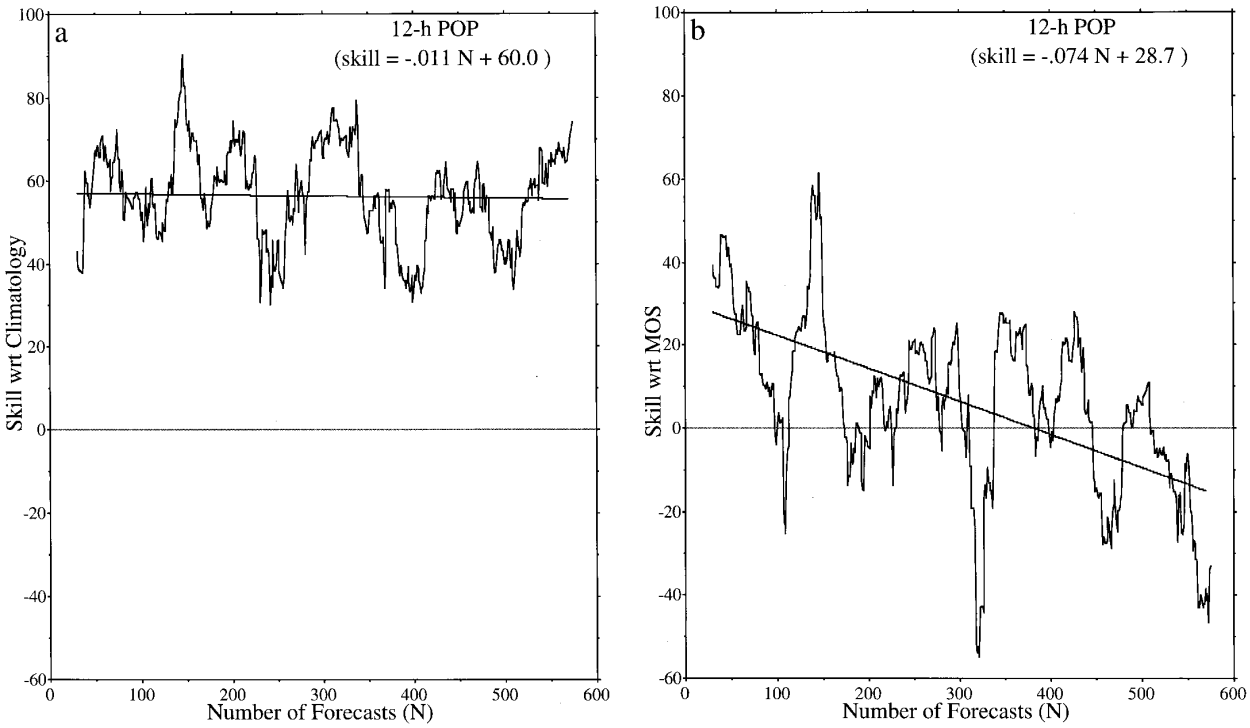

FIG. 11. As in Fig. 9 but for 12-h POP.

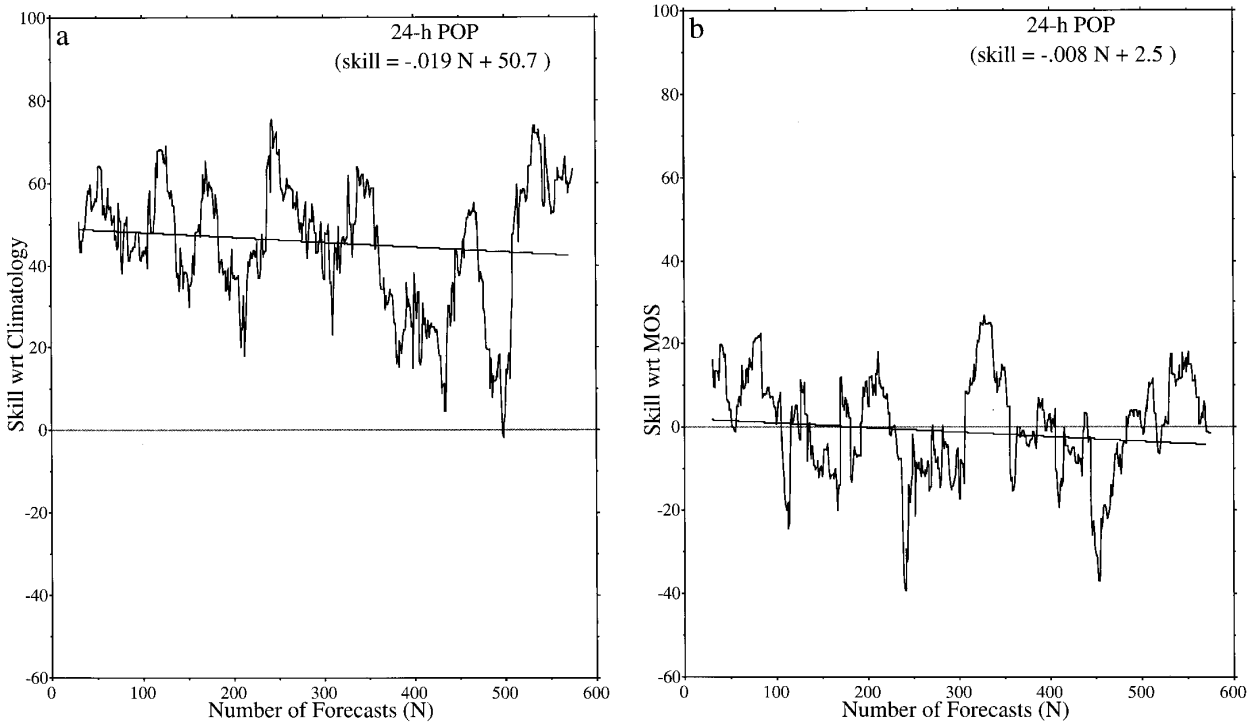

FIG. 12. As in Fig. 9 but for 24-h POP.

chooses to interpret this result as continued degeneration or progress seems to the authors a matter of philosophy; nonetheless, it is abundantly clear that the days of *routine* human forecasting advantage are all but gone. It will be interesting to see whether increased data availability on the mesoscale, such as with Doppler radar, will reintroduce the human advantage for short-range precipitation forecasts. Further discussion of these issues is presented in the next section.

4. Conclusions

The analysis of nine semesters of temperature and precipitation forecasts for Albany, New York, has demonstrated that forecast skill is largely determined by experience. The relative advantage of experienced forecasters appears to be the result of successfully implementing a simple strategy of both maintaining a high level of linear consistency between the information that forms the basis of the forecast (the cues) and the forecast itself and also recognizing those instances in which those linear relationships do not apply. Composite analysis suggests that, at least for Albany, the instances when such nonlinear departures are warranted can often be recognized using simple rules.

Because of the high correlation between forecast cues, it was not possible to isolate how linear emphasis on particular cues might lead to higher skill. Another approach to isolating this information would be to perform a full factor analysis of the cues. This approach would reduce the set of forecast cues from the original 13 for minimum temperature to a smaller set of orthogonal factors that are a linear combination of the original cues. The constraint of orthogonality removes the correlation between the forecast cues (which would now be the factors), and one could then perform the skill decompositions [(5)] using these factors. Such transformations retain the ''coarse-grain'' predictive information contained in the data and discard the ''finescale'' detail that many times can be regarded simply as noise. However, in some circumstances, this finescale detail may actually contain important predictive information, and the transformation will not be useful.

With these ideas in mind, such an approach was attempted for the minimum temperature data. The factor analysis captured most but not all of the variance in the original 13 cues (since about 6.5% of the described variance was lost in the factor analysis). This lost information is unique for each cue; in other words, some of the discarded finescale information is of some predictive use in this application. As a result, the residual term increased, reflecting the loss of the unique cue information that was removed by the factor analysis. Accordingly, we conclude that it is precisely this (unique) finescale cue information that accounts for much of the difference in skill between the high- and low-experience forecasters.

The analysis of section 3b indicated that the growth of skill with experience for initially inexperienced forecasters progresses fairly rapidly through clearly definable stages and reflects their increased ability to implement the basic forecasting strategy (maintain linear consistency except where significant nonlinear departures are warranted). However, the skill advantage of human forecasters with respect to numerical guidance continues to diminish and now largely reflects the human ability to recognize these relatively infrequent (about 10% of the cases studied) departures.

Given this continuing erosion of human forecast skills relative to the MOS forecasts (a surrogate for continually improving numerical weather prediction models), an important problem is how to profitably retain human involvement in the forecast process. One strategy might be to have human forecasters focus only on those weather situations that pose a significant threat to public safety. In this approach groups of highly experienced forecasters at a handful of national forecasting centers can use their expertise to monitor the forecast problem(s) of the day and to provide guidance as needed to public safety officials and other concerned users. Another strategy, complementary to the first, might be to emphasize the importance of nowcasting on the local and regional scale using forecasters highly knowledgeable and experienced with local weather and climate idiosyncracies related to physiographic features. An assumption behind this strategy is that the learning curve for human forecasters will be steeper than for the machine when it comes to the application of new technologies (e.g., Doppler radars, profilers, onsite high-resolution local models with complex terrain) to the practical forecast problem. An additional assumption, again based upon considerable personal experience, is that the number of humans actually capable of interpreting numerical weather forecasts correctly is only a relatively small percentage of the people capable of generating such model forecasts, so that there will always be a need for the human in the forecast process.

Finally, we note the likelihood that human forecaster skills will atrophy with time unless they are used on a regular basis. The continuing convergence of human and machine skill levels appears to be inevitable; nonetheless, means must be found to enable humans to polish and extend their forecasting skills by taking advantage of the very technology that is driving this convergence. To do otherwise is to put the entire forecast enterprise on automatic pilot, a system in which human accountability and responsibility must surely vanish.

Acknowledgments. The authors would like to thank Dr. Thomas Stewart of SUNY/Albany's Rockefeller College for helpful discussions relevant to this paper. We also thank the numerous SUNYA students, past and present, who regularly participated in the daily forecast contests, making research studies such as this one possible. Partial support for this research was provided by National Science Foundation Grant ATM-9114598. Ideas for this research began to germinate through the

continued involvement of the authors in the practical forecast problem, including most recently the STORM-FEST field program of 1992.

REFERENCES

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

——, 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 weather forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

DiMego, G. J., K. E. Mitchell, R. A. Petersen, J. E. Hoke, J. P. Gerrity, J. J. Tuccillo, R. L. Wobus, and H.-M. H. Juang, 1992: Changes to NMC's regional analysis and forecast system. *Wea. Forecasting,* **7,** 185–198.

Firestone, J. K., 1979: Comment on ''Forecasting skill of beginners.'' *Bull. Amer. Meteor. Soc.,* **60,** 1206–1207.

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

——, 1979: Reply to ''Comment on forecasting skill of beginners'' and ''Rebuttal to forecasting skill of beginners.'' *Bull. Amer. Meteor. Soc.,* **60,** 1208–1209.

Glahn, H. R., 1985: Yes, precipitation forecasts have improved. *Bull. Amer. Meteor. Soc.,* **66,** 820–830.

——, and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

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

Mass, C. F., H. J. Edmon, H. J. Friedman, N. R. Cheney, and E. E. Recker, 1987: The use of compact discs for storage of large meteorological and oceanographic data sets. *Bull. Amer. Meteor. Soc.,* **68,** 1556–1558.

Murphy, A. H., 1988: Skill scores based on their mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.,* **116,** 2417–2424.

——, and R. W. Katz, 1985: *Probability, Statistics and Decision Making in the Atmospheric Sciences.* Westview Press, 545 pp.

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

Ramage, C. S., 1978: Bring back analysis. Preprints, *Conf. on Weather Forecasting and Analysis and Aviation Meteorology,* Silver Spring, MD, Amer. Meteor. Soc., 273–274.

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

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

Stewart, T. R., 1990: A decomposition of the correlation coefficient and its use in analyzing forecasting skill. *Wea. Forecasting,* **5,** 661–666.

——, W. R. Moninger, J. Grassia, R. H. Brady, and F. H. Merrem, 1989: Analysis of expert judgment in a hail forecasting experiment. *Wea. Forecasting,* **4,** 24–34.

Trenberth, K. E., 1978: On the interpretation of the diagnostic quasi-geostrophic omega equation. *Mon. Wea. Rev.,* **106,** 131–137.

Wernly, D. R., 1979: Rebuttal to ''Forecasting skill of beginners.'' *Bull. Amer. Meteor. Soc.,* **60,** 1207–1208.