The Regime Dependence of Degree Day Forecast Technique, Skill, and Value

PAUL J. ROEBBER

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

(Manuscript received 16 July 1997, in final form 30 January 1998)

ABSTRACT

An investigation into the manner in which forecasters adjust their reliance on particular pieces of forecast information as the large-scale flow pattern evolves into different regimes, and the relationship between those adjustments and forecast skill and value is presented. For the cold season months (December–February) of the period 1 January 1973 through 31 December 1992, a total of three regime types (identified through cluster analysis) comprising 63% of the days were identified. A framework for investigating the weighting of pieces of forecast information, based upon multiple regression techniques, was applied to National Weather Service (NWS) degree day forecasts (constructed from the 12–24-h minimum and 24–36-h maximum temperature forecasts) for this period. It was determined that substantial changes in the usage of Model Output Statistics (MOS) by NWS forecasters have occurred with the advent of the improved numerical model guidance represented by the Limited Fine Mesh (LFM) MOS, and that these changes occurred in response to improvements in the longer-range forecasts (validating 24–36 h from the initial time). However, it was also shown that this increased weighting of MOS was situation dependent and that forecast skill and value were maintained under large-scale flow regimes in which MOS was less useful through significant adjustment of forecast technique. Overall, skills were found to be lowest for flows in which either the variability of the MOS weight was highest (reflecting uncertainty in its reliability) or in which limitations of that guidance were evident. These results are then related to earlier investigations concerning the relationship between forecast skill and experience.

1. Introduction

The study of meteorological forecast skill has received considerable attention in the literature. Large-scale studies (Kalnay et al. 1990; Shuman 1989) have shown a steady increase in skill in predicting the upper-level flow over the past 30 years. Recent studies by Roebber and Bosart (1996a and 1996b, hereafter denoted by RBa and RBb, respectively) have shown that experienced forecasters can take advantage of regional knowledge to further improve forecasts of both temperature and precipitation and that, in specific circumstances, this human intervention can add considerably to the value of those forecasts. While these and other studies have helped to clarify some outstanding questions regarding the forecast process, an issue that has not yet been addressed is the large-scale regime dependence of forecast technique; that is, given a profile of forecast information (e.g., 850-hPa temperature, dewpoint temperature, surface wind speed, etc., hereafter called cues), what will be the relative weight that a forecaster will assign to a particular cue under different meteorological flows? Such a study has particular relevance under current operational constraints, in which human forecasts are overall only marginally more skillful than that of the numerical–statistical guidance (e.g. RBa), the Model Output Statistics (MOS). Under these conditions, human forecast skill is strongly controlled by the ability to recognize those circumstances in which significant deviations from MOS are warranted. A common example of this process is found in National Weather Service (NWS) forecast discussions, in which forecasters describe the rationale for choosing a ‘‘model of the day’’ and the reasons for accepting or deviating from the guidance of that model.

In this paper, the particular forecast parameter selected for study is the degree day (defined by the 65°F threshold) forecast for Albany, New York (ALB), as derived from routine NWS forecasts of minimum and maximum temperature [obtained from the long-term verification dataset described by Carter and Polger (1986) and Landis (1994)]. This choice was brought about by the observations that temperature forecasts have considerably higher skill than corresponding forecasts of precipitation [a limitation that apparently arises from the mesoscale nature of the precipitation process; see Roebber and Bosart (1998) for an example] and that the common techniques used to measure the individual components of forecast skill in relation to basic cues have been most successfully applied to forecasts of temperature (Stewart et al. 1997). Furthermore, RBb found

Corresponding author address: Dr. Paul J. Roebber, Department of Geosciences, University of Wisconsin—Milwaukee, Lapham Hall 352, P.O. Box 413, Milwaukee, WI 53201.
E-mail: roebber@csd.uwm.edu

that human forecasts of wintertime temperature for use by gas utilities (an application of degree day forecasts) possess considerable value above that of the numerical–statistical guidance, in effect amplifying under critical circumstances the importance of deviating correctly from MOS. Section 2 will describe the data and methodologies employed to study this issue. Section 3 will provide the results of the analysis and section 4 will contain a concluding discussion.

2. Data and methodology

In order to conduct a study of the regime dependence of forecast technique, skill, and value, it is necessary to first define what is meant by a regime. Once this definition is established, a means of quantifying the skill and value of a set of forecasts made under such conditions is required. Finally, it will be necessary to measure the relative weight (with error limits) that a forecaster has ascribed to a particular cue. In this section, the procedure for defining the regimes, the techniques for quantifying forecast skill and value, and the method for measuring the relative weight of cues will be outlined. The data employed in these analyses are also identified.

a. Flow regimes

There are many ways of defining flow regimes. Teleconnection indices may be used to define a characteristic pattern, for example, by identifying the orthogonal index upon which the flow field projects most strongly. Alternatively, empirical orthogonal function analysis [also referred to as principal components analysis; see Peixoto and Oort (1992) for a review] can be used to partition the variance of the flow pattern time series into orthogonal spatial patterns or eigenvectors (representing a series of coefficients that describes the time evolution of the particular spatial mode). Each successive eigenvector, when ordered, accounts for the maximum amount of the remaining variance in the data, and those patterns that are associated with the largest fraction of accountable variance (typically the first eigenvector) are related to the physics. However, Stoss and Mullen (1995, hereafter referred to as SM95), have outlined a clustering technique for defining large-scale regimes that has the advantage of unambiguously assigning a map to a particular regime (including the case where the map does not belong to a regime) and that, from the large-scale perspective, seems to coincide well with the actual regime recognition process of operational forecasters. Several observational studies have shown that three to five regimes can account for as much as one-half the total number of Northern Hemispheric winter maps (SM95; Cheng and Wallace 1993; Mo and Ghil 1988). The SM95 scheme assigns flow classifications based on mutual spatial correlations (R) among the entire set of maps. Outlier events, those maps that do not correlate well with any of the clusters, are collected into a separate (unclassified) category. In this study, a modified SM95 clustering approach is applied to 500-hPa height data [obtained from the Northern Hemisphere 381-km octagonal gridded dataset; Shuman and Hovermale (1968)] for the cold season months (December–February) for the period 1 January 1973 through 31 December 1992. The procedure is as follows.

1) The correlation coefficient matrix, based on 500-hPa height deviations from the cold season climatology, is computed for once-daily maps in the region indicated in Fig. 1. A nine-point low-pass filter (Haltiner and Williams 1980, 397) is successively applied to the daily 500-hPa height fields to damp the shortest waves prior to this calculation.
2) The map that has the highest number of correlation coefficients above the threshold value $R = 0.50$ is identified. This individual map, designated as a seed map, and all those maps that correlate with it above the threshold are removed from the sample. While the specific value of this threshold is somewhat arbitrary, past research (e.g., Hollingsworth et al. 1980) has suggested that maps with such correlations present ''synoptically useful'' information. Although correspondence in precipitation patterns appears to defy such thresholds owing to the inherently mesoscale nature of the process (e.g., Roebber and Bosart 1998), temperature patterns except in regions of substantial physiographic features are better tuned to the large-scale signal.
3) Step two is repeated successively until all seeds with associated members of at least 7.5% of all days are identified.
4) Final cluster assignments are made by assigning each map to a particular category based on the seed map with which it is most strongly correlated. Any map that falls below a correlation of 0.40 with each seed is considered unclassifiable and is not assigned to any cluster.

This procedure led to the identification of three distinct regimes comprising 63% of the total number of days in the period of study (1307 days when cases with missing cue or forecast data were excluded). Figure 1 shows the 500-hPa composite maps and deviations from climatology for these three regimes. The potential impact of these regimes on temperatures at ALB are readily apparent and are discussed next.

The most frequent regime (24% of all days) is shown in Fig. 1a. This pattern resembles the ''high index'' or HI pattern of SM95 and produces strong zonal flow across North America. Since the main axis of the zonal jet is positioned north of ALB, one might expect relatively mild (compared to the climatological average) conditions at the forecast site, punctuated by the passage of mobile synoptic systems. The next most frequent regime (20.5% of all days) is shown in Fig. 1b. This pattern is the ''Pacific blocking'' or PB pattern of SM95,

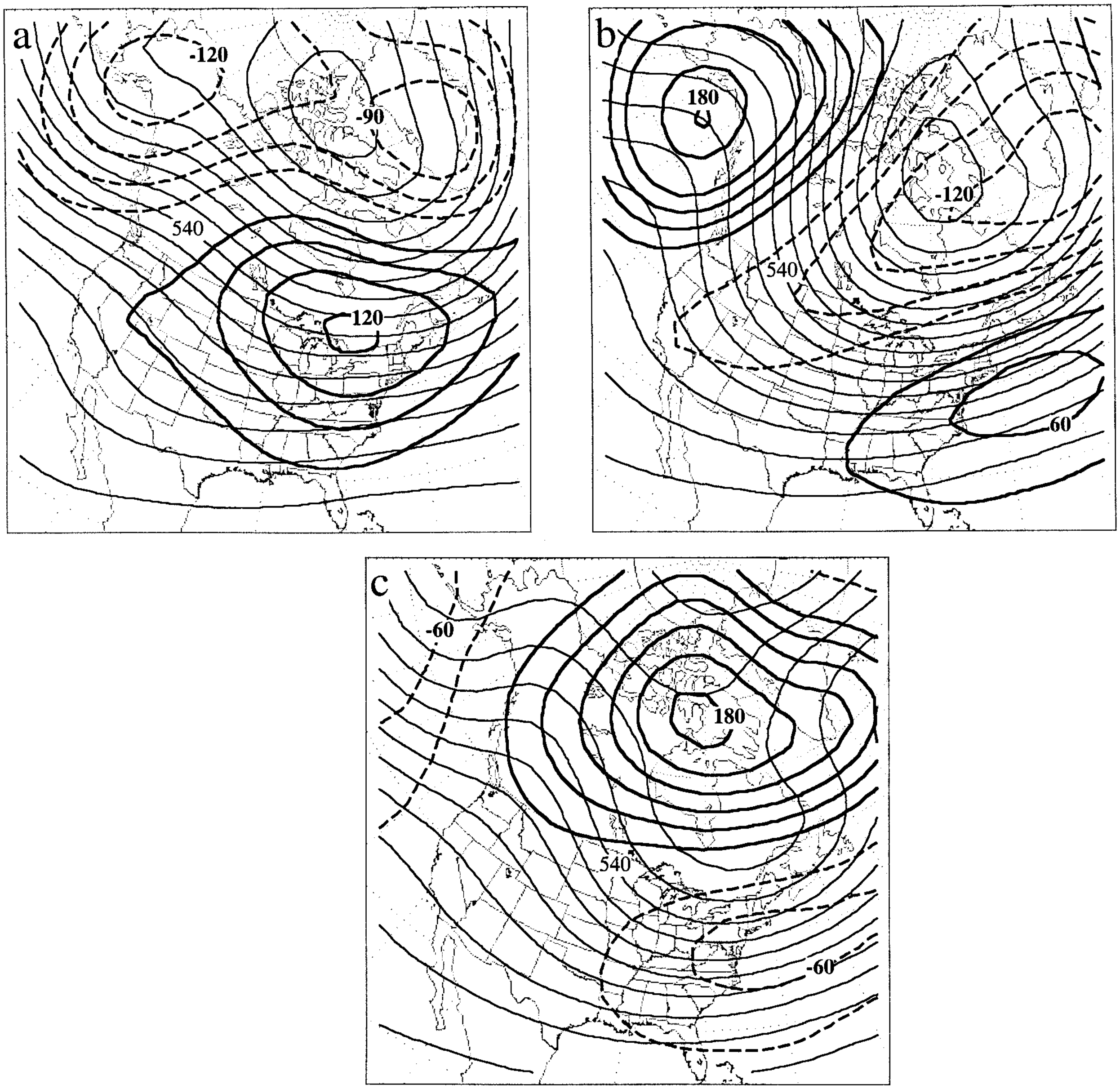

FIG. 1. The 500-hPa composite geopotential height (gray lines, 6-dam interval) and deviation from cold season climatology (black lines, dashed negative, 3-dam interval, zero line omitted) for (a) the high index (HI), (b) Pacific blocking (PB), and (c) low index (LI) regimes.

who noted that this regime closely resembles the negative phase of the Pacific–North American (PNA) teleconnection pattern of Wallace and Gutzler (1981). The negative PNA pattern has been associated with poorer skill in medium-range forecasts (O'Lenic and Livezey 1989; Palmer 1988). SM95 found that at short range, the PB pattern is associated with operational numerical model forecast biases of like sign, acting to reinforce the actual anomalies. Since this pattern itself is associated with an enhanced storm track across the Great Lakes, as evidenced by an amplified southwesterly geostrophic flow across this region at 500 hPa, and since SM95 have shown that the variance of the PB flow is dominated in the northeast by the random component (rather than bias), one might expect that forecasts would be strongly affected by initial condition uncertainty related to the storm track and that skill levels would be relatively low in this region. However, since ALB lies to the southeast of this track, it is unclear whether such influences would be measurable at that location. The final regime, which was identified on 18.5% of all the days, appears to represent a pattern distinct from the four identified by SM95. The key feature of this regime is anomalous easterly geostrophic flow extending from the mid-Atlantic states north through Baffin Island, suggesting weaker westerlies and persistent climatologically cool conditions in the northeastern United States. This pattern will be identified as ‘‘low index’’ or LI.

b. Forecast skill and value

Quantification of forecast skill can be achieved through a variety of measures. The standard skill score (SS), defined with respect to the mean square error (mse) of a reference forecast (the mean of the variable being forecast, which collapses to climatology given a sufficiently long time series; denoted here as $\mathrm{mse_c}$), can be written as

$$\mathrm{SS} = 1 - \frac{\mathrm{mse_f}}{\mathrm{mse_c}}, \qquad (1)$$

where the $\mathrm{mse_f}$ is

$$\mathrm{mse_f} = \frac{1}{N}\sum_{i=1}^{N}(f_i - O_i)^2 \qquad (2)$$

and f_i and O_i are the ith forecast and observation, respectively. Correspondingly, the $\mathrm{mse_c}$ is defined such that

$$\mathrm{mse_c} = \frac{1}{N}\sum_{i=1}^{N}(\overline{O} - O_i)^2, \qquad (3)$$

where the overbar denotes the mean. Since N observations may not conform to the long-term climatological normals because of differing reference periods, sample size, or nonstationarity in climate, this mean will be termed ‘‘period climatology’’ to distinguish it from standard climatology. The skill score defined by (1) is 1.0 for perfect forecasts and 0.0 (negative) for forecasts that are only as accurate as (less accurate than) the reference forecast.

Degree day (DD) forecasts were constructed from the NWS forecasts of minimum and maximum temperature (denoted $T_{\min}$ and $T_{\max}$, respectively) issued following the 1200 UTC forecast cycle (representing forecast ranges of 12–24 h and 24–36 h for $T_{\min}$ and $T_{\max}$, respectively). Thus, the DD forecast is taken to be

$$\mathrm{DD} = 65 - \frac{T_{\min} + T_{\max}}{2}, \qquad (4)$$

where $T_{\min}$ and $T_{\max}$ are reported in degrees Fahrenheit.

The value of the total set of DD forecasts will be considered in the context of forecasts during winter heating extremes for gas utilities, as outlined by Suchman et al. (1979) and employed by RBb. In that approach, the costs resulting from additional demand during peak wintertime heating periods are assessed. In the case of an overforecast (beyond some error tolerance) of the observed DD in excess of a critical value (denoted the critical point), unneeded gas is generated at additional expense, while in the case of an underforecast (beyond the tolerance), costly alternative supplies are tapped to make up the shortfall. Costs are also incurred for days in which the observed DD falls below the critical point but the forecast exceeds that value and the forecast error is in excess of the tolerance (again resulting in excess generation of gas). For this analysis, the critical point, tolerance, gas usage and cost factor are set to 53 DD, 1°F (0.6°C), 53.1×10^3 m^3 DD^{-1}, and \$123.60 $(10^3\ \mathrm{m}^3)^{-1}$, respectively; values representative of a moderate-sized utility in the northeastern United States (Suchman et al. 1979).

c. Forecast cues and the evaluation of forecast technique

RBa and Stewart et al. (1997) have identified a series of cues that correlate well with temperature forecasts and observations. The procedure for defining and quantifying the cues used in this study was discussed by RBa; essentially, this process involves identifying, through physical reasoning, those parameters that are likely to influence forecasts of temperature, and replacing the cues available to the forecasters of the time (which were not recorded but were presumably obtained from numerical model charts, gridded data, and analyses) by the actual observations through a kind of ‘‘perfect prog’’ approach (Klein et al. 1959).

To estimate these data, it is 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, actual observations valid during the time period of the forecast verification derived from a combination of surface and sounding data at ALB are used to simulate the numerical modeling data that cannot easily be reconstructed. Although this assumption represents an upper limit on the accuracy of the data actually available to the forecasters, support for this approach is provided by statistical evaluations of biases and standard deviation errors (SDE) of the National Centers for Environmental Prediction (NCEP, formerly the National Meteorological Center) Regional Analysis and Forecast System (RAFS; Dimego et al. 1992). The biases and SDE at 0 h are quite comparable to typical radiosonde errors, indicating that the uncertainty in the RAFS initialized fields is comparable to the observations on which the RAFS analysis and initialization procedure is based. Furthermore, the 24-h forecast SDE bands remain within a factor of 2 of 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 associated with typical synoptic-scale transient weather regimes. The cues so identified for forecasts of $T_{\min}$ and $T_{\max}$ are shown in Table 1, along with their mean values and standard deviations as a function of identified regime. These data are consistent with the expectations stated above for temperature conditions at ALB during the HI (LI) regime, averaging 4.0°C above (2.1°C below) climatology.

A number of approaches are possible for studying forecast methodology. One approach that could be used would be to make the default assumption that the fore-

TABLE 1. Composite statistics of observations and cues for the three identified regimes of the period 1973–92: TD is the dewpoint temperature, T850 is the 850-hPa temperature, and PW is the precipitable water. Day 1 refers to the first overnight period and following day.

Cue	High index	Pacific blocking	Low index
Minimum-observed (°C)	−4.1 ± 6.4	−7.3 ± 6.9	−8.8 ± 7.4
Maximum-observed (°C)	+3.9 ± 5.6	+ 0.3 ± 6.2	−1.3 ± 5.8
Minimum-MOS (°C)	−5.3 ± 5.1	−8.6 ± 6.2	−10.1 ± 6.1
Maximum-MOS (°C)	+3.8 ± 4.6	+0.7 ± 5.4	−0.9 ± 5.0
Minimum-persistence (°C)	−5.1 ± 6.8	−7.2 ± 7.1	−8.3 ± 7.4
Maximum-persistence (°C)	+3.4 ± 5.6	+0.6 ± 6.1	−0.9 ± 6.2
Minimum-climatology (°C)	−8.9 ± 1.9	−9.2 ± 1.7	−8.9 ± 2.1
Maximum-climatology (°C)	+0.7 ± 1.6	+0.6 ± 1.4	+0.9 ± 1.7
Temp-0000 UTC Day 1 (°C)	+0.7 ± 5.2	−2.7 ± 6.1	−4.4 ± 6.1
TD-0000 UTC Day 1 (°C)	−3.9 ± 6.2	−8.6 ± 7.2	−9.6 ± 7.1
TD-1200 UTC Day 1 (°C)	−5.3 ± 6.9	−9.3 ± 7.7	−11.1 ± 8.2
TD-0000 UTC Day 2 (°C)	−4.1 ± 6.1	−8.9 ± 7.3	−9.7 ± 6.8
T850-0000 UTC Day 1 (°C)	−2.7 ± 6.2	−7.8 ± 7.7	−9.0 ± 7.1
T850-1200 UTC Day 1 (°C)	−2.8 ± 6.3	−8.1 ± 7.6	−9.4 ± 6.9
T850-0000 UTC Day 2 (°C)	−2.6 ± 6.1	−8.3 ± 7.5	−9.4 ± 6.9
Wind-0000 UTC Day 1 (ms^{-1})	4.1 ± 2.4	4.4 ± 2.5	4.7 ± 2.9
Wind-1200 UTC Day 1 (ms^{-1})	3.6 ± 2.4	3.6 ± 2.3	3.9 ± 2.7
Wind-0000 UTC Day 2 (ms^{-1})	4.1 ± 2.5	4.6 ± 2.8	4.7 ± 2.9
PW-0000 UTC Day 1 (mm)	10.4 ± 5.8	8.7 ± 5.9	7.6 ± 4.8
PW-1200 UTC Day 1 (mm)	10.7 ± 6.0	8.7 ± 5.8	7.3 ± 4.8
PW-0000 UTC Day 2 (mm)	10.4 ± 5.6	8.3 ± 5.7	7.1 ± 4.4

cast baseline is provided by MOS. Under this assumption, one could perform a regression analysis between the cues and the forecast deviations from MOS rather than using MOS itself as a predictor. In fact, the basis for the present study was provided by just such an analysis as performed by RBa, who found that skill in predicting observed deviations from MOS is closely tied to overall skill ranking. Since the MOS approach seeks to minimize the mse by correcting for the systematic bias in the numerical model forecasts taken as a whole, 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. For example, in a case where strong radiative cooling might be expected (clear skies, light winds), forecasters recognize that dewpoint temperatures are an important indicator of the likely minimum temperature. In contrast, under conditions where the skies are overcast or the winds are high, the minimum temperature is not strongly governed by surface dewpoints. As a result, the optimal weighting of the dewpoint temperature varies between these cases, although the MOS equations cannot easily account for such differences when taken over the entire sample of cases (since the net effect of dewpoint temperatures is ‘‘on’’ or ‘‘off’’ depending on the other variables).

This suggests a problem with applying an ‘‘MOS-baseline’’ methodology a priori, most particularly in the case where regime constraints are to be considered: the forecast may in some regimes be formulated independent of information from MOS. In this circumstance, deviations from MOS could appear essentially random (unless the effects of that regime resulted in a consistent MOS bias) and an analysis predicated on such an assumption would necessarily fail. Accordingly, the analysis of the relationship between forecast cues and the forecasts will be conducted directly as outlined below.

A useful analysis tool for the study of the forecast judgment process is the lens model (Stewart 1990; RBa; Stewart et al. 1997). Here, we shall consider the analysis in the context of multiple linear regression (MLR) models of the relationship between the cues and the forecasts and the cues and the observations. Thus, the forecast (observed) event is partitioned into a MLR equation expressing this relationship:

$$\mathrm{f} = \mathrm{M}_{\mathrm{f}|\mathrm{x}}(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N) + \mathrm{E}_{\mathrm{f}|\mathrm{x}} \qquad (5)$$

$$\mathrm{O} = \mathrm{M}_{\mathrm{o}|\mathrm{x}}(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N) + \mathrm{E}_{\mathrm{o}|\mathrm{x}}, \qquad (6)$$

where the x_N are the cues, $\mathrm{M}_{\mathrm{f}|\mathrm{x}}$ ($\mathrm{M}_{\mathrm{o}|\mathrm{x}}$) represent the MLR models for the forecasts (observations), and $\mathrm{E}_{\mathrm{f}|\mathrm{x}}$ ($\mathrm{E}_{\mathrm{o}|\mathrm{x}}$) are the MLR model residuals. The lens model equation, which decomposes the correlation (denoted r_{fo}) between the forecasts (f) and observations (O), is then written as

$$r_{\mathrm{fo}} = GR_{\mathrm{f}|\mathrm{x}}R_{\mathrm{o}|\mathrm{x}} + C\sqrt{1 - R^2_{\mathrm{f}|\mathrm{x}}}\sqrt{1 - R^2_{\mathrm{o}|\mathrm{x}}}, \qquad (7)$$

where G is the correlation between $\mathrm{M}_{\mathrm{f}|\mathrm{x}}$ and $\mathrm{M}_{\mathrm{o}|\mathrm{x}}$, $R_{\mathrm{f}|\mathrm{x}}$ is the correlation between the forecasts (f) and $\mathrm{M}_{\mathrm{f}|\mathrm{x}}$, $R_{\mathrm{o}|\mathrm{x}}$ is the correlation between the observations (O) and $\mathrm{M}_{\mathrm{o}|\mathrm{x}}$, and C is the correlation between $\mathrm{E}_{\mathrm{f}|\mathrm{x}}$ and $\mathrm{E}_{\mathrm{o}|\mathrm{x}}$. Murphy (1988) has shown that the skill score defined by (1) can be written in terms of r_{fo} and two additional terms representing conditional and unconditional bias. Stewart (1990) then expressed the skill score in terms of the lens model decomposition shown in (7). However, since the bias of temperature forecasts has previously been shown to be low (RBa), we will not consider bias in

TABLE 2. Summary of lens model measures of forecast performance (after Stewart et al. 1997).

Symbol	Measure	Interpretation
r_{fo}	Correlation between the forecasts (f) and the observations (O)	Potential skill (when squared, equals the skill score for unbiased forecasts)
G	Correlation between MLR models of the forecasts ($\mathrm{M}_{\mathrm{f\|x}}$) and the observations ($\mathrm{M}_{\mathrm{o\|x}}$)	Optimal (linear) weighting of cues
$R_{\mathrm{f\|x}}$	Correlation between the forecasts (f) and the MLR model of the forecasts ($\mathrm{M}_{\mathrm{f\|x}}$)	Forecast consistency or reliability (identical cues lead to identical forecasts)*
$R_{\mathrm{o\|x}}$	Correlation between the observations (O) and the MLR model of the observations ($\mathrm{M}_{\mathrm{o\|x}}$)	Task predictability (maximum predictability of the observations for given set of cues)**
C	Correlation between residuals of the MLR models of the forecasts ($\mathrm{E}_{\mathrm{f\|x}}$) and the observations ($\mathrm{E}_{\mathrm{o\|x}}$)	Validity of the nonlinear component of judgment (second term of lens model equation measures nonlinear contribution to skill)

* Assumes that the forecast model ($\mathrm{M}_{\mathrm{f|x}}$) exhausts the systematic variance in the forecast (f).

** Assumes that the environmental model ($\mathrm{M}_{\mathrm{o|x}}$) exhausts the systematic variance of the observations (O).

this analysis. Interpretation of the lens model decomposition terms has been presented by Stewart et al. (1997) and is summarized in Table 2. As will be shown in section 3, key terms in this study will be G, $R_{\mathrm{f|x}}$, and $R_{\mathrm{o|x}}$, since they will demonstrate the impact on forecast skill of nonoptimal weighting of cues, forecast consistency, and task predictability, respectively.

A number of techniques have been identified in the judgmental research literature for quantifying the relative weight that a judge (forecaster) will ascribe to a particular cue. For instances in which the cues exhibit considerable intercorrelation (collinearity), the typical case in a natural setting such as studied here, no single choice is clearly indicated. The problem that arises in such instances is that much of the information content of one cue is not unique and can be replaced by one or more of the other cues (e.g., surface dewpoint temperature and precipitable water). Correspondingly, care must be taken to assure that the weight ascribed to a particular cue relative to the others is not artificially inflated by this shared information. To assess this issue, intercorrelations for the cues listed in Table 1 were examined. Although most of the cue intercorrelations are below critical levels ($R \sim 0.9$), some collinearity exists (e.g., minimum-climatology vs maximum-climatology, $R = 0.963$; temp-0000 UTC day 1 vs T_{d}-0000 UTC day 1, $R = 0.904$; seven pairs above 0.8).

As a means of dealing with the problem of collinearity, the ‘‘usefulness coefficient’’ methodology for estimating cue weights described by Cooksey (1996) is followed. In this approach, the amount of variance in judgment that can be uniquely attributed to a particular cue is given by the squared semipartial correlation (sr^2), which is measured by the difference between the squared multiple correlation (R^2) from a regression using all available cues and a second R^2 based upon a regression with all cues save the one of interest. For example, the sr^2 for minimum-MOS would be determined by the difference of the R^2 of the DD forecasts using all of the cues listed in Table 1 and a second R^2 using the same set of cues with minimum-MOS removed. Under the constraint of simultaneous regression, where no particular prespecified order of the cues is supposed (and thus no clear partitioning of shared variance is indicated), the overlapping variance among the cues is simply removed. Thus, the sum of the sr^2 for the set of cues will always be less than the R^2 for the regression using all the predictors (see section 3). The squared semipartial correlation of the ith cue so defined can then be converted to a relative weight (rw_i) by

$$\mathrm{rw}_i = 100 \frac{\mathrm{sr}_i^2}{\sum_{i=1}^{k} \mathrm{sr}_i^2}, \tag{8}$$

where the k cue relative weights sum to 100 units. In order to provide estimates of error margins and the significance of assigned weights, the technique of statistical bootstrapping (with case replacement) is employed. Freedman (1981), Freedman and Peters (1984), and Cooksey (1996) have demonstrated the relevance of this technique to the interpretation of multiple regression models. Further details concerning this technique are provided in section 3.

3. Results

It is well known that the numerical–statistical guidance has improved over the period of study in this paper. Since forecasters are quick to recognize and exploit useful information, an issue that needs to be addressed is whether changes in the skill of MOS have led to changes in use of that cue by forecasters. Clearly, the use of other cues relative to MOS would also be affected by such changes. Consequently, we have split the sample of forecasts, observations, and cues into two subsets: 1 December 1973 through 29 February 1980, a period in which MOS was based upon the primitive equation (PE) model, and the period from 1 December 1980 through 31 December 1992, during which time the Limited Fine Mesh (LFM) MOS was available.

Figure 2 presents a box-plot summary of the absolute error (AE) for the MOS and NWS DD forecasts for the PE-MOS and LFM-MOS periods. These data confirm

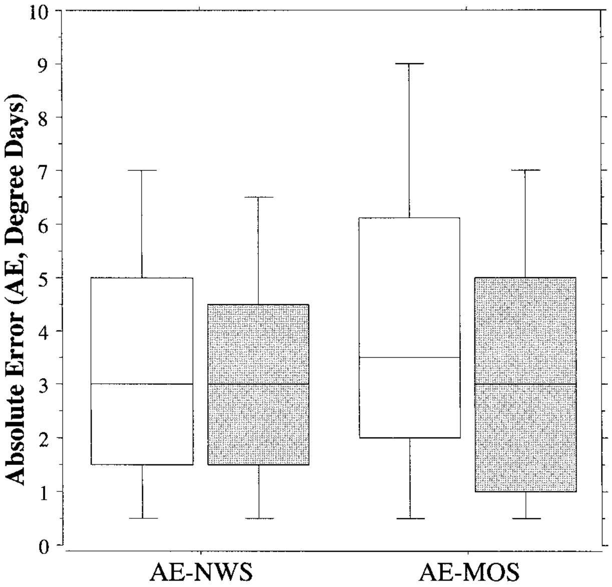

FIG. 2. Box-plot diagram of the absolute error (AE) for the MOS and NWS DD forecasts for the PE-MOS (white) and LFM-MOS (shaded) periods. The horizontal lines denote the 90th, 75th, 50th, 25th, and 10th percentile levels.

the findings of others that MOS has improved substantially during this period, through reductions in error in the midpoint and the extremes of the distribution. These improvements have also been associated with gains in the NWS forecasts, primarily through the reduction in the frequency of outlier forecasts (see tables and discussion below for an assessment of skill and value changes during these periods).

Before proceeding further with this analysis, one must answer the question: ‘‘Might not forecasters have used different sets of cues whose combined effects mimic those of the particular cues examined here?’’ The fact that 94%–96% of the forecast variance is accounted for in both periods by regression with all the selected cues clearly indicates that the most important predictors have been identified. Furthermore, the use of cues that are in some sense redundant (collinearity) is not problematic, since 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, 305–308). In order to assess changes in cue use during these periods, relative weights based on usefulness coefficients [Eq. (8)] were generated (Fig. 3). This analysis reveals that a total of only 3%–4% of the forecast variance is unique to individual cues. However, Stewart et al. (1997) showed that for maximum and minimum temperature forecasts at ALB, the difference in observed variance accounted for by highly skilled and lesser skilled forecasters was of this same magnitude. Thus, these small differences are critical to understanding differences in the skill and value of such forecasts.

For the data of Fig. 3, a statistically significant increase (at the 90% level) in the use of MOS following the introduction of LFM-based guidance is detectable, with resultant statistically significant decreases in the use of persistence, dewpoint temperature, and precipitable water. All other cues also exhibited evidence of decreased utilization, but these changes were not statistically significant at the 90% level. Most of these changes occurred in the variables reflecting the 24–36-h time frame (e.g., maximum-MOS), with relatively little change occurring for the 12–24-h forecast interval. These results provide strong justification of the view that the human role for routine forecasts has changed to one of a ‘‘supervisory’’ nature, where substantial intervention may occur on a relatively few but critical occasions. Such a view is reinforced by a calculation of the skill [defined according to Eq. (1)] and value (based on the gas utility application discussed in section 2) of the NWS and MOS DD forecasts with respect to climatology during the PE and LFM eras (Table 3). Although the relative differences in the forecasts of

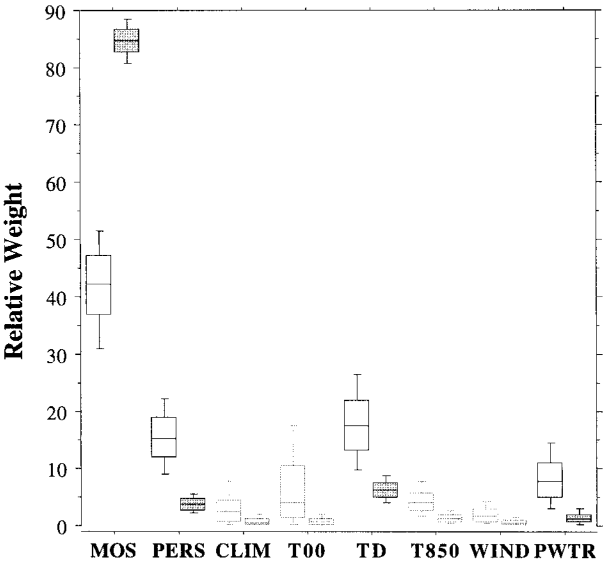

FIG. 3. Box plot of relative weights (based on usefulness coefficients) of the cues for the PE-MOS (white) and LFM-MOS (shaded) periods. General cue categories (e.g., all dewpoint temperature cues) have been combined. Cues are model output statistics (MOS), persistence (PERS), climatology (CLIM), temperature at 0000 UTC of day 1 (T00), dewpoint temperature (TD), 850-hPa temperature (T850), wind speed (WIND), and the logarithm of the precipitable water (PWTR). Changes in cue weights between the two periods that are not statistically significant at the 90% level are grayed.

TABLE 3. Analysis of forecast skill and value (with respect to climatology) of NWS and MOS DD forecasts for ALB for the PE (1 Dec 1973–29 Feb 1980) and LFM-MOS (1 Dec 1980–31 Dec 1992) periods. Value is expressed in constant 1979 dollars on a per heating season basis, with an assumed average of 10.5 events per season.

Period	NWS skill	MOS skill	NWS value ($)	MOS value ($)
PE	0.8030	0.7182	754,054.25	535,789.96
LFM	0.8694	0.8542	862,110.72	827,984.95

TABLE 4. Analysis of forecast skill and value (with respect to period climatology) of NWS DD forecasts for ALB as a function of regime for the LFM-MOS (1 Dec 1980–31 Dec 1992) period. Value is expressed in constant 1979 dollars on a per heating season basis, normalized to an average of 10.5 events per season. MOS skill is also shown for reference.

Regime	Days	NWS skill	MOS skill	Value events	NWS value ($)
Unclassified	298	0.8481	0.8343	42	854,726.31
High index	228	0.8140	0.7985	4	585,703.13
Pacific blocking	209	0.8605	0.8497	24	834,159.37
Low index	147	0.9099	0.8793	30	931,663.69
All cases	882	0.8719	0.8570	100	862,110.72

MOS and the NWS have lessened considerably during this 20-yr period, the latter forecasts remain more skillful and add substantial value to that provided by MOS alone. Presumably, the enhanced skill and value arise from guided intervention by NWS forecasters during critical periods (see RBa for an example; also further discussion below). These findings cast considerable doubt on the reliability of cue weight estimates based on a mixture of the PE- and LFM-guidance eras. Consequently, in the remainder of the paper, calculations will be presented using data only from the period 1 December 1980 through 31 December 1992 (LFM MOS), representing 882 cue profiles.

Table 4 presents the DD skill scores and the value of those forecasts as a function of regime. Although skill scores are expectedly high, some notable variations emerge. The highest skill results from the cold temperature LI regime, representing a departure from climatology. However, the lowest skill results in conjunction with the HI pattern, in which regime temperatures depart the most significantly from climatology (4.8°C and 3.2°C above normals for $T_{\min}$ and $T_{\max}$, respectively). Interestingly, if the *30-yr climatological normals* are used as the reference forecast for skill computations, the skill of the HI regime is much higher (0.8850), while the skill of the other regimes is not substantially altered. This suggests that the 30-yr normals result in an artificial inflation of skill levels during the HI regime; presumably it would be fairly obvious that temperatures would be above normal at these times and the critical forecast issue would be the magnitude of these positive deviations from climatology, given the occasional complication of frontal passages. The value of the DD forecasts (normalized to identical event frequency) validating during the cold LI regime was also considerably higher than those made during the warm HI pattern. The source of these differences appears to be tied to the sensitivity of gas utility DD forecasts to forecast errors once critical levels have been reached; it is important to note that all of the HI value events did occur during significant transitions to colder conditions and thus reflect the difficulty of forecasting during rapid synoptic adjustments. These results also suggest that comparison of matched regime sets such as HI/LI may prove insightful and will be pursued below.

Figure 4 presents a box-plot summary of the relative weights of the cues for each of the three regimes. Qualitative ideas concerning the nature of cue reliance within these regimes are reflected in the results. For example, one might expect unclassifiable flows in the mean to tend toward climatological normals, and enhanced reliance on the climatological cues is suggested for this

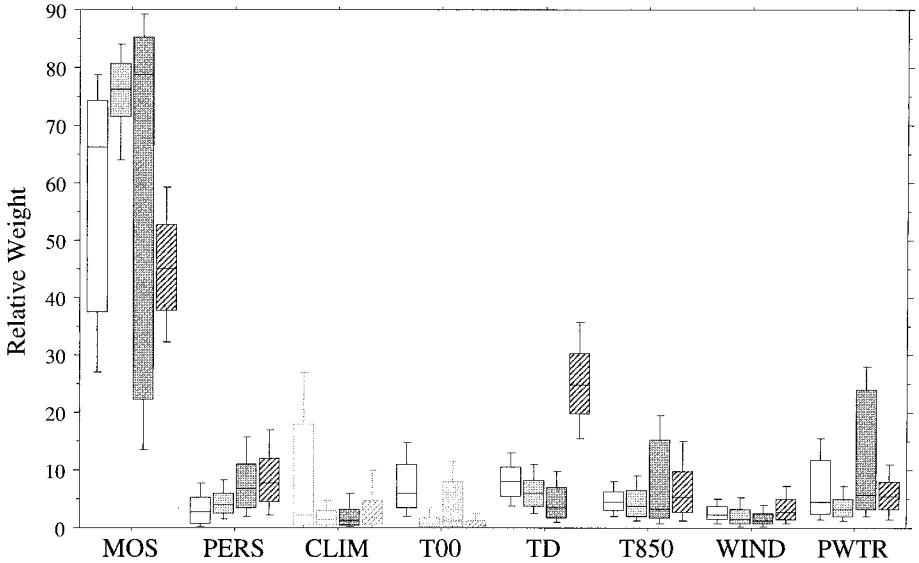

FIG. 4. Box plot of relative weights (based on usefulness coefficients) of the cues for the unclassified cases (white) and the high index (light grey), Pacific blocking (dark grey), and low index (hatched) regimes. General cue categories have been combined and are defined as in Fig. 3. Weights that are not statistically significant at the 90% level are grayed.

regime. Similarly, under the LI regime, an expected increased reliance on persistence is evident. Most important, however, are the obvious distinctions in cue usage between the LI and the other regimes, with the former flow generally requiring much less reliance on model guidance and an increased dependence on dewpoint temperatures, techniques reminiscent of the PE-MOS era (cf. Fig. 4 to Fig. 3) and reflecting the dominance of radiative versus synoptic-scale advective mechanisms (note also increases in the weights assigned to persistence, wind, and precipitable water).

Snellman (1977) warned of the dangers of excessive reliance on automated guidance (''meteorological cancer'') shortly after its introduction and widespread use in forecasting. Since this relative abandonment of MOS in LI flows occurs simultaneously with the attainment of the highest overall skill and value of any of the regimes (Table 4), these results put recent forecast practice in a more positive light. The suggestion is that forecasters have *implicitly* developed adaptive strategies that maximize the skill and value of their forecasts by taking into account the relative utility of forecast information during particular meteorological flows. Such strategies are necessarily implicit since forecasters tend to adjust MOS explicitly on the basis of recent history (a crude surrogate for flow regime), documented biases connected to the position of synoptic-scale features relative to the forecast site, and diurnal errors associated with winds and moisture. Accordingly, in an LI regime, the persistence of colder than normal temperatures and excessive MOS forecasts of the diurnal temperature range (Table 1) might be sufficient to convince forecasters to place greater emphasis on variables related to radiative processes.

However, the strong and consistent reliance on MOS in the HI regime appears to have been ill advised, given the relatively low forecast skill of these events. There are two plausible explanations for this relative forecast failure. It may be that forecasters simply did not recognize the limitations of the guidance during periods in which rapid synoptic transitions were occurring. On the other hand, the forecasters may have recognized these limitations but elected to use the information in the absence of any clear alternatives. Similarly, the drop in skill associated with flows in which the variability of the MOS weight is large (unclassified, PB) suggests that uncertainty in the utility of MOS can also be damaging to forecast skill. In the PB regime, this uncertainty is likely tied to variability in the forecast storm track as discussed in section 2, while for the unclassified patterns, it is likely owing to the lack of a clear large-scale signal through which the forecaster can focus past experience.

Although the regime-dependent cue reliances discussed above satisfy the important constraint that they make physical sense, it would be helpful to apply a rigorous statistical test to provide further confidence that these differences are not simply the result of sampling errors (i.e., the weight differences of Fig. 4 truly reflect differences in the underlying forecast strategies). Consequently, we wish to test the null hypothesis that the regression model for a given regime simply represents a random sample from an underlying population model that is the same for all events. Rejection of this hypothesis allows us to conclude at some specified level of confidence that the underlying forecast strategy for that regime is distinct. This hypothesis is not addressed by pair-wise between-regime comparisons of the individual weights, but rather requires a test of the entire regression model (i.e., the profiles of the cue weights). Furthermore, such a test should not be applied to the usefulness coefficients, since these weights have regime-dependent variance/covariance matrices induced by the correlations between cues. Instead, the test should be applied to the raw MLR weights (the 19 weights directly obtained, e.g., by regressing the cues of Table 1 against the NWS DD forecasts), which, to account for measurement scale differences in the cues and the forecasts (e.g., wind and temperature), have been normalized according to

$$\mathrm{b}' = \mathrm{b}\frac{\mathrm{x}_{\mathrm{max}} - \mathrm{x}_{\mathrm{min}}}{\mathrm{f}_{\mathrm{max}} - \mathrm{f}_{\mathrm{min}}}, \qquad (9)$$

where b is the MLR weight obtained from the regression of the cues (x) against the forecast (f) and the subscripts max and min refer to the maximum and minimum values of the sample data. A measure that has been frequently employed in the judgmental analysis literature (see Cooksey 1996) for assessing dissimilarity in cue weight profiles is the Euclidean distance (D_{ij}):

$$D_{ij} = \sum_{h=1}^{k} (w_{ih} - w_{jh})^2, \qquad (10)$$

where the summation is over the k weights w_{ih} and w_{jh} of the two profiles i and j. Thus, a D_{ij} value of 0.0 results for identical profiles and increases with dissimilarity. Since two weight profiles, if monotonically identical in direction of movement about some mean value, will be perfectly correlated even if the profiles exhibit distinctly different variance about that mean, the Euclidean distance measure, which takes the profile variability and shape into account, is the preferred measure for this application.

An ad hoc method for performing the test of the above hypothesis was constructed in collaboration with T. R. Stewart of the State University of New York at Albany (1997, personal communication). The procedure is as follows: 1) compute D_{ij} for the cue weight profiles of the two samples to be tested; 2) pool the cue profile data for the two samples; 3) bootstrap from the pooled data, forming one test set with the same number of cue profiles as the first sample, and a second test set the same size as the second sample; 4) compute D_{ij} for the cue weight profiles derived from the two test sets; 5) repeat steps 3–4 1000 times to generate a Euclidean

TABLE 5. Results of the test of null hypothesis that the regression model for a given regime represents a random sample from an underlying population model that is the same for all events.

Regime	Sample size	D_{ij}	P value	Conclusion (90% level)
Unclassified	298	0.0343	69.6%	Not distinct
High index	228	0.0168	64.8%	Not distinct
Pacific blocking	209	0.0217	68.6%	Not distinct
Low index	147	0.0959	97.1%	Distinct

distance sampling distribution $f(D_{ij})$, which is ordered from smallest to largest value; and 6) compare D_{ij} to $f(D_s)$ to determine the probability value (p value) of the test result. For example, in step 6, if D_{ij} exceeds the 900th (but not the 901st) value of $f(D_s)$ from a sampling distribution of 1000 members, one could conclude at the 90% confidence level (p value is 90%) that the underlying forecast strategy of that regime is distinct. This method was applied to the regime forecast data by forming one sample of the regime to be tested and a second sample composed of the remaining cue profiles. Thus, for the LI regime, the first (second) sample consisted of 147 (735) members.

The results of this analysis are displayed in Table 5. Although the probability that each of the identified regimes represents a distinctly different underlying forecast regression model (and thus indicates a distinctly different reliance on the available cues) is relatively high, such a conclusion can only be supported with a very high degree of confidence (well above 95% level) for the LI regime. Returning to Fig. 4, the signatory features of forecasting temperatures under this regime are a reduced reliance on MOS and an enhanced emphasis of persistence and dewpoint temperatures.

To further examine the idea of guided intervention by NWS forecasters during critical periods, a bivariate contour analysis of the percentage contribution of forecasts within particular error ranges to the overall skill advantage of NWS forecasts with respect to MOS forecasts (see Table 4) was constructed (Fig. 5). Thus, in Fig. 5a, which shows the results of the analysis for all forecasts (882 cases), the percentages sum to 100%. The expected relationship of low NWS–high MOS error (high NWS–low MOS error) to increased (decreased) skill differential is modified by the strong weighting toward increased skill across a range of MOS errors for moderate NWS errors. Figure 5b shows that this weighting is in large part the result of forecasts issued during the LI regime. As this regime is also associated with a disproportionately large number of value events (Table 4), the origin of the high value of these forecasts is made clear. Thus, the distinct adjustment of forecast strategy during this type of regime connects directly to increases in both skill and value and is the result of forecaster recognition of when to depart substantially from the MOS guidance and rely more on traditional forecast cues such as dewpoint temperatures and persistence.

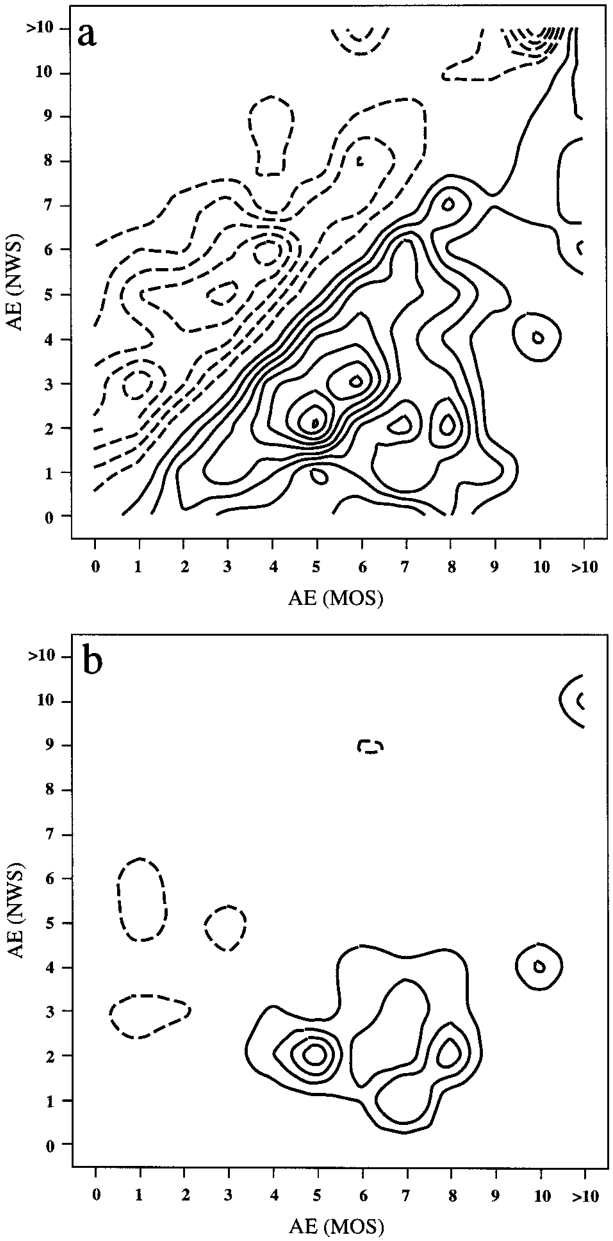

FIG. 5. The percentage contribution of forecasts within particular AE ranges (degree days) to the overall skill advantage of NWS forecasts with respect to MOS forecasts in the LFM-MOS period for (a) all forecasts (882 cases) and (b) LI regime forecasts (147 cases). The contour interval is 2%, with positive (negative) contributions denoted by solid (dashed) lines and the zero line omitted. Contour values are omitted for clarity.

A lens model analysis of forecasting skill, in which the LI regime cases were separated out from the remaining forecasts, was performed and the results are displayed in Table 6. These results indicate that task predictability ($R_{o|x}$) is higher for the LI regime than the other cases and accounts for approximately 50% of the

TABLE 6. Lens model equation analysis of forecast skill.

Regime	Sample size	r_{fo}	Linear component* of r_{fo}	Nonlinear component* of r_{fo}	G	$R_{f\|x}$	$R_{o\|x}$	C
Low index	147	0.955	0.950	0.005	0.976	0.986	0.987	0.187
Other cases	735	0.931	0.922	0.009	0.968	0.982	0.970	0.196
All cases	882	0.937	0.929	0.008	0.971	0.982	0.973	0.194

* Here, r_{fo} = linear component ($G\, R_{f|x}\, R_{o|x}$) + nonlinear component ($C \sqrt{1 - R^2_{f|x}} \sqrt{1 - R^2_{o|x}}$).

skill advantage of forecasts made in that regime. An additional 32% of the skill difference is accounted for by more optimal weighting (G) of the cues. Most of the remainder of the skill differences is accounted for by interactions between $R_{o|x}$ and G; in other words, differences in forecast consistency ($R_{f|x}$, the tendency for a forecaster when given the same set of cues to make the same forecast) and nonlinear skill components are small and make no significant contribution to differences in forecast skill between the LI regime and the other cases. This suggests that some of the relatively better skill obtained in the LI regime is an inherent property governed by the predictability of that flow; however, relatively large gains in forecast skill in other flows relative to the LI regime could be obtained by better use of the available information. For example, as noted above, excessive reliance on MOS during HI regimes appears to have damaged skill; better use could have been made of other forecast information to elevate skill in these instances.

4. Discussion

This study has sought to investigate the manner in which forecasters adjust their reliance on particular pieces of forecast information as the large-scale flow pattern evolves into different regimes, and the relationship between those adjustments and forecast skill and value. The framework for investigating these questions was based upon multiple regression techniques applied extensively in the judgmental analysis literature. Significant findings of the study are as follows.

- Substantial changes in the usage of MOS have occurred with the advent of improved numerical model guidance. These changes have occurred primarily together with improvements in the longer-range forecasts (validating 24–36 h from the initial time).
- The increased reliance on MOS was situation dependent; skill and value were sometimes maintained under regime situations in which MOS was less skillful through significant adjustment of forecast technique.
- Skills were lower for flows in which either the variability of the MOS weight was high, reflecting the relative difficulty of the forecast task when the reliability of a particular piece of critical information is unknown, or in which the limitations of the guidance were evident (as, for example, in high index flows in which rapid synoptic transitions sometimes occur).

RBa investigated the relationship between forecast skill and experience and found that experienced forecasters are better able to recognize those instances when simple forecast strategies do not apply. Roebber et al. (1996) extended this finding by showing that distance from the forecast site affected the skill of experienced forecasters, most probably by eliminating information about regional controls on forecasts that can be used to determine when to deviate from these simple forecast strategies. This paper has made explicit the nature of these procedures. For all flows, a considerable portion of skill is obtained through the judicious use of MOS. However, the large-scale flow regime strongly governs this procedure: under some flows, MOS is less useful and greater reliance is or should be placed upon other forecast measures (most notably surface dewpoint temperatures in the case of the LI regime) to exercise the maximum degree of forecast skill. Experience likely teaches forecasters under which circumstances these adjusted weighting profiles best apply. Continued improvements in MOS have not eliminated the relative advantage that human forecasters display (see Table 4). However, if MOS were further subdivided into equations split by large-scale regime as well as season, it seems likely that an additional portion of that advantage would disappear. Given that both man and machine temperature forecasts are already highly skilled, the recommendation of RBa can only be reiterated: ‘‘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. Lance F. Bosart has been instrumental in kindling my interest in the rigorous study of forecast verification. Tom Stewart is to be thanked for introducing me to the voluminous judgmental analysis literature and drawing my interest to the powerful but not widely used (by meteorologists) techniques of that discipline. Thanks are also extended to the reviewers

for their several suggestions toward clarifying the presentation of these results.

REFERENCES

Carter, G. M., and P. D. Polger, 1986: A 20 year summary of National Weather Service verification results for temperature and precipitation. NOAA Tech. Memo. NWS FCST-31, National Weather Service, NOAA, 50 pp. [Available from National Weather Service, NOAA, Office of Meteorology, W/OM21, Program Requirements and Development Division, 1325 East–West Highway, Silver Spring, MD 20910-3283.]

Cheng, X., and J. M. Wallace, 1993: Cluster analysis of the Northern Hemisphere wintertime 500-hPa height field: Spatial patterns. *J. Atmos. Sci.,* **50,** 2674–2696.

Cooksey, R. W., 1996: *Judgment Analysis: Theory, Methods and Applications.* Academic Press, 407 pp.

DiMego, G. I., 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.

Freedman, D. A., 1981: Bootstrapping regression models. *Ann. Stat.,* **9,** 1218–1228.

——, and S. C. Peters, 1984: Bootstrapping a regression equation: Some empirical results. *J. Amer. Stat. Assoc.,* **79,** 97–106.

Haltiner, G. J., and R. T. Williams, 1980: *Numerical Prediction and Dynamic Meteorology.* John Wiley and Sons, 477 pp.

Hollingsworth, A., K. Arpe, M. Capaldo, and H. Savijari, 1980: The performance of a medium range forecast model in winter—Impact of physical parameterizations. *Mon. Wea. Rev.,* **108,** 1736–1773.

Kalnay, E., M. Kanamitsu, and W. E. Baker, 1990: Global numerical weather prediction at the National Meteorological Center. *Bull. Amer. Meteor. Soc.,* **71,** 1410–1428.

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

Landis, R. C., 1994: Comments on ''Forecasting in Meteorology.'' *Bull. Amer. Meteor. Soc.,* **75,** 823–827.

Mo, K. C., and M. Ghil, 1988: Cluster analysis of multiple planetary flow regimes. *J. Geophys. Res.,* **93D,** 10 927–10 952.

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.

O'Lenic, E. A., and R. E. Livezey, 1989: Relationships between systematic errors in medium range numerical forecasts and some principle modes of low-frequency variability of the Northern Hemisphere 700 mb circulation. *Mon. Wea. Rev.,* **117,** 1262–1280.

Palmer, T. N., 1988: Medium and extended range predictability and stability of the Pacific/North American mode. *Quart. J. Roy. Meteor. Soc.,* **114,** 691–713.

Peixoto, J. P., and A. H. Oort, 1992: *Physics of Climate.* American Institute of Physics, 520 pp.

Roebber, P. J., and L. F. Bosart, 1996a: The contributions of education and experience to forecast skill. *Wea. Forecasting,* **11,** 21–40.

——, and ——, 1996b: The complex relationship between forecast skill and forecast value: A real-world analysis. *Wea. Forecasting,* **11,** 544–559.

——, and ——, 1998: The sensitivity of precipitation to circulation details. Part I: An analysis of regional analogs. *Mon. Wea. Rev.,* **126,** 437–455.

——, ——, and G. J. Forbes, 1996: Does distance from the forecast site affect skill? *Wea. Forecasting,* **11,** 582–589.

Shuman, F. G., 1989: History of numerical weather prediction at the National Meteorological Center. *Wea. Forecasting,* **4,** 286–296.

——, and J. B. Hovermale, 1968: An operational six-layer primitive equation model. *J. Appl. Meteor.,* **7,** 525–547.

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

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

——, P. J. Roebber, and L. F. Bosart, 1997: The importance of the task in analyzing expert judgment. *Org. Behav. Hum. Decis. Proc.,* **69,** 205–219.

Stoss, L. A., and S. L. Mullen, 1995: The dependence of short-range 500-mb height forecasts on the initial flow regime. *Wea. Forecasting,* **10,** 353–368.

Suchman, D., B. A. Auvine, and B. H. Hinton, 1979: Some economic effects of private meteorological forecasting. *Bull. Amer. Meteor. Soc.,* **60,** 1148–1156.

Wallace, J. M., and D. S. Gutzler, 1981: Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.,* **109,** 784–812.