Comparison of Methodologies for Probabilistic Quantitative Precipitation Forecasting*

SCOTT APPLEQUIST,+ GREGORY E. GAHRS, AND RICHARD L. PFEFFER

Geophysical Fluid Dynamics Institute and Meteorology Department, The Florida State University, Tallahassee, Florida

XU-FENG NIU

Department of Statistics, The Florida State University, Tallahassee, Florida

(Manuscript received 8 May 2001, in final form 27 November 2001)

ABSTRACT

Twenty-four-hour probabilistic quantitative precipitation forecasts (PQPFs) for accumulations exceeding thresholds of 0.01, 0.05, and 0.10 in. are produced for 154 meteorological stations over the eastern and central regions of the United States. Comparisons of skill are made among forecasts generated using five different linear and nonlinear statistical methodologies, namely, linear regression, discriminant analysis, logistic regression, neural networks, and a classifier system. The predictors for the different statistical models were selected from a large pool of analyzed and predicted variables generated by the Nested Grid Model (NGM) during the four cool seasons (December–March) from 1992/93 to 1995/96. Because linear regression is the current method used by the National Weather Service, it is chosen as the benchmark by which the other methodologies are compared. The results indicate that logistic regression performs best among all methodologies. Most notable is that it performs significantly better at the 99% confidence limits than linear regression, attaining Brier skill scores of 0.413, 0.480, and 0.478 versus 0.378, 0.440, and 0.457 for linear regression, at thresholds of 0.01, 0.05, and 0.10 in., respectively. Attributes diagrams reveal that linear regression gives a greater number of forecast probabilities closer to climatology than does logistic regression at all three thresholds. Moreover, these forecasts are more biased toward lower-than-observed probabilities and are further from the ''perfect reliability'' line in almost all probability categories than are the forecasts made by logistic regression. For the other methodologies, the classifier system also showed significantly greater skill than did linear regression, and discriminant analysis and neural networks gave mixed results.

1. Introduction

The steady improvement of numerical forecast models and initialization procedures over recent decades has resulted in a corresponding improvement in weather forecasting (Olson et al. 1995). However, owing to the fact that precipitation is highly dependent on small-scale processes and local geography, quantitative precipitation forecasting (QPF) is still not nearly as accurate as predictions of synoptic-scale fields of pressure, temperature, humidity, and wind velocity. To date, equitable threat scores (Schaefer 1990) for QPF reported in the literature and in practice by the National Weather Service do not, in general, exceed 0.5 on a scale from 0 to 1.0 (Hydrometeorological Prediction Center Website: http://www.hpc.ncep.noaa.gov/html/hpcverif.html).

Operational meteorologists typically examine products of different numerical models, especially when such models do not agree with one another on details of the forecast, and make judgments about the potential accuracy of each, depending on how consistent the initial analysis is with the available observations from various sources. Subjective forecasts made by professionals using the numerical prediction products, as well as their own experience, improve QPFs over and above the skill provided by numerical models themselves (Funk 1991; Olson et al. 1995).

A more quantitative approach to improving QPF skill is through statistical guidance using the products of numerical prediction models. The two methods that have been employed for this purpose are model output statistics (MOS) and perfect prog (PP) (Glahn and Lowry 1972; Brunet et al. 1988; Wilks 1995; Antolik 1995; Charba 1998). The MOS approach uses statistical relationships among observed and model variables based on past runs of a particular numerical model. The PP

* Geophysical Fluid Dynamics Institute Contribution Number 422.

+ Current affiliation: Air Force Weather Agency, Offutt Air Force Base, Nebraska.

Corresponding author address: Gregory E. Gahrs, Geophysical Fluid Dynamics Institute, The Florida State University, 18 Keen Bldg., Tallahassee, FL 32306-4360.
E-mail: gahrs@gfdi.fsu.edu

approach derives statistical relationships among atmospheric variables from the historical record and applies them to whatever numerical prediction model is in use at the time. The relative advantages of the two methodologies are well known (see Wilks 1995). The MOS approach was developed and implemented in an operational environment by the Meteorological Development Laboratory (MDL, formerly known as the Techniques Development Laboratory) of the National Oceanic and Atmospheric Administration (NOAA; Glahn and Lowry 1972). MOS forecasts have also been used in a semioperational mode, together with other forecast products, as a basis for subjective probabilistic precipitation forecasts for selected river basins (Krzysztofowicz et al. 1993). Marginal improvements over the conventional MOS approach in predicting probability of precipitation (POP) have been found by using both the MOS and PP products as predictors (Vislocky and Young 1989), and significant improvements have been achieved by using a combination of the MOS forecasts corresponding to two different numerical models (Vislocky and Fritsch 1995b).

Both MOS and PP are generally based on linear regression (with a set of predictors that usually includes nonlinear quantities such as moisture and temperature advection). There are other statistical methods that might be tried, and it is not clear a priori that linear regression is necessarily the best among them. Nonlinear methodologies, such as logistic regression (Walker and Duncan 1967; Vislocky and Young 1989), generalized additive modeling (Vislocky and Fritsch 1995a), and neural networks (Koizumi 1999; Hall et al. 1999; Lindner and Krein 1993) have been used successfully in various prediction problems. Hall et al. reported very good results using neural networks when applied to POP forecasts. Koizumi found that neural networks gave better predictions of precipitation coverage than did linear regression; Lindner and Krein obtained forecasts for the maximum observed 24-h precipitation within an area, using neural networks, that were better than those using linear regression, and Vislocky and Young found marginal improvements over linear regression using logistic regression for MOS POP forecasts. The purpose of this paper is to explore the use of several different methods for statistical guidance for probabilistic quantitative precipitation forecasting (PQPF) at thresholds of 0.01, 0.05, and 0.10 in. and to compare the results obtainable from each. Higher thresholds were not included because their frequency of occurrence in the 4-yr cool season data record that was available to us was, in our judgment, too low to provide a reliable statistical analysis. For the warm season, however, which we have not investigated, there is generally a much higher frequency of greater amounts of precipitation accumulation, so the Nested Grid Model (NGM) dataset should be adequate for dealing with such thresholds. As longer data records become available, it should be possible to do the same for the cool season. Moreover, regionalization, as discussed in section 9d, can significantly increase the data available for training.

The methods employed were linear regression, discriminant analysis, logistic regression, neural networks, and a classifier system. Generalized additive modeling (GAM) with a smoothing spline, which is a nonparametric regression model, was also tried, but we abandoned it when we found it to be less skillful than a simpler class of GAM, namely logistic regression. We believe that this is due to overfitting of a higher-order relationship on the training data.

The plan of this paper is as follows. Section 2 defines the forecast problem undertaken and discusses the use of the data for training and verification, as well as the method employed for evaluating the relative skills of the different methodologies. Section 3 describes the statistical methods used. In section 4 the selection and normalization of predictors are discussed. Section 5 details the implementation of the different statistical forecast methods and the procedures used for selecting predictors for each method. The Brier skill scores and their statistical significance are presented in section 6. In section 7, attributes diagrams are presented and interpreted. Section 8 discusses the most frequently selected predictors for each statistical model. Section 9 is devoted to determining the sensitivity of the results to the choice of predictors. In section 10 we estimate the upper limit of predictability attainable using the entire pool of predictors from the NGM. Our conclusions are found in section 11.

2. The forecast problem, data, predictors, training, and scoring

The problem we consider is that of comparing different methodologies for making 24-h predictions, initialized at 1200 UTC, of the probability of precipitation exceeding 0.01, 0.05, and 0.10 in. for 154 stations distributed in the central and eastern parts of the United States. Figure 1 shows the stations utilized in this study. The forecast period spans 4 yr during the cool season (December–March, DJFM) from December 1992 to March 1996. The data used for this purpose were NGM operational synoptic surface and upper-air analyses and 6-hourly precipitation forecasts up to 24 h over North America on the polar projection 41 × 38 Limited-Area Fine Mesh (LFM) grid with approximate grid spacing of 190.5 km, obtained from the NGM archive maintained by the National Center for Atmospheric Research (NCAR) (http://dss.ucar.edu/datasets/ds069.5/). With the exception of model-predicted precipitation, the archive contains analyzed rather than predicted variables. In a real forecast situation, only model-predicted variables for times other than the initial or past times would be available as predictors. The substitution of analyzed data in place of predictions where the latter are unavailable should not, however, detract from the basic conclusions of the present study because the focus here

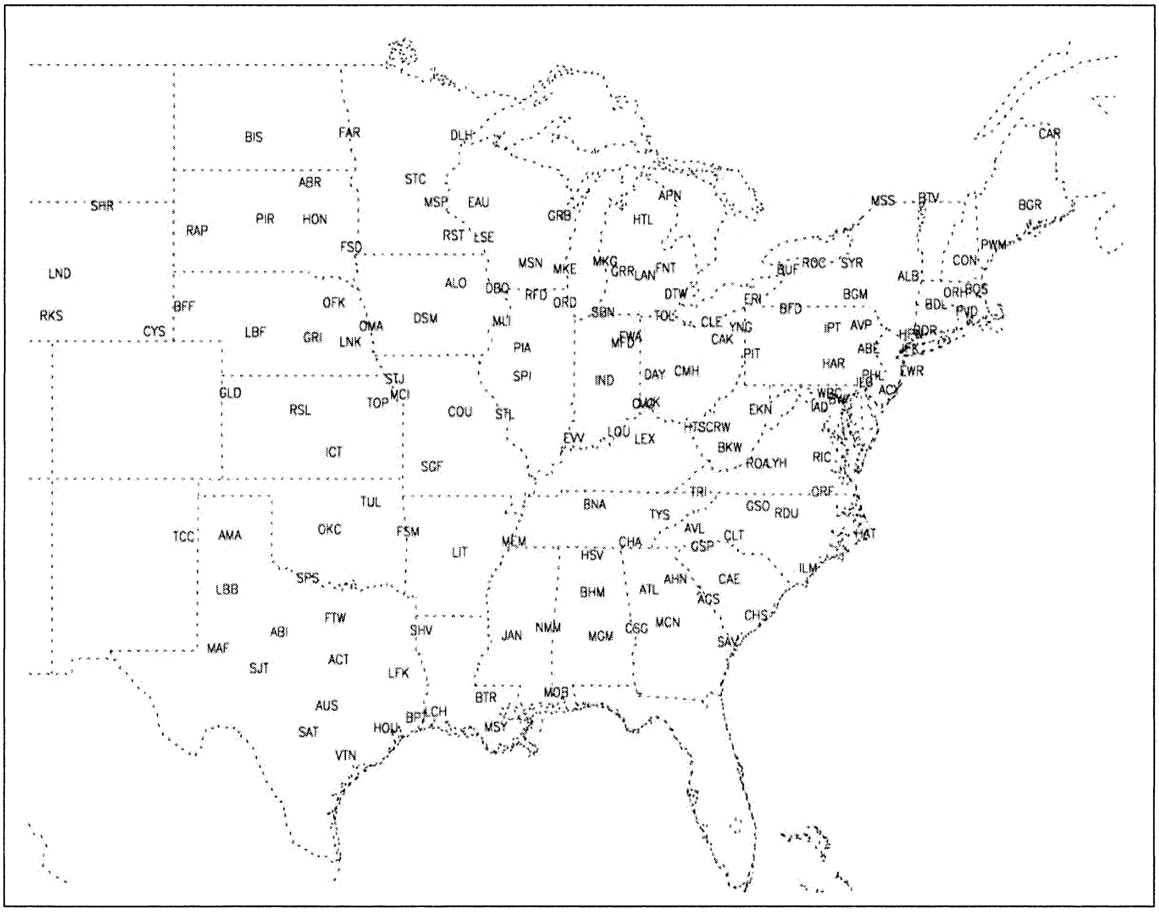

FIG. 1. The 154 stations used for this study.

is on the comparative skill of different methodologies, all of which utilize the same data for prediction and verification.

The pool of potential predictors is shown in Table 1. There are 78 different quantities at three different time periods (0, +12, +24 h), resulting in 234 unique, but not independent, predictors. The predictor pool for each station is formed by interpolating the NGM gridded data to that station's location. It includes predictors suggested by Antolik (1997), Hall et al. (1999), and by the third author of the present paper. In addition, some variables (e.g., relative humidity) were combined into vertical means, which we believe serve as more robust predictors than the values at individual levels. Such averages are indicated in Table 1 by brackets around the levels at which data were available to form the mean. Gridded binary values of some predictors were calculated, as has been done in the past at MDL. For example, choosing a threshold magnitude for average relative humidity above 70%, we assign a value of unity to all points on the NGM grid that have a relative humidity greater than 70%, and a value of zero to all other points. Then we interpolate these values to each station location, with the result that the values at the station locations vary from zero to one. The last five entries in Table 1 define the predictors for which gridded binaries were used.

The rain gauge data were taken from the U.S. Control Cooperative Hourly Precipitation archive maintained by NCAR (http://dss.ucar.edu/datasets/ds505.0/). For verification purposes we made areal averages encompassing all gauges within 60 km of each station.

Since statistical models require a dataset for training that is independent of the verification dataset on which the skill of each model can be evaluated, cross validation is employed (Elsner and Schmertmann 1994). In particular, the coefficients or rules corresponding to each model are determined using 3 years of training data and are verified on the fourth. By choosing different combinations of 3 years among the four for training, verification forecasts can be produced for 4 years of independent data.

We use the Brier skill score (BSS; Brier 1950),

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{CS}}, \qquad (1)$$

which measures the percent improvement of the forecasts over climatology, to assess the skill of each forecast methodology. Here BS is the Brier score given by

TABLE 1. List of the variables in the predictor pool with the pressure level (hPa) at which they were taken. Bracketed terms on the right-hand side indicate vertical averages, and the last five entries represent gridded binaries with the threshold values of each one denoted in brackets on the left-hand side.

Variable	Level (hPa except where noted)
0–24-h precipitation forecast	
0–12-h precipitation forecast	
12–24-h precipitation forecast	
Precipitable water	
East–west wind	10 m, [950, 850, 700, 500],* 300
North–south wind	10 m, [950, 850, 700, 500], 300
Sea level pressure	
Geopotential height	1000, 950, 850, 700, 500, 300
Temperature	[1000, 950, 850, 700, 500]
Specific humidity	[950, 850, 700, 500]
Relative humidity	[950, 850, 700, 500]
Model mean relative humidity	surface to 490
K index	
Lapse rate	Between 700 and 500
Convective instability	850–500
Thickness	From 850 to 300
Temperature advection	[850, 700, 500]
Specific humidity advection	[850, 700, 500]
Vertical velocity	[950, 850, 700, 500], 300
Relative vorticity	[950, 850, 700, 500]
Vorticity advection by geostrophic wind	[700, 500]
Vorticity advection by thermal wind	[950, 850, 700, 500]
Divergence of specific humidity	[950, 850, 700, 500]
Precipitable water times vertical velocity	[950, 850, 700, 500]
Equivalent potential temperature	[850, 700]
Equivalent potential temperature advection	[850, 700]
Q-vector divergence	[850, 700, 500]
Divergence	300
Differential divergence	Between 850 and 300
Gridded binaries	
Vertical velocity [1, 2, 3, 5, 9 cm s^{-1}]**	950, 850, 700, 500
Relative humidity [70%, 90%]	900, 850, 800, 750, 700
K index [20, 30, 35, 40]	
Mean relative humidity [50%, 70%, 90%]	surface to 490
12-h precipitation accumulation [0.01, 0.05, 0.10, 0.25]	

* [ ] denotes vertical range over which mean value calculated.

** [ ] denotes thresholds used to create binary (yes/no) predictors.

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(f_k - o_k)^2, \qquad (2)$$

where n is the number of forecasts in the verification dataset, f_k is the probability forecast ranging from zero to unity, and o_k is the observation, which is equal to unity when the precipitation amount is greater than or equal to the threshold, or zero if it is below the threshold. Parameter CS is the climatology score given by (2) with the climatological forecast based on the training dataset substituted in place of the statistical model forecast. The range of scores for the BSS is from negative infinity to unity, with unity representing a perfect forecast, zero indicating the same skill as a climatological forecast, and a negative value representing a score worse than climatology.

After computing the BSS over the 4-yr verification dataset for each statistical model at each station, the null hypothesis $E(\mathrm{BSS}_1) = E(\mathrm{BSS}_2)$ is tested to determine whether the score for one model is significantly different from that for another. Here, E is the expected value and the subscripts are used to distinguish between the two models being compared. To accomplish this, a paired t test is used (Weiss and Hassett 1991):

$$t = \frac{\overline{d}}{s/\sqrt{m}}. \qquad (3)$$

Here, $\overline{d}$ is the mean difference (averaged over all stations) between the Brier skill scores for the two statistical models being compared, s is the sample standard deviation of the difference, and m is the number of stations for which forecasts are made. In this study, we assume that the differences between the Brier skill scores for any two statistical methods at each station are independent of the differences at another station. It follows that the number of degrees of freedom for the t statistic defined in (3) is the sample size, m. When $m > 30$, as in our application (in which $m = 154$), the t distribution is closely approximated by the standard normal distribution. The critical values for the paired t test in this case are found from the standard normal curve. The null hypothesis will be untrue (which means that the two scores can be considered to be significantly different from each other) when t has an absolute value greater than 1.645 corresponding to the 90% significance level, 1.960 for 95%, and 2.576 for 99%.

3. Statistical methods

In PQPF, the predictand is taken as the probability of precipitation exceeding a defined threshold value. The observed values are either 100% or 0%, corresponding to whether or not the event occurred. Using the training data to determine the coefficients or rules that give the best fit between the predictors and the predictands for each statistical model, we obtain forecast probabilities within a continuous range of possible values of the predictand. The statistical methods used for producing PQPFs are linear regression, discriminant analysis, logistic regression, neural networks, and the classifier system. The former two methods are linear; the latter three are nonlinear.

Linear regression involves fitting a linear function of independent variables (predictors) onto dependent variables (predictands) by minimizing the least square difference between the predicted and observed values of the predictands (Draper and Smith 1966).

Discriminant analysis is designed to discriminate between two mutually exclusive events (in our case, rain exceeding or not exceeding a given threshold) on the basis of a set of variables x_k that define the state of the system (in our case, the predictors). The analysis is accomplished by finding coefficients λ_k in the equation

$$z = \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_j x_j \quad (j = \text{number of predictors}) \tag{4}$$

such that a hyperplane is defined in j-dimensional space of the predictors in which the values of z are divided into two groups (rain exceeding or not exceeding a given threshold) and are separated as widely as possible with respect to the variance of the z values within each group. As detailed by Hoel (1962) and Miller (1962), the analysis maximizes the function G:

$$G = \frac{(\overline{z}_1 - \overline{z}_2)^2}{\sum_{l=1}^{n_1} (z_{1l} - \overline{z}_1)^2 + \sum_{l=1}^{n_2} (z_{2l} - \overline{z}_2)^2}, \tag{5}$$

which is the ratio of the squared difference of the means of z between the rain (z_1) and no-rain (z_2) cases to the sum of the variances of z for these cases. Here, n_1 represents the number of cases in which the threshold precipitation is achieved and n_2 represents the number of cases in which it is not. Parameters $\overline{z}_1$ and $\overline{z}_2$ represent the means of z over the n_1 and n_2 cases, respectively.

Logistic regression, also known as a specialized class of GAM, can be considered a degenerate case of a neural network with no hidden layer, in which the predictand y is expressed in terms of the predictors x_k by the formula

$$y = \left\{1 + \exp\left[-\left(\alpha_0 + \sum_{k=1}^{j} \alpha_k x_k\right)\right]\right\}^{-1}. \tag{6}$$

Here, the α_k are the coefficients to be determined from the training data and j is the number of predictors. An optimal set of α_k is sought that minimizes the squared difference between the predicted and observed values of the predictand. The exponential acts as a ''squashing function,'' which limits the forecasts to lie between zero and unity (Wilks 1995). Since the optimal coefficients cannot be determined analytically, iterative techniques, such as the gradient descent method, are utilized to estimate these values. The gradient descent method commences with a random set of initial weights (α_k) and, through a series of iterations, seeks new weights, which minimize the squared error for the predictand. In order to improve the chances of finding the global minimum, rather than a local minimum, we performed the method 50 times with different initial weights.

A neural network is an artificial intelligence system that models the behavior of neurons in the brain (Lawrence 1991; McCann 1992). It comprises a specified number of layers, each consisting of a specified number of nodes (representing neurons). The first layer contains the input nodes (the predictors) and the last contains the output node (the predictand). The layers in between are called hidden layers. Each node in a layer is connected to all the nodes in the next layer via a prescribed expression, similar to Eq. (6), involving the values at the nodes and a set of weights. The nodes in the hidden layer pass information to the output node, which then becomes the forecast.

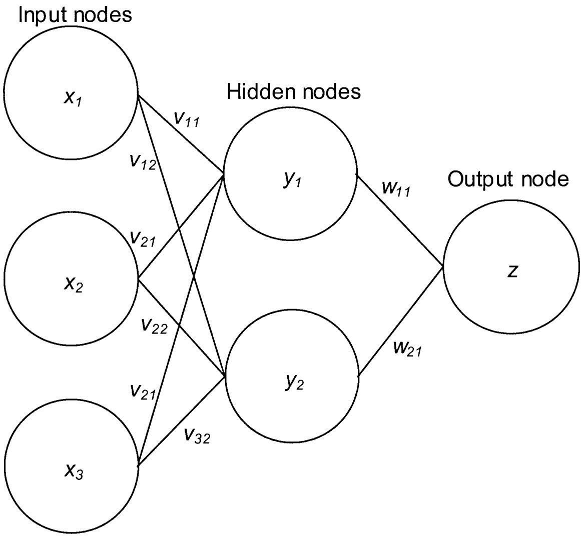

FIG. 2. An example of a neural network having three input nodes, two hidden nodes, and one output node.

Figure 2 illustrates a simple neural network with three input nodes and one hidden layer with two nodes. Here the x_k are the nodal values in layer 1 (the predictors), the y_l are the nodal values in the hidden layer, and z is the predictand. The v_{kl} and w_{l1} are the weights connecting the nodes in the hidden layer to the predictors and the predictands, respectively. The functional relationships among the nodes are given by

$$y_l = \left\{1 + \exp\left[-\left(v_0 + \sum_{k=1}^{j} v_{kl} x_k\right)\right]\right\}^{-1} \quad \text{and} \tag{7}$$

$$z = \left\{1 + \exp\left[-\left(w_0 + \sum_{l=1}^{k} w_{l1} y_l\right)\right]\right\}^{-1}, \tag{8}$$

where k is the number of nodes in the hidden layer. As in the case of logistic regression, the training data are used to find the optimal weights utilizing the gradient descent method with various sets of initial weights.

The classifier system using a genetic algorithm is another form of artificial intelligence, employing the concept of ''survival of the fittest.'' The methodology determines a set of ''if–then'' rules (called a chromosome) relating a predictand to a prescribed set of predictors. Each rule consists of a specified number of ''genes,'' typically three in applications such as weather prediction, which contain information about the predictor to be used, the threshold value of the predictor that must be exceeded in order to make a change in the forecast, and the corresponding increment to be added to or subtracted from the forecast. The number of bits used to define each of these must be specified in advance. The choices of predictors, threshold, and increment are determined by training the classifier system on the dependent data using a genetic algorithm. Detailed information about genetic algorithms employed in classifier systems can be found in Goldberg (1989). An example of a rule for PQPF would be, ‘‘if the average relative humidity between 1000 and 850 hPa is greater than 50%, increase the PQPF by 25%.’’

The procedure begins with a randomly generated population of chromosomes (rules). In the present study, we started with a population of 48 chromosomes, which is within the recommended range from 20 to 90 (Davis 1987), and specified that each one would have four rules. It was felt that too few rules would fail to provide sufficient information, and that too many could result in overfitting the data in the dependent sample.

The first gene in each rule indicated the predictor to be used. Since we limited the maximum number of predictors to eight, specification of the predictor required three binary bits of information. The next gene determined the threshold value of the predictor. As discussed in the following section, all predictors were normalized and the values were subdivided into 1024 parts within the range −3.0 to 3.0, thereby requiring 10 bits of information. The third gene, requiring seven bits, indicated how much to add to the forecast. A population of 48 such chromosomes (with the resulting 80 bits of information laid out in a binary string) was tested on the training data. The 24 chromosomes that yielded the highest BSS, when applied to these data, were allowed to mate with one another to form the next generation of chromosomes to be tested. In particular, the top eight chromosomes (sets of four rules) mated three times, the next eight, ranked by BSS, mated twice, and the last eight each mated once. In this way the characteristics of the most fit chromosomes appeared in the next generation. Mating between a pair of chromosomes is accomplished by splitting each one into two parts at a randomly chosen location in the binary string (the same location in both) and exchanging the information in one of the two parts between the two chromosomes. No chromosome was allowed to mate with a copy of itself. Moreover, in order to allow for greater diversity in successive generations, each newly formed chromosome was given a 10% chance of mutating by having one randomly chosen bit changed from a one to a zero, or vice versa.

Once again, the 24 best chromosomes in the new generation, selected on the basis of the BSS applied to the training data, were allowed to mate in the same way to produce the next generation. This process was repeated to produce 200 successive generations. Since the methodology is nonlinear, there is no guarantee that carrying out this procedure for 200 or more generations from a given randomly generated starting population would produce a chromosome whose rules would give the optimal BSS on an independent dataset. To increase our chances of finding the optimal set of rules, we repeated the procedure starting with 100 different populations with randomly chosen initial values of the bits within each chromosome. We then obtained Brier skill scores for each of the chromosomes in all of the generations produced by all of the initial populations (20 000 in all) and selected for use on the independent data the one whose rules gave the highest BSS on the in-sample (or training) data.

4. Procedure for creating a model

a. Normalization

The predictors used for creating the statistical models (e.g., relative humidity, model-predicted precipitation amount) have different physical units and sometimes range over different orders of magnitude. In this work, each potential predictor x_k was normalized using the expression

$$x_k(\text{normalized}) = \frac{(x_k - \overline{x}_k)}{s_d}, \tag{9}$$

where $\overline{x}_k$ is the mean and s_d is the standard deviation for each potential predictor estimated from the 3-yr training dataset. For normal distributions, the normalized values of the predictors range in magnitude typically from −3 to +3.

b. Stepwise selection

For each meteorological station, the initial pool of 234 normalized potential predictors includes model variables and derived quantities (such as advections), as detailed in Table 1, from analyses at time 0 and at times +12 and +24 h. A stepwise selection method is employed to choose the best predictors from among this large set. Using linear regression, logistic regression, or discriminant analysis, as described in section 3, the stepwise methodology begins by choosing a single predictor that gives the highest BSS for the given station over the 3-yr dependent dataset. Having chosen the first predictor in this way, the methodology proceeds to choose from among the remaining predictors the one that scores best when combined with the first predictor. This process is continued until the in-sample BSS score fails to increase by a prescribed percentage when adding another predictor.

When linear regression is used in stepwise selection, the in-sample BSS will never decrease with the addition of more predictors (because unrelated predictors would have zero coefficients in the linear model) and is most likely to increase with the addition of more predictors. As noted by Lorenz (1977), however, the use of too many predictors typically leads to overfitting of the in-sample data, with the result that the BSS on the *verification* data decreases as more predictors are added. An example is shown in Fig. 3, which displays the in-sample BSS (triangles) and verification BSS (asterisks) for

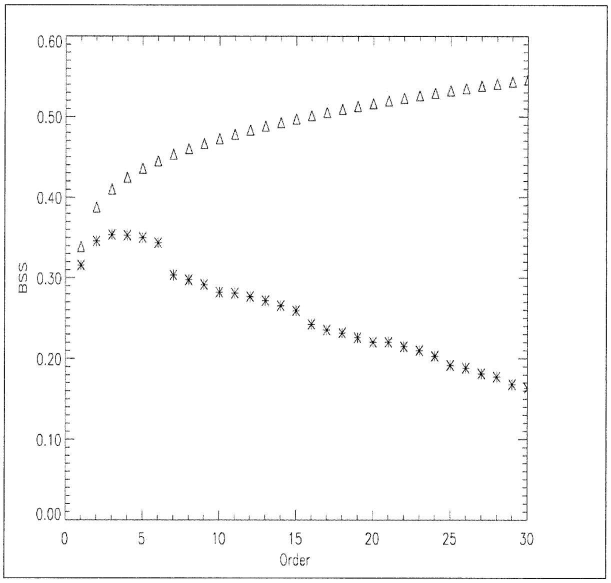

FIG. 3. Average BSS vs the model order (i.e., number of predictors) for the 0.01-in. threshold for the linear regression model. Triangles denote training sample score and asterisks denote verification data score.

linear regression models as a function of the model order (i.e., number of predictors chosen). One can see that, at a certain point, adding more predictors yields a worse verification BSS.

In order to avoid overfitting of the in-sample data, we used a ‘‘stopping rule’’ to determine at what point to stop adding predictors. By trial and error, it was found that the verification scores for choosing predictors via linear regression and discriminant analysis were optimized when stepwise selection was terminated at the point at which the BSS on the training data failed to increase by 5% with the addition of another predictor. We call this a 5% stopping rule. For stepwise selection using logistic regression, a 50% stopping rule gave the optimal forecasts when applied to the independent dataset.

By their very nature, some statistical methodologies, such as linear regression, can yield forecasts of probabilities greater than 100% or less than 0%. To correct for this, we have considered two different approaches. One is called clipping, in which all forecasts of probabilities greater than 100% are set to 100%, and all less than 0% are set to 0%. The other is called transforming or forecast calibration (Murphy and Winkler 1987), in which the forecast probabilities in the *training* dataset are ordered from the lowest to the highest and divided into deciles. Within each decile, the percentage of cases in which the observed precipitation exceeded the prescribed threshold is determined and used as the forecast probability for all forecasts in the independent dataset that lie in the same decile range. Thus, for example, if precipitation exceeding a certain threshold was observed in the training dataset 92% of the time that the methodology predicted a probability of precipitation between 97% and 115%, the probability forecast for all cases that the statistical methodology predicted to be in the range 97% to 115% would be adjusted to 92%. Transformation not only adjusts predictions of greater than 100% and less than 0%, but also modifies the forecasts in all deciles. In our research we found that clipping worked best on forecasts made with linear regression and with the classifier system, and that transformation worked best on those made with logistic regression, discriminant analysis, and neural networks.

c. Principal component predictor selection

In addition to selecting predictors from among a large pool of relevant meteorological variables, we investigated whether appropriate combinations of these predictors determined by principal component (PC) analysis (Wilks 1995) could serve as a better predictor set. Principal component analysis yields a set of orthogonal basis functions in which most of the variance of the original variables is contained in a very small number of functions (i.e., the leading PCs). As a result, a given number of PCs contains at least as much, and usually more, information as the same number of predictors in the original pool. This method was tested with only with one statistical prediction methodology, namely logistic regression, because logistic regression proved to be superior to the other statistical methodologies.

5. Forecast models and implementation

Six different statistical forecast models, each consisting of a statistical methodology and a predictor selection method, were investigated, as discussed below.

a. Linear regression with predictors chosen by linear regression (LiLi)

Predictors for the linear regression model were chosen from the entire pool of 234 potential predictors by applying stepwise selection using linear regression. The stopping rule employed was 5%. In addition, clipping was used to limit the forecast probabilities to the range of 0% to 100%.

b. Discriminant analysis with predictors chosen by discriminant analysis (DADA)

Predictors for the discriminant analysis model were chosen from the entire pool of 234 potential predictors by applying stepwise selection using discriminant analysis. The stopping rule employed was 5%. The z values were divided into deciles and the forecast probabilities were determined by the percentage of cases in which precipitation exceeding the selected threshold occurred within each decile range in the training record.

c. *Logistic regression with predictors chosen by logistic regression (LoLo)*

Predictors for one version of the logistic regression model were chosen from the entire pool of 234 potential predictors by applying stepwise selection using logistic regression. The stopping rule employed was 50%. Although logistic regression yields a probability between 0% and 100%, we found that the BSS could be improved by transforming the probabilities obtained by logistic regression on the training data as described earlier. That is, they were ordered from the lowest to the highest percent probability and divided into deciles. For all forecasts within a given decile range, the final probability predicted was given by the percentage of cases in which precipitation exceeding the selected threshold occurred within that decile range in the training record.

d. *Logistic regression with predictors chosen by principal component analysis (LoPC)*

Predictors for a second version of the logistic regression model were chosen from among the leading 10 principal components (ordered by decreasing variance) derived from a screened set of potential predictors. Since the original pool of potential predictors was too large to allow for the calculation of the PCs with the available computer resources, the number of potential predictors in the pool was first reduced by eliminating those for which the correlation coefficient between the predictor and the precipitation observed in the training data did not exceed 0.35. From the remaining pool of potential predictors, the 10 leading PCs were determined using the proprietary S-Plus 2000 statistical software (MathSoft 1999). Stepwise logistic regression was then employed to reduce further the number of PCs to be used as predictors in making forecasts.

e. *Neural network with the predictor pool reduced by discriminant analysis (NNDA)*

Since stepwise selection with a neural network is quite tedious when applied to a large number of variables, the pool of potential predictors was first reduced from 234 to a smaller number for each station by stepwise selection using discriminant analysis with a 5% stopping rule. Stepwise selection with a neural network was then used to determine the final predictor set. The maximum number of nodes and hidden nodes in the neural networks employed in this study was 10. In order to determine whether a neural network with fewer than 10 nodes should be used in place of one with 10 nodes for each station, the in-sample scores for all model orders up to 10 were calculated and the model with the best in-sample score was chosen as the final one for use on the independent data. It was found that the skill scores for neural network forecasts could be improved significantly by transforming the probabilities in the same way as was done with the logistic regression forecasts.

f. *Classifier system with predictors chosen by discriminant analysis (CSDA)*

As was the case with neural networks, the pool of potential predictors was first reduced from 234 to a smaller number for each station by stepwise selection using discriminant analysis with a 5% stopping rule, since it is not computationally feasible to apply a genetic algorithm to all the predictors. The genetic algorithm was then used to select the final number of predictors from this more limited pool. In order to limit the final probability forecasts to the range from 0% to 100%, we clipped the predictions made by the classifier system by setting all forecasts of probabilities greater than 100% to 100% and all less than 0% to 0%.

The following is an example of the set of rules that was used for Abilene, Texas, for the 0.10-in. threshold. The order of these rules is of no consequence, because the process gives the same result no matter what order is followed.

Rule No.	Predictor	Threshold	Add to prediction (%)
1	Average RH at +24 h	≥80%	21.9%
2	24-h model rainfall	≥0.08 in.	51.6%
3	24-h model rainfall	≥0.30 in.	18.8%
4	24-h model rainfall	≥0 in.	−7.8%

Rule 4 is basically an offset, equivalent to a constant in a regression equation. These rules, with clipping for probabilities less than zero or greater than unity, can be interpreted as providing the following probabilistic forecasts under the following circumstances:

- If the mean RH $<$ 80% and the model rainfall $P <$ 0.08 in., clip the prediction of -7.8% chance of rain exceeding 0.10 in., made by the rules, to 0% chance.
- If the mean RH $<$ 80% and 0.08 in. $\leq P <$ 0.3 in., predict 43.8% chance of rain exceeding 0.10 in. (namely, 51.6% $-$ 7.8%).
- If the mean RH $<$ 80% and the model rainfall $P \geq$ 0.3 in., predict 62.6% chance of rain exceeding 0.10 in. (namely, 51.6% $+$ 18.8% $-$ 7.8%).
- If the mean RH $\geq$ 80% and the model rainfall $P <$ 0.08 in., predict 14.1% chance of rain exceeding 0.10 in. (namely, 21.9% $-$ 7.8%).
- If the mean RH $\geq$ 80% and 0.08 in. $\leq P <$ 0.37 in., predict 65.7% chance of rain exceeding 0.10 in. (namely, 21.9% $+$ 51.6% $-$ 7.8%).
- If the mean RH $\geq$ 80% and the model rainfall $P \geq$ 0.3 in., predict 84.5% chance of rain exceeding 0.10 in. (namely, 21.9% + 51.6% + 18.8% $-$ 7.8%). This is the maximum probability that the rules for this station allow.

TABLE 2. Brier skill scores for the six statistical models at the three precipitation thresholds. The DA in parentheses indicates that stepwise selection using discriminant analysis was employed first to reduce the size of the predictor pool; PC in parentheses indicates that the predictors were selected from among the leading principal components.

Precipitation threshold	0.01 in.	0.05 in.	0.10 in.
Linear regression	0.378	0.440	0.457
Logistic regression	0.384	0.472	**0.478***
Logistic regression (PC)	**0.413***	**0.480***	0.473
Discriminant analysis	0.364	0.459	0.469
Neural network (DA)	0.365	0.458	0.467
Classifier system (DA)	0.371	0.460	0.475

* Bold score is best result in each column.

6. Brier skill scores

The BSS for the six different statistical models and three precipitation thresholds applied to the independent dataset are presented in Table 2. These scores are computed by averaging the individual Brier skill score from all of the 154 stations.

For all thresholds, every method performs better than a climatological forecast, with scores ranging from 0.364 to 0.480, and typically increasing progressively as the threshold gets higher. The numbers in the table reveal, too, that logistic regression scores better than all other methodologies.

At the lowest threshold, 0.01 in., linear regression, which is the method used by the National Weather Service in MOS forecasts, scored better than discriminant analysis, neural networks, and the classifier system, but not as well as logistic regression, particularly with the use of predictors chosen by PC analysis. At the two other thresholds, linear regression gives the lowest BSS among all the methods. At all three thresholds, logistic regression gives the highest score.

The results of the paired *t* tests representing the comparisons among all models are shown in Tables 3–5. The numbers in the body of these tables are the *t* values calculated using (3), with a positive sign indicating that the method listed in the left-hand column has a higher mean BSS than the method listed along the top row, and a negative sign indicating the reverse. A value greater than 2.576 in the table indicates 99% confidence that the method with the higher mean BSS is significantly better than the one with the lower mean score. In each table, the numbers to the lower left of the main diagonal are mirror images of those on the upper right, with the signs reversed. All the numbers are retained here in order to facilitate comparisons between a single method and all the rest.

TABLE 3. Significance scores for the six statistical models at the threshold of 0.01 in. The table is read as Model 1 vs Model 2 with Model 1 on the left-hand side and Model 2 across the top. Negative scores represent a better performance by Model 2, and positive scores indicate a better performance by Model 1. The threshold values for 90%, 95%, and 99% confidence levels are 1.645, 1.960, and 2.576, respectively. Abbreviations are for models described in sections 5a–5f.

	Paired difference scores for 0.01 in.					
	Lili	LoLo	LoPC	DADA	NNDA	CSDA
Lili	0.000	−1.217	−11.523	2.821	2.598	1.791
LoLo	1.217	0.000	−6.325	5.198	4.188	3.884
LoPC	11.523	6.325	0.000	11.589	12.098	11.587
DADA	−2.821	−5.198	−11.589	0.000	−0.593	−2.106
NNDA	−2.598	−4.188	−12.098	0.593	0.000	−1.526
CSDA	−1.791	−3.884	−11.587	2.106	1.526	0.000

TABLE 4. Same as in Table 3 but for precipitation threshold of 0.05 in.

	Paired difference score for 0.05 in.					
	Lili	LoLo	LoPC	DADA	NNDA	CSDA
Lili	0.000	−5.656	−11.044	−3.379	−3.236	−4.305
LoLo	5.656	0.000	−1.641	3.080	3.009	3.126
LoPC	11.044	1.641	0.000	4.418	4.521	4.591
DADA	3.379	3.080	−4.418	0.000	0.323	−0.475
NNDA	3.236	−3.009	−4.521	−0.323	0.000	−0.707
CSDA	4.305	−3.126	−4.591	0.475	0.707	0.000

Focusing attention on the first row of numbers in Table 3 for the 0.01-in. threshold, we see that, at the 99% level, the BSS for linear regression in Table 2 is significantly better than that for both discriminant analysis and neural networks, and significantly worse than that for logistic regression using predictors based on PC analysis. Perhaps the most significant result shown in this table is found by focusing on the third row of numbers. Here we find that, for the 0.01-in. threshold, logistic regression using predictors based on PC analysis is significantly better than all the other methodologies at the 99% level.

The main conclusions to be drawn from inspection of Table 4 for the 0.05-in. threshold is that, at the 99% level, the scores in Table 2 for logistic regression with either choice of predictors (rows 2 and 3) are significantly better than those for any of the other three methods, and that the score for linear regression (row 1) is significantly worse than those of all four of the other methods.

Table 5 reveals that, at the 0.10-in. threshold, the score for linear regression is significantly worse at the

TABLE 5. Same as in Table 3 but for precipitation threshold of 0.10 in.

	Paired difference scores for 0.10 in.					
	Lili	LoLo	LoPC	DADA	NNDA	CSDA
Lili	0.000	−3.134	−3.809	−1.792	−1.623	−3.427
LoLo	3.134	0.000	0.600	1.692	2.110	0.353
LoPC	3.809	−0.600	0.000	0.664	0.958	−0.369
DADA	1.792	−1.692	−0.664	0.000	0.525	−1.398
NNDA	1.623	−2.100	−0.958	−0.525	0.000	−1.770
CSDA	3.427	−0.353	0.369	1.398	1.770	0.000

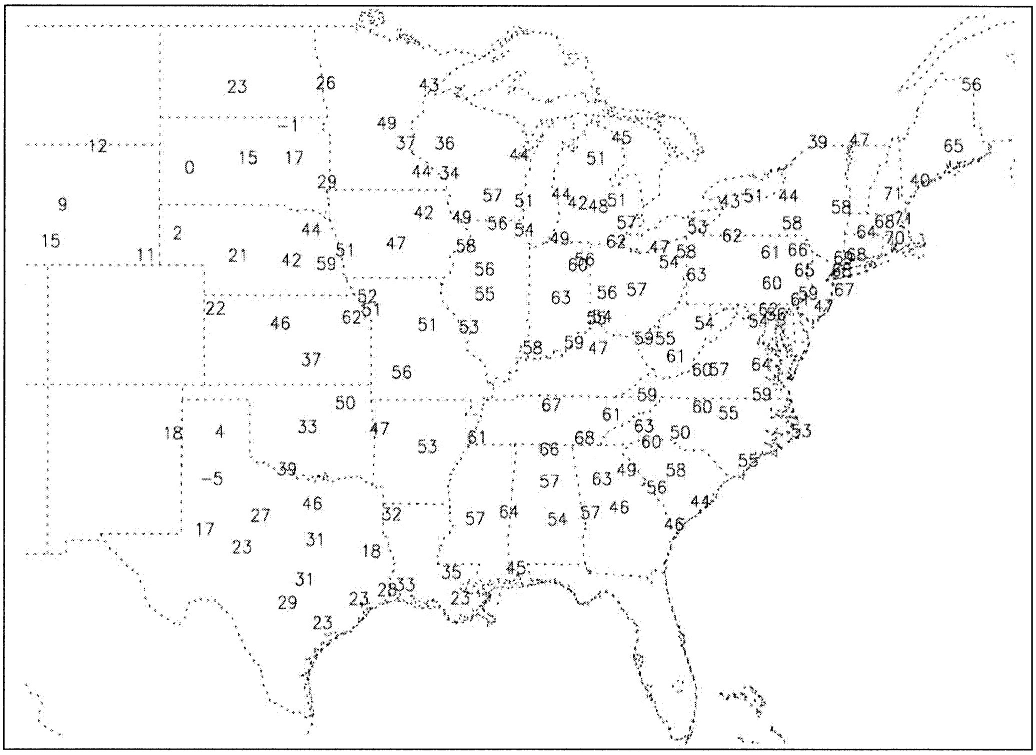

FIG. 4. BSS (%) for the 154 individual stations for the logistic regression model using logistic regression stepwise selection for the 0.10-in. threshold over the entire 4-yr verification period.

99% confidence level than those for the classifier system and logistic regression with either choice of predictors. The differences among all other scores in Table 2 are not significantly different from one another at the 99% level.

Taken together, the results in Tables 2–5 suggest that methods other than linear regression, particularly logistic regression, hold promise of improving statistical guidance in predicting the probability of precipitation exceeding different thresholds.

Figure 4 displays the Brier skill score for the logistic regression model forecasts using predictors selected by logistic regression at each individual station for the 0.10-in. threshold averaged over the four verification years. The higher scores over the eastern portion of the forecast area and the lower scores over the western portion are common to all methodologies and all three thresholds. While there is some evidence (Junker et al. 1992) to suggest that the NGM model forecasts of precipitation amount, which is the primary predictor in our statistical models, are better in the eastern portion of our region than in the western portion, further study would be required to determine the full reason for such a pattern of skill. There is reason to believe, however, that, at some stations around the country, low skill scores could be due to errors in reporting the actual precipitation accumulations. After completing the calculations and examining the results in Fig. 4, we became suspicious that the BSS of 0.40 at Portland, Maine, which is substantially lower than the scores at all surrounding stations, reflected such errors, rather than low skill among all statistical models. Accordingly, we compared the precipitation reported at Portland with that reported at the nearest station, Concord, New Hampshire (where the BSS is 0.71), and found significant discrepancies. In particular, no precipitation was reported in the NCAR archive for Portland from 25 December 1994 through 23 January 1995, whereas there were six precipitation events at Concord during this period. A similar situation occurred from 21 December 1995 through 19 January 1996. Subsequently, we checked the records for Portland from their Web site (http://205.156.54.206/er/gyx/climate.shtml) and found that there had, indeed, been precipitation recorded there during these periods, which did not find its way into the NCAR archive.

7. Attributes diagrams

For each forecast probability (f_i) of precipitation equaling or exceeding a given threshold, there is an observed frequency of occurrence ($\overline{o}_i$) of precipitation equaling or exceeding that threshold,

$$\overline{o}_i \equiv \frac{1}{N_i} \sum_{k \in N_i} o_k. \qquad (10)$$

Here, N_i is the number of times f_i was the predicted probability in the verification dataset and o_k is defined as in (2). The sample climatological frequency of precipitation is given by

$$\overline{o} \equiv \frac{1}{n}\sum_{k=1}^{n} o_k, \qquad (11)$$

where n is the total number of forecasts.

With some algebraic manipulation (Murphy 1973; Wilks 1995), the Brier skill score can be rewritten in terms of f_i, $\overline{o}_i$, and $\overline{o}$ in the form

$$\mathrm{BSS} = \frac{\displaystyle\sum_{i=1}^{I} \frac{N_i}{n}(\overline{o}_i - \overline{o})^2 - \sum_{i=1}^{I} \frac{N_i}{n}(f_i - \overline{o}_i)^2}{\overline{o}(1 - \overline{o})}, \qquad (12)$$

where I is the total number of different forecast probabilities (f_i) in the verification dataset.

The first term in the numerator of (12) is called the resolution term. It measures the degree to which the forecasts sort the observational events into groups that are different from one another. In order to attain a high BSS, it is necessary that the forecast methodology succeed in identifying subsample forecast periods in which the frequencies of occurrence differs substantially from the climatological frequency and, therefore, substantially from one another.

The second term in the numerator is called the reliability term. This term measures the closeness of the observed frequencies to the predicted probabilities for the different categories of probabilities predicted. Forecasts exhibiting good reliability are ones for which this term is small. A forecast methodology is said to have perfect reliability when, for each predicted probability, the observed frequency of occurrence is identical to the predicted probability. Perfect reliability does not, however, ensure a high BSS, unless the resolution is high for a large number of forecasts. For example, forecasts of probabilities close to the climatological frequency of occurrence will have low Brier skill scores, even when they exhibit perfect reliability.

The term in the denominator of (12) is called the uncertainty term. This term depends only on the climatological frequency of occurrence of the event (in our case precipitation equaling or exceeding a given threshold). The uncertainty of a forecast is greatest when the climatological frequency is 50% and least when it is near 0% or 100%. Clearly, we can be much more certain that it will not rain on most days if the climatological frequency is close to 0% than if it is close to 50%. To attain a high BSS the resolution term must exceed the reliability term in the numerator by a greater amount at locations in which the uncertainty is greater.

Attributes diagrams (Hsu and Murphy 1986), as shown in Figs. 5a, 5b, and 5c for the 0.01-, 0.05-, and 0.10-in. thresholds, respectively, encompassing all 154 stations and all 4 yr, are used to display the contributions of the terms in (12) to the BSS and to provide a more complete picture of the comparative performances of different forecast methodologies. Here, the abscissa is the forecast probability of precipitation equaling or exceeding each threshold, and the ordinate is the observed frequency of occurrence for each forecast probability. In these figures, the heavy solid diagonal line represents perfect reliability and the horizontal dotted–dashed line (at the climatological probability) represents zero resolution. The heavy dashed line in the bisector of the angle between perfect reliability and no resolution represents zero skill (BSS = 0). Data points that fall below this line to the right of the vertical dotted–dashed line (representing the climatological frequency of occurrence), or above this line to the left of the vertical line, have a negative Brier skill score representing forecasts that have less skill than a climatological forecast.

In order to interpret the information on these diagrams, we note that, for each forecast probability, data points that lie close to the heavy solid diagonal line exhibit good reliability and those that lie close to the horizontal line exhibit low resolution. Thus, points that are close to the diagonal line and far from the horizontal line represent forecasts that contribute toward a higher Brier skill score. Moreover, data points that lie above the diagonal line represent a bias in the forecast methodology such that the predicted probabilities of precipitation equaling or exceeding the defined threshold are generally too low. Those that lie below the diagonal line represent predicted probabilities that are too high. The magnitude of the contribution of each data point to the BSS is determined not only by its distance from the perfect reliability and no resolution lines, but also by the frequency of occurrence of each forecast probability. It is customary to list the numerical value of this frequency next to each data point. For greater ease of interpreting, we chose instead to represent this information in the form of histograms (Figs. 5d, 5e, and 5f, respectively) below the attributes diagrams.

The circles connected by a thin solid line in Figs. 5a, 5b, and 5c, are the data points for linear regression (the benchmark methodology); the squares connected by a thin dashed line are the data points for logistic regression (the methodology that generally gave the highest BSS). In our application, these data points are actually plotted at the midpoints of the following ranges of forecast probabilities: $0.0 \leq f_i < 0.05$, $0.05 \leq f_i < 0.15$, $0.15 \leq f_i < 0.25$, $0.25 \leq f_i < 0.35$, $0.35 \leq f_i < 0.45$, $0.45 \leq f_i < 0.55$, $0.55 \leq f_i < 0.65$, $0.65 \leq f_i < 0.75$, $0.75 \leq f_i < 0.85$, $0.85 \leq f_i < 0.95$, and $0.95 \leq f_i \leq 1.0$. The frequencies of occurrence of different forecast probabilities in these ranges are shown in Figs. 5d, 5e, and 5f by dark and light bars for logistic regression and linear regression, respectively.

The data in Figs. 5a, 5b, and 5c indicate that both linear regression and logistic regression exhibit good reliability and good resolution at all three thresholds, which accounts for their skill in providing forecasts that are superior to climatological forecasts (Table 2). The data points for linear regression do not, however, lie as

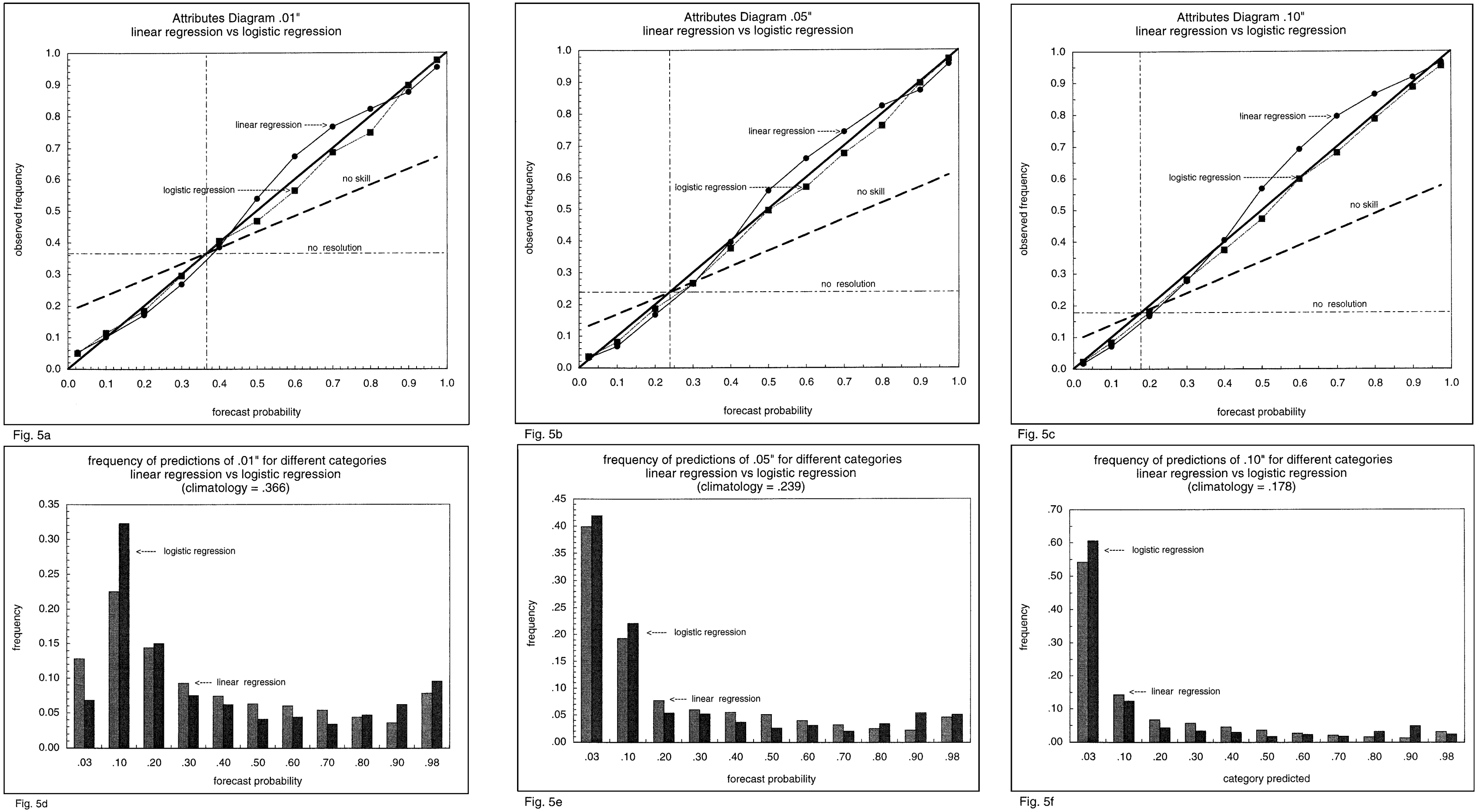

FIG. 5. (a) Attributes diagram for all 154 stations for the 0.01-in. threshold for linear regression and logistic regression. The heavy solid line represents perfect reliability. The dotted–dashed line represents "no resolution," and the heavy dashed line represents "no skill." (b) Same as (a) but for the 0.05-in. threshold. (c) Same as (a) but for the 0.10-in. threshold. (d) Histogram of the frequency of the forecasts made by linear regression and logistic regression for all 154 stations for the 0.01-in. threshold. (e) Same as (d) but for the 0.05-in. threshold. (f) Same as (d) but for the 0.10-in. threshold.

close to the perfect reliability curve as do those for logistic regression. Moreover, they display a systematic bias toward underpredicting the probability of precipitation exceeding each of the three thresholds in the midrange from about 50% to 80% probability. While logistic regression forecasts have a bias toward overpredicting these probabilities, this bias is comparatively small.

With the exception of the lowest probability for the 0.01-in. threshold, the histograms in Figs. 5d, 5e, and 5f reveal that linear regression predicts more probabilities closer to climatology than does logistic regression, which gives a greater number of highly reliable forecasts at the extremes, where the resolution is high. Owing to the fact that the climatological frequencies for all three thresholds are rather low, the smaller number of forecasts of very high probabilities contributes more to the higher Brier skill score for logistic regression than does the larger number of forecasts of very low probabilities, because of the very large resolution of highly reliable forecasts at probabilities ≥ 0.8. These characteristics give insight into why the BSSs for logistic regression are significantly higher than those for linear regression.

8. Predictors

Some insight into the prediction process can be gained by examining the predictors that are most commonly chosen by stepwise selection using linear regression, discriminant analysis, and logistic regression, and by the genetic algorithm used with the classifier system. Typically, with linear regression, using the 5% stopping rule, the methodology selects between three and six predictors, many fewer than have been used by the NOAA Meteorological Development Laboratory with comparable success (Su 1993; Dagostaro and Dallavalle 1997). With discriminant analysis, using the same stopping rule, two to four predictors are typically selected, and with logistic regression with a 50% stopping rule no more than two predictors are almost always chosen. Since the pool of potential predictors for the classifier system was limited to those chosen by stepwise regression using discriminant analysis, the numbers in that case were comparable.

The predictors chosen most frequently for each of the aforementioned methodologies are displayed in Table 6. Here we list only those that were selected at least 5% of the time out of 1848 possible combinations (154 stations times 4 yr of training data times three different thresholds of precipitation). Note that the classifier system actually has 7392 possibilities that can be chosen because there are four rules, and a predictor can be chosen up to four times for a particular station, year, and threshold. For all methodologies, the predictor that was chosen with the greatest frequency was the 24-h accumulated precipitation. This variable was selected in 51.5% of the cases with linear regression, 93% with discriminant analysis, 85.2% with logistic regression, and 64% with the classifier system. Clearly, the numerical model forecasts of precipitation accumulation, while not sufficiently accurate to be used without statistical correction, contain an overwhelming amount of information needed to improve the forecast.

TABLE 6. Frequency of predictors chosen at least 5% of the time by linear regression, discriminant analysis, logistic regression, and the classifier system. GB designates gridded binary variables.

	Frequency	Percentage
Linear regression predictors		
Precipitation forecast (0–24 h)	952	51.5
GB rain > 0.01 (+0 to 12 h)	881	47.7
GB rain >0.01 (+12 to 24 h)	848	45.9
GB avg RH > 70% at +12 h	244	13.2
GB avg RH > 70% at +0 h	239	12.9
Avg RH at +12 h	193	10.4
Avg RH at +0 h	106	5.7
GB avg RH > 70% at +24 h	99	5.4
Discriminant analysis predictors		
Precipitation forecast (0–24 h)	1733	93.8
Avg RH at +12 h	122	6.6
Logistic regression predictors		
Precipitation forecast (0–24 h)	1574	85.2
Avg RH at +12 h	136	7.5
Classifier system predictors		
Precipitation forecast (0–24 h)	4730	64.0

Among the seven other predictors chosen more than 5% of the time with linear regression, two are gridded binaries (GB in Table 6) of 12-h accumulated precipitation and four are measures of relative humidity (RH). In the case of discriminant analysis and logistic regression, only one other predictor (namely, the layer-averaged relative humidity from 950 to 500 hPa at +12 h) was chosen more than 5% of the time. With the classifier system, no predictors, other than 24-h predicted precipitation accumulation, appear more than 5% of the time.

9. Sensitivity studies

In this section we examine the sensitivity of the results to the number of predictors used in logistic regression (the method that yielded the highest BSS) and to the number of rules specified in the classifier system. We investigate also whether the scores can be improved by regionalization, that is, the use of data from a number of observing stations in a region collectively, rather than individually, to determine the coefficients for each statistical prediction model applied to each station in the region. Regionalization is a standard procedure that has been used by the NOAA Meteorological Development Laboratory to increase the number of observations upon which the model can be trained.

a. *Sensitivity of logistic regression to the number of predictors*

The stopping rule determines the number of predictors that will be used in the model. For logistic regres-

TABLE 7. BSS for logistic regression at each of the three thresholds for varying stopping rules. The column on the left indicates the stopping rule and the one on the right gives the typical number of predictors used in each case.

Stopping rule (%)	Precipitation threshold 0.01 in.	 0.05 in.	 0.10 in.	Typical No. of predictors
0	0.349	0.378	0.361	8–10
1	0.356	0.396	0.371	6–8
5	0.386	0.431	0.412	3–5
10	0.392	0.455	0.439	2–4
15	**0.394***	0.467	0.461	2–3
25	0.390	0.472	0.470	2–3, mainly 2
50	0.384	**0.472***	**0.478***	2

* Bold score is best result in each column.

sion, in which stepwise selection was terminated at the point at which the BSS on the training data failed to increase by 50% with the addition of another predictor, two predictors were typically chosen. The smaller the percentage used as a stopping rule, the greater would be the number of predictors required for use in the model and the greater would be the computational expense of running the model. Table 7 shows the typical number of predictors selected and the verification BSS corresponding to the use of seven different stopping rules from 0% to 50%. The highest scores at the 0.05- and 0.10-in. thresholds are found with the 50% stopping rule. At the 0.01-in. threshold the scores are higher with the use of a 15% stopping rule. Using the paired *t* test, we have determined that the BSS using the 15% stopping rule for the 0.01-in. threshold is significantly higher at the 99% significance level than the scores using either 25% or 50%. This suggests that the 15% stopping rule (yielding more predictors) is more appropriate for this threshold. Tests with additional datasets should be done to confirm this conclusion.

At the 0.05-in. threshold, the paired *t* test reveals that the scores for 15%, 25%, and 50% are not significantly different from one another either at the 99% or the 95% significance level. In order to avoid overfitting of in-sample data and obtain a robust result that would hold over many different datasets, it is advisable with any methodology to use the smallest number of predictors among the choices that give the best scores when these scores are not significantly different from one another. Moreover, the use of fewer predictors renders the methodology more computationally efficient. Accordingly, it is appropriate to use the 50% stopping rule (yielding two predictors) at this threshold, as we did in obtaining the results in Table 2. At the 0.10-in. threshold, the BSS for 50% is significantly higher than that for 25% and 15% at the 95% and 99% significance levels, respectively, once again suggesting the use of the 50% stopping rule.

b. *Sensitivity of logistic regression to the choice of predictors*

Stepwise selection by discriminant analysis and by logistic regression yields the same two dominant predictors, namely 24-h model rainfall and layer-averaged relative humidity (Table 6). While each method sometimes uses additional predictors at different locations, it is tempting to determine whether, when applying logistic regression to PQPF, predictors selected by the much less computationally demanding discriminant analysis would serve equally well as those selected by stepwise logistic regression. Accordingly, logistic regression forecasts were made for each station with the former set of predictors and the results yielded Brier skill scores of 0.382, 0.475, and 0.475 for the thresholds 0.01, 0.05, and 0.10 in., respectively. Use of the paired *t* test showed that these scores are not significantly different from the scores in the third row of Table 2 (namely, 0.384, 0.472, and 0.478, respectively). It follows that predictors chosen by stepwise discriminant analysis can be used effectively with logistic regression models.

TABLE 8. (a) BSS for the classifier system at the three different precipitation thresholds using different number of rules. (b) The *t* test scores corresponding to the use of five rules for the 0.01-in. threshold, and four rules for the 0.05- and 0.10-in. thresholds.

(a)	No. of rules 1	 2	 3	 4	 5	 6	 7	 8	 9	 10
0.01 in.	0.243	0.327	0.353	0.358	0.363	0.358	0.357	0.354	0.350	0.349
0.05 in.	0.327	0.415	0.448	0.453	0.450	0.448	0.451	0.448	0.441	0.443
0.10 in.	0.365	0.442	0.466	0.472	0.469	0.471	0.467	0.465	0.460	0.455

(b)	No. of rules 1	 2	 3	 4	 5	 6	 7	 8	 9	 10
					0.01 in.; 5 rules					
	27.90	11.36	3.91	2.01	0.00	2.06	2.31	3.46	5.36	4.72
					0.05 in.; 4 rules					
	25.52	10.89	1.62	0.00	1.11	1.94	0.67	1.62	3.61	3.68
					0.10 in.; 4 rules					
	20.03	9.57	2.47	0.00	1.03	0.32	1.57	2.10	3.89	4.81

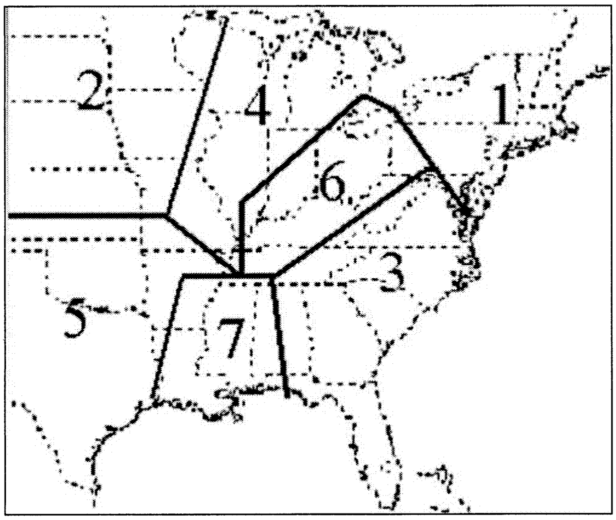

FIG. 6. The seven subregions for PQPFs determined by factor analysis.

c. Sensitivity of the classifier system to the number of rules

Tables 8a and 8b show the results of a study to determine the sensitivity of the BSS to the number of rules specified in the classifier system. In this study, the genetic algorithm was started with 25 different initial conditions and, for each generation, was terminated after 20 generations. Accordingly, we cannot expect to achieve Brier skill scores as high as are shown in Table 2, where 100 initial conditions and 200 generations were used. This study is intended, however, to determine only the relative skill achievable with the use of different numbers of rules. The numbers in Table 8a are the Brier skill scores obtained for forecasts on the independent dataset using from 1 to 10 rules. The highest BSS for the 0.01 in. threshold occurs with the use of five rules. For each of the other two thresholds, the highest BSS occurs with the use of four rules.

The results of the paired *t* test for determining whether the highest scores are significantly better than the others for each threshold are given in Table 8b. This table reveals that the score for five rules at the 0.01-in. threshold in Table 8a is significantly higher at the 99% confidence level than those for 1, 2, 3, 8, 9, and 10 rules. Moreover, it is significantly higher at the 95% confidence level than those for four, six, or seven rules. This suggests that it would be better to make predictions based on the classifier system for the 0.01-in. threshold using five rules instead of the four that were specified in obtaining the results shown in Table 2. Tests with additional datasets, however, would be needed to confirm this conclusion.

The scores for four rules at both the 0.05- and 0.10-in. thresholds are significantly higher at the 99% confidence level than those for 1, 2, 9, or 10 rules and are not significantly different at this level from any of the other scores in Table 8a at either threshold. For the 0.10-in threshold, the scores for four rules are also significantly higher than those for three and eight rules at the 95% confidence level. In keeping with our philosophy of using the smallest number of predictors or rules possible, we recommend the use of only three rules for PQPF at the 0.05-in. threshold, since the results using three rules are not significantly different from those using four for this threshold. Tests with additional datasets are recommended to determine whether this conclusion is correct.

TABLE 9. Same as in Table 7 except these are scores that result from regionalizing the stations into seven subgroups.

Stopping rule (%)	Precipitation threshold 0.01 in.	0.05 in.	0.10 in.	Typical No. of predictors
0	0.389	0.454	0.455	8–10
1	0.389	0.445	0.472	6–9
5	0.392	0.439	0.474	4–7
10	0.375	0.436	0.457	4–7
15	0.377	0.420	0.433	3–6
25	0.374	0.416	0.403	3–5
50	0.320	0.396	0.340	2–3

d. Regionalization

One way of obtaining a larger sample of training data is to use the data from a number of stations in a region collectively, rather than individually, to determine the coefficients for a statistical model. The model can then be used for PQPF at each station in the region. In the present study, factor analysis, as used by Carter and Elsner (1996), was employed to group the stations into regions (Fig. 6) in which the climate is similar, and the logistic regression model was trained using the data for all stations in each region. The Brier skill scores corresponding to the use of different stopping rules with regionalization are shown in Table 9, which may be compared with those in Table 7 corresponding to the use of the data for each station individually to train the model for prediction at that station. Table 9 shows that, with regionalization, the highest scores are obtained with the use of the 0% stopping rule for the 0.05-in. threshold and 5% for the other thresholds, requiring from 8 to 10 predictors for the former threshold and from 4 to 7 for the other two. The fact that the highest scores in Table 9 are lower than the highest scores in Table 7, and that more predictors would be required to achieve the scores in Table 9, suggests that there is no advantage to regionalization when there are sufficient data at individual stations available for statistical analysis. Regionalization can, however, be useful for prediction at higher thresholds where there are many fewer instances of precipitation exceeding such thresholds.

10. The upper limits of accuracy for statistical forecasts

Even the large number of potential predictors we have chosen for this study cannot completely account for the occurrence and amount of precipitation observed (because precipitation also depends on variations of parameters over spatial and temporal scales smaller than those measured). Thus, if we relate all the potential predictors to the precipitation probabilities within the dependent data sample, say, for example, by linear regression, and use the in-sample Brier skill score as a measure of the degree to which we have accounted for the observed precipitation, we cannot expect to get a perfect score (i.e., BSS = 1). The upper limit of predictability of a statistical forecast on independent data must, therefore, be bounded by the score we get from such a calculation. To get an estimate of this limit using predictors from the NGM model, we fit all of the predictors in the original pool to the in-sample precipitation probabilities using linear regression. The Brier skill scores for the thresholds 0.01, 0.05, and 0.10-in. were 0.724, 0.792, and 0.816, respectively. If other statistical relationships or other numerical models were used, the scores would, no doubt, differ somewhat from these, but they can be expected to be in the same ballpark. While these scores are significantly below unity, they are still significantly greater than the scores we have achieved thus far. Although we can never hope to achieve such scores in forecasts on independent data, we believe there is much room for improvement. In the future we plan to determine if significant improvements can be achieved through the use of predictors from an ensemble of numerical prediction models.

11. Summary and conclusions

PQPFs for 24-h precipitation exceeding thresholds of 0.01, 0.05, and 0.10 in. were made for 154 stations across the central and eastern United States during four cool seasons using six statistical models. The predictors for each model were chosen by stepwise selection from the output of the NGM. The Brier skill score was used as a measure of the forecast skill. The benchmark for comparison with all other methodologies was linear regression, since it is the method currently employed by NOAA in MOS forecasts. Perhaps the most important finding is that, at all three thresholds, logistic regression scored significantly higher than linear regression at the 99% confidence level. This conclusion would be unchanged even if we regarded only one-third of the stations, rather than all the stations, as independent in the calculation of the *t* statistic, except that the confidence level would be 95% for the 0.10-in. threshold. Attributes diagrams reveal that this is because linear regression gives a greater number of forecast probabilities closer to the climatological frequency of occurrence, and that these forecasts are further from ‘‘perfect reliability’’ in almost all forecasts categories than are the forecasts made by logistic regression. Moreover, they are more biased toward lower-than-observed probabilities of precipitation.

With regard to the methodologies, the classifier system also showed significantly greater skill than did linear regression at the two higher thresholds, and discriminant analysis and neural networks gave mixed results.

Since the logistic regression forecasts exhibited the greatest skill among all methodologies tested, an effort was made to ascertain whether further improvements in forecast skill could be achieved with this method by choosing a stopping rule that selects more predictors or by creating a single forecast model for each region in which the stations have a similar climatology (using the data for all stations in the region to train the model). It was found that regionalization did not improve the skill and that a 15% stopping rule (requiring more predictors than we used in this study) gave a significantly higher BSS than did the 50% rule only for the 0.01-in. threshold. A further finding was that logistic regression could be made more computationally efficient, without loss of skill, by replacing stepwise selection of predictors using logistic regression by the faster stepwise selection using discriminant analysis, since both methods choose the same predictors for this model.

The two most frequently selected predictors are found to be model-predicted precipitation amounts and relative humidities. Parameters such as vertical velocity, moisture, and temperature advections are not chosen. This might be due, in part, to the fact that this information is implicit in the dynamical model prediction of precipitation.

We intend to explore, in the future, whether further improvement in PQPF can be achieved by the use of predictors from an ensemble of numerical models.

Acknowledgments. This research was supported jointly by NOAA and NSF under USWRP Grant ATM 9714414, in part by NOAA CSTAR Grant NA17WA1010, and AFOSR AASERT Grant F49620-93-1-0531. The authors thank the Geophysical Fluid Dynamics Institute at The Florida State University for supplying the computer resources necessary for implementing and running the statistical models, Dr. James Elsner for instruction of the first author in the techniques used to initiate this work, and Dr. Jon Ahlquist for some very insightful discussions. We are also grateful for the suggestions made by three anonymous reviewers, which greatly improved the paper. Additionally, the data were provided by the Data Support Section, Scientific Computing Division at the National Center for Atmospheric Research. NCAR is supported by grants from the National Science Foundation.

REFERENCES

Antolik, M. S., 1995: NGM-based quantitative precipitation forecast guidance: Performance tests and practical applications. Preprints, *14th Conf. on Weather Analysis and Forecasting,* Dallas, TX, Amer. Meteor. Soc., 182–187.

——, 1997: NGM-based statistical quantitative precipitation forecast guidance for the contiguous United States and Alaska. NWS Tech. Procedures Bull. 461, 28 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

Brunet, N., R. Verret, and N. Yacowar, 1988: An objective comparison of model output statistics and ''perfect prog'' systems in producing numerical weather element forecasts. *Wea. Forecasting,* **3,** 273–283.

Carter, M. M., and J. B. Elsner, 1996: Convective rainfall regions of Puerto Rico. *Int. J. Climatol.,* **16,** 1033–1043.

Charba, J. P., 1998: The LAMP QPF products. Part I: Model development. *Wea. Forecasting,* **13,** 934–962.

Dagostaro, V. J., and J. P. Dallavalle, 1997: AFOS–ERA verification of guidance and local aviation/public weather forecasts—No. 23 (October 1994–March 1995). TDL Office Note 97-3, 52 pp.

Davis, L., 1987: *Genetic Algorithms and Simulated Annealing.* Pitman Publishing, 216 pp.

Draper, N. R., and H. Smith, 1966: *Applied Regression Analysis.* John Wiley and Sons, 709 pp.

Elsner, J. B., and C. P. Schmertmann, 1994: Assessing forecast skill through cross validation. *Wea. Forecasting,* **9,** 619–624.

Funk, T. W., 1991: Forecasting techniques utilized by the Forecast Branch of the National Meteorological Center during a major convective rainfall event. *Wea. Forecasting,* **6,** 548–564.

Glahn, H. R., and D. A. Lowry, 1972: Use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Goldberg, D. E., 1989: *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, 412 pp.

Hall, T., H. E. Brooks, and C. A. Doswell, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting,* **14,** 338–345.

Hoel, P. G., 1962. *Introduction to Mathematical Statistics.* John Wiley and Sons, 427 pp.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

Junker, N. W., J. E. Hoke, B. E. Sullivan, K. F. Brill, and F. J. Hughes, 1992: Seasonal and geographic variations in the quantitative precipitation prediction by NMC's Nested-Grid Model and the Medium-Range Forecast Model. *Wea. Forecasting,* **7,** 410–429.

Koizumi, K., 1999: An objective method to modify numerical model forecasts with newly given weather data using an artificial neural network. *Wea. Forecasting,* **14,** 109–118.

Krzysztofowicz, R., W. J. Drzal, T. R. Drake, J. C. Weyman, and L. A. Giordano, 1993: Probabilistic quantitative precipitation forecasts for river basins. *Wea. Forecasting,* **8,** 424–439.

Lawrence, J., 1991: *Introduction to Neural Networks.* California Scientific Software, 213 pp.

Lindner, A. J., and A. S. Krein, 1993: A neural network for forecasting heavy precipitation. Preprints, *13th Conf. on Weather Analysis and Forecasting,* Vienna, VA, Amer. Meteor. Soc., 612–615.

Lorenz, E. N., 1977: An experiment in nonlinear statistical weather forecasting. *Mon. Wea. Rev.,* **105,** 590–602.

MathSoft, 1999: *S-Plus 2000 Guide to Statistics.* Vol. 2. Data Analysis Products Division, MathSoft, Inc., Seattle, WA, 582 pp.

McCann, D. W., 1992: A neural network short term forecast of significant thunderstorms. *Wea. Forecasting,* **7,** 525–534.

Miller, R. G., 1962: *Statistical Prediction by Discriminant Analysis. Meteor. Monogr.,* No. 25, Amer. Meteor. Soc., 53 pp.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

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

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at NMC. *Wea. Forecasting,* **10,** 498–511.

Schaefer, J. T., 1990: Critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Su, J. C., 1993: NGM-based MOS guidance for the probability of precipitation (PoP). NWS Tech. Procedures Bull. 409, 14 pp.

Vislocky, R. L., and G. Y. Young, 1989: The use of perfect prog forecasts to improve model output statistics forecasts of precipitation probability. *Wea. Forecasting,* **4,** 202–209.

——, and J. M. Fritsch, 1995a: Generalized additive models versus linear regression in generating probabilistic MOS forecasts of aviation weather parameters. *Wea. Forecasting,* **10,** 669–680.

——, and ——, 1995b: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.,* **76,** 1157–1164.

Walker, S. H., and D. B. Duncan, 1967: Estimation of the probability of an event as a function of several independent variables. *Biometrika,* **54,** 167–179.

Weiss, N. A., and M. J. Hassett, 1991: *Introductory Statistics.* Addison-Wesley, 834 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.