Improved Results for Probabilistic Quantitative Precipitation Forecasting*

GREGORY E. GAHRS, SCOTT APPLEQUIST,+ 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 19 March 2002, in final form 21 November 2002)

ABSTRACT

As a follow-up to a recent paper by the authors in which various methodologies for probabilistic quantitative precipitation forecasting were compared, it is shown here that the skill scores for linear regression and logistic regression can be improved by the use of alternative methods to obtain the model order and the coefficients of the predictors. Moreover, it is found that an even simpler, and more computationally efficient, methodology, called binning, yields Brier skill scores that are comparable to those of logistic regression. The Brier skill scores for both logistic regression and binning are found to be significantly higher at the 99% confidence level than the ones for linear regression.

In response to questions that have arisen concerning the significance test used in the authors' previous study, an alternative method for determining the confidence level is used in this study and it is found that it yields results comparable to those obtained previously, thereby lending support to the conclusion that logistic regression is significantly more skillful than linear regression.

1. Introduction

In a recent paper (Applequist et al. 2002, hereafter referred to as AGPN), the authors examined the relative skills of various statistical methodologies for making probabilistic quantitative precipitation forecasts (PQPFs) at each of 154 stations over the central and eastern regions of the United States (Fig. 1). This was done for 24-h precipitation accumulations (0–24 h initialized at 1200 UTC) exceeding thresholds of 0.01, 0.05, and 0.10 in. during the cool season (December–March). The predictors used were Nested Grid Model (NGM) gridded analyses and predictions of various meteorological quantities for the period December 1992–March 1996 obtained from the National Center for Atmospheric Research (NCAR) archive (available online at http://dss.ucar.edu/datasets/ds069.5/). This method is known as model output statistics (MOS; Glahn and Lowry 1972). The statistical methodologies included linear regression, discriminant analysis, logistic regression, neural networks, and a classifier system with a genetic algorithm. For each methodology, the coefficients that relate the predictand to the predictors were determined from a training dataset consisting of NGM model output and observed precipitation accumulations for different combinations of three cool seasons. The skill of each methodology was tested on an independent dataset consisting of similar data for the other cool season. The primary finding was that, at all three thresholds, logistic regression had a significantly higher Brier skill score (Brier 1950) than linear regression (the benchmark methodology used by the National Weather Service) at the 99% confidence limits. The other methods, while generally better than linear regression, showed mixed results.

For each methodology, the coefficients assigned to the different predictors in our previous work were determined from the training dataset by maximizing the Brier skill score (BSS) and using a stopping rule to determine the model order at which the BSS failed to increase by a specified percentage. The present paper contains improved scores for linear regression and logistic regression obtained by choosing the model order using the generalized information criterion (GIC). In this work, the choice of coefficients for logistic regression is based on the use of the maximum likelihood method with a Fisher scoring algorithm. In addition, a third statistical methodology is examined in which the

* Geophysical Fluid Dynamics Institute Contribution Number 433.

+ Current affiliation: Air Force Combat Climatology Center, Asheville, North Carolina.

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

© 2003 American Meteorological Society

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

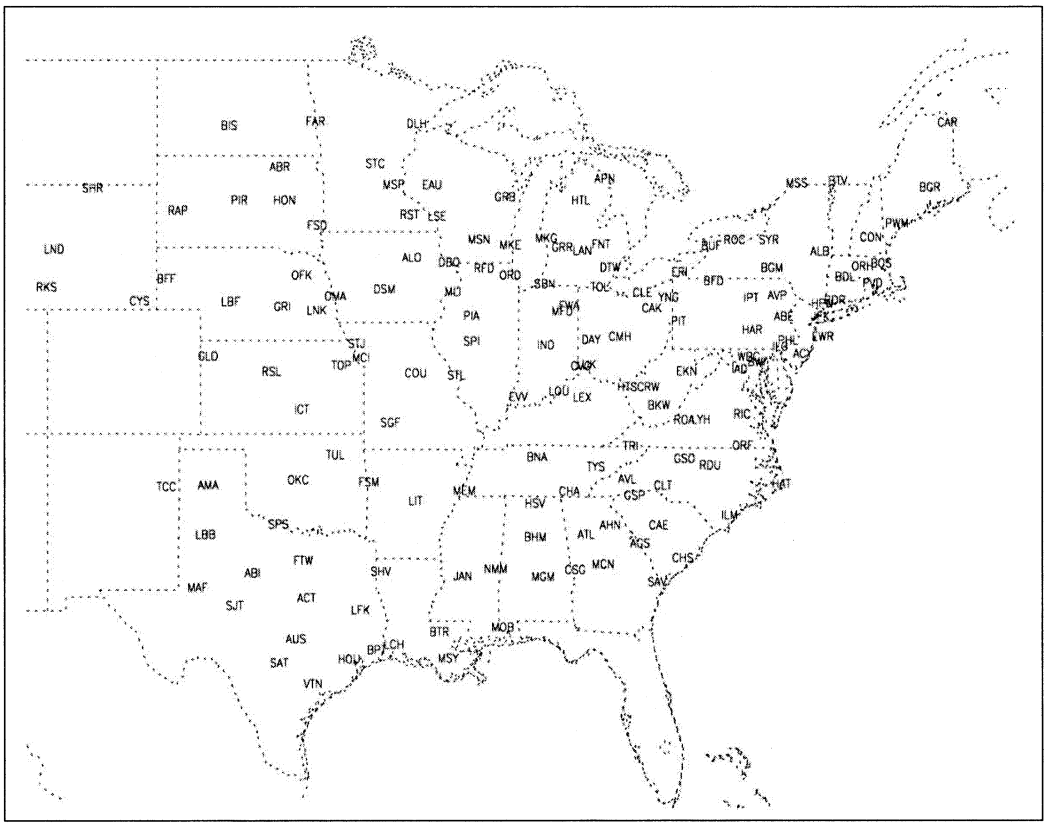

FIG. 1. The 154 stations used for this study.

training data are used to pair the numerical model predictions of precipitation accumulation with observed accumulations to determine the probability of precipitation exceeding a given threshold for a given model-predicted accumulation. Finally, in response to questions raised about the independence of the scores at different stations used to determine the significance of our results, we present here an alternative significance test, the results of which support our earlier conclusion concerning the greater forecast skill of logistic regression over linear regression.

2. Statistical methods

The statistical methods for PQPF to be compared in this paper are linear regression, logistic regression, and the binning method. The reader is referred to AGPN for the details of the first two methods. In both, the predictors (X_k) were normalized by the following expression:

$$x_k(\text{normalized}) = \frac{(x_k - \overline{x}_k)}{s_k}, \qquad (1)$$

where $\overline{x}_k$ is the mean, k is the predictor index, and s_k is the estimated standard deviation for the kth predictor estimated from the 3-yr training dataset. The normalized values typically range from -3 to 3. Here, we discuss the differences between our present and previous approaches and we give the details of the binning method as we applied it.

a. Screening of predictors

In AGPN the programs for training and application of linear regression and logistic regression were written by the first author of that paper in such a way that they could handle large numbers of potential predictors. Rather than rewrite these programs to incorporate the generalized information criterion for determining model order, and the maximum likelihood approach for choosing the coefficients, we purchased the S-Plus statistical software package (MathSoft Inc. 1999), which already incorporates these methodologies. This package is, however, more limited in the number of potential predictors it can handle. Accordingly, we reduced the set of 234 potential predictors (see Table 1) to a pool of 20 predictors by keeping only those that correlated best with the actual rainfall amount in the training dataset for each of the 154 stations, 4 yr, and three thresholds (a total of 1848 sets with 20 predictors in each set). This drastic reduction in the initial pool of predictors is judged to be inconsequential to the final results: first because many of the predictors are highly dependent on one another, and second because, in our previous results, we

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

TABLE 1. List of the variables in the predictor pool with the pressure level (in 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 value 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 to 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	[850, 700]
Advection	
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]	

found that our statistical models rarely required more than 10 predictors and that the same leading predictors, namely, model-predicted precipitation and the vertical average of relative humidity, were selected in almost all models at almost all stations.

b. Choice of coefficients for logistic regression

In AGPN, the coefficients for logistic regression were determined by maximizing the Brier skill score. In the present investigation we use the maximum likelihood method in conjunction with Fisher scoring (Agresti 1990) for this purpose. In the maximum likelihood approach (Agresti 1990), one seeks to maximize the function

$$L(\mu|y) = \sum_{i=1}^{n} [y_i \log(\mu_i) + (1 - y_i) \log(1 - \mu_i)]. \quad (2)$$

Here, y_i is the observation (either equal to 1 if the observed precipitation accumulation is greater than or equal to the threshold value, or zero if it is below that value); μ_i, which ranges from 0 to 1, is the expected value or forecast probability of the predictand on the ith day; $\mu = (\mu_1, \ldots, \mu_n)$ is the expected value vector of the observations $y = (y_1, \ldots, y_n)$; and n is the number of days in our training data. The expected value μ_i in our application is given by the logistic regression expression

$$\mu_i = \left\{1 + \exp\left[-\left(\alpha_0 + \sum_{j=1}^{k} \alpha_j x_{ij}\right)\right]\right\}^{-1}. \quad (3)$$

Here x_{ij} is the value of the jth predictor on the ith day, the α_j are the coefficients to be determined from the training data, and k is the number of predictors.

Inspection of (2) reveals that each term can range in value from negative infinity to zero. Hence, the maximum possible value for the function L is zero, which represents

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

a perfect forecast (i.e., the predictions μ_i are equal to the observations y_i). Since L is nonlinear, the maximum value of this function for a particular set of observed and predicted values of precipitation cannot be determined analytically. Inasmuch as L is a concave function in the $(k + 1)$-dimensional space $(\alpha_0, \alpha_1, \ldots, \alpha_k) \in \Re^{k+1}$ at fixed y_i (Wedderburn 1976), the optimum values of the α_j that maximize L can be found using iterative techniques, particularly if the initial guesses for these coefficients are close enough to the values that maximize it. In the present application we set the initial values of the α_j to zero, which is equivalent to starting with an initial guess of a 50% chance of precipitation exceeding a selected threshold, and we use the Fisher scoring technique to find the α_j by iteration. This technique is similar to the more widely known Newton–Raphson method, with the exception that the second derivative matrix in the latter is replaced by the expected value of this matrix. At each iteration the new values of the α_j are checked to see the extent to which they differ from the previous ones. When the changes are smaller than a prescribed amount, the process is stopped. This occurs when the deviance (defined as $-L$) of the new iteration is not significantly different from the deviance of the previous iteration. This can expressed using the following criterion,

$$\left| \frac{-L_{m-1} - L_m}{-L_{m-1} + \varepsilon} \right| \leq \varepsilon, \tag{4}$$

where the subscript m represents the iteration and ε is a parameter that is taken here as 0.0001. This process then yields an estimate for μ_i.

c. Determination of model order

In our previous work, we utilized a rule for choosing the model order in which we stopped adding predictors when the BSS failed to increase by a given percentage when adding another predictor. In the present study, we use the GIC, which is the likelihood version of Mallow's C_p statistic (Weisberg 1985) and has the expression

$$\mathrm{GIC}(p) = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{s_i^2} + Ap. \tag{5}$$

Here, y_i and n are defined as in (2), $\hat{\mu}_i$ is the estimate of the expected value μ_i, s_i^2 is the estimated variance of the forecasts, p is the number of predictors in the model, and A is a penalty weight on this number. The model order chosen is the one that yields the lowest value of the GIC(p) statistic. The penalty function is added to prevent overfitting of the training data by penalizing a model that has too many predictors. In the present investigation, we compared three different choices of A namely $A = \sqrt{n}$, $A = \ln(n)$ [Bayesian information criterion; Schwarz (1978)] and $A = 2$ [Akaike information criterion; Akaike (1973)]. Since $n = 360$ in our case $\sqrt{n} > \ln(n) > 2$, so the first choice yields the lowest model order (smallest number of predictors) and the last yields the highest model order. It was found that $A = 2$ gave the best results for linear regression and $A = \ln(n)$ gave the best results for logistic regression. This is consistent with the results of AGPN in which it was found that linear regression required a significantly larger number of predictors than logistic regression.

d. The binning method

In this methodology, the model-predicted precipitation interpolated from the model grid to each station is the only predictor, R, was divided into six bins as follows: $R = 0.0$ in., $0.0 < R < 0.01$ in., $0.01 \leq R < 0.05$ in., $0.05 \leq R < 0.10$ in., $0.10 \leq R < 0.25$ in., and 0.25 in. $\leq R$. Other precipitation ranges and numbers of bins were tried and it was found that the results were not sensitive to the details. If, in the training dataset, any bin had less than five forecasts of precipitation falling within its range, we combined this bin with the next lower one. This situation occurred at some of the stations with drier climatologies.

The probability forecasts were constructed by taking the entire 3-yr training set of model predicted precipitation (R) and calculating the frequency of the observed rainfall (y) exceeding each threshold (in our case, 0.01, 0.05, or 0.10 in.) for each of the six bins. If, for example, on a given day in the training dataset, the observed rainfall is 0.07 in., the corresponding values of y for 0.01, 0.05, and 0.10 in. would be 1, 1, and 0, respectively. When R falls within a given bin (e.g., $0.05 \leq R < 0.10$ in.), the predicted probability of precipitation ($\hat{p}$) exceeding a given threshold (e.g., 0.01 in.) is taken as the ratio

$$\hat{p} = \frac{\text{number of days in the training data with } y = 1 \text{ and } 0.05 \leq R < 0.10 \text{ in.}}{\text{number of days in the training data with } 0.05 \leq R < 0.10 \text{ in.}}. \tag{6}$$

This ratio is determined from the training data. Suppose, for example, that in the training dataset there were 88 cases in which the rainfall was predicted in the range $0.05 \leq R < 0.10$ in. If, in 44 of these cases, the observed rainfall exceeded 0.01 in., the forecast probability $\hat{p}$ for this threshold would be taken as 50%. Moreover, if in 22 of the 88 cases, the observed rainfall exceeded 0.05 in., the forecast probability $\hat{p}$ for this threshold would

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

TABLE 2. Brier skill scores for the six models (two linear regression models, three logistic regression models, and binning) at the three precipitation thresholds. The symbols BSS and S+ within the parentheses indicate the method of choosing the model order (Brier skill score or generalized information criterion, respectively). Here, PC represents the use of principal components as predictors.

	Precipitation thresholds		
	0.01 in.	0.05 in.	0.10 in.
Linear regression (BSS)	0.378	0.440	0.457
Linear regression (S+)	0.389	0.462	0.472
Logistic regression (BSS)	0.384	0.472	0.478
Logistic regression (S+)	0.407	0.491	0.503
Logistic regression (PC)	**0.413**	0.480	0.473
Binning	0.400	**0.492**	**0.510**

be taken as 25%. This methodology represents a form of conditional probability (Wilks 1995). Conditional probability has also been used in ensemble forecasting, where a probability forecast is made by taking the ratio of the number of ensemble members that forecast an event to the total members of the ensemble (e.g., Du et al. 1997; Hamill and Colucci 1998; Buizza et al. 1999; Ebert 2001).

3. Results

In order to make maximum use of our dataset, we employed cross validation (Elsner and Schmertmann 1994), where the coefficients corresponding to each model are determined using three years of training data and verified on the fourth. By choosing different combinations of three years among the four for training, verification forecasts were produced for four years of independent data. Table 2 compares the Brier skill scores for the methods used by AGPN [viz., linear and logistic regression using a stopping rule based on the Brier skill score to limit the number of predictors, and logistic regression using predictors chosen by principal component (PC) analysis] with the new methods (viz., linear regression and logistic regression using S-Plus software, and binning). Each of these scores is the mean of the individual scores for the 154 stations over the four independent cool seasons. Every method is seen to do better than climatology, with scores ranging from 0.378 to 0.510 and (with the exception of logistic regression using PC analysis) increasing as the threshold is increased from 0.01 to 0.10 in. However, it should be noted that Brier skill scores in other studies have been found to decrease at thresholds above 0.10 in. (Ebert 2001). At the lowest threshold, logistic regression using predictors determined from principal component analysis gives the highest BSS, while, at the higher thresholds, the binning method, which removes the bias from the numerical predictions of precipitation accumulation, scores best. The use of the S-Plus software, which incorporates the generalized information criterion for determining the model order, as well as the maximum likelihood approach for choosing the coefficients, improves the scores at all thresholds for both linear regression and logistic regression over those obtained by the use of a stopping rule. At the 0.01-in. threshold, however, logistic regression using predictors chosen by PC analysis still exhibits the greatest skill.

Following the approach in AGPN, we made a paired *t* test (Weiss and Hassett 1991) to determine, for each threshold, the extent to which the differences among the Brier skill scores corresponding to the different methodologies are significantly different from one another. The results of this test are shown in Tables 3–5. The numbers in the body of these tables are the *t* values, positive indicating that the method in the left-hand column has a higher mean BSS than the one listed along the top row, and negative indicating the reverse. A magnitude greater than 2.576 in the table indicates with at least 99% confidence that the method with the higher mean BSS is significantly more skilled than the one with the lower mean score. For the 95% and 90% confidence limits the corresponding criteria are 1.960 and 1.645, respectively. In each table, the numbers to the lower left of the main diagonal are mirror images of those to 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 gives the *t* values for the 0.01-in. threshold. We compare first the results for linear and logistic regression based on the use of the S-Plus software with those for the corresponding methodologies using a stopping rule based on the BSS reported in AGPN. The boldfaced number in the second row indicates that the BSS of 0.389 for linear regression using S+ (Table 2)

TABLE 3. Paired difference *t* scores for the six models at the threshold of 0.01 in. 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 the 90%, 95%, and 99% confidence levels are 1.645, 1.960, and 2.576, respectively.

	Paired difference *t* scores for 0.01 in.					
	Li (BSS)	Li (S+)	Lo (BSS)	Lo (S+)	Lo (PC)	Binning
Li (BSS)	0.000	−4.716	−1.217	−10.956	−11.523	−5.872
Li (S+)	**4.716**	0.000	1.086	−9.138	−10.311	−3.138
Lo (BSS)	1.217	−1.086	0.000	−5.749	−6.325	−5.215
Lo (S+)	10.956	9.138	**5.749**	0.000	−**2.085**	2.702
Lo (PC)	11.523	10.311	6.325	2.085	0.000	3.472
Binning	5.872	3.138	5.215	−2.702	−3.472	0.000

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

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

	Paired difference *t* scores for 0.05 in.					
	Li (BSS)	Li (S+)	Lo (BSS)	Lo (S+)	Lo (PC)	Binning
Li (BSS)	0.000	−10.292	−5.656	−14.176	−11.044	−13.006
Li (S+)	**10.292**	0.000	−1.694	−9.288	−6.686	−7.760
Lo (BSS)	5.656	1.694	0.000	−4.675	−1.641	−7.059
Lo (S+)	14.716	9.288	**4.675**	0.000	**3.556**	−0.385
Lo (PC)	11.044	6.686	1.641	−3.556	0.000	−3.121
Binning	13.006	7.760	7.059	0.385	3.121	0.000

is significantly higher at the 99% confidence level than the score of 0.378 obtained using our earlier method. The boldfaced numbers in the fourth row of Table 3 indicate that the BSS of 0.407 in Table 2 for logistic regression using S-Plus is significantly higher at the 99% level than is the score of 0.384 for the previous method (which uses a stopping rule based on changes in the BSS as predictors are added), but it is significantly lower at the 95% level than the score of 0.413 for logistic regression using the leading principal components as predictors.

The last row of Table 3 reveals that the score of 0.400 in Table 2 for the 0.01-in. threshold using the binning method is significantly higher at the 99% level than both scores for linear regression and the score for logistic regression using a stopping rule to limit the number of predictors, but significantly lower than the scores for logistic regression using the top 10 predictors from a PC analysis or using S-Plus. As indicated by the comparisons in the fifth row, logistic regression using PC analysis is clearly the method with the greatest overall skill at the 0.01-in. threshold, while the comparisons in the first two rows indicate that linear regression is the method with the lowest overall skill.

The *t* values for the 0.05- and the 0.10-in. thresholds are given in Tables 4 and 5, respectively. The boldfaced numbers in the second and fourth rows of each of these tables reveal that the Brier skill scores for both linear and logistic regression at this threshold corresponding to the use of the S-Plus software are significantly higher at the 99% level than their counterparts using our previous methodologies. Moreover, the comparisons in the last row of each table reveal that the Brier skill scores of 0.492 for 0.05 in. and 0.510 for 0.10 in. (Table 2) corresponding to the use of binning are significantly higher at the 99% level than those for all methodologies other than logistic regression using S-Plus software. Since the skill scores corresponding to binning and logistic regression using S-Plus software are not significantly different from each other at either the 99% or 95% level binning must be regarded as the method of choice for practical forecasting because it is computationally much more efficient than logistic regression using S-Plus or any other software. It should be emphasized that these conclusions are based on the use of the NGM model over the eastern half of the United States during the winter season. For other regions, other seasons, or the use of other numerical models, studies such as this one and the one by AGPN should be repeated.

Attributes diagrams (Murphy 1973; Wilks 1995), such as those shown in Figs. 2–4, provide insight into the relative performances of different methodologies. In particular, they display graphically the reliability of forecasts of different probabilities and the frequency with which they are made by each methodology.

A forecast methodology is considered perfectly reliable if, corresponding to each range of predicted probabilities, the observed frequencies of precipitation equaling or exceeding the threshold are equal to the predicted probabilities. The closer the points on a plot of observed frequency versus forecast probability are to the diagonal from (0, 0) to (1, 1), the more reliable are the forecasts. The upper graphs in Figs. 2–4 reveal that logistic regression forecasts on an independent dataset using S-Plus software are systematically more reliable at all three thresholds than the linear regression forecasts on the same dataset using the same software. While not shown here, the curves for binning are almost identical to those for logistic regression.

Good or perfect reliability is not, however, a sufficient condition for a forecast methodology to achieve a high BSS. It is necessary that the methodology also make a

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

	Paired difference *t* scores for 0.10					
	Li (BSS)	Li (S+)	Lo (BSS)	Lo (S+)	Lo (PC)	Binning
Li (BSS)	0.000	−5.766	−3.134	−10.704	−3.809	−11.042
Li (S+)	**5.766**	0.000	−0.756	−8.070	−0.285	−8.765
Lo (BSS)	3.134	0.756	0.000	−5.040	0.600	−7.206
Lo (S+)	10.704	8.070	**5.040**	0.000	**7.372**	−1.758
Lo (PC)	3.809	0.285	−0.600	−7.372	0.000	−7.246
Binning	11.042	8.765	7.206	1.758	7.246	0.000

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

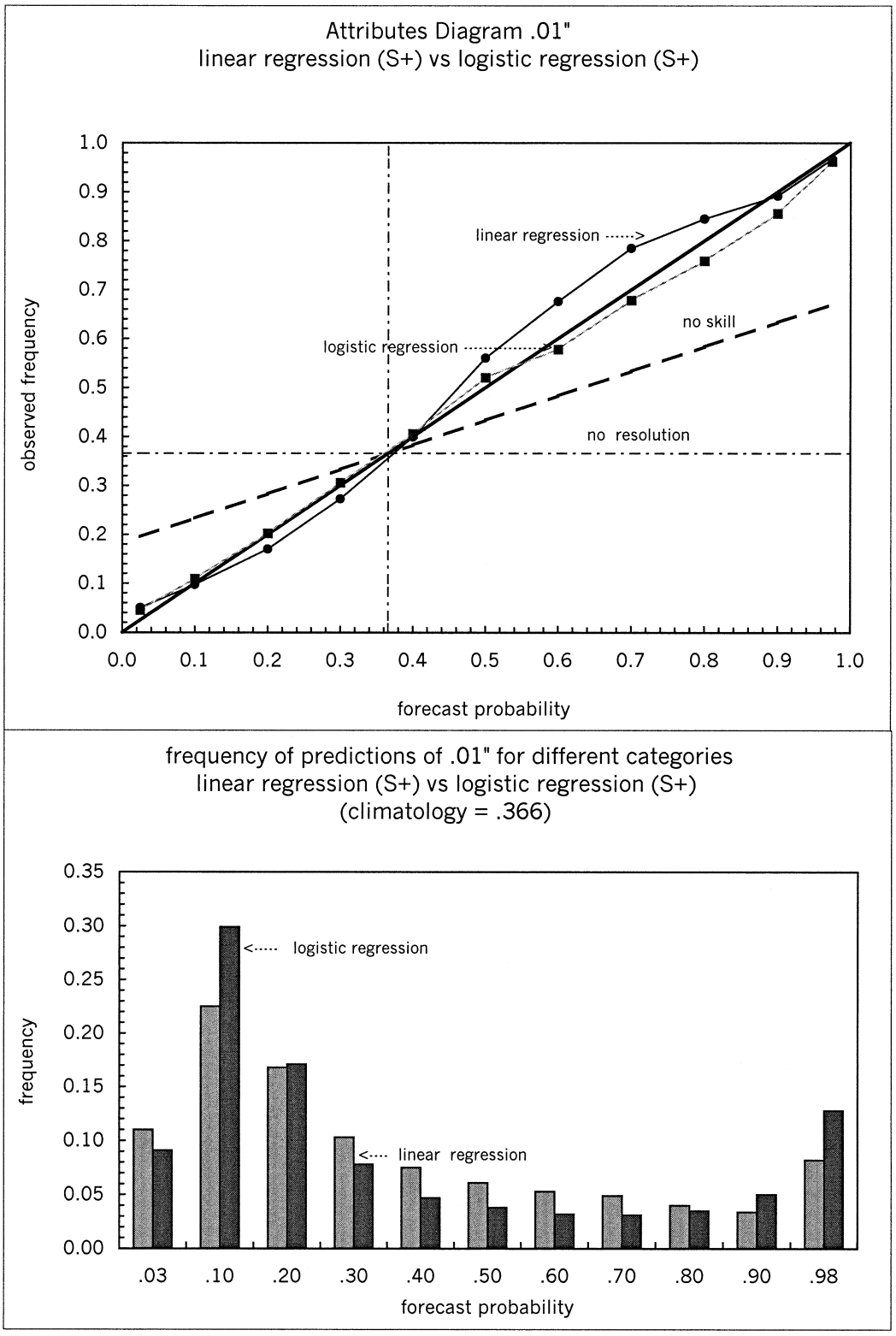

FIG. 2. Attributes diagram for all 154 stations for the 0.01-in. threshold for linear regression and logistic regression using the S-Plus software. The heavy solid line represents perfect reliability. The dotted–dashed line represents ''no resolution,'' and the heavy dashed line represents ''no skill.'' The histogram below shows the frequency of the different probability forecasts made by linear and logistic regression for all 154 stations at the 0.01-in. threshold.

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

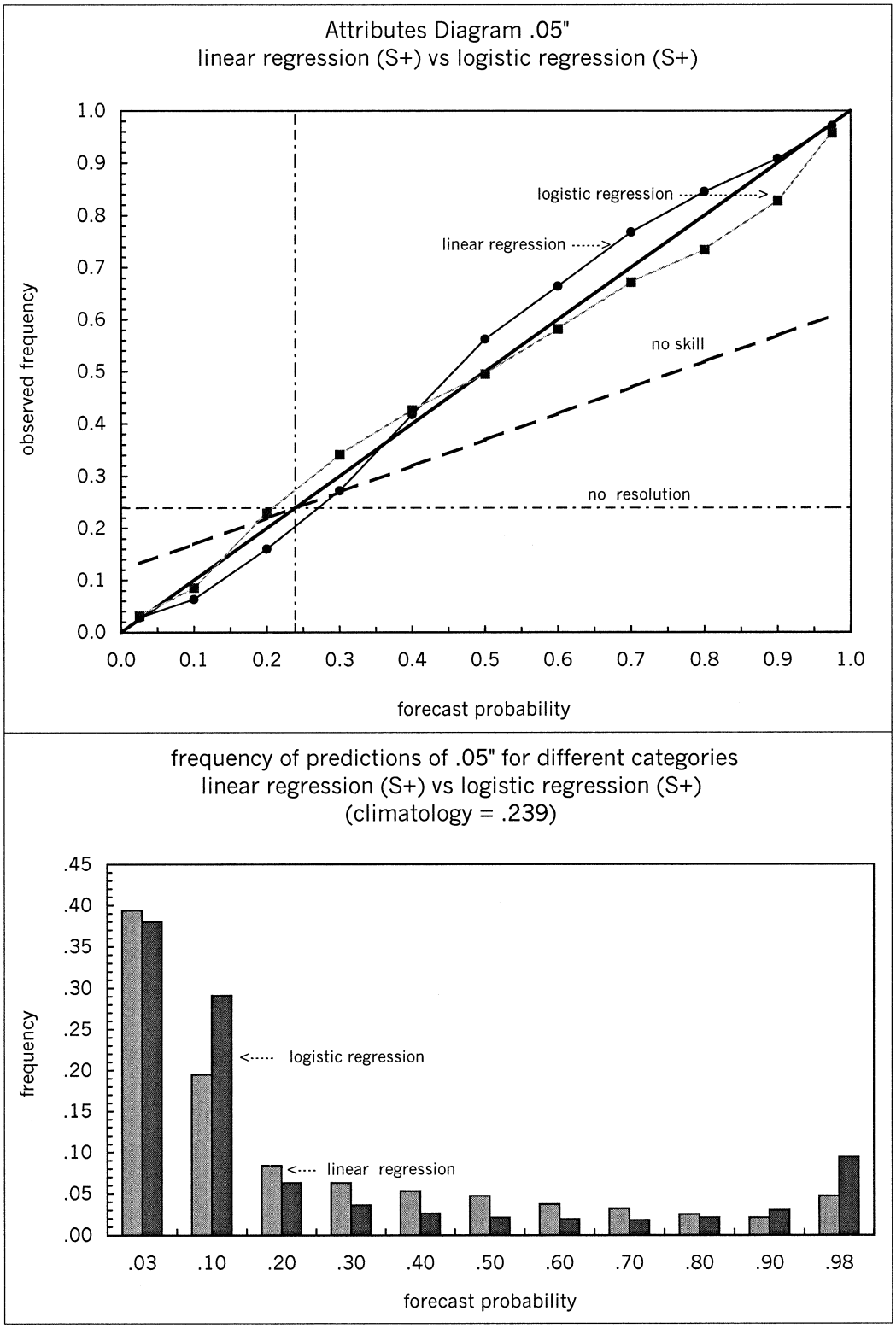

FIG. 3. Same as in Fig. 2 except for at the 0.05-in. threshold.

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

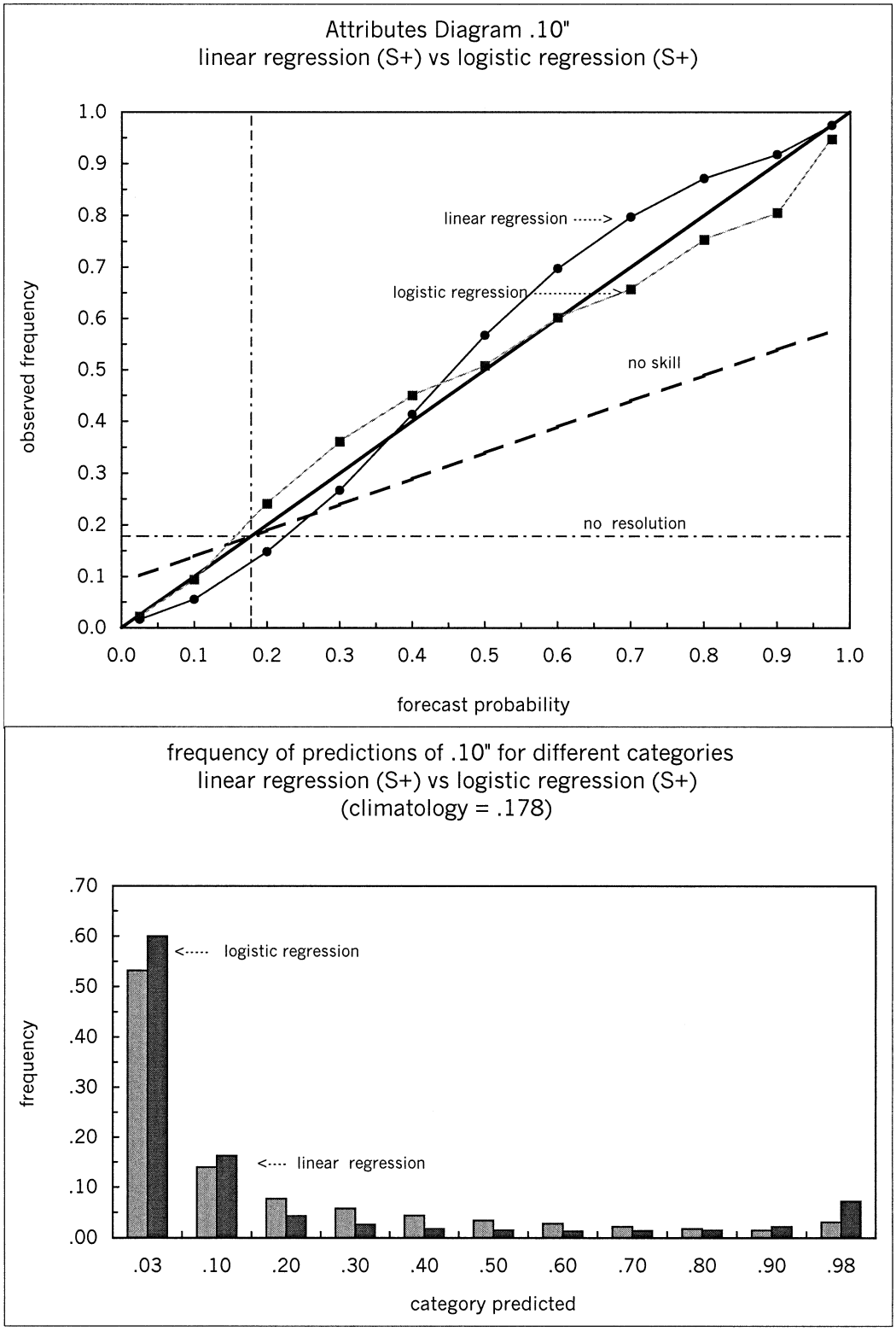

FIG. 4. Same as in Fig. 2 except for at the 0.10-in. threshold.

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

TABLE 6. BSSs for the three different thresholds for the linear regression and logistic regression models using three different values for the generalized information criterion [$A = 2$, $A = \ln(n)$, $A = \sqrt{n}$].

	Precipitation thresholds		
	0.01 in.	0.05 in.	0.10 in.
Linear regression ($A = 2$)	**0.389**	**0.462**	**0.472**
Linear regression [$A = \ln(n)$]	0.378	0.453	0.470
Linear regression ($A = \sqrt{n}$)	0.350	0.425	0.457
Logistic regression ($A = 2$)	0.404	0.483	0.483
Logistic regression [$A = \ln(n)$]	**0.407**	**0.491**	0.503
Logistic regression ($A = \sqrt{n}$)	0.379	0.477	**0.505**

TABLE 7. Paired difference t scores for the 0.01-in. threshold for logistic regression and linear regression using the generalized information criterion [$A = 2$, $A = \ln(n)$, $A = \sqrt{n}$].

	Linear regression		
	$A = 2$	$A = \ln(n)$	$A = \sqrt{n}$
Linear regression ($A = 2$)	0.000	**7.731**	**13.855**
Linear regression [$A = \ln(n)$]	−7.731	0.000	12.922
Linear regression ($A = \sqrt{n}$)	−13.855	−12.922	0.000
	Logistic regression		
	$A = 2$	$A = \ln(n)$	$A = \sqrt{n}$
Logistic regression ($A = 2$)	0.000	−2.216	8.494
Logistic regression [$A = \ln(n)$]	**2.216**	0.000	**10.729**
Logistic regression ($A = \sqrt{n}$)	−8.494	−10.729	0.000

significant number of forecasts of probabilities that are not close to, or equal to, the climatological probability for each location. This is because the BSS is zero (i.e., no skill) when the forecast probability is equal to the climatological frequency of the event. A methodology with good reliability that gives probability forecasts far removed from the climatological frequency will have a significantly higher BSS than one with equally good reliability that gives probability forecasts close to the climatological frequency. The histograms at the bottom of Figs. 2–4 reveal that logistic regression gives many more forecasts far removed from the climatological frequency and that the linear regression forecast probabilities tend to be much closer to the climatological frequencies at all three thresholds.

The conclusions reached here about the greater reliability and greater frequency of forecasts far removed from climatology for logistic regression are similar to those reached in AGPN. The Brier skill scores are, however, higher, and the frequencies of logistic regression forecasts of 98% chance of precipitation equaling or exceeding the threshold are significantly greater with the use of the S-Plus software.

4. Sensitivity to the choice of the penalty weight A

In the statistics literature, two choices are recommended for the value of the penalty function, A, in (5). One is based on the Akaike information criterion ($A = 2$) and the other on the Bayesian information criterion [$A = \ln(n)$]. As noted earlier, the former yields a higher-order model with more predictors than the latter. For a given application, the value that gives the best results depends on the nature of the problem and the structure of the data and must be determined empirically. In the present study, with $n = 360$, we compared both of these choices with a third choice (viz., $A = \sqrt{n}$) that yields an even smaller number of predictors than the Bayesian information criterion. The Brier skill scores for each of these values are shown in Table 6 and the t values giving the significance of the differences in the scores for each threshold are shown in Tables 7–9. The boldfaced numbers in Tables 7 and 8 indicate that, for the 0.01- and 0.05-in. thresholds, the Brier skill scores in Table 6 for linear regression with $A = 2$ and for logistic regression with $A = \ln(n)$ are significantly higher at the 99% level than the other scores for these thresholds. The boldfaced numbers in Table 9 indicate that, for the 0.10-in. threshold, the BSS of 0.472 for linear regression with $A = 2$ is not significantly different at the 99%, 95%, or 90% confidence level from the score of 0.470 for linear regression with $A = \ln(n)$, and, at the same levels of confidence, the BSS of 0.505 for logistic regression with $A = \sqrt{n}$ is not significantly different from the score of 0.503 for logistic regression with $A = \ln(n)$. The table also reveals that the scores for linear regression with either $A = 2$ or $A = \ln(n)$ are significantly higher at the 99% confidence level than that with $A = \sqrt{n}$, and the scores for logistic regression with either $A = \ln(n)$ or $A = \sqrt{n}$ are significantly higher at the same level of confidence than the score with $A = 2$. The fact that linear regression is optimized at all thresholds with smaller values of A, and therefore a greater number of predictors, and that logistic regression is optimized with a larger value of A, and therefore fewer predictors, is consistent with the results reported by AGPN, where it was found that logistic regression required fewer predictors than linear regression.

5. Significance test revisited

Since the completion of AGPN, some questions have arisen concerning our assumption that the differences

TABLE 8. Same as in Table 7 except for the 0.05-in. threshold.

	Linear regression		
	$A = 2$	$A = \ln(n)$	$A = \sqrt{n}$
Linear regression ($A = 2$)	0.000	**6.553**	**11.726**
Linear regression [$A = \ln(n)$]	−6.553	0.000	11.434
Linear regression ($A = \sqrt{n}$)	−11.726	−11.434	0.000
	Logistic regression		
	$A = 2$	$A = \ln(n)$	$A = \sqrt{n}$
Logistic regression ($A = 2$)	0.000	−4.194	1.578
Logistic regression [$A = \ln(n)$]	**4.194**	0.000	**5.267**
Logistic regression ($A = \sqrt{n}$)	−1.578	−5.267	0.000

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

TABLE 9. Same as in Table 7 except for the 0.10-in. threshold.

	Linear regression		
	$A = 2$	$A = \ln(n)$	$A = \sqrt{n}$
Linear regression ($A = 2$)	0.000	**1.096**	**4.174**
Linear regression [$A = \ln(n)$]	−1.096	0.000	**5.014**
Linear regression ($A = \sqrt{n}$)	−4.174	−5.014	0.000
	Logistic regression		
	$A = 2$	$A = \ln(n)$	$A = \sqrt{n}$
Logistic regression ($A = 2$)	0.000	−5.651	−4.970
Logistic regression [$A = \ln(n)$]	**5.651**	0.000	**−0.886**
Logistic regression ($A = \sqrt{n}$)	**4.970**	−0.886	0.000

between the Brier skill scores for linear regression and logistic regression at each of the 154 meteorological stations are independent of those at all the other stations. This is because of the expected spatial dependence of the predictors and the predictand associated with organized atmospheric disturbances. The assumption of independence led to the use of $m = 154$ in the paired t test for significance in that paper and in the earlier sections of the present paper. In order to address these concerns, we revisit the subject here using, for each methodology, only one score for each of the seven regions defined in AGPN. It will be recalled that the stations within each region were determined, with the use of factor analysis, as having meteorological characteristics that were similar to one another and different from those in the other regions. There can be no question, therefore, that the differences between the Brier skill scores for linear regression and logistic regression for each of the seven regions are independent of all the others. Moreover, to eliminate any concerns about the possible lack of independence of two nearby stations in adjacent regions, we eliminated all stations that are closer than 500 km to a station in an adjacent region. This reduced the total number of stations used for the new significance test to 88.

A Brier skill score was then computed for each of the seven regions using all of the remaining stations in the region, and a t test was performed on the paired differences of the scores for linear regression and logistic regression using the S-Plus software with $m = 7$. We regard the use of $m = 7$ as the most stringent possible significance test, since, in reality, there must be a larger number of stations over the eastern half of the United States that can be considered independent of one another. If significance at a sufficiently high level can be found using $m = 7$, we would feel even more confident in the conclusions.

TABLE 10. Averages of the seven regional BSSs, the corresponding paired difference t scores, and the confidence levels based on the limited sample of 88 stations.

	Precipitation thresholds		
	0.01 in.	0.05 in.	0.10 in.
BSS linear regression (S+)	0.390	0.467	0.474
BSS logistic regression (S+)	0.407	0.491	0.501
Paired difference t score	3.127	2.343	2.690
Confidence level	>95%	>90%	>95%

The averages of the Brier skill scores for linear regression and logistic regression over the seven regions are given in Table 10. Since 88 stations were used to obtain seven regional scores, which were then averaged to get the values in Table 10, whereas 154 individual scores were averaged to get the scores in Table 2, the corresponding scores in the two tables are not identical. Nevertheless, as in Table 2, the results in Table 10 reveal higher skill scores for logistic regression at all three thresholds. For $m = 7$, the 90%, 95%, and 99% confidence levels are now 1.943, 2.447, and 3.707. The t values for the three thresholds and the corresponding confidence levels are given in the last two rows of Table 10. While the confidence levels are below 99%, they are sufficiently high in this most stringent test to support the conclusion that logistic regression is significantly more skillful in PQPF using predictors derived from the NGM over the eastern half of the United States at all three thresholds.

6. Conclusions

The results of this follow-up study suggest the following for probability forecasting of cool season precipitation accumulations at thresholds of 0.01, 0.05, and 0.10 in. using NGM model analyses and forecasts.

1) The BSS for linear regression forecasts can be improved with the use of the Akaike information criterion to determine the model order.
2) The BSS for logistic regression can also be improved by use of the Bayesian information criterion to determine the model order and the maximum likelihood method with a Fisher scoring algorithm to determine the coefficients. At the lowest threshold, however, the use of the leading principal components determined from the original predictor set of meteorological variables still gives the highest BSS when applied with logistic regression.
3) For PQPF, logistic regression yields significantly

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC

higher Brier skill scores than does linear regression. This is not surprising, since standard linear regression is based on the assumption that the predictand is normally distributed, whereas the predictand in this case (precipitation exceeding or not exceeding a prescribed threshold) has a binomial distribution (0 or 1).

4) At the 0.01-in. threshold, logistic regression with predictors chosen from among the leading PCs as predictors gives significantly higher Brier skill scores than does binning. At the two higher thresholds, the differences between the Brier skill scores for binning and for logistic regression using the Bayesian information criterion and the maximum likelihood function with a Fisher scoring algorithm (which yields the highest skill score for logistic regression) are not significantly different from each other. Since binning is much less computationally intensive than logistic regression, it is suggested that this is the method of choice at these thresholds.
5) Further significance testing supports the notion that logistic regression is significantly more skillful than linear regression at very high confidence levels.

Acknowledgments. This research was supported by NOAA under CSTAR Grant NA17WA1010, and in part by NSF and NOAA under USWRP Grant ATM 9714414, and by AFOSR AASERT Grant F49620-93-1-0531. The authors would like to thank the Geophysical Fluid Dynamics Institute at The Florida State University (FSU) for supplying the computer resources necessary for implementing and running the statistical models, and Dr. Jon Ahlquist of the FSU Meteorology Department for some very insightful discussions. Additionally, the data were provided by the Data Support Section, Scientific Computing Division, at the National Center for Atmospheric Research, which is supported by grants from the National Science Foundation.

REFERENCES

Agresti, A., 1990: *Categorical Data Analysis.* John Wiley and Sons, 558 pp.

Akaike, H., 1973: Information theory and an extension of the maximum likelihood principle. *Proc. Second Int. Symp. on Information Theory,* Budapest, Hungary, Akademiai Kaido, 267–281.

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting,* **17,** 783–799.

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

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting,* **14,** 168–189.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.,* **125,** 2427–2459.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.,* **129,** 2461–2480.

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

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

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of the Eta/RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

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

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

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.,* **6,** 461–464.

Wedderburn, R. W. M., 1976: On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika,* **63,** 27–32.

Weisberg, S., 1985: *Applied Linear Regression.* John Wiley and Sons, 324 pp.

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.

Unauthenticated | Downloaded 05/07/21 02:12 AM UTC