

## A Case Study of the Use of Statistical Models in Forecast Verification: Precipitation Probability Forecasts

ALLAN H. MURPHY\*

*Prediction and Evaluation Systems, Corvallis, Oregon*

DANIEL S. WILKS

*Atmospheric Science Group, Cornell University, Ithaca, New York*

(Manuscript received 19 May 1997, in final form 27 January 1998)

### ABSTRACT

The traditional approach to forecast verification consists of computing one, or at most very few, quantities from a set of forecasts and verifying observations. However, this approach necessarily discards a large portion of the information regarding forecast quality that is contained in a set of forecasts and observations. Theoretically sound alternative verification approaches exist, but these often involve computation and examination of many quantities in order to obtain a complete description of forecast quality and, thus, pose difficulties in interpretation. This paper proposes and illustrates an intermediate approach to forecast verification, in which the multifaceted nature of forecast quality is recognized but the description of forecast quality is encapsulated in a much smaller number of parameters. These parameters are derived from statistical models fit to verification datasets. Forecasting performance as characterized by the statistical models can then be assessed in a relatively complete manner. In addition, the fitted statistical models provide a mechanism for smoothing sampling variations in particular finite samples of forecasts and observations.

This approach to forecast verification is illustrated by evaluating and comparing selected samples of probability of precipitation (PoP) forecasts and the matching binary observations. A linear regression model is fit to the conditional distributions of the observations given the forecasts and a beta distribution is fit to the frequencies of use of the allowable probabilities. Taken together, these two models describe the joint distribution of forecasts and observations, and reduce a 21-dimensional verification problem to 4 dimensions (two parameters each for the regression and beta models). Performance of the selected PoP forecasts is evaluated and compared across forecast type, location, and lead time in terms of these four parameters (and simple functions of the parameters), and selected graphical displays are explored as a means of obtaining relatively transparent views of forecasting performance within this approach to verification.

### 1. Introduction

The practice of forecast verification usually consists of calculating one or two measures of overall forecasting performance based on a particular sample of forecasts and observations. For example, forecasters might calculate the probability of detection (POD) and the false alarm rate (FAR), given a set of yes–no precipitation occurrence forecasts summarized in a  $2 \times 2$  contingency table. However, the fact that method A's forecasts are more accurate or more skillful than method B's forecasts, according to some single measure of overall per-

formance, is no guarantee that all aspects of A's performance are superior to all aspects of B's performance. This traditional practice of relying on only one or two scores or summary statistics has been described as *measures-oriented* forecast verification (Murphy 1997). It is relatively easy to show that the measures-oriented approach to verification problems (VPs) is inadequate in general to demonstrate unambiguous superiority (i.e., superiority in terms of the economic value of the forecasts to all users and in terms of all relevant aspects of forecast quality), even in simple settings such as that described above. The inadequacy of the measures-oriented approach is even more pronounced in VPs involving multicategory variables and/or probabilistic forecasts.

The information content in a dataset of forecasts and their corresponding observations can be organized and displayed as a joint frequency distribution of the forecasts and observations. The familiar  $2 \times 2$  contingency table is the simplest possible example of this joint dis-

---

\* Deceased.

---

Corresponding author address: Dr. Daniel S. Wilks, Atmospheric Science Group, Cornell University, 1113 Bradfield Hall, Ithaca, NY 14853.  
E-mail: dsw5@cornell.edu

tribution. Under specific but not unduly restrictive assumptions (see section 2), the joint distribution of forecasts and observations contains all of the information required for a complete assessment of the various aspects of forecast quality and, thus, provides the basis for a sound approach to VPs (Murphy and Winkler 1987). Verification approaches based on this joint distribution can be described as *distributions oriented* (Murphy 1997). Comparison of measures- and distributions-oriented approaches for a particular verification problem has been described recently by Brooks and Doswell (1996). A distributions-oriented approach is diagnostic in the sense that it places particular emphasis on the assessment of basic strengths and weaknesses in forecasting performance, with the ultimate goal of guiding efforts to improve forecasting methods and models.

A basic concept within the distributions-oriented approach is the notion of the dimensionality of VPs (Murphy 1991). In this context, dimensionality refers to the number of parameters (e.g., joint probabilities) that must be determined in order to reconstruct the underlying empirical joint distribution. Consideration of traditional practices from the perspective of the distributions-oriented approach reveals that evaluation methods associated with these practices generally fail to respect the dimensionality of VPs. As a result, measures-oriented practices frequently overlook important aspects of forecast quality, thereby yielding potentially misleading results regarding absolute and relative forecasting performance, and possibly even producing incorrect orderings of competing forecasts in terms of their economic value to particular users (e.g., Brooks and Douglas 1998; Murphy and Ehrendorfer 1987; Murphy 1997).

When the distribution-oriented approach to VPs is applied to empirical joint distributions of forecasts and observations (and/or to the associated conditional and marginal empirical distributions into which the joint distribution may be factored), the dimensionality of these problems can become quite large (see section 2). That is, a complete description of forecast quality may require the determination of a relatively large number of parameters. This fact alone may deter practitioners from adopting approaches that fully respect the underlying dimensionality of VPs.

Murphy (1991) suggested that it might be possible to reduce the dimensionality of VPs in an efficient and effective manner by fitting statistical models to the basic joint, conditional, and/or marginal distributions. Evaluation or comparison of forecasting performance would then be based on the parameters of the statistical models. It should be noted that the use of parametric statistical models to describe forecast quality is not a new concept. For example, Anders Angstrom used a Gaussian distribution to model forecast errors more than 75 years ago (see Liljas and Murphy 1994). Statistical models have been used in the context of forecast verification by Clemen and Winkler (1987), Mason (1982), and Wilks and

Shen (1991). In addition, several decision-analytic studies of the economic value of weather/climate forecasts (e.g., Katz et al. 1982; Wilks 1991; Wilks and Murphy 1986; Wilks et al. 1993) have employed statistical models of forecast quality to facilitate, among other things, the assessment of relationships between the quality and economic value of forecasts. Such models have also been used in studies in which the sufficiency relation, which can in some cases reveal unambiguous superiority of one set of forecasts over another for all forecast users (this is more fully described in section 5), was applied to problems involving the comparative evaluation of forecasts (e.g., Krzysztofowicz 1992; Krzysztofowicz and Long 1991a,b). What is novel in the results to be presented here is the use of the parameters of the statistical models as measures of aspects of forecast quality in the context of a diagnostic approach to VPs.

The purposes of this paper are to describe the motivation and background for the application of statistical models to the problem of forecast verification, and to report some results of a study that illustrates this approach for the case of selected samples of historical precipitation probability forecasts. Section 2 examines deficiencies in the traditional verification practices with respect to two issues: (a) the dimensionality of VPs and (b) the sampling variability of verification results. The statistical models used here to fit verification data samples, consisting of probability of precipitation (PoP) forecasts and the matching dichotomous observations, are identified in section 3. This section also discusses the goodness of these fits. The use of the parameters of these models to evaluate and compare PoP forecasting performance is illustrated in section 4. Section 5 presents the results of a companion study in which the use of statistical models to reduce the effects of sampling variability is investigated. This exploratory study addresses the sampling variability issue in the context of an assessment of the unambiguous superiority of alternative PoP forecasts. Section 6 contains a short discussion of some implications of these results and several outstanding issues in this area, as well as some concluding remarks.

## 2. The practice of forecast verification: Two issues

As noted in section 1, traditional practices in forecast verification usually involve calculating one or two measures of overall performance for a verification data sample (VDS) consisting of matched pairs of forecasts and observations. Although this measures-oriented approach may be adequate to identify gross features of forecasting performance such as overall accuracy or skill, it is inadequate in at least two important ways. First, measures-oriented approaches generally fail to respect the underlying dimensionality of VPs. As a result, verification studies frequently overlook important aspects of forecasting performance and, thereby, may produce misleading results concerning the quality of alternative

forecasting methods or models. Second, traditional practices usually ignore the problem of sampling variability. As a consequence, it may be difficult to judge to what extent the results of a specific verification study are representative of the results that would have been obtained if the study had been based on another VDS (e.g., forecasts and observations for the same weather element at the same location, but for a different month, season, or year).

Under the assumption that a VDS represents a stationary bivariate time series consisting of independent pairs of forecasts ( $f$ ) and observations ( $x$ ), the empirical joint relative frequencies of  $f$  and  $x$  represent estimates of the probabilities that constitute the joint distribution of forecasts and observations. This joint distribution, denoted here by  $p(f, x)$ , contains all of the information in the VDS relevant to the assessment of the various aspects of forecast quality. Moreover, it is possible to factor  $p(f, x)$  into conditional and marginal distributions in two ways:

$$p(f, x) = \begin{cases} q(x|f)s(f), & (1a) \\ r(f|x)t(x), & (1b) \end{cases}$$

where  $q(x|f)$  represents the conditional distributions of the observations given the forecasts,  $r(f|x)$  represents the conditional distributions of the forecasts given the observations,  $s(f)$  represents the marginal distribution (the frequencies of use) of the forecasts, and  $t(x)$  represents the marginal distribution of the observations (the sample climatological distribution). In brief, these factorizations reveal that forecast quality can be fully described by specifying  $p(f, x)$ ,  $q(x|f)$ , and  $s(f)$ , or  $r(f|x)$  and  $t(x)$  (Murphy and Winker 1987). These three specifications are most appropriately viewed as complementary (as opposed to alternative) descriptions of forecast quality.

From the perspective of the VDS, the dimensionality of a VP can be defined as the number of joint relative frequencies that must be specified in order to recover the underlying empirical joint distribution. Equivalent definitions of dimensionality can be formulated in terms of conditional and/or marginal relative frequencies. Since the joint relative frequencies must sum to one, the dimensionality ( $d$ ) can be defined as

$$d = n_f n_x - 1, \quad (2)$$

where  $n_f$  and  $n_x$  are the number of distinct values that may be taken on by the forecasts and observations, respectively. That is, it takes a minimum of  $d$  independent parameters (e.g., joint relative frequencies) to specify fully the joint distribution  $p(f, x)$ .

The dimensionality of some common VPs involving nonprobabilistic or probabilistic forecasts is indicated in Table 1. For example, the case of nonprobabilistic yes-no forecasts for precipitation occurrence yields the simplest possible joint distribution, with  $n_x = 2$ ,  $n_f = 2$ , and  $d = 3$ . In this problem, three independent pa-

TABLE 1. Dimensionality of some common verification problems involving (a) nonprobabilistic forecasts and (b) probabilistic forecasts, when the underlying distributions are modeled in terms of empirical joint, conditional, and/or marginal relative frequencies of forecasts and observations.

No. of observations $n_x$	No. of forecasts $n_f$	Dimensionality $d$
(a) Nonprobabilistic forecasts		
2	2	3
3	3	8
5	5	24
10	10	99
20	20	399
(b) Probabilistic forecasts*		
2	11	21
3	66	197
4	286	1143

\* The number of distinct probabilistic forecasts is based on the assumption that 11 permissible probabilities (e.g., 0.0, 0.1, 0.2, . . . , 1.0) are available for use in the forecasts.

rameters must be specified in order to determine  $p(f, x)$ . These three parameters could be any three joint relative frequencies, two independent conditional relative frequencies and one marginal relative frequency, or three independent verification measures. For example, for a given sample size the four entries in the  $2 \times 2$  contingency table can be expressed algebraically in terms of the  $d = 3$  parameters: hit rate, POD, and FAR. It is evident that VPs involving nonprobabilistic forecasts with more than two possible forecasts and observations, as well as problems involving probabilistic forecasts, are of considerably greater dimensionality.

The contents of Table 1 indicate that one or two measures of overall performance are inadequate to reconstruct the underlying joint distribution in all VPs. Thus, a verification study based on such an approach necessarily overlooks potentially important aspects of forecast quality. Viewed from this perspective, it is clear that approaches that fully respect the underlying dimensionality of these problems generally require the determination of a relatively large number of parameters. Obviously, it would be desirable to find a conceptually sound way to reduce the dimensionality of VPs and yet retain the important characteristics of forecasting performance reflected in the underlying empirical joint distribution.

Another issue that arises when absolute or relative forecasting performance is evaluated on the basis of specific VDS is the problem of sampling variability. Obviously, complete reliance on results obtained from a particular data sample is seldom warranted, unless the sample size is very large and otherwise representative. It may be possible to assess the effects of sampling variability on such results by applying various well-known procedures of classical statistical inference, but the assumptions underlying these procedures are not sat-

ified in many situations. An inferential approach based on modern computer-intensive methods such as resampling would provide a means of assessing the effects of sampling variability without invoking these assumptions.

As an alternative to these inferential procedures, the effects of sampling variability could be reduced by modeling the empirical distributions of forecasts and/or observations derived from the VDS. The basic notion here is that parametric models would smooth out at least some of the variability associated with particular data samples. At the same time, it is essential that the modeled distributions retain the important features of the forecasting performance that are embodied in the underlying empirical distributions.

### 3. Precipitation probability forecasts and binary observations: Statistical models

#### a. Distributions and models

The VPs of interest here involve PoP forecasts and binary observations. If these PoP forecasts employ 11 possible probability values (e.g.,  $F = 0.0, 0.1, 0.2, \dots, 0.9, 1.0$ ) and the corresponding observations consist of only the occurrence ( $X = 1$ ) or nonoccurrence ( $X = 0$ ) of measurable precipitation, then  $n_f = 11$ ,  $n_x = 2$ , and the dimensionality of VPs of this kind is  $d = 11 \times 2 - 1 = 21$  [see (2) and Table 1]. That is, it takes at least 21 parameters (e.g., joint, conditional, and/or marginal relative frequencies) to describe forecast quality completely in the context of this VP.

From the factorizations of the joint distribution in (1a) and (1b), it is evident that this bivariate distribution can be described by modeling  $p(f, x)$ , by modeling  $q(x|f)$  and  $s(f)$ , or by modeling  $r(f|x)$  and  $t(x)$ . We choose to model the components of the calibration-refinement factorization (Murphy and Winkler 1987) of  $p(f, x)$  in (1a), namely,  $q(x|f)$  and  $s(f)$ . Specifically, we model the conditional distributions  $q(x|f)$  with a linear regression equation and the marginal distribution  $s(f)$  with a beta distribution.

In situations in which  $X \in \{0, 1\}$ ,  $q(X = 1|f) + q(X = 0|f) = 1$  and  $E(X|f) = q(X = 1|f)$ , where  $E(X|f)$  denotes the expectation (or mean) of  $X$  given  $F = f$ . Thus, modeling the conditional means  $E(X|f)$  is equivalent to modeling the conditional distributions  $q(x|f)$ . Here, we consider a simple but natural model for such conditional expectations, namely, a linear regression model in which the forecasts are regressed on the observations. This model takes the following form:

$$E(X|F = f) = b_0 + b_1 f, \quad (3)$$

where  $b_0$  and  $b_1$  are estimates of the (unknown) regression coefficients. These estimates are determined by minimizing the sum of squared deviations of  $E(X|F = f)$  in (3) from the corresponding observed values of  $X$  (i.e., the estimates of  $b_0$  and  $b_1$  are least squares estimates of the regression coefficients).

The beta distribution is a plausible candidate model for the marginal distribution of the PoP forecasts,  $s(f)$ . This distribution is defined over the closed unit interval (recall that  $0 \leq f \leq 1$ ), and it can assume a variety of shapes that at least approximate the relative-frequency-of-use of probability values in various samples of PoP forecasts. If the probability density function of the beta distribution for  $F$  is denoted by  $h(f)$ , then

$$h(f) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} f^{a-1} (1-f)^{b-1}, \quad (4)$$

$$(0 \leq f \leq 1, a > 0, b > 0),$$

where  $a$  and  $b$  are the parameters of the beta distribution and  $\Gamma(z)$  is the gamma function of  $z$  (e.g., Wilks 1995, p 96). When the values of  $a$  and  $b$  are both greater than one, the distribution is unimodal. For  $a < 1$  probability tends to be concentrated near  $F = 0$  and for  $b < 1$ , probability tends to be concentrated near  $F = 1$ , so that the distribution is U-shaped when  $a$  and  $b$  are both less than one. Expressed in terms of the parameters  $a$  and  $b$ , the mean of the beta distribution is  $E(F) = a/(a+b)$  and its variance is  $V(F) = ab/[(a+b)^2(a+b+1)]$ .

The linear regression model of the conditional distributions  $q(X = 1|f) = E(X|F = f)$  and the beta model of the marginal distribution  $s(f)$  each involve two parameters. These parameters must be estimated from the data in a VDS. Taken together, these two models describe the joint distribution of forecasts and observations  $p(f, x)$  [see (1a)]. Use of these models to describe forecasting performance, instead of the empirical joint, conditional, and/or marginal relative frequencies, reduces the dimensionality of the underlying VP from  $d = 21$  to  $d = 4$ . Evaluation of the quality of example sets of PoP forecasts in terms of the parameters of these models is described in section 4.

#### b. Fitting distributions to verification data samples

The VDSs considered here consist of numerical-statistical PoP forecasts, local (subjectively formulated) PoP forecasts for the same times, and the corresponding observations for two locations in the United States during the period October 1983 through March 1987. The numerical-statistical forecasts are based on the Model Output Statistics (MOS) approach (Glahn and Lowry 1972). These historical MOS PoP forecasts were derived from the Limited Fine Mesh (LFM) model, and generally were available as guidance to the National Weather Service forecasters who formulated the local (LCL) PoP forecasts (Carter et al. 1989). Both the MOS and LCL PoP forecasts were formulated twice each day (0000 and 1200 UTC cycles) for three consecutive 12-h periods, with lead times of 12–24 h, 24–36 h, and 36–48 h. In this paper, we restrict our attention to forecasts from the 1200 UTC cycle, for lead times of 12–24 and 36–48 h. These two lead times were chosen so

that forecast quality can be compared across different lead times for exactly the same set of forecast valid periods.

As indicated in section 3a, beta distributions are fit to the marginal distributions of forecasts (i.e., the relative frequencies with which the forecast probabilities are used) and regression models are fit to the conditional distributions of observations given each of the permissible forecast probabilities. The former involve only the forecasts, whereas the latter involve both the forecasts and observations. For this reason, it is convenient to discuss the fitting of the marginal distributions  $s(f)$  first and the fitting of the conditional distributions  $q(x|f)$  second.

### 1) BETA MODEL FITS

The fits of the beta model to the marginal distributions of forecasts  $s(f)$  for the cool season (October–March) at Syracuse, New York, and the warm season (April–September) at Tucson, Arizona, are depicted in Figs. 1 and 2, respectively. These distributions have been fit using the method of moments. Examination of beta model fits for Syracuse and Tucson permits consideration of VDSs with very different climatological probabilities of precipitation occurrence. Sample sizes for these two VDSs are  $n = 646$  for Syracuse and  $n = 473$  for Tucson. The following permissible probability values were used: 0.00, 0.02 (MOS only), 0.05, 0.10, 0.20, . . . , 0.90, and 1.00. The rounding to these values is a consequence of the format in which these historical forecasts were archived (Carter and Polger 1986). In the fitting process, the empirical distributions of forecasts were represented by histograms involving  $n_f = 11$  distinct intervals (or bins) of probability values. These bins were defined as follows: 0.000–0.049, 0.050–0.149, 0.150–0.249, . . . , 0.750–0.849, 0.850–0.949, and 0.950–1.000. Each MOS LCL PoP forecast was assigned to the appropriate probability bin, and the bins were represented by their respective midpoints. The heights of the histogram bars in Figs. 1 and 2 have been rescaled so that the areas in the rectangles sum to one, which results in equal integrals and thus equivalent vertical scales for both the smooth beta densities and the histograms, allowing the two to be compared visually.

Qualitatively, the beta distributions in Figs. 1 and 2 appear to represent fairly good fits to the empirical distributions (i.e., the histograms). In the case of Syracuse, the most noticeable differences between the beta models and the histograms occur in the 0.950–1.000 probability interval for the 12–24-h forecasts. Relatively large differences between the models and the histograms are also evident in the 0.050–0.149 interval for both lead times in the case of Tucson. These apparent discrepancies might be the result of relatively small numbers of extreme forecasts, artifacts associated with the rounding process (particularly near  $F = 0$  and  $F = 1$ ), funda-

mental inadequacies of the beta model for these forecasts, or some combination of these.

The chi-square statistic was computed as a quantitative measure of the overall fit of the beta models to the empirical distributions, and the values of the statistic for the various combinations of location, forecast type, and lead time appear in Figs. 1 and 2 (the chi-square statistics are also recorded in Table 2). These statistics are highly significant for all four forecast type–lead time combinations at Tucson and they are significant for three of the four combinations at Syracuse. In the strict statistical sense, then, relatively large differences exist between the empirical and modeled relative frequencies in most of the cases considered. While these differences suggest caution for quantitative application of the present results for these datasets, the qualitative reasonableness of the fits suggests that these beta distributions are adequate at least to illustrate the technique proposed here.

### 2) REGRESSION MODEL FITS

Some results of fitting regression models to the conditional distributions  $q(x|f) = E(X|f)$  for cool season VDSs from Syracuse, New York, and warm season VDSs from Tucson, Arizona, are shown in Figs. 3 and 4, respectively (the regression statistics are also recorded in Table 3). Each figure contains fitted distributions for MOS and LCL forecasts at 12–24-h and 36–48-h lead times. In fitting these distributions, a weighted linear regression model was used, to take account of the differences in sample sizes (indicated by the areas of the histogram bars in the corresponding parts of Figs. 1 and 2) among the points that define the empirical reliability curve. As in the case of the beta model, examination of regression model fits for Syracuse and Tucson permits consideration of VDSs for locations with very different climatological probabilities of precipitation occurrence.

In a qualitative sense, the regression models appear to provide reasonable fits to the empirical reliability curves defined by the points  $[q(X = 1|f), f]$ . Of course, this simple linear model subjects the irregular behavior of the reliability curves to considerable smoothing during the model fitting process. In an effort to identify a relatively simple model that might provide a better fit to the empirical reliability curves, two types of nonlinear regression models were considered. First, a logistic regression was applied, in which the observation variable  $X$  was subjected to a log-odds transformation. Second, a double-logistic regression was employed, in which both the forecast variable  $F$  and observation variable  $X$  were subjected to log-odds transformations. In the cases of these LFM-MOS and corresponding LCL PoP forecasts for Syracuse and Tucson, these alternative models did not yield appreciably better fits to the empirical reliability data.

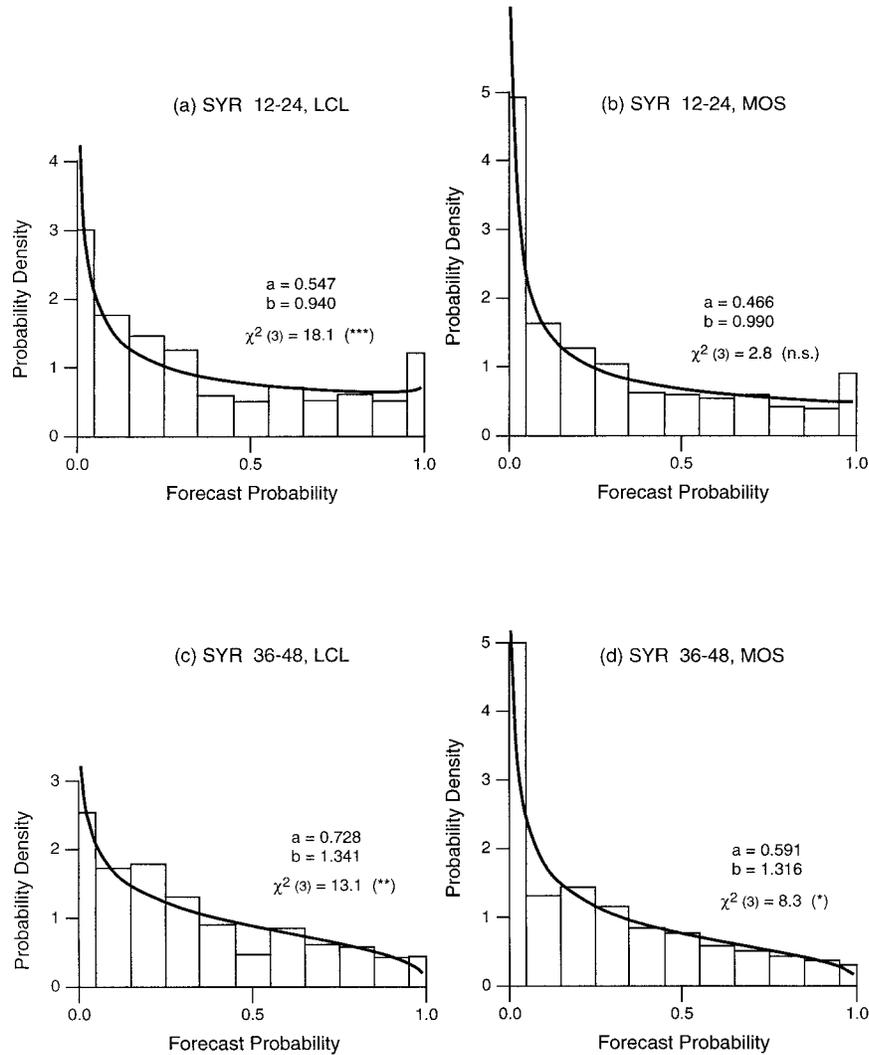


FIG. 1. Beta models fit to the marginal distribution of PoP forecasts,  $s(f)$ , for the cool season at Syracuse, NY: (a) 12–24-h LCL forecasts, (b) 12–24-h MOS forecasts, (c) 36–48-h LCL forecasts, and (d) 36–48-h MOS forecasts.

3) JOINT BETA AND REGRESSION MODEL FITS

As a further check on the model fits, we calculated the terms in the following decomposition (Murphy 1973) of the Brier score (Brier 1950):

$$BS = UNC + REL - RES, \quad (5)$$

where  $UNC [= \bar{x}(1 - \bar{x})]$  is the variance of the binary observations,  $REL$  is a measure of the reliability of the forecasts, and the  $RES$  is a measure of the resolution of the forecasts (see Murphy 1973, 1997). The Brier score itself, and the terms on the right-hand side (rhs) of (5), were computed from both the empirical data (data-based results) and the statistical models (model-based results). The expressions for the terms on the rhs of (5) for the sums involving the empirical data and the integrals involving the statistical models are reproduced

here to illustrate the differences between these two approaches to VPs. In the case of the empirical data.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (6)$$

$$REL = \frac{1}{n} \sum_{j=1}^{11} n_j (f_j - \bar{x}_j)^2, \quad (7)$$

and

$$RES = \frac{1}{n} \sum_{j=1}^{11} n_j (\bar{x}_j - \bar{x})^2, \quad (8)$$

in which  $\bar{x}_j$  is the relative frequency of measurable precipitation when  $F = f_j$  and  $n_j$  is the number of forecasts for which  $F = f_j$  ( $\sum_j n_j = n$ ;  $j = 1, \dots, 11$ ). In the case of the statistical models,

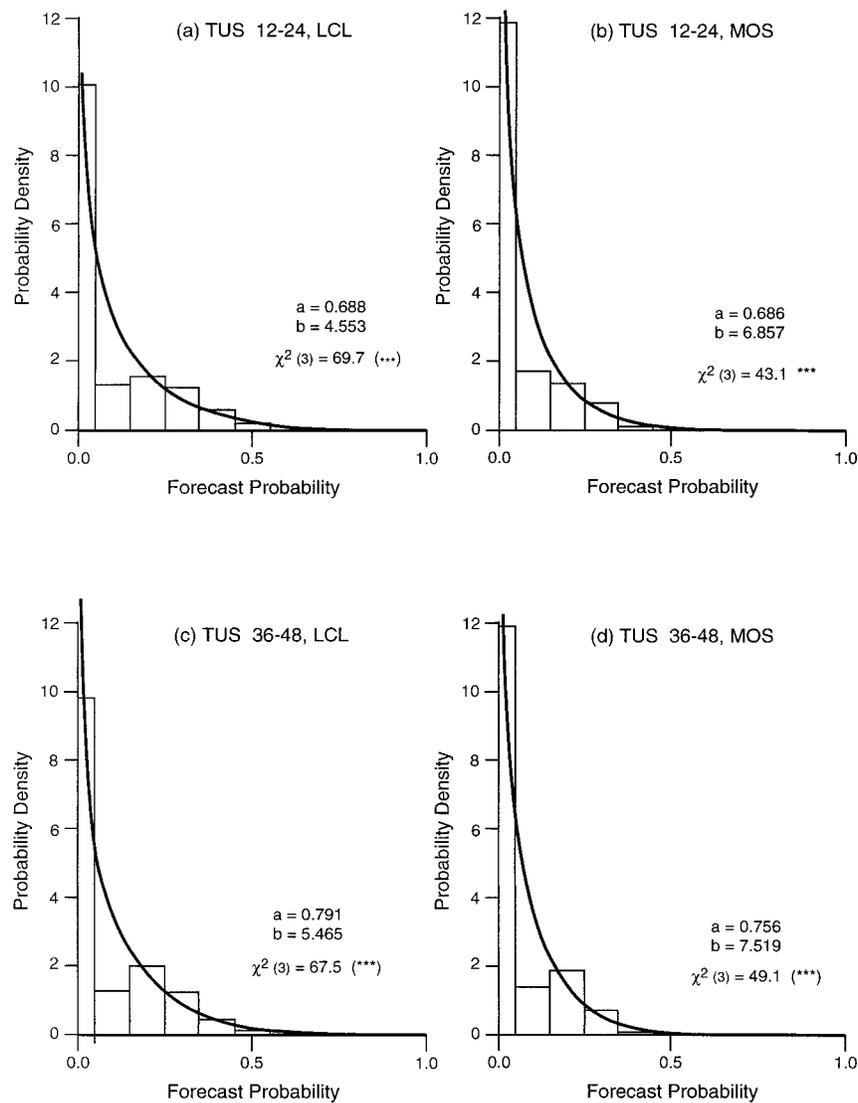


FIG. 2. Beta models fit to the marginal distribution of PoP forecasts,  $s(f)$ , for the warm season at Tucson, AZ: (a) 12–24-h LCL forecasts, (b) 12–24-h MOS forecasts, (c) 36–48-h LCL forecasts, and (d) 36–48-h MOS forecasts.

$$\bar{x} = \int_0^1 (b_0 + b_1 f) s(f) df, \quad (9)$$

$$REL = \int_0^1 [f - (b_0 + b_1 f)]^2 s(f) df, \quad (10)$$

and

$$RES = \int_0^1 [(b_0 + b_1 f) - \bar{x}]^2 s(f) df. \quad (11)$$

Since the expressions in (9), (10), and (11) involve both the regression and beta models, comparison of the data-based and model-based values of these terms and their sum represents a joint check on the fits of these models.

The values of  $E(X)$  ( $=\bar{x}$ ), UNC, REL, RES, and BS are included in Table 4 for the various combinations of location, forecast type, and lead time. Comparison of the model-based and data-based values of  $E(X)$  indicates that the models reproduce the mean value of the binary observations very closely: only in the case of the 12–24-h MOS forecasts at Syracuse is the difference between these means greater than 0.002. This result suggests that the fitting of these VDSs with the statistical models is consistent in an overall sense. The close correspondence between model-based and data-based values of  $E(X)$  also implies that the respective values of the UNC term are in very good agreement.

With regard to the REL and RES terms, the model-based results consistently underestimate the correspond-

TABLE 2. Parameter estimates, chi-square statistics, and means and variances of forecasts for beta models of marginal distributions  $s(f)$ : (a) Syracuse, NY, in the cool season; and (b) Tucson, AZ, in the warm season.

Type of forecast	Lead time (h)	Parameter estimates		Chi-squared statistic $\chi^2$	Mean $E(F)$	Variance $V(F)$
		$a$	$b$			
(a) Syracuse, NY—cool season						
MOS	12–24	0.466	0.990	2.8 <sup>a</sup>	0.320	0.089
LCL	12–24	0.547	0.940	18.1 <sup>d</sup>	0.368	0.093
MOS	36–48	0.591	1.316	8.3 <sup>b</sup>	0.310	0.074
LCL	36–48	0.728	1.341	13.1 <sup>c</sup>	0.352	0.074
(b) Tucson, AZ—warm season						
MOS	12–24	0.686	6.857	43.1 <sup>d</sup>	0.091	0.010
LCL	12–24	0.688	4.553	69.7 <sup>d</sup>	0.131	0.018
MOS	36–48	0.756	7.519	49.1 <sup>d</sup>	0.091	0.009
LCL	36–48	0.791	5.465	67.5 <sup>d</sup>	0.126	0.015

<sup>a</sup> Chi-square value not statistically significant.  
<sup>b</sup> Chi-square value statistically significant at 5% level.  
<sup>c</sup> Chi-square value statistically significant at 1% level.  
<sup>d</sup> Chi-square value statistically significant at 0.1% level.

ing terms for the data-based results. This underestimation is a consequence of the smoothing that occurs in the model fitting process and the squared-error nature of these terms. For example, in the case of RES, the relatively higher contributions to the integrals (for model-based results) when the empirical data points fall be-

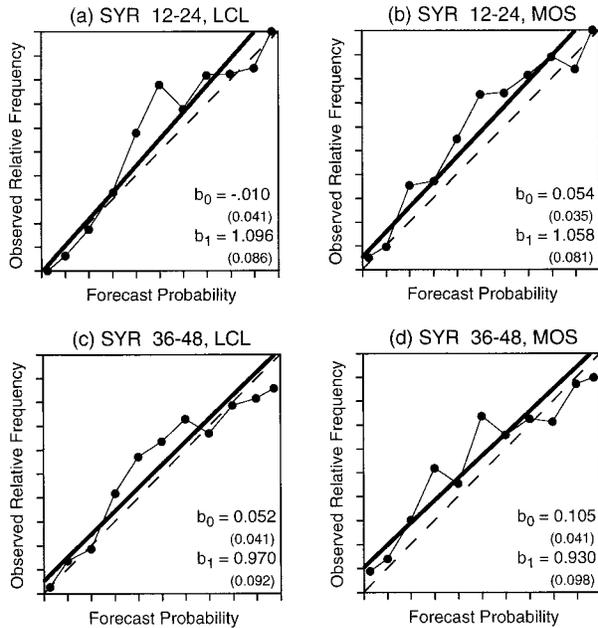


FIG. 3. Regression models fit to the conditional distributions of observations given PoP forecasts.  $q(x|f)$ , for the cool season at Syracuse, NY: (a) 12–24-h LCL forecasts, (b) 12–24-h MOS forecasts, (c) 36–48-h LCL forecasts, and (d) 36–48-h MOS forecasts.

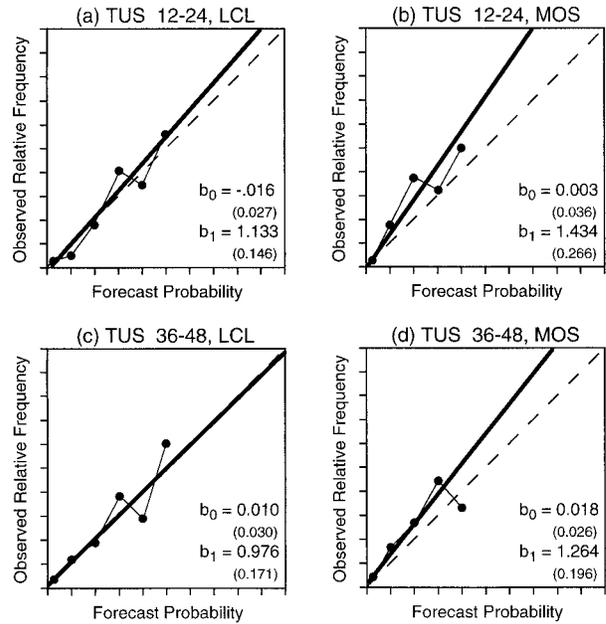


FIG. 4. Regression models fit to the conditional distributions of observations given PoP forecasts,  $q(x|f)$ , for the warm season at Tucson, AZ: (a) 12–24-h LCL forecasts, (b) 12–24-h MOS forecasts, (c) 36–48-h LCL forecasts, and (d) 36–48-h MOS forecasts.

tween the regression line and the horizontal lines  $E(X)$  are given less weight than the relatively lower contributions to the integrals when the regression line falls between these two points and the line  $E(X)$ . The effect of these VDSs when BS itself is calculated, although a slight tendency exists for the model-based estimates of BS to be less than the corresponding data-based estimates.

TABLE 3. Parameter estimates, standard errors, and measure  $R$  for the linear regression models of the conditional distributions  $q(x|f)$ : (a) Syracuse, NY, in the cool season, and (b) Tucson, AZ, in the warm season.

Type of forecast	Lead time (h)	Parameter estimates		Standard errors		Measure $R$
		$b_0$	$b_1$	$b_0$	$b_1$	
(a) Syracuse, NY—cool season						
MOS	12–24	0.054	1.058	0.035	0.081	0.112
LCL	12–24	-0.010	1.096	0.041	0.086	0.106
MOS	36–48	0.105	0.930	0.041	0.098	0.175
LCL	36–48	0.052	0.970	0.041	0.092	0.082
(b) Tucson, AZ—warm season						
MOS	12–24	0.003	1.434	0.036	0.266	0.437
LCL	12–24	-0.016	1.133	0.027	0.146	0.149
MOS	36–48	0.018	1.264	0.026	0.196	0.282
LCL	36–48	0.010	0.976	0.030	0.171	0.034

TABLE 4. The Brier score (BS), the terms in its decomposition, and  $E(X)$ , estimated on the basis of both empirical data and statistical models: (a) Syracuse, NY, in the cool season, and (b) Tucson, AZ, in the warm season.

Forecast type	Lead time (h)	Estimation method	$E(X)$	UNC	REL	RES	BS
(a) Syracuse, NY—cool season							
MOS	12–24	Data	0.393	0.239	0.011	0.104	0.145
		Model	0.388	0.237	0.005	0.091	0.151
LCL	12–24	Data	0.393	0.239	0.008	0.118	0.128
		Model	0.391	0.238	0.001	0.106	0.133
MOS	36–48	Data	0.393	0.239	0.014	0.070	0.182
		Model	0.391	0.238	0.007	0.062	0.183
LCL	36–48	Data	0.393	0.239	0.007	0.076	0.171
		Model	0.393	0.239	0.002	0.069	0.171
(b) Tucson, AZ—warm season							
MOS	12–24	Data	0.133	0.115	0.006	0.023	0.099
		Model	0.133	0.116	0.004	0.019	0.100
LCL	12–24	Data	0.133	0.115	0.003	0.026	0.092
		Model	0.134	0.116	0.000	0.023	0.094
MOS	36–48	Data	0.133	0.115	0.004	0.016	0.104
		Model	0.133	0.115	0.002	0.014	0.104
LCL	36–48	Data	0.133	0.115	0.002	0.017	0.101
		Model	0.133	0.116	0.000	0.014	0.101

4. Assessment of forecasting performance in terms of model parameters

In this section various samples of PoP forecasts are evaluated and compared using the parameters of the regression and beta models as measures of aspects of forecasting performance. Since each model possesses two parameters, the dimensionality of the underlying verification problem has been reduced from the 21 parameters of the empirical conditional and marginal probabilities to the four parameters of the regression and beta models of these distributions. Model-based estimates of reliability and sharpness are also compared with corresponding data-based estimates.

a. Beta model parameters: Measures of sharpness

In the case of the beta model, the beta distribution in (4) is fit to the marginal distribution of forecasts,  $s(f)$ . In general, probabilistic forecasts are perfectly sharp if probability values of zero or one only are used in the forecasts. Relatively sharp forecasts exhibit a bimodal, U-shaped distribution, whereas (unbiased) forecasts lacking in sharpness exhibit a unimodal distribution centered near the climatological probability of the event of interest. Since the shape of the beta distribution is determined by the values of the parameters  $a$  and  $b$  (see section 3a), these parameters individually and jointly provide insight into the sharpness of the forecasts.

To assess the overall sharpness of the modeled forecasts in terms of a single number, a measure of this attribute of the forecasts is defined in terms of the parameters  $a$  and  $b$ . In view of the relationship between

the shape of the beta distribution and the values of its parameters, the variance of the distribution  $V(F)$ —a statistic that depends on both  $a$  and  $b$ —appears to be a reasonable choice for such a measure. A U-shaped distribution possesses relatively large variance, whereas a unimodal distribution possesses relatively small variance. In interpreting the results of this assessment of forecasting performance in terms of the parameters of beta models, it should be kept in mind that the values of both  $a$  and  $b$  are required for a complete description of the sharpness of the modeled forecasts. The variance of the beta distribution represents a one-dimensional or scalar quantity, in which the aspects of sharpness characterized by  $a$  and  $b$  are combined (and confounded) in a particular manner.

In the case of the modeled forecasts for Syracuse, the values of  $a$  and  $b$  are both less than one for the 12–24-h forecasts, indicating at least a slight tendency toward bimodality. Comparison of the parameters for the MOS and LCL forecasts indicates that the former are definitely sharper at the lower end of the probability scale (i.e., near zero) and the latter are somewhat sharper at the upper end of the probability scale (i.e., near one). The parameter values reveal a substantial decrease in sharpness from 12–24 h to 36–48 h, with both types of modeled forecasts possessing unimodal distributions at the longer lead time. At 36–48 h, the MOS forecasts are somewhat sharper than the LCL forecasts near zero, with the two types of modeled forecasts exhibiting comparable sharpness near one.

The modeled forecasts for Tucson exhibit somewhat different characteristics than the modeled forecasts for Syracuse. For all four combinations of forecast type and lead time at Tucson, the values of  $a$  and  $b$  are such that the distribution  $s(f)$  is strongly unimodal with the mode at zero. At the lower end of the probability scale, little difference in sharpness exists between the MOS and LCL forecasts for Tucson, but the values of  $b$  indicate that the LCL forecasts are sharper than the MOS forecasts at the upper end of the scale (for Tucson, in the vicinity of  $F = 0.50$ ). As in the case of Syracuse, comparison of parameter values for the 12–24- and 36–48-h lead times reveals that the modeled forecasts for Tucson are sharper at the shorter lead time.

In terms of the values of the parameters  $a$  and  $b$ , the modeled forecasts for Syracuse are sharper than the modeled forecasts for Tucson. This difference in sharpness is especially noticeable at the upper end of the probability scale, and is a consequence of the large differences between the climatological probabilities of measurable precipitation at the two locations. As indicated by the overall measure of sharpness  $V(F)$  (see Table 2), both types of forecasts are considerably sharper at Syracuse than at Tucson (the low climatological probability of measurable precipitation largely precludes the realization of sharp distributions at Tucson). The LCL forecasts are somewhat sharper than the MOS forecasts for three of the four location–lead time com-

binations (the exception is the 36–48-h lead time at Syracuse). According to  $V(F)$ , sharpness decreases as lead time increases, although the decrease in sharpness from 12–24 h to 36–48 h is very modest at Tucson, possibly as a consequence of relatively few precipitation events in this dataset at Tucson.

Of course, none of the above conclusions are unexpected, but they serve to illustrate that this statistical model is capable of reflecting various important aspects of the empirical distributions  $s(f)$  that would not be evident from examination of a single verification measure.

#### *b. Regression model parameters: Measures of reliability*

The regression model in (3) fits a line to the empirical reliability data. This linear model describes the relationship between the conditional mean observation given a forecast,  $E(X|F = f)$ , and the forecast  $f$ . In general, forecasts are perfectly reliable when the values of  $E(X|F = f)$  and  $f$  are equal over all values of  $f$ . In terms of the model parameters, perfect reliability is achieved only when  $b_0 = 0$  and  $b_1 = 1$ . On the other hand, reliability is less than perfect when  $b_0 \neq 0$  or  $b_1 \neq 1$ , or both. In qualitative terms, the reliability of the modeled forecasts is determined by the degree to which the values of  $b_0$  and  $b_1$  jointly approach (or depart from) the ideal values of zero and one, respectively. As a scalar measure of the overall reliability of the modeled forecasts, we introduce the measure  $R$ , where

$$R = |b_0| + |b_1 - 1|. \quad (12)$$

The modeled forecasts are completely reliable when  $R = 0$ , and reliability decreases (according to this measure in which the effects of  $b_0$  and  $b_1$  are combined and confounded) as  $R$  increases. Note that the measure  $R$  is but one of many such measures that could be devised. It is introduced not because it is necessarily best, or indeed even better than REL in (10), but rather to underline the fact that different one-dimensional measures of the same aspect of quality can and will yield different results.

The values of the parameters  $b_0$  and  $b_1$  for the PoP forecasts of interest here are presented in Table 3. This table also contains standard errors of these parameter estimates. In this discussion of the reliability of the modeled forecasts, the standard errors serve simply as rough points of reference against which to compare differences between the estimated values  $b_0$  and  $b_1$  and their respective ideal (i.e., completely reliable) values of zero and one. In this regard, with only one exception the estimated values of both parameters lie within two standard errors of the ideal values (the exception is the value of  $b_0$  for 36–48-h MOS forecasts for Syracuse).

In the case of Syracuse, comparison of the parameter estimates  $b_0$  and  $b_1$  with the ideal values of zero and one suggests that the LCL forecasts are more reliable

than the MOS forecasts for the 36–48-h lead time. However, the analogous comparison for the 12–24-h lead time yields a mixed result (the LCL value of  $b_0$  is closer to zero but the MOS value of  $b_1$  is closer to one). In the case of Tucson, comparison of  $b_0$  and  $b_1$  with the ideal values indicates that the LCL forecasts are more reliable than the MOS forecasts for both lead times. Comparison of the parameter values for Syracuse and Tucson reveals that the  $b_0$  values are generally closer to zero at Tucson and that the  $b_1$  values are generally closer to one at Syracuse, a mixed result.

With regard to overall reliability, as determined by the scalar measure  $R$  (Table 3), the LCL forecasts are more reliable than the MOS forecasts at both lead times at both locations. These differences in reliability are particularly large in the case of the PoP forecasts for Tucson. It is also interesting to note that, according to the measure  $R$ , reliability generally improves as lead time increases (the MOS forecasts at Syracuse are an exception). Except in the case of the LCL forecasts for the 36–48-h lead time, the measure  $R$  indicates that PoP forecasts for Syracuse are more reliable than PoP forecasts for Tucson.

The quantity REL (Table 4) can also serve as an overall scalar measure of reliability. In terms of the model-based estimates of REL, PoP forecast reliability is better at Tucson than at Syracuse, better for the LCL forecasts than for the MOS forecasts, and generally better for the 12–24-h (36–48-h) h forecasts than the 36–48-h (12–24-h) forecasts at Syracuse (Tucson). The differences between these results and the results based on the measure  $R$  emphasize the ambiguities that can arise when comparative verification is based on scalar or one-dimensional measures.

As before, these results are not surprising from a forecasting standpoint, especially considering that the forecasters producing the LCL forecasts generally had the corresponding MOS forecasts available to them as guidance. Rather, the point here is that manipulation and examination of the parameters of the two statistical models can bring out important aspects of forecast quality that may be hidden by particular one-dimensional measures.

#### *c. Regression and beta model parameters: Joint depictions of reliability and sharpness*

Separate evaluations of the sharpness and reliability of the PoP forecasts in terms of the beta model parameters  $a$  and  $b$  and the regression model parameters  $b_0$  and  $b_1$  served as the foci of sections 4a and 4b, respectively. In this section, we turn our attention to the joint characterization of the sharpness and reliability of these forecasts in terms of both the basic parameters themselves and simple functions or transformations of these parameters. Although the model-based approach introduced here has reduced the dimensionality of the PoP verification problem from  $d = 21$  to  $d = 4$ , the

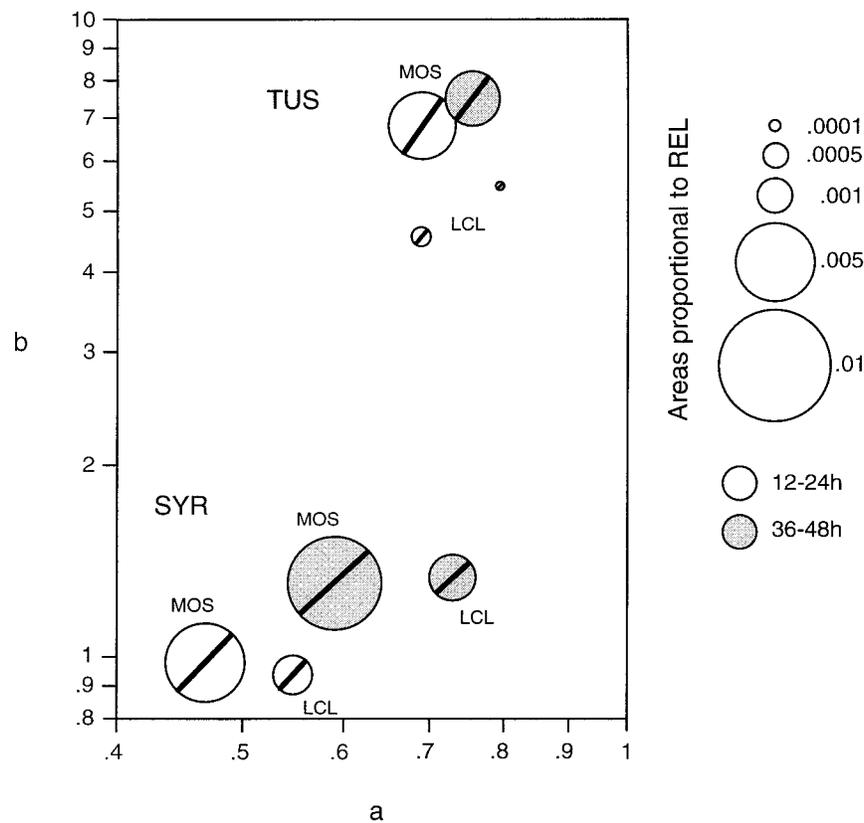


FIG. 5. Depiction of PoP forecast quality in terms of the parameters  $a$ ,  $b$ ,  $b_1$ , and REL.

best way to represent or display even this four-dimensional set of information is not obvious. Several solutions to this representation problem are presented here. We do not claim that any of these alternatives is necessarily superior to the many other possible displays that could have been constructed. Rather, they are offered as examples of the ways in which information related to these four parameters might be depicted. Hopefully, consideration of these and other displays will identify representations that constitute improvements, in terms of both intuition and insight, over four-variate tabulations.

Figure 5 depicts forecasting performance of the modeled PoP forecasts at Syracuse and Tucson in terms of four parameters. These parameters are  $a$  and  $b$  (the parameters of the beta model),  $b_1$  (a parameter of the regression model), and REL [Eq. (10), a one-dimensional measure of reliability and a function of all four basic parameters]. The parameters  $a$  and  $b$  serve as the axes of the two-dimensional diagrams, with logarithmic scales used for both quantities. The measure REL is depicted by a circle of area proportional to its numerical value, and the parameter  $b_1$  is depicted within this circle by a line (or chord) with slope depicting its numerical value.

Roughly speaking, “good” forecasts would be represented in Fig. 5 by small circles (REL small) con-

taining chords with approximately 45° slopes ( $b_1 = 1$ ) located in the lower left-hand portion of the diagram ( $a < 1$  and  $b < 1$ ). Goodness decreases as circle size increases, chord slope departs from 45°, and chord location moves upward and/or to the right. The differences between PoP forecasting performance at Syracuse and Tucson are immediately evident in Fig. 5. In terms of aspects of sharpness (as indicated by the magnitudes of  $a$  and  $b$ ), the forecasts at Syracuse exhibit a substantially greater tendency toward bimodality than the forecasts at Tucson. As noted previously, this result is presumably due largely to the difference between the climatological probabilities of precipitation at the two locations. [It should also be noted that comparisons between forecasting performance at Syracuse and Tucson are compromised by the fact such comparisons are necessarily based on unmatched comparative verification; see Murphy (1991).] Careful study of the chords reveals that their slopes are closer to the ideal 45° (i.e.,  $b_1$  closer to one) at Syracuse than Tucson. On the other hand, the sizes of the circles indicate that the overall reliability of the PoP forecasts as measured by REL is better at Tucson than at Syracuse. This “disagreement” between results based on  $b_1$  and results based on REL raises the issue of seemingly conflicting results. However, no real conflict exists here, since one-dimensional measures of reliability need not yield similar results: since forecast

quality is multidimensional, superiority with respect to one aspect of quality is no guarantee of superiority with respect to other aspects of quality.

In terms of the results within locations, Fig. 5 reveals that the 12–24-h forecasts are clearly sharper (circles closer to lower-left corner) than the 36–48-h forecasts at both Syracuse and Tucson. On the other hand, the indicators of reliability (i.e., circle size and chord slope) yield mixed results, with (for example) REL indicating better reliability at the short (longer) lead time at Syracuse (Tucson). Comparison of the two types of forecasts suggests that the MOS forecasts are somewhat sharper than the LCL forecasts at Syracuse (reflected mostly in the respective values of  $a$ , indicating more frequent use of very small probabilities), whereas this ordinal relationship is reversed at Tucson (reflected mostly in the respective values of  $b$ , indicative of differences in the frequency of use of relatively high probabilities). In the case of reliability, relatively large differences in this aspect of quality (as reflected in circle size and chord slope) favor the LCL forecasts over the MOS forecasts at Tucson. At Syracuse, these differences are smaller; nevertheless, they suggest that the LCL forecasts are somewhat more reliable than MOS forecasts (this difference is especially noticeable in terms of circle size). Perhaps it should be noted here that one signature in Fig. 5 for overall bias in a VDS is a relatively large value of REL when  $b_1$  is approximately equal to one.

An alternative depiction of performance in terms of three parameters is shown in Fig. 6, in which the quantities  $b_0$ ,  $b_1$ , and  $[V(F)/V(F^*)]^{1/2}$  are used to characterize the different aspects of quality. The quantity  $V(F^*)$  is the variance of perfect forecasts, which is an upper bound on the variance of all (unbiased) imperfect PoP forecasts. In this figure, good forecasts are represented by large circles near the center of the diagram, with goodness decreasing as circle size decreases or distance from the center increases. Differences in reliability (as reflected in the values of  $b_0$  and  $b_1$ ) are relatively transparent in this figure. For example, the differences in reliability between the two types of forecasts (i.e., LCL versus MOS) are associated largely with better values of  $b_0$  in the case of Syracuse (Fig. 6b) and better values of  $b_1$  in the case of Tucson (Fig. 6a). Also, differences in reliability as a function of lead time appear to be larger at Tucson and smaller at Syracuse. With regard to sharpness, it is clear that the variability or sharpness of the forecasts at Syracuse is a much greater fraction of the variability of perfect forecasts than the corresponding variability or sharpness of the forecasts at Tucson. The relative size of the circles also indicates that the LCL forecasts are sharper than the MOS forecasts at Tucson and that differences in sharpness as a function of forecast type at Syracuse are much more modest. Finally, it is of some interest to note that circle size (i.e., the one-dimensional measure of sharpness employed here) suggests that although sharpness decreases as lead

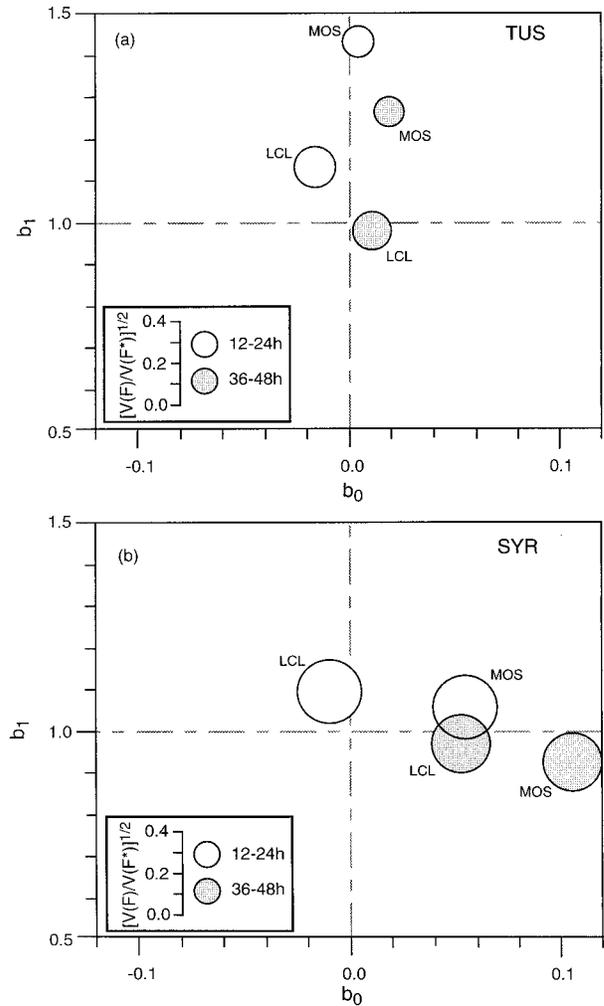


FIG. 6. Depiction of PoP forecast quality in terms of the parameters  $b_0$ ,  $b_1$ , and  $[V(F)/V(F^*)]^{1/2}$ : (a) Tucson, AZ, for the warm season, and (b) Syracuse, NY, for the cool season.

time increases for both forecast types at both locations, these differences in sharpness are relatively small.

A third depiction of the performance of these PoP forecasts is shown in Fig. 7, in which the one-dimensional quantities REL and  $V(F)/V(F^*)$  are used to characterize overall reliability and sharpness, respectively. In this figure, good forecasts are represented by circles or squares located in the upper-left portion of the diagram [small REL and large  $V(F)/V(F^*)$ ], with goodness decreasing as these circles–squares move to the right or downward. The differences between PoP forecasting performance at Syracuse and Tucson are emphasized in this figure (but recall that comparison across locations involve unmatched VDSs). Forecasts for Tucson are more reliable (smaller REL), whereas forecasts for Syracuse are sharper [larger  $V(F)/V(F^*)$ ]. In this depiction, the LCL forecasts are more reliable than the MOS forecasts at both locations. Differences in sharpness as a function of forecast type are smaller, with a clear ad-

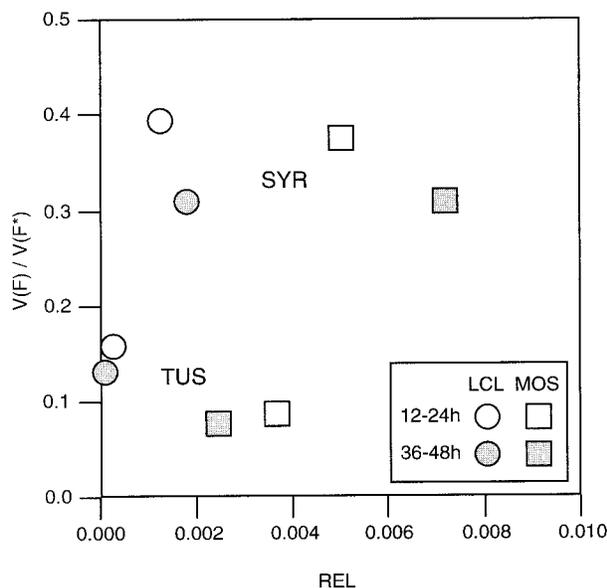


FIG. 7. Depiction of PoP forecast quality in terms of the parameters REL and  $V(F)/V(F^*)$ .

vantage for the LCL forecasts over the MOS forecasts at Tucson. With regard to lead time, the 12–24-h forecasts are more reliable and sharper than the 36–48-h forecasts at Syracuse. On the other hand, only a relatively small advantage in sharpness can be detected for the shorter-range forecasts over the longer-range forecasts at Tucson, with the measure REL indicating that the 36–48-h forecasts are actually more reliable than 12–24-h forecasts at this location.

While these (and other) results generally support the familiar notion that shorter-range forecasts perform better than longer-range forecasts, lead time differences in performance for these PoP forecasts clearly vary across aspects of quality, as well as across different measures of the same aspect of quality. Such results underline the need to diagnose and assess forecasting performance using an approach that respects the multidimensional nature of forecast quality.

**5. Sampling variability, statistical models, and unambiguous superiority**

In this section we investigate the use of parametric statistical models as means of reducing the effects of sampling variability, in the context of assessments of the unambiguous superiority of one set of forecasts over another set of forecasts. We note that this problem has been addressed in a somewhat different context by Krzysztofowicz and Long (1991a). The forecasts compared here are the PoP forecasts introduced in section 3. Unambiguous superiority, as defined in this paper, implies that all rational decision makers whose activities are sensitive to the forecasts would realize greater economic benefit from the superior forecasts than from the

inferior forecasts (assuming that the decision makers must choose between the two sources of information). When the conditions for unambiguous superiority are met, the superior forecasts are said to be *sufficient* for the inferior forecasts (Ehrendorfer and Murphy 1988, 1992; Clemen et al. 1995). It is important to note that the sufficiency relationship may be indeterminate for a particular pair of forecast sets, so that in many cases it is not possible to declare that one forecaster or forecasting system is unambiguously superior to another. (Otherwise, comparative forecast verification would be greatly simplified.) In such cases some decision makers would derive greater economic benefit from one forecast source, while other decision makers should prefer the alternative source.

Sufficiency imposes conditions on the relationship between the respective joint distributions of forecasts and observations. In general, one-dimensional measures of quality (or its aspects) are inadequate to determine sufficiency. Failure to respect the dimensionality of VPs can lead to what are known as quality/value reversals, in which forecasts that are judged to be inferior (according to one or more one-dimensional measures) may be of greater economic value to some decision makers (e.g., Brooks and Douglas 1998; Murphy and Ehrendorfer 1987; Murphy 1997).

One way to investigate unambiguous superiority of forecasting system A over forecasting system B is to determine if the condition

$$\int_0^u S_A(t) dt - \int_0^u S_B(t) dt \geq 0$$

for all  $u$      $(0 \leq u \leq 1)$     (13)

is met, where

$$S(t) = \int_0^{f^*} s(v) dv$$

(14)

is the cumulative distribution function of the calibrated forecasts  $f^* = E(X = 1 | f)$  (DeGroot and Eriksson 1985; Clemen et al. 1995).

To investigate the efficacy of using statistical models to reduce the effects of sampling variability in determinations of unambiguous superiority, the distributions in (13) are evaluated using (a) empirical relative frequencies and (b) model-based (or “smooth”) probabilities. In the case of evaluation based on empirical relative frequencies,  $s(f)$  is estimated using the appropriate histogram in Figs. 1 or 2 and  $q(s | f) = E(X | f)$  is estimated using the corresponding empirical reliability data in Figs. 3 or 4. In the case of evaluation based on smooth probabilities,  $s(f)$  is represented by the beta model in Figs. 1 or 2 and  $q(x | f) = E(X | f)$  is represented by the corresponding regression model in Figs. 3 or 4.

Figure 8 contains an example of the application of these two approaches to the assessment of unambiguous superiority. In this diagram, the sufficiency of the 12–

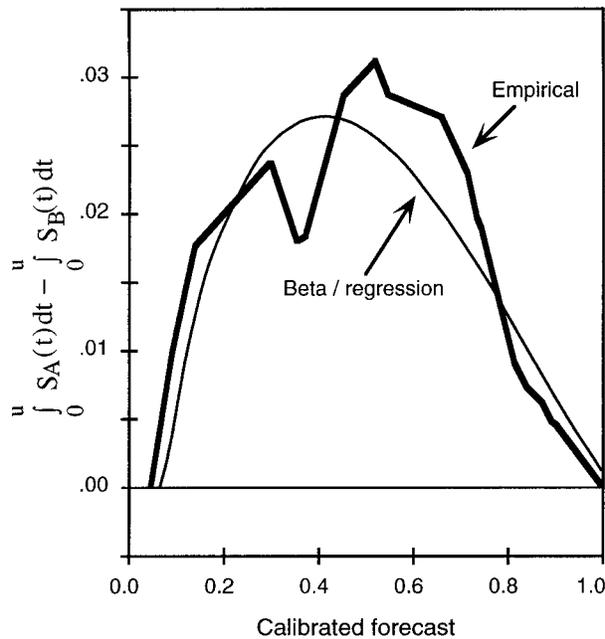


FIG. 8. Sufficiency of 12–24-h PoP forecasts for 36–48-h PoP forecasts for the cool season at Syracuse, NY, as indicated by the difference of the two integrals in Eq. (13).

24-h MOS forecasts for the 36–48-h MOS forecasts at Syracuse is determined using both methods described in the previous paragraph. The heavy irregular curve represents Eq. (13) as a function of the integration variable  $u$  for the empirical verification data. The numerical values of this difference are everywhere nonnegative, indicating that the 12–24-h forecasts are sufficient for the 36–48-h forecasts. The irregularity of this curve reflects the sampling variability in the empirical data. The light, smooth curve represents Eq. (13) evaluated using the beta/regression models of the calibration-refinement components of the joint distribution  $p(f, x)$ . This curve also declares that the 12–24-h forecasts are sufficient for the 36–48-h forecasts. In addition, this model-based curve apparently smooths out the sampling variations evident in the data-based results.

The ability of parametric models of the joint distribution  $p(f, x)$  to reduce the effects of sampling variability should be most evident for smaller sample sizes. In Fig. 9 we present some results regarding sufficiency for repeated subsamples drawn from the verification data underlying Fig. 8. Specifically, the original VDS consisting of  $n = 646$  forecast–observation pairs has been sampled (without replacement) 1000 times each, for sample sizes ranging from 50 to 600, in steps of 50. For each resample, sufficiency was evaluated using Eq. (13) on the basis of both the empirical data and the beta/regression models fit to the data. The proportion of samples in which a correct determination of sufficiency was made by each method is plotted against sample size in Fig. 9. Both methods correctly declare the sufficiency

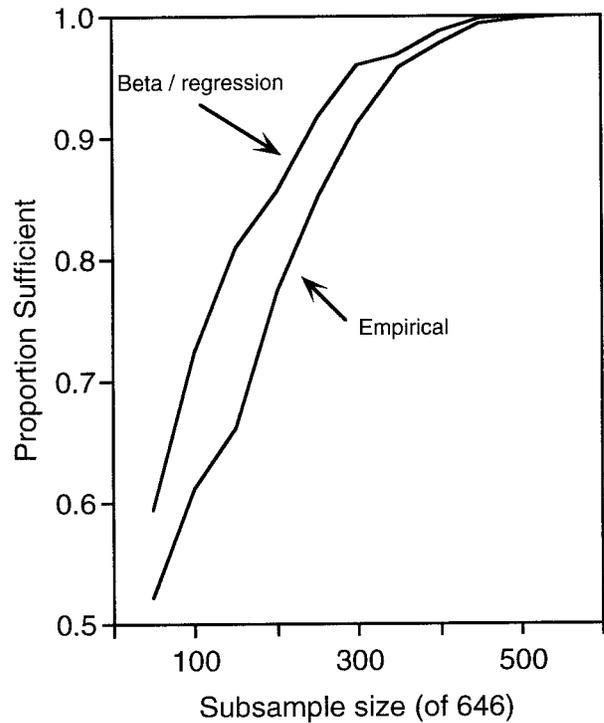


FIG. 9. Proportion of 1000 samples (drawn without replacement) for which the 12–24-h PoP forecasts are declared sufficient for the 36–48-h PoP forecasts for the cool season at Syracuse, NY, as a function of sample size.

of the 12–24-h forecasts for the 36–48-h forecasts with high probability for the larger sample sizes. In the case of small sample sizes, however, the sufficiency declarations are more often correct when based on the model-based calculations [from Eq. (13)]. Since the differences between the two curves in Fig. 9 are fairly modest, the results of this exploratory study should be viewed with some caution. Nevertheless, parametric modeling of the joint distributions of VDSs appears to offer some promise as a means of reducing the effects of sampling variability, and possibly enhancing the likelihood that sufficiency can be detected in the context of comparative verification.

## 6. Discussion and conclusions

A conceptually and methodologically sound approach to forecast verification must respect the dimensionality of VPs. When such an approach is based on empirical verification data, it frequently suffers from the “curse of dimensionality”; that is, it requires the calculation of a relatively large number of parameters to describe forecasting performance in a complete manner. This paper has presented a two-stage approach to VPs designed to address dimensionality-related considerations. In this approach, parametric statistical models are fit to samples of matched pairs of forecasts and observations, and then both absolute and relative forecasting performance are

assessed in terms of the parameters of these models. The approach is illustrated here by (a) modeling samples of PoP forecasts and the corresponding binary observations and (b) evaluating and comparing forecast quality across different types of PoP forecasts, different locations, and different lead times.

The joint distributions of PoP forecasts and binary observations considered here were modeled by fitting linear regression equations to the conditional distributions of observations given forecasts, and beta distributions to the marginal distributions of forecasts. By fitting statistical models to these data, the dimensionality of the underlying verification problem was reduced substantially. Twenty-one-dimensional data-based problems were reduced to four-dimensional model-based problems. Although such a reduction in dimensionality may conceal some of the information that is contained in the original high-dimensional joint distribution, inspection of Figs. 1–4 suggests that the model fits are qualitatively reasonable and that most if not all prominent features in the empirical relative frequencies have been captured, at least for these VDSs.

It then proved to be possible to interpret the four parameters determining the regression and beta models (two parameters for each model) in terms of basic aspects of the quality of the PoP forecasts. Specifically, the two regression parameters related to aspects of reliability and the two beta parameters related to aspects of sharpness. Various samples of PoP forecasts were then evaluated and compared using the model parameters as measures of these aspects of quality. In a companion study, the efficacy of using the model-based approach as a means of reducing the effects of sampling variability in comparative evaluation of PoP forecasts was also investigated.

In summary, the two-stage, model-based approach to forecast verification appears to provide a potentially attractive alternative to the traditional data-based approach, at least in the context of the VP considered here. The model-based approach achieved a substantial reduction in the dimensionality of the underlying VP, apparently without appreciable loss of information in the underlying VDS concerning aspects of forecast quality. Moreover, it proved to be possible to describe basic aspects of forecasting performance in terms of model parameters and to use these parameters to evaluate and compare samples of PoP forecasts. This result supports the notion that model parameters can serve as verification measures. Taken together, these two results suggest that the model-based approach can substantially increase the degree of completeness and comprehensibility of the verification process in high-dimensional VPs. In addition, it appears that the model-based approach is also effective in reducing the impact of sampling variability on comparative evaluations.

In evaluating and comparing the data-based and model-based approaches to forecast verification, several considerations should be kept in mind. First, in the data-

based approach, the basic inputs to the verification process are the empirical relative frequencies of forecasts and/or observations (i.e., the raw data that constitutes the underlying VDS). In the model-based approach, on the other hand, the basic inputs are relative frequencies of forecasts and/or observations derived from one or more models of the VDS. In effect, it is modeled forecasts rather than raw forecasts that are verified when the model-based approach is applied to VPs. A potential pitfall of this approach could be that interesting or important details of the empirical joint distribution might be smoothed over by the statistical modeling process.

Second, the difference between qualitative assessments of forecasting performance and fully quantitative assessments of forecasting performance should be kept in mind. In the data-based approach to verification problems involving PoP forecasts, quantitative assessments are usually limited to the calculation of at most two or three measures of aspects of quality. Clearly, this approach fails to respect the dimensionality of all but the simplest verification problems. The data-based approach may also include the visual inspection of reliability and/or sharpness diagrams, which taken together depict the PoP verification problem considered here in its full dimensionality. However, this inspection constitutes only a qualitative assessment of forecasting performance in terms of these aspects of quality. The model-based approach, on the other hand, provides a parsimonious framework in which forecasting performance can be evaluated quantitatively in its full (modeled) dimensionality.

In this paper attention has been focused on the use of model parameters as measures of aspects of forecast quality. A related issue involves the relationship between common measures of forecasting performance and model parameters in the context of model-based verification. The Brier score, and the terms in its decomposition [see Eq. (5)], provide some insight into the nature of such relationships. When the terms on the rhs of Eq. (5) are expressed in the form of the model-based integrals in Eqs. (9)–(11), direct relationships can be seen to exist between the respective terms, thus between the Brier score itself and the parameters of the regression and beta models. Since modeled forecast quality is four-dimensional in this context, the fact that the relationships *in general* between these parameters and one-dimensional measures of aspects of quality such as BS, REL, and RES are complex should come as no surprise. In some cases, when model parameters assume particular values, simpler relationships may arise. For example, when  $b_0 = 0$  and  $b_1 = 1$  (perfectly reliable modeled forecasts), it follows that  $E(X) = E(F)$ ,  $REL = 0$ ,  $RES = V(F)$ , and  $BS = E(F) - E(F^2)$ . In this case, RES equals the variance of the forecasts and the Brier score is the difference between the first and second moments of the distribution of forecasts  $s(f)$ .

With regard to future work in this area, methodological studies ranging from efforts to improve models in

the context of PoP verification problems to applications of the model-based approach to other verification problems could be undertaken. Also, PoP verification problems could be approached by modeling the conditional distribution  $r(f|x)$  and the marginal distribution  $t(x)$  [instead of  $q(x|f)$  and  $s(f)$ ]. As an example of the general applicability of this technique to other weather elements, forecasts and observations of continuous variables such as temperature could be modeled using bivariate Gaussian distributions. Applications of the model-based approach to VPs involving bivariate Gaussian models would reduce these relatively high-dimensional problems to five-dimensional problems (two means, two variances, and a covariance or correlation coefficient). Further studies related to the utilization and interpretation of model parameters as measures of aspects of forecast quality, as well as the development of insightful displays of the information embodied in these parameters, should be undertaken.

To be truly useful, the practice of forecast verification must be based on a conceptually and methodologically sound approach and it should facilitate the identification of basic strengths and weaknesses in forecasting performance. After all, assessment of these strengths and weaknesses represent key steps in the process of improving forecast quality.

## REFERENCES

- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.
- Brooks, H. E., and C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288–303.
- , and A. P. Douglas, 1998: Value of weather forecasts for electric utility load forecasting. Preprints, *16th Conf. on Weather Analysis and Forecasting*, Phoenix, AZ, Amer. Meteor. Soc., 361–364.
- Carter, G. M., and P. D. Polger, 1986: A 20-year summary of National Weather Service verification results for temperature and precipitation. NWS Tech. Memo. NWS FCST 31, NOAA/NWS, 50 pp. [Available from NTIS, 5285 Port Royal Road, Springfield, VA 22161.]
- , J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, **4**, 401–412.
- Clemen, R. T., and R. L. Winkler, 1987: Calibrating and combining precipitation probability forecasts. *Probability and Bayesian Statistics*, R. Viertl, Ed., Plenum, 97–110.
- , A. H. Murphy, and R. L. Winkler, 1995: Screening probability forecasts: Contrasts between choosing and combining. *Int. J. Forecasting*, **11**, 133–146.
- DeGroot, M. H., and E. A. Eriksson, 1985: Probability forecasting, stochastic dominance, and the Lorenz curve. *Bayesian Statistics*, J. Bernardo et al., Eds., North Holland, 99–118.
- Ehrendorfer, M., and A. H. Murphy, 1988: Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. *Mon. Wea. Rev.*, **116**, 1757–1770.
- , and —, 1992: Evaluation of prototypical climate forecasts: The sufficiency relation. *J. Climate*, **5**, 876–887.
- Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.
- Katz, R. W., A. H. Murphy, and R. L. Winkler, 1982: Assessing the value of frost forecasts to orchardists: A dynamic decision-making approach. *J. Appl. Meteor.*, **21**, 518–531.
- Krzysztofowicz, R., 1992: Bayesian correlation score: A utilitarian measure of forecast skill. *Mon. Wea. Rev.*, **120**, 208–219.
- , and D. Long, 1991a: Forecast sufficiency characteristic: Construction and application. *Int. J. Forecasting*, **7**, 39–45.
- , and —, 1991b: Beta likelihood models of probabilistic forecasts. *Int. J. Forecasting*, **7**, 47–55.
- Liljas, E., and A. H. Murphy, 1994: Anders Angstrom and his early papers on probability forecasting and the use/value of forecasts. *Bull. Amer. Meteor. Soc.*, **75**, 1227–1236.
- Mason, I. B., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.
- Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.
- , 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601.
- , 1997: Forecast verification. *Economic Value of Weather and Climate Forecasts*, R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 19–74.
- , and M. Ehrendorfer, 1987: On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Wea. Forecasting*, **2**, 243–251.
- , and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.
- Wilks, D. S., 1991: Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models. *Mon. Wea. Rev.*, **119**, 1640–1662.
- , 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 464 pp.
- , and A. H. Murphy, 1986: A decision-analytic study of the joint value of seasonal precipitation and temperature forecasts in a choice-of-crop problem. *Atmos.–Ocean*, **24**, 353–368.
- , and K. W. Shen, 1991: Threshold relative humidity duration forecasts for plant disease prediction. *J. Appl. Meteor.*, **30**, 463–477.
- , R. E. Pitt, and G. W. Fick, 1993: Modeling optimal alfalfa harvest scheduling using short-range weather forecasts. *Agric. Syst.*, **42**, 277–305.