Distributions-Oriented Verification of Probability Forecasts for Small Data Samples

A. ALLEN BRADLEY AND TEMPEI HASHINO

IIHR—Hydroscience and Engineering, and Department of Civil and Environmental Engineering, University of Iowa, Iowa City, Iowa

STUART S. SCHWARTZ

University of North Carolina Water Resources Research Institute, North Carolina State University, Raleigh, North Carolina

(Manuscript received 10 September 2002, in final form 19 February 2003)

ABSTRACT

The distributions-oriented approach to forecast verification uses an estimate of the joint distribution of forecasts and observations to evaluate forecast quality. However, small verification data samples can produce unreliable estimates of forecast quality due to sampling variability and biases. In this paper, new techniques for verification of probability forecasts of dichotomous events are presented. For forecasts of this type, simplified expressions for forecast quality measures can be derived from the joint distribution. Although traditional approaches assume that forecasts are discrete variables, the simplified expressions apply to either discrete or continuous forecasts. With the derived expressions, most of the forecast quality measures can be estimated analytically using sample moments of forecasts and observations from the verification data sample. Other measures require a statistical modeling approach for estimation. Results from Monte Carlo experiments for two forecasting examples show that the statistical modeling approach can significantly improve estimates of these measures in many situations. The improvement is achieved mostly by reducing the bias of forecast quality estimates and, for very small sample sizes, by slightly reducing the sampling variability. The statistical modeling techniques are most useful when the verification data sample is small (a few hundred forecast–observation pairs or less), and for verification of rare events, where the sampling variability of forecast quality measures is inherently large.

1. Introduction

Many forecasting systems produce probability forecasts of weather, climate, or hydrologic events. A probability forecast is one that assigns a probability (or likelihood) to the occurrence of a specific event. A familiar example is a probability-of-precipitation forecast. In this case, there are only two possible outcomes—either the event (precipitation) occurs or it does not.

To assess the quality of probability forecasts, a comparison of many forecasts and observations is necessary (Wilks 1995). Among the first attempts at verification of probability forecasts is the work of Brier (1950), which introduces a basic framework for comparing probability forecasts with observations. More recently, Murphy and Winkler (1987) proposed a verification framework called the distributions-oriented (DO) approach. In this framework, the correspondence between forecasts and observations is modeled explicitly by their joint probability distribution. Aspects of forecast quality of interest in verification are derived from the joint distribution. For a comprehensive review of the DO approach, see Murphy (1997). For an introduction to the DO approach and other commonly used forecast verification approaches, see Wilks (1995). Applications of the DO approach include Murphy and Winkler (1992) for probability-of-precipitation forecast verification, Brooks and Doswell (1996) for temperature forecast verification, and Wilks (2000) for climate forecast verification, among others. These applications demonstrate the advantage of the DO approach for diagnostic verification of forecasting systems.

An essential element of forecast verification using the DO approach is the estimation of the joint distribution using a verification data sample. As originally presented by Murphy and Winkler (1987), forecasts and observations are discrete random variables, each defined by a finite set of values (or categories). Therefore, the primitive model of the joint distribution is a contingency table, where the elements are relative frequencies of each forecast–observation pair. Hence, the parameters of the joint distribution model are the relative frequencies and are estimated from the verification data sample by the empirical relative frequencies. The dimensionality (D) of the verification problem refers to the number of model parameters that must be estimated to describe the joint distribution (Murphy 1991). For example, if a forecast of a dichotomous (or binary) event can take on

Corresponding author address: Allen Bradley, IIHR—Hydroscience and Engineering, University of Iowa, Iowa City, IA 52242.
E-mail: allen-bradley@uiowa.edu

one of M distinct values, the minimum number of empirical relative frequencies that must be estimated to define all the elements of the contingency table is $2M - 1$ (since the relative frequencies in the table must sum to 1). For dichotomous events, high dimensionality results when the forecasts can take on a large number of discrete values, or are essentially continuous random variables.

For small verification data samples, high dimensionality is a serious obstacle to forecast verification. In such situations, the empirical relative frequencies of the primitive model would have large sampling variability. Derived measures of forecast quality would also have large uncertainties. This situation could lead to incorrect inferences in diagnostic verification if estimation uncertainty is ignored, which is usually the case in forecast verification (Seaman et al. 1996; Kane and Brown 2000; Stephenson 2000).

One approach for reducing high dimensionality is to reformulate the verification problem. Essentially, a reduced set of discrete forecasts and observations is defined, and data are reclassified (or binned) into this set. Verification is then carried out using the primitive model applied to the reduced set of forecasts and observations. Examples of this approach are shown by Brooks and Doswell (1996) and Wilks (2000). Another approach for reducing dimensionality is to replace the primitive model with a statistical model of the joint distribution. This approach is usually implemented by assuming distributional models for conditional and/or marginal distributions of the joint distribution. Examples of this approach for forecast verification include Clemen and Winkler (1987), Wilks and Shen (1991), and Murphy and Wilks (1998), among others. Statistical models have also been used extensively to represent forecast systems in the context of decision making with forecast information (see Katz et al. 1982; Krzysztofowicz and Long 1991; Wilks 1991; Wilks and Murphy 1986; Wilks et al. 1993).

With either approach, dimensionality and sampling variability are reduced because the assumed model requires fewer parameters. However, whenever the assumed model is a reformulation or simplification of the joint distribution, sample estimates of forecast quality measures may be biased. The issues of bias and sampling variability are discussed by Murphy and Wilks (1998) in their case study of a statistical modeling approach for forecast verification. Parametric models were used to reduce the dimensionality for verification of probability-of-precipitation forecasts from 21 to 4. Despite some minor lack of fit for the assumed statistical models, model biases did not adversely impact inferences of forecast quality. At the same time, the reduced sampling variability was shown to enhance comparative verification of alternate forecasting systems.

In this study, we examine the sampling characteristics of forecast quality measures for probability forecasts of dichotomous events. Our motivation is to identify efficient techniques for DO verification for situations with high dimensionality and small verification data samples (a common situation for probability forecasts from ensemble prediction systems). First, simplified expressions for the joint distribution and forecast quality measures are derived for probability forecasts of dichotomous events. Next, sample estimators are presented for forecast quality measures and the joint distribution. As will be seen, most forecast quality measures can be estimated without reformulation or statistical modeling. For the remaining measures, alternate statistical modeling approaches are considered for estimation. Monte Carlo simulations are then carried out to assess the uncertainty of forecast quality estimators. Two forecast verification examples—one with continuous probability forecasts and the other with discrete forecasts—are used to evaluate the sampling characteristics for various verification data sample sizes. These examples provide some guidance on when to use various estimation techniques depending on the nature of the verification problem and the data sample.

2. Distributions-oriented verification

In describing the DO approach for forecast verification, our presentation follows the notation and summarizes concepts presented by Murphy (1997). Let f be a random variable that denotes a forecast. Let x be random variable that denotes an observation of the quantity being forecast. Forecast verification using the DO approach focuses the joint distribution of forecasts and observations $p(f, x)$ (Murphy 1997). If forecast–observation pairs (f, x) at different times are independent and identically distributed, then any aspect of forecast quality can be determined directly from $p(f, x)$.

The joint distribution can be factored in two ways (Murphy and Winkler 1987). The calibration–refinement (CR) factorization is

$$p(f, x) = q(x \mid f)s(f), \tag{1}$$

where $q(x \mid f)$ is the conditional distribution of the observations given the forecast and $s(f)$ is the marginal distribution of the forecasts. The likelihood-base rate (LBR) factorization is

$$p(f, x) = r(f \mid x)t(x), \tag{2}$$

where $r(f \mid x)$ is the conditional distribution of the forecasts given the observation and $t(x)$ is the marginal distribution of the observations. Both factorizations are employed to examine aspects of forecast quality. Specifically, moments of the joint distribution, its marginal distributions, and the conditional distributions for the two factorizations describe many aspects of the forecast quality of interest in verification (Murphy 1997).

a. Probability forecasts for dichotomous events

Consider a forecasting system that produces a probability forecast for a dichotomous event. Let x be a

Bernoulli random variable that takes on a value of 1 if a particular event occurs and 0 if it does not. Let f be a probability forecast of the occurrence of the event. That is, f is the forecast probability that the event occurs (i.e., that $x = 1$).

For the case of probability forecasts of dichotomous events, simplified expressions for the distributions shown in Eqs. (1) and (2) can be derived. Since the observations x can only take on the values of 0 and 1, the marginal distribution $t(x)$ has the following property:

$$t(x = 0) + t(x = 1) = 1. \tag{3}$$

Since x is a Bernoulli random variable, it follows that

$$t(x = 0) = 1 - \mu_x \quad \text{and} \tag{4}$$

$$t(x = 1) = \mu_x, \tag{5}$$

where μ_x is expected value of the observations. Note that μ_x is equivalent to the climatological probability of the event occurrence. Likewise, the conditional distribution $q(x \mid f)$ has the following property:

$$q(x = 0 \mid f) + q(x = 1 \mid f) = 1, \tag{6}$$

with

$$q(x = 0 \mid f) = 1 - \mu_{x \mid f} \quad \text{and} \tag{7}$$

$$q(x = 1 \mid f) = \mu_{x \mid f}, \tag{8}$$

where $\mu_{x \mid f}$ is the conditional expected value of x given the forecast f.

Aspects of forecast quality are described by the joint distribution, its conditional and marginal distributions, and moments of these distributions. Murphy (1997) defines these aspects and their relevance to forecast verification, and presents summary measures of each. In the following sections, we present expressions for these summary measures of forecast quality for the case of probability forecasts of dichotomous events. Note that some of these expressions are identical to those presented in Murphy (1997). However, others are simplified expressions that apply to dichotomous events only. For completeness, expressions for all the summary measures described in Murphy (1997) are presented.

b. Bias, accuracy, and association

A measure of bias is the mean error (ME). The mean error is defined as

$$\text{ME} = \mu_f - \mu_x, \tag{9}$$

where μ_f is the expected value of the forecasts and μ_x is the expected value of the observations.

A measure of accuracy is the mean square error (MSE). For a probability forecast of a dichotomous event, the mean square error can be written as

$$\text{MSE}(f, x) = E(f - x)^2 \tag{10}$$

$$= (\sigma_f^2 + \mu_f^2) + \mu_x(1 - 2\mu_{f \mid x=1}), \tag{11}$$

where σ_f^2 is the variance of the forecasts and $\mu_{f \mid x=1}$ is a conditional expected value of the forecasts. Alternate expressions for the MSE, based on the CR and LBR factorizations, are given in subsequent sections.

A measure of association is the correlation coefficient (ρ_{fx}). The correlation coefficient is

$$\rho_{fx} = \frac{\text{cov}(f, x)}{\sigma_f \sigma_x} \tag{12}$$

$$= \sqrt{\frac{\mu_x}{1 - \mu_x}}\left(\frac{\mu_{f \mid x=1} - \mu_f}{\sigma_f}\right). \tag{13}$$

c. Calibration refinement measures

Given a specific probability forecast f, certain aspects of the distribution of observations x are desirable. The calibration–refinement factorization, which conditions on the forecast, can be used to explore these aspects of forecast quality.

The reliability describes the bias of the observations given a forecast f. Forecasts that are conditionally unbiased are desirable. One measure of this conditional bias (B_1) is

$$B_1 = E_f(\mu_{x \mid f} - f)^2, \tag{14}$$

where E_f is the expected value with respect to the distribution of the forecasts and $\mu_{x \mid f}$ is the expected value of the observations conditioned on the forecast. For a probability forecast of a dichotomous event,

$$B_1 = E_f(\mu_{x \mid f}^2) - 2\mu_x\mu_{f \mid x=1} + \sigma_f^2 + \mu_f^2. \tag{15}$$

The resolution describes the degree to which the observations for a specific forecast f differ from the unconditional mean (or climatological probability). Forecasts with large differences (high resolution) are more desirable. One measure of the resolution (RES) is

$$\text{RES} = E_f(\mu_{x \mid f} - \mu_x)^2. \tag{16}$$

For a probability forecast of a dichotomous event,

$$\text{RES} = E_f(\mu_{x \mid f}^2) - 2\mu_x\mu_f + \mu_f^2. \tag{17}$$

The relationship between the accuracy of the forecasts, and the reliability and resolution of the forecasts, can be seen through a decomposition of the MSE into its components. Conditioning on the forecast leads to the CR decomposition:

$$\text{MSE}_{\text{CR}}(f, x) = \sigma_x^2 + B_1 - \text{RES}, \tag{18}$$

where the variance of the observation σ_x^2 is a measure of the inherent uncertainty of the variable being forecast. For binary variables, the uncertainty is

$$\sigma_x^2 = \mu_x(1 - \mu_x). \tag{19}$$

d. Likelihood-base rate measures

Given a specific observation x (i.e., the event occurs or it does not), certain aspects of the distribution of the

probability forecasts f are desirable. The likelihood-base rate factorization, which conditions on the observation, can be used to explore these aspects of forecast quality.

The type 2 conditional bias describes the bias of the forecasts given the observation x. Forecasts that are conditionally unbiased are desirable. One measure of this conditional bias (B_2) is

$$B_2 = E_x(\mu_{f|x} - x)^2, \tag{20}$$

where E_x is the expected value with respect to the distribution of the observations and $\mu_{f|x}$ is the expected value of the forecasts conditioned on the observation. For a probability forecast of a dichotomous event,

$$B_2 = (1 - \mu_x)\mu_{f|x=0}^2 + \mu_x(\mu_{f|x=1} - 1)^2. \tag{21}$$

The discrimination describes the degree to which the forecasts differ for a different observations (in this case, for $x = 0$ and $x = 1$). Forecasts with large differences (high discrimination) are more desirable. One measure of the discrimination (DIS) is

$$\text{DIS} = E_x(\mu_{f|x} - \mu_f)^2. \tag{22}$$

For a probability forecast of a dichotomous event,

$$\text{DIS} = (1 - \mu_x)(\mu_{f|x=0} - \mu_f)^2 + \mu_x(\mu_{f|x=1} - \mu_f)^2. \tag{23}$$

The relationship between the accuracy of the forecasts, and the type 2 conditional bias and discrimination of the forecasts, can be seen through another decomposition of the MSE into its components. Conditioning on the observation leads to the LBR decomposition:

$$\text{MSE}_{\text{LBR}}(f, x) = \sigma_f^2 + B_2 - \text{DIS}, \tag{24}$$

where the variance of the forecasts σ_f^2 is a measure of the sharpness of the forecasts. The sharpness is the degree to which probability forecasts are close to 0 and 1. Forecasts with high sharpness are desirable.

As is suggested by (24), there is a trade-off between sharpness, discrimination, and type 2 conditional bias. At one extreme, if the same forecast is always made, the measures of sharpness and discrimination are zero, but the type 2 conditional bias is maximized. Hence, sharp forecasts are required for good forecast quality (i.e., high discrimination and low type 2 conditional bias).

3. Estimation from sample data

A sample of forecasts and observations for the forecast system is used for verification. Let x_i be the observation at time i. Let f_i be the probability forecast of the event at time i. A goal of DO forecast verification is to estimate the joint distribution $p(f, x)$ and attributes of forecast quality using the sample $\{f_i, x_i, i = 1, \ldots, N\}$.

For the forecast quality measures shown in sections 2b–d, most depend only on the moments (unconditional and conditional) of the joint distribution. Therefore, traditional sample estimators of these quantities can be used for estimation (see appendix A). Using the sample of observations $\{x_i, i = 1, \ldots, N\}$, the sample mean can be used to estimate μ_x. Likewise, using the sample of forecasts $\{f_i, i = 1, \ldots, N\}$, the sample mean and variance can be used to estimate μ_f and σ_f^2. The forecast quality measures also have terms involving the conditional means of the forecasts given the observation. These are estimated by partitioning the forecasts f_i into two sets, one for the case with $x_i = 0$ and the other for $x_i = 1$. The conditional means $\mu_{f|x=0}$ and $\mu_{f|x=1}$ are then estimated by their respective sample means.

Two forecast quality attributes for the CR factorization, the reliability B_1 and the resolution RES, also require estimates of $E_f(\mu_{x|f}^2)$. A closed-form analytical expression in terms of moments of the joint distribution is not possible for this term (if the form of the distribution is unknown). For the primitive model of the joint distribution $p(f, x)$ based on a contingency table, estimation of the conditional mean $\mu_{x|f}$ for each discrete forecast value is straightforward. However, for situations where a large number of discrete forecasts are permissible, or where forecasts are essentially continuous variables, small verification samples can produce unreliable estimates of the conditional mean due to high sampling variability and bias. Therefore, a stronger assumption regarding the form of $\mu_{x|f}$ is needed for estimation. Two alternative statistical modeling approaches are described in the following sections.

a. Logistic regression approach

In the statistical literature, a logistic regression model is often used to represent the form of the relationship between the relative frequency of binary outcomes and other variables (Cox and Snell 1989; Hosmer and Lemeshow 1989). The classic logistic regression model assumes a linear relationship between the logistic transformation of $\mu_{x|f}$ and f, or

$$\ln\left(\frac{\mu_{x|f}}{1 - \mu_{x|f}}\right) = \beta_0 + \beta_1 f, \tag{25}$$

where β_0 and β_1 are model parameters. In essence, this model assumes a parametric form for the distribution $q(x \mid f)$ from the CR factorization, because of its relationship with $\mu_{x|f}$ [see Eqs. (5) and (6)]. Solving for the conditional mean gives

$$\mu_{x|f} = \frac{e^{\beta_0 + \beta_1 f}}{1 + e^{\beta_0 + \beta_1 f}}. \tag{26}$$

The model parameters β_0 and β_1 can be estimated from sample data using the method of maximum likelihood (see appendix B). The sample estimator $\hat{\mu}_{x|f}$ is obtained by substituting $\hat{\beta}_0$ and $\hat{\beta}_1$, the estimated parameters from the logistic regression, into (26). Using this relationship, the term in the reliability B_1 and resolution R equations can be estimated by

$$\hat{E}_f(\mu^2_{x|f}) = \frac{1}{N}\sum_{i=1}^{N} \hat{\mu}^2_{x|f_i}, \tag{27}$$

where $\hat{\mu}_{x|f_i}$ is the estimated value from the logistic regression evaluated at forecast f_i.

b. Kernel density estimation

An alternate way of estimating $\mu_{x|f}$ uses the LBR factorization. From Eqs. (1) and (2), the conditional distribution $q(x \mid f)$ may be written as

$$q(x|f) = \frac{t(x)r(f|x)}{s(f)}. \tag{28}$$

Using the properties of forecasts for dichotomous events discussed in section 2a, the conditional mean $\mu_{x|f}$ simplifies to

$$\mu_{x|f} = \mu_x \frac{r(f|x=1)}{s(f)}. \tag{29}$$

The marginal distribution $s(f)$ may also be written as

$$s(f) = (1-\mu_x)r(f|x=0) + \mu_x r(f|x=1). \tag{30}$$

Hence, the conditional mean $\mu_{x|f}$ can be estimated using the conditional distribution $r(f \mid x)$ from the LBR decomposition.

One approach would be to choose a parametric distribution for $r(f \mid x = 0)$ and $r(f \mid x = 1)$. For example, Clemen and Winkler (1987) and Krzysztofowicz and Long (1991) used beta distributions to model the conditional distributions of forecasts. However, if the appropriate form of the distribution is not known a priori, nonparametric methods offer a means for estimation (Scott 1992). We examined using a kernel density estimation technique to estimate $r(f \mid x = 0)$ and $r(f \mid x = 1)$ from the sample of forecasts $\{f_i, i = 1, \ldots, N\}$. Although kernel estimators for discrete distributions may be used when f takes on a finite set of discrete values (Titterington 1980; Wang and Van Ryzin 1981; Rajagopalan and Lall 1995, among others), we will assume that f is a continuous random variable.

The first step in estimation is to partition the forecasts f_i into two sets, one for the case with $x_i = 0$ and the other for $x_i = 1$. Kernel density estimation is then used to estimate $r(f \mid x)$ with the respective subsamples. The continuous kernel density estimator of $r(f \mid x)$ is given by

$$\hat{r}(f|x) = \frac{1}{N_x h_x}\sum_{i=1}^{N_x} K\left(\frac{f - f_i^x}{h_x}\right), \tag{31}$$

where $K(\,)$ is the kernel, N_x is the size of the particular subsample ($x \in 0, 1$), h_x is the bandwidth parameter, and f_i^x is the ith forecast in the subsample. The approach used to estimate the bandwidth parameter and density function is outlined in appendix C.

Since f is assumed to be continuous, the term in the reliability B_1 and resolution RES equations can be written as

$$E_f(\mu^2_{x|f}) = \int_0^1 \mu^2_{x|f} s(f)\, df. \tag{32}$$

An estimator $\hat{E}_f(\mu_{x|f})$ is obtained by substituting kernel density estimators for $r(f \mid x = 0)$ and $r(f \mid x = 1)$ into (30) for $\mu_{x|f}$ and (31) for $s(f)$ and integrating numerically.

c. Joint distribution

With the estimators described so far, all the forecast quality measures shown in sections 2b–d can be estimated. In our evaluation in subsequent sections, we will focus primarily on the sampling characteristics of these estimators. Still, estimates of the joint distribution $p(f, x)$ and its factorizations are extremely valuable in diagnostic verification. For example, reliability and discrimination diagrams use graphical representations of the marginal and conditional distributions to visualize aspects of forecast quality (Wilks 1995). When the kernel density estimation method is used to estimate $r(f \mid x)$, the joint distribution $p(f, x)$ is completely defined by the LBR factorization in (2). In contrast, when the logistic regression method is used to estimate $q(x \mid f)$, an estimate of the marginal distribution $s(f)$ would still be required to define $p(f, x)$ by the CR factorization in (1). One approach for estimating $s(f)$ would be to use the nonparametric kernel density estimation method outlined in appendix C.

4. Examples

Two forecast verification examples are used to evaluate the sampling uncertainty of the forecast quality measure estimators. In each example, the joint distribution $p(f, x)$ and the true values of the forecast quality measures are known (either analytically or by Monte Carlo simulation). Sets of forecast–observation pairs are then randomly generated. One thousand verification data samples are generated for sample sizes ranging from 50 to 1000 pairs. For each verification sample, verification approaches are applied to estimate forecast quality measures. The CR measures of reliability B_1 and resolution RES are estimated using both the logistic regression (LR) and kernel density (KD) methods. The remaining measures are estimated analytically using the derived expressions involving sample estimators of the moments and conditional moments (for simplicity, we will use the term ''analytical expression'' to describe one of these estimators). In addition, the primitive model (PM) of the joint distribution is used to estimate the forecast quality measures.

a. Example 1: Continuous forecasts

This example examines continuous probability forecasts of drought occurrence generated by an ensemble streamflow forecasting technique (Day 1985; Smith et

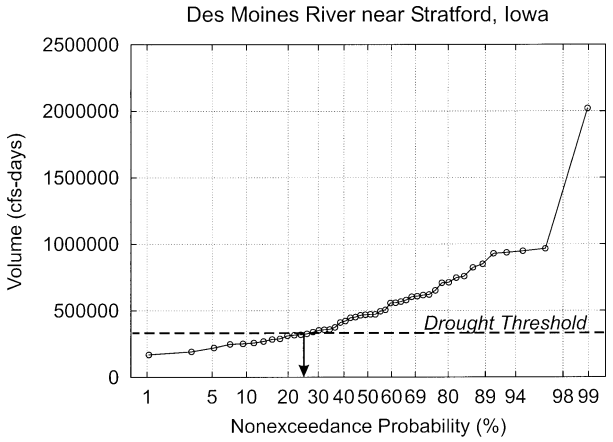

FIG. 1. Determination of the probability forecast f for a drought event from an ensemble forecast. A specific threshold defines a drought event. The nonexceedence probability is determined for the threshold from the forecast empirical distribution. The example is based on the ensemble streamflow prediction of seasonal flow volume for the Des Moines River at Stratford, IA (Hashino et al. 2002).

al. 1992). The ensemble forecasts were originally made using an experimental system for the Des Moines River at Stratford, Iowa. A stochastic model that mimics the ensemble forecasts of monthly (September) streamflow volume was then developed for the Monte Carlo simulation. For a detailed description of the stochastic model of the Des Moines River system used in this example, see Hashino et al. (2002).

The Monte Carlo simulations are carried out by first randomly generating a monthly flow volume, then randomly generating a synthetic ensemble forecast (conditioned on the monthly volume). If the monthly volume is less than a specified threshold, we say a ''drought'' occurs and the observation x is 1. If the volume is greater than the threshold (i.e., no drought occurs), the observation x is 0. The probability forecast f is simply the probability that the volume is less than the threshold, which is found directly from the synthetic ensemble forecast (see Fig. 1). Two cases are examined. In one, the volume threshold is set so that drought occurrence is a relatively common event ($\mu_x = 0.25$). In the other, the threshold is set so that drought occurrence is a relatively rare event ($\mu_x = 0.05$). The true values of the forecast quality measures are shown in Table 1.

For each of 1000 verification data samples, forecast quality measures are estimated. Since the actual forecasts are continuous random variables, the forecasts must be reformulated to apply the PM approach. In this example, we selected a set of 11 discrete forecast values, {0, 0.1, 0.2, . . . , 1}, for a dimensionality D of 21. Binning of the continuous forecast is done by assigning the forecast to the nearest discrete value. With the 1000 estimates, the uncertainty in the estimators for each approach is quantified.

First we examine the uncertainty in the estimators for summary measures of bias (ME) and accuracy (MSE). Note that both of these measures can be estimated with an analytical expression. Figure 2 shows ME for forecasts of the common and the rare drought events for sample sizes ranging from 50 to 1000. Figure 3 shows similar results for MSE. To facilitate comparisons among difference cases and examples, relative errors are presented. The relative errors are found by normalizing by the (true) uncertainty σ_x^2, or its square root in the case of ME (see Table 1). The symbols in the figures indicate the mean relative error (a measure of the bias of the estimator). The error bars indicates the standard error of the estimator (a measure of the sampling variability).

As can be seen in Figs. 2 and 3, the uncertainty in the measures based on the analytical expressions and the PM approach are virtually the same. The reason is that the sample estimators based on the two approaches are mathematically equivalent. The only difference is that the forecast values for the PM approach have been recoded to discrete values. Still, in the case of continuous (or nearly continuous) forecasts, the analytical expressions have a clear advantage in that there is no need to select bin sizes and recode forecasts to estimate the forecast quality measures.

Figures 2 and 3 also illustrate two other general results. First, the uncertainty in the estimates decreases sharply as the sample size increases. For all the analytical expressions, the standard error for a sample size of 50 is roughly 4.5 times larger than for a sample size of 1000, about what one would expect if the standard error decreases as $N^{-1/2}$. Second, the relative uncertainty is larger for the rare events ($\mu_x = 0.05$). Fewer occurrences of the forecast event leads to higher uncertainty for the same sample size. For the ME estimates (Fig. 2), the differences in uncertainty for the two cases are not as great as for MSE estimates (Fig. 3). The reason is that ME depends on the first moments of the joint distribution, whereas MSE depends on the second mo-

TABLE 1. Relative measures of forecast quality.

	MSE/σ_x^2	ME/σ_x	B_1/σ_x^2	RES/σ_x^2	B_2/σ_x^2	DIS/σ_x^2
			Example 1: Continuous forecasts			
Common events	0.558	0.025	0.028	0.470	0.201	0.305
Rare events	1.293	0.072	0.514	0.220	0.325	0.189
			Example 2: Discrete forecast			
PoP	0.539	0.040	0.00603	0.467	0.286	0.217

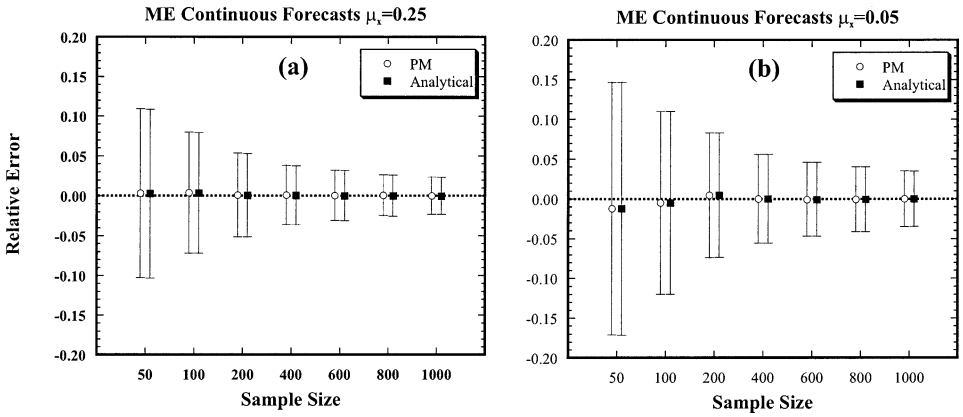

FIG. 2. Relative errors of estimates of ME for example 1 (continuous forecasts) for (a) the common drought event ($\mu_x = 0.25$) and (b) the rare drought event ($\mu_x = 0.05$). Results are shown for PM and the derived analytical expressions. The symbols indicate the mean error (bias) and the error bars indicate ± one standard error.

ments. Since all the other analytical expressions involve second-order moments, the differences for rare and common events cases are more similar to those seen for MSE. As a result, for forecasts of rare events, it is clear that the uncertainty of many forecast quality measures will be quite large for the smallest sample sizes shown.

Since the results for all the analytical expressions are similar to those illustrated in Figs. 2 and 3, the remainder of this section will focus on estimators for CR measures of reliability (B_1) and resolution (RES), which depend on the method for estimating $\mu_{x|f}$. Figure 4 shows the relative errors for B_1 for forecasts of the common and rare drought events. Figure 5 shows similar results for RES. The figures show results for the LR and KD statistical modeling methods, and the PM approach applied to the recoded set of discrete forecasts.

For forecasts of the common event (Figs. 4a and 5a), the PM estimator of B_1 and RES is the worst estimator for the smaller sample sizes due to large biases and high standard errors. The LR estimator tends to be the best for most sample sizes; it has slightly lower standard error than the KD estimator, and it tends to have low bias. The KD estimator is best only at the smallest sample sizes, when its bias is the lowest. For forecasts of the rare event (Figs. 4b and 5b), the LR estimator has much lower bias and slightly lower sampling variability than the other estimators for all sample sizes. The KD estimator has the highest uncertainty. Its poor performance for the rare event is due to the partitioning of the verification data sample into two subsamples. In the case of a sample size of 50, on average there are only 2.5 events in the subsample for $x = 1$ for kernel density estimation. Even for larger samples, the uncertainty for $r(f \mid x = 1)$ will be much higher than for $r(f \mid x = 0)$ because its subsample is much smaller relative to the other. This uncertainty results in large biases and higher standard errors for the KD method over the range of sample sizes.

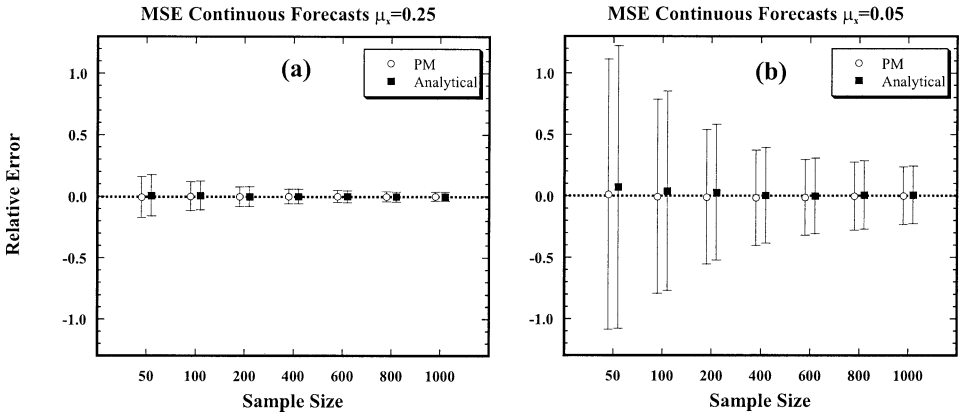

FIG. 3. Relative errors of estimates of MSE for example 1 (continuous forecasts) for (a) the common drought event ($\mu_x = 0.25$) and (b) the rare drought event ($\mu_x = 0.05$). Results are shown for PM and the derived analytical expressions. The symbols indicate the mean error (bias) and the error bars indicate ± one standard error.

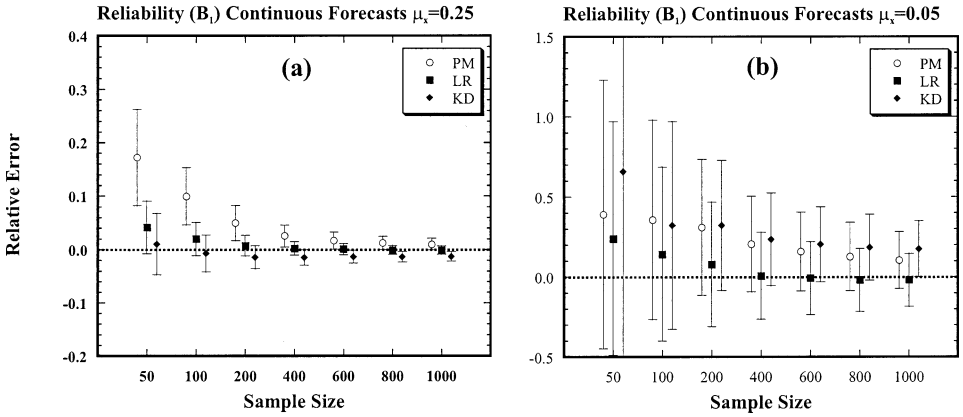

FIG. 4. Relative errors of estimates of reliability (B_1) for example 1 (continuous forecasts) for (a) the common drought event ($\mu_x = 0.25$) and (b) the rare drought event ($\mu_x = 0.05$). Results are shown for PM, and the LR and KD methods. The symbols indicate the mean error (bias) and the error bars indicate ± one standard error.

The performance of the two CR estimators can be better understood by examining the distributions in the CR factorization. A reliability diagram is a graphical representation of components of the CR factorization (Wilks 1995). Figure 6 shows the reliability diagram, $\mu_{x|f}$ versus the forecast f, for a sample size of 50. Forecasts that have perfect reliability (or are conditionally unbiased) follow the 1:1 line. Forecasts have resolution if $\mu_{x|f}$ differs from the unconditional mean μ_x. The true relation for $\mu_{x|f}$ (displayed in all the panels) shows that the drought forecasting system has resolution, although there are some conditional biases. The mean (filled circles) and one standard deviation (error bars) for the estimators are also shown for each method.

For the PM estimator, the results are very different for forecasts of the common and the rare event. For the common event ($\mu_x = 0.25$), the estimates of $\mu_{x|f}$ usually have the lowest bias, but much higher variability than the statistical modeling approaches. In contrast, for the rare event ($\mu_x = 0.05$), the estimates are significantly biased, but have lower variability than the other approaches. With only 2.5 events (on average) with $x = 1$ for the rare event with a sample size of 50, the PM finds that most discrete forecast bins in the contingency table will have no observations of this outcome (i.e., the empirical relative frequency is 0). This results in a mean estimated value of $\mu_{x|f}$ near zero for all discrete forecast values.

Due to the shape of the true curves near $f = 0$ and 1, the assumed LR model does not fit the true curve perfectly. Still, the lower variability of the estimates and fairly low bias lead to its good performance. Although the KD method performs well for the common event, it has the highest variability for the rare event. So even with its lower biases over much of the range of forecasts, its estimators of B_1 and RES have higher uncertainty than even those for the PM.

b. Example 2: Discrete forecasts

In the first example, the forecasts f are issued as continuous numbers between 0 and 1. To use the primitive model, the forecasts were recorded into a discrete set of forecast values. However, for a forecasting system that issues forecasts as discrete values, the primitive model would in fact be the correct model of the joint distribution $p(f, x)$. How well would the statistical modeling approaches work in this case?

This example examines discrete forecasts of the probability of precipitation (PoP). The example is drawn from the subjective 12–24-h PoP forecasts for the United States that were verified by Wilks (1995) using the DO approach. The verification results were used in this example to define the true joint distribution $p(f, x)$. The true values of the forecast quality measures are shown in Table 1. For the Monte Carlo experiments, verification datasets were created by randomly generating forecast–observation pairs from the joint distribution model. The forecasts were issued as a set of discrete values $\{f = 0, 0.05, 0.1, 0.2, \ldots, 0.9, 1\}$. Hence, $p(f, x)$ has dimensionality D of 23. This same set of discrete forecasts was used to estimate forecast quality measures using the PM. The LR and KD approaches were applied as before for the two CR measures, even though these approaches assume that f is a continuous variable.

Figure 7 shows the relative errors for B_1 and RES for sample sizes ranging from 50 to 1000. For the PoP forecasts, the mean observation μ_x is 0.162 (i.e., precipitation occurs 16.2% of the time). This is in between the common and rare event cases in example 1. Still, the sampling uncertainty of the relative measures is only slightly higher than for the common event case.

Even for forecasts issued as discrete values, where the primitive model is the correct model, the forecast quality estimators for the statistical modeling approaches tend to have lower uncertainty. For both B_1 and RES,

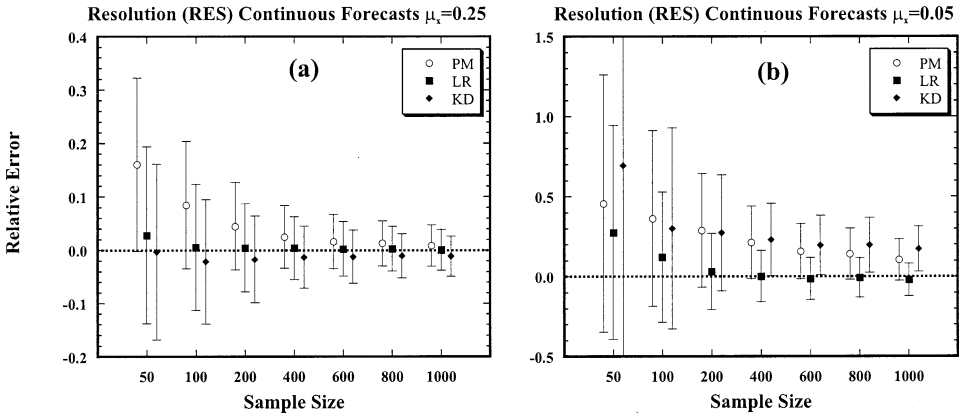

FIG. 5. Relative errors of estimates of resolution (RES) for example 1 (continuous forecasts) for (a) the common drought event ($\mu_x = 0.25$) and (b) the rare drought event ($\mu_x = 0.05$). Results are shown for PM, and the LR and KD methods. The symbols indicate the mean error (bias) and the error bars indicate ± one standard error.

the KD and LR estimators have lower biases than the PM estimator for the smallest sample sizes. For B_1, the standard errors are also much lower for the KD and LR estimators. Unlike example 1, the KD method performs best for both B_1 and RES for all sample sizes due to its lower biases. At sample sizes of 600 or greater, the PM estimator is slightly better than the LR estimator for both B_1 and RES due to lower biases.

Figure 8 shows the reliability diagram for the three methods for sample sizes of 50 and 1000. The true curve shows a nearly linear relationship (low conditional bias). For a sample size of 50, the PM estimates have large (negative) biases for f of around 0.5 and greater, and much higher variability than the other methods. The KD method is more flexible, and fits the true curve better than the LR, resulting in its better overall performance for the resolution and reliability measures. At a sample size of 1000, the PM estimates are unbiased, but with higher variability than that for the LR and KD. Again, KD fits the true curve better than LR, yielding estimates with lower uncertainty, even at this larger sample size.

5. Discussion

The DO approach offers a sound statistical theory for forecast verification problems. As originally presented, the DO approach assumes that both the forecast and the observation are discrete random variables. The techniques presented here essentially relax this assumption for the special case of probability forecasts for dichotomous events; the derived expressions for forecast quality measures are applicable to either discrete or continuous forecasts. For this special forecasting case, most of the forecast quality measures do not depend on the distributional form of the forecasts. These measures are easily estimated using analytical expressions involving sample moment estimators, even in cases with high dimensionality (i.e., a large set of discrete forecast or essentially continuous forecasts). Only the CR factorization measures of reliability and resolution depend on distributional assumptions, and they are dealt with using a statistical modeling approach.

a. Analytical expressions

From the standpoint of uncertainty, the analytical expressions offer little advantage over the application of the primitive model to a reformulated set of forecasts with discrete values. However, from a practical standpoint, the analytical expressions simplify the calculations and remove subjective variability from the results. Specifically, the selection of discrete forecast values and bin sizes is a subjective decision when forecasts must be recoded to apply the primitive model. Yet different choices will produce slightly different forecast quality estimates for the sample. Such problems can simply be avoided using the analytical expressions.

b. Statistical modeling approach

For the CR measures, the distributional assumptions implied by the statistical modeling approaches greatly improve the estimates of forecast quality, especially for small sample sizes. The improvement is achieved mostly by reducing the bias of the estimators. However, for the smallest sample sizes, the standard errors are often lower as well.

We presented two approaches for estimating the CR measures. The logistic regression model essentially assumes a parametric model for the CR conditional distribution $q(x \mid f)$. The kernel density estimation assumes a nonparametric model for the LBR conditional distribution $r(f \mid x)$. An advantage of the KD method is that the parametric form of the distribution is not required. This allows for greater flexibility in modeling the joint distribution, which the results for example 2

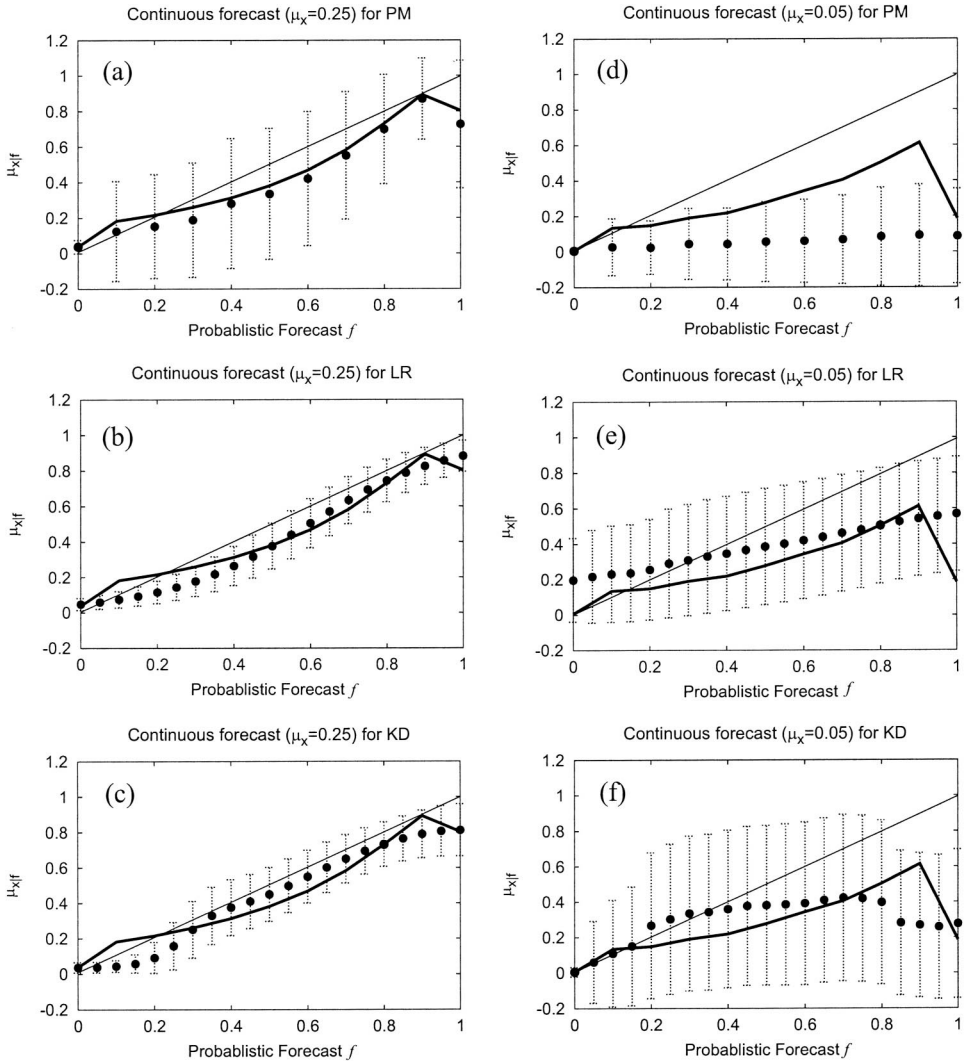

FIG. 6. Estimates of $\mu_{x|f}$ for example 1 for common events ($\mu_x = 0.25$) for (a) PM, and the (b) LR and (c) KD methods, and for rare events ($\mu_x = 0.05$) for (d) PM and the (e) LR and (f) KD methods. The solid line (thick) shows the true $\mu_{x|f}$. A one-to-one line (thin) for perfectly reliable forecasts is also shown for comparison. The symbols indicate the mean estimate and the error bars indicate $\pm$ one standard error.

show can lead to lower biases and overall uncertainty in some forecasting situations. In contrast, the LR method makes a stronger assumption regarding the relationship between forecasts and observations, which results in lower sampling variability, but also higher biases in some situations. An advantage of the LR method is that the model is fitted using the entire verification data sample. In contrast, the KD method requires that the forecast sample be split into two subsamples. When the size of one subsample is much smaller than the other, which happens when one of the binary outcomes occurs infrequently, the results show that the KD method performs poorly.

The Monte Carlo simulations revealed cases where logistic regression and kernel density estimation techniques will fail in forecast verification applications. For example, when all the forecasts f_i for the case of $x_i = 0$ are less than those when $x_i = 1$, the logistic regression cannot find an optimal set of parameters by the method of maximum likelihood. This situation can occur when there are just a few observations for one of the cases (rare events). Also, for kernel density estimation, a subsample must contain at minimum two values for estimation. Of course, this is only a limitation for forecasts of rare events, when the KD method performs poorly anyway.

c. Alternate statistical modeling methods

There are many variations of the two statistical modeling methods presented that could be explored. A nonparametric logistic regression technique (Bowman and

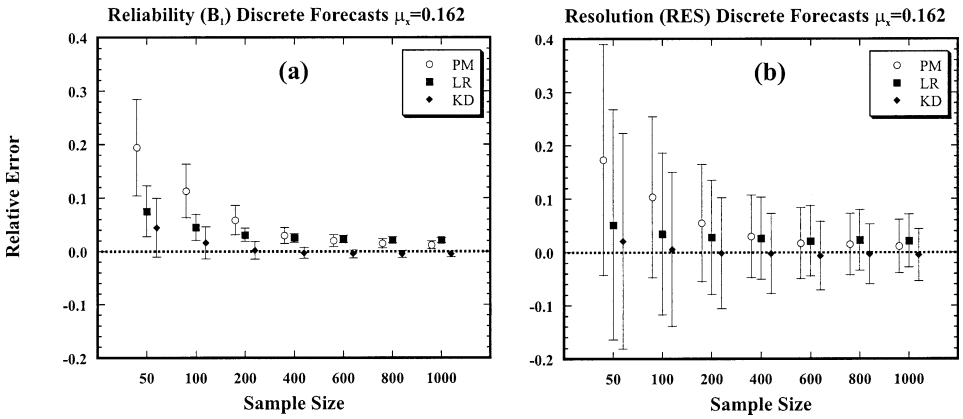

FIG. 7. Relative errors of estimates of CR measures for example 2 (discrete forecasts) for P. P forecasts (Wilks 1995): (a) reliability (B_1) and (b) resolution (RES). Results are shown for PM, and the LR and KD methods. The symbols indicate the mean error (bias) and the error bars indicate ± one standard error.

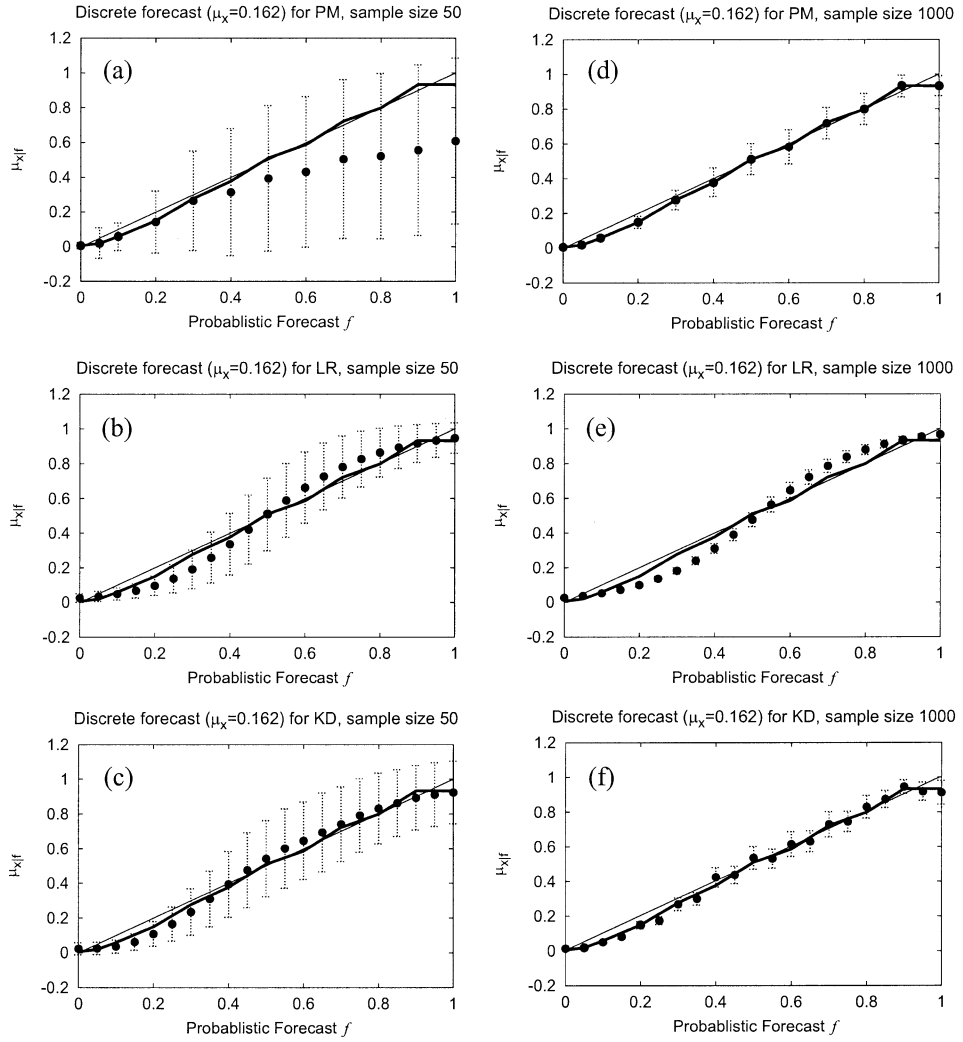

FIG. 8. Estimates of $\mu_{x|f}$ for example 2 for a sample size of 50 for (a) PM, and the (b) LR and (c) KD methods, method, and for a sample size for 1000 for (d) PM and the (e) LR and (f) KD methods. The solid line (thick) shows the true $\mu_{x|f}$. A one-to-one line (thin) for perfectly reliable forecasts is also shown for comparison. The symbols indicate the mean estimate and the error bars indicate ± one standard error.

Azzalini 1997; Nottingham et al. 2000) would provide a more flexible model for estimating $q(x \mid f)$. The resulting estimators would likely be less biased but more variable. Likewise, the reduced variability associated with a parametric model for $r(f \mid x)$ could be more efficient in some situations. Still, the poor performance for forecasts of rare events is not likely to disappear with parametric models. Murphy and Wilks (1998) utilized a linear regression model for $q(x \mid f)$ and a beta distribution model for $s(f)$. Although a linear model may not be the best choice for use with the derived expressions, since it often produces values of $q(x \mid f)$ outside the 0 to 1 range for certain values of f, one advantage of the model was that its parameters could be directly interpreted as alternate measures of forecast quality. In applications of the statistical modeling approach, careful study of the distributional form of forecasts and observations may be warranted to help in selecting the best model—either parametric for its lower sampling variability, or nonparametric for its better fit and lower biases.

We have already looked at some other minor variations (Hashino et al. 2002). For instance, we have found that using the sample estimator $E_f(\mu^2_{x|f})$ in (27) may be better than the integral estimator in (33) for the KD method for the smallest sample sizes. We have also explored using the integral estimator with the LR method [which would require an estimate of $s(f)$], but the results typically are worse than with the sample estimator.

d. Guidelines and recommendations

Although more comprehensive study of the sampling characteristics of forecast quality estimators is needed, we offer some rough guidelines for application of the methods studied. Obviously, the analytical expressions offer the simplest means for estimating most forecast quality measures for probability forecasts (continuous or discrete). For the CR measures and the joint distribution, either the reformulated PM, the LR method, or the KD method could be selected. For small sample sizes of around 500 forecast–observation pairs or less, we would recommend using the LR method for both continuous and discrete forecasts. It is clearly the best of the approaches examined for forecasts of rare events and still performs about as well as the KD method for common events. When using the LR method, an estimate of $s(f)$ is required to completely define the joint distribution $p(f, x)$. If estimates of the forecast quality measures are sufficient for verification, then there is no need to develop an estimate $s(f)$. Still, the complete description of $p(f, x)$ is an essential element of the DO philosophy (and is addressed in section 3). We would also recommend the KD method, but only for verification of forecasts of frequently occurring events. Roughly speaking, the climatological probability of the forecast event should be between about 0.1 and 0.9 for application of the KD method. Unlike the LR method, the joint distribution $p(f, x)$ is completely defined using the KD method. For continuous forecasts, where subjective decisions on bin size are needed to apply the PM, the two statistical modeling approaches may still be useful for sample sizes greater than 500. However, due to inherent biases due to lack of fit for the statistical models, it is important to recognize that at some point, the PM may become the most efficient method, even for continuous forecasts.

Finally, in most applications of forecast verification, the uncertainty in verification measures is rarely considered (some notable exceptions include Woodcock 1976; Seaman et al. 1996; Hamill 1999; Kane and Brown 2000; Stephenson 2000; Connor and Woodcock 2000; Thornes and Stephenson 2001). In this work, the uncertainty of the estimators is central to our evaluation. For probability forecasts of events that occur relatively frequently, with a verification data sample consisting of many hundreds of forecast–observation pairs, the standard errors for the CR estimates with the primitive model may already be acceptably low (usually much less than 10%). However, the uncertainty of forecast quality estimates is much larger for forecasts of rare events, where there are fewer observations of the event itself. Obviously, for rare events, verification should be carried out with the longest possible data sample. However, in many situations, the size of the data sample is severely constrained. For example, with streamflow volume forecasts, there may be only one forecast per year for a particular season. In this case, a verification dataset created based on 50 yr of historical streamflow observations will only contain 50 forecast–observation pairs. Even in the case of PoP forecasts issued on a daily basis, the forecasting system would need to be run for almost 3 yr to obtain 1000 observations for a single site. In many decision problems, rare events are also high-impact events (e.g., droughts, severe weather), and users are quite interested in the forecast quality for these events. Clearly, verification of rare event forecasts using small sample sizes needs to consider the uncertainty of forecast quality estimates as part of the process.

6. Summary and conclusions

The distributions-oriented (DO) approach to forecast verification defines aspects of forecast quality using the joint distribution of forecasts and observations. As originally proposed by Murphy and Winkler (1987), the primitive model of the joint distribution is a contingency table, where the elements of the table are relative frequency of each combination of forecast and observation. Aspects of forecast quality are evaluated using empirical relative frequencies estimated using a verification data sample.

An alternative DO approach is presented for verification of probability forecasts of dichotomous events. For this special case, one can derive simplified expressions for summary measures of forecast quality. Unlike

the original DO framework, which assumes that forecasts are discrete random variables, the simplified expressions are valid for either discrete or continuous forecasts.

Using the simplified expressions, most of the forecast quality measures are estimated analytically using sample moments from a verification data sample. Although the sample estimators for these measures are mathematically equivalent to those for the primitive model, slight differences in estimates can arise when continuous forecasts must be recoded to discrete values for use of the primitive model. This step in the process introduces subjective variability to the estimates, but is unnecessary with the analytical expressions.

Two other forecast quality measures from the calibration–refinement (CR) factorization depend on the distributional form of the forecasts and cannot be estimated from sample moments alone. Two statistical modeling approaches, one a parametric approach (logistic regression) and the other nonparametric (kernel density estimation), were examined for estimating these measures. Monte Carlo experiments for two forecasting examples show that use of the statistical modeling approach can significantly improve estimates of the CR forecast quality measures for small sample sizes—about 500 forecast–observation pairs or less. The statistical modeling approach improves estimates mostly by reducing the large bias of the primitive model estimates, although in some cases, the sampling variability is also reduced. Even when forecasts are issued as discrete values, where the primitive model is the correct model of the joint distribution, the statistical modeling approaches yield better estimates of the CR measures for small sample sizes. The logistic regression method performs well in most situations; the kernel density estimation method also works well, but only for probability forecasts of frequently occurring events.

As a practical matter, the improvements in the estimates of the CR measures afforded by the statistical modeling approach are most critical for forecasts of rare events and for very small verification data samples. In these cases, estimates of forecast quality using the primitive model are highly biased and have large sampling variability, which can lead to incorrect inferences in diagnostic verification. Use of the statistical modeling approach does not eliminate the problem, but it does help to make a better assessment of forecast quality. However, even with improved methods for estimating forecast quality measures, techniques are still needed to evaluate the uncertainty of forecast quality measures as an integral part of the verification process.

Acknowledgments. This work was supported in part by the National Oceanic and Atmospheric Administration (NOAA) Office of Global Programs under Grants NA86GP0365 and NA16GP1569, as part of the GEWEX Continental-Scale International Project (GCIP) and the GEWEX Americas Prediction Project (GAPP). We gratefully acknowledge this support. We would also like to thank the two anonymous reviewers for their thoughtful comments and suggestions.

APPENDIX A

Sample Moment Estimators

Let x_i be the observation at time *i.* Let f_i be the probability forecast of the event at time *i.* The verification data sample is then $\{f_i, x_i, i = 1, \ldots, N\}$. The traditional sample moment estimators for the mean are

$$\hat{\mu}_x = \frac{1}{N}\sum_i^N x_i \quad \text{and} \tag{A1}$$

$$\hat{\mu}_f = \frac{1}{N}\sum_i^N f_i. \tag{A2}$$

Since x is a Bernoulli random variable, the sample estimator for the variance of the observations σ_x^2 is simply

$$\hat{\sigma}_x^2 = \hat{\mu}_x(1 - \hat{\mu}_x). \tag{A3}$$

The sample estimator for the variance of the forecast σ_f^2 is

$$\hat{\sigma}_f^2 = \frac{1}{N}\sum_i^N (f_i - \hat{\mu}_f)^2. \tag{A4}$$

To estimate the conditional means $\mu_{f|x=0}$ and $\mu_{f|x=1}$, the verification data sample is partitioned into two sets. Let $\{f_j^0, j = 1, \ldots, N_0\}$ be the subsample of forecasts for the case where the event does not occur ($x = 0$). Let $\{f_k^1, k = 1, \ldots, N_1\}$ be the subsample of forecasts for the case where the event occurs ($x = 1$). The conditional means are then estimated using the sample means:

$$\hat{\mu}_{f|x=0} = \frac{1}{N_0}\sum_f^{N_0} f_j^0 \quad \text{and} \tag{A5}$$

$$\hat{\mu}_{f|x=1} = \frac{1}{N_1}\sum_k^{N_1} f_k^1. \tag{A6}$$

The analytical expressions for the forecast quality measures in sections 2b–d can then be estimated by replacing the moments with the sample moment estimates shown above. (It is worth noting that the forecast quality measures could be computed in other ways. For example, the MSE can be estimated more directly by the mean squared difference between the sample forecasts f_i and observations x_i.)

APPENDIX B

Logistic Regression

For the logistic regression model shown in (25), the model parameters β_0 and β_1 can be estimated from sample data using the method of maximum likelihood (see Cox and Snell 1989). For the data sample $\{f_i, x_i, i =$

$1, \ldots, N\}$, the probability for each outcome is given by

$$P\{X = x_i | \beta_0, \beta_1\} = \frac{e^{x_i(\beta_0+\beta_1 f_i)}}{1 + e^{\beta_0+\beta_1 f_i}}. \quad \text{(B1)}$$

The log-likelihood function is then simply

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{N} \ln P\{X = x_i | \beta_0, \beta_1\} \quad \text{(B2)}$$

$$= \sum_{i=1}^{N} \{x_i(\beta_0 + \beta_1 f_i) - \ln[1 + \exp(\beta_0 + \beta_1 f_i)]\}. \quad \text{(B3)}$$

Maximizing the log-likelihood with respect to the parameters yields the parameter estimates. This can be done numerically using an optimization procedure. We utilized the modified Newton's method, initialized using the best parameters from a grid search of the parameter space. The sample estimator $\hat{\mu}_{x|f}$ is then obtained by substituting $\hat{\beta}_0$ and $\hat{\beta}_1$, the estimated parameters from the logistic regression, into (25).

APPENDIX C

Kernel Density Estimation

The continuous kernel density estimator of a density function $f(x)$ is

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \quad \text{(C1)}$$

where $K(\,)$ is the kernel, h is the bandwidth, and x_i are the sample data. A common choice for $K(\,)$ is the biweight kernel, given by

$$K(t) = \frac{15}{16}(1 - t^2)^2. \quad \text{(C2)}$$

The bandwidth h acts as a smoothing parameter for the density estimation. The bandwidth is often estimated from sample data by cross validation or other approaches that attempt to find the best fit to the data. Equivalent bandwidth scaling is often used to convert the bandwidth obtained from a normal kernel cross-validation rule into one for another kernel (Scott 1992). Here, we use the bandwidth estimated through the normal reference rule, which minimizes the asymptotic mean integrated error between a normal distribution and a normal kernel estimate, multiplied by the equivalent bandwidth scaling factor for a biweight kernel, or

$$h = 2.623(4/3)^{1/5}\sigma N^{-1/5}, \quad \text{(C3)}$$

where σ is estimated by the sample variance from the data sample.

The difficulty in using kernel estimation to estimate the conditional or marginal distribution of forecasts is that the forecast f is bounded at 0 and 1. As a result, a kernel density estimator may not be a consistent estimator near 0 and 1 (Zhang et al. 1999). An approach for dealing with the boundary effect is the reflection method. This method literally reflects original samples against a boundary and then applies the ordinary kernel to the reflected samples (Scott 1992). The final density estimates are obtained by tripling the estimates between 0 and 1, since the original sample size has been tripled.

REFERENCES

Bowman, A. W., and A. Azzalini, 1997: *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations.* Oxford University Press, 208 pp.

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, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting,* **11,** 288–303.

Clemen, R. T., and R. L. Winkler, 1987: Calibrating and combining precipitation probability forecasts. *Probability and Bayesian Statistics,* R. Viertl, Ed., Plenum Press, 97–110.

Connor, G. J., and F. Woodcock, 2000: The application of synoptic stratification to precipitation forecasting in the trade wind regime. *Wea. Forecasting,* **15,** 276–297.

Cox, D. R., and E. J. Snell, 1989: *Analysis of Binary Data.* 2d ed. Chapman and Hall, 236 pp.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plan. Manage.,* **111,** 157–170.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2002: Verification of probabilistic streamflow forecasts. IIHR Rep. 427, IIHR—Hydroscience and Engineering, Iowa City, IA.

Hosmer, D. W., Jr., and S. Lemeshow, 1989: *Applied Logistic Regression.* John Wiley and Sons, 307 pp.

Kane, T. L., and B. G. Brown, 2000: Confidence intervals for some verification measures—A survey of several methods. Preprints, *15th Conf. on Probability and Statistics in the Atmospheric Sciences,* Asheville, NC, Amer. Meteor. Soc., 46–49.

Katz, R. W., A. H. Murphy, and R. L. Winkler, 1982: Assessing the value of frost forecast to orchardists: A dynamic decision-making approach. *J. Appl. Meteor.,* **21,** 518–531.

Krzysztofowicz, R., and D. Long, 1991: Beta likelihood models of probabilistic forecasts. *Int. J. Forecasting,* **7,** 47–55.

Murphy, A. H., 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.

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

——, and ——, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7,** 435–455.

——, and D. S. Wilks, 1998: A case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Wea. Forecasting,* **13,** 795–810.

Nottingham, Q. J., J. B. Birch, and B. A. Bodt, 2000: Local logistic regression: An application to Army penetration data. *J. Stat. Comput. Simulat.,* **66** (1), 35–50.

Rajagopalan, B., and U. Lall, 1995: A kernel estimator for discrete distributions. *J. Nonparametric Stat.,* **4,** 409–426.

Scott, D. W., 1992: *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, 317 pp.

Seaman, R., I. Mason, and F. Woodcock, 1996: Confidence intervals for some performance measures of yes–no forecasts. *Aust. Meteor. Mag.,* **45,** 49–53.

Smith, J. A., G. N. Day, and M. D. Kane, 1992: Nonparametric

framework for long-range streamflow forecasting. *J. Water Resour. Plan. Manage.*, **118,** 82–92.
Stephenson, D. B., 2000: Use of the ''odds ratio'' for diagnosing forecast skill. *Wea. Forecasting*, **15,** 221–232.
Thornes, J. E., and D. B. Stephenson, 2001: How to judge the quality and value of weather forecast products. *Meteor. Appl.*, **8,** 307–314.
Titterington, D. M., 1980: A comparative study of kernel-based density estimates for categorical data. *Technometrics*, **22,** 259–268.
Wang, M.-C., and J. Van Ryzin, 1981: A class of smooth estimators for discrete distribution. *Biometrika*, **68,** 301–309.
Wilks, D. S., 1991: Representing serial correlation of meteorological events and forecasts in dynamic decision analytic model. *Mon. Wea. Rev.*, **119,** 1640–1662.
——, 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.
——, 2000: Diagnostic verification of the Climate Prediction Center Long-Lead Outlooks, 1995–98. *J. Climate*, **13,** 2389–2403.
——, 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.
Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.*, **104,** 1209–1214.
Zhang, S., R. J. Karunamuni, and M. C. Jones, 1999: An improved estimator of the density function at the boundary. *J. Amer. Stat. Assoc.*, **94,** 1231–1241.