Sampling Uncertainty and Confidence Intervals for the Brier Score and Brier Skill Score

A. ALLEN BRADLEY

IIHR-Hydroscience and Engineering, The University of Iowa, Iowa City, Iowa

STUART S. SCHWARTZ

Center for Urban Environmental Research and Education, University of Maryland, Baltimore County, Baltimore, Maryland

TEMPEI HASHINO

Department of Atmospheric and Ocean Sciences, University of Wisconsin—Madison, Madison, Wisconsin

(Manuscript received 4 June 2007, in final form 5 December 2007)

ABSTRACT

For probability forecasts, the Brier score and Brier skill score are commonly used verification measures of forecast accuracy and skill. Using sampling theory, analytical expressions are derived to estimate their sampling uncertainties. The Brier score is an unbiased estimator of the accuracy, and an exact expression defines its sampling variance. The Brier skill score (with climatology as a reference forecast) is a biased estimator, and approximations are needed to estimate its bias and sampling variance. The uncertainty estimators depend only on the moments of the forecasts and observations, so it is easy to routinely compute them at the same time as the Brier score and skill score. The resulting uncertainty estimates can be used to construct error bars or confidence intervals for the verification measures, or perform hypothesis testing.

Monte Carlo experiments using synthetic forecasting examples illustrate the performance of the expressions. In general, the estimates provide very reliable information on uncertainty. However, the quality of an estimate depends on both the sample size and the occurrence frequency of the forecast event. The examples also illustrate that with infrequently occurring events, verification sample sizes of a few hundred forecast–observation pairs are needed to establish that a forecast is skillful because of the large uncertainties that exist.

1. Introduction

Forecast verification provides critical information on the quality of forecasts and can help guide forecasters in their efforts to improve forecasting systems (Jolliffe and Stephenson 2003; Welles et al. 2007). Although forecasters routinely report verification results, they often do not consider the inherent uncertainty of the verification measures (Jolliffe 2007). However, computing verification measures requires a *sample* of forecasts and their corresponding observations. The resulting verification measures are therefore *sample estimates*. The sampling uncertainty of verification measures depends on both the sample size and the statistical characteristics of the forecasts and observations. In situations where the sample size is relatively small, the sampling uncertainty can be significant (Bradley et al. 2003).

One approach to assessing the sampling uncertainty of verification measures is a resampling method (Efron 1981; Wilks 2006). Forecast–observation pairs are randomly selected from the verification dataset to construct synthetic data samples, and the verification measures are computed; the process is repeated many times to derive an empirical probability distribution for the sample measures. Examples of a resampling approach include Mason and Mimmack (1992), Wilks (1996), Hamill (1999), Zhang and Casey (2000), Doblas-Reyes et al. (2003), Accadia et al. (2003), Ebert et al. (2004), Accadia et al. (2005), Jolliffe (2007), and Ferro (2007), among others. Although the approach is straightfor-

Corresponding author address: A. Allen Bradley, IIHR-Hydroscience and Engineering, The University of Iowa, 107 C. Maxwell Stanley Hydraulics Laboratory, Iowa City, IA 52242.
E-mail: allen-bradley@uiowa.edu

DOI: 10.1175/2007WAF2007049.1

ward and robust, it adds significantly to the computational effort associated with verification.

Another approach is to employ sampling theory to derive analytical expressions for sampling uncertainty. Most examples in the literature are for categorical forecast verification measures, such as the probability of detection or hit rate (Kane and Brown 2000; Jolliffe 2007), odds ratio (Stephenson 2000), false alarm rate and skill scores (Thornes and Stephenson 2001), the correlation coefficient (Jolliffe 2007), and relative operating characteristics (Mason and Graham 2002). For probability forecasts, sampling theory has been applied to evaluate the sampling uncertainty of the bias (Schwartz 1992; Carpenter and Georgakakos 2001) and accuracy measures (Ferro 2007). Sampling theory has also been used to develop confidence intervals (Seaman et al. 1996) and hypothesis tests (Hamill 1999; Briggs and Ruppert 2005; Briggs 2005). Although the assumptions associated with deriving analytical expressions can be limiting, the effort required to evaluate the expressions is minimal.

In this paper, we use sampling theory to develop analytical expressions for the sampling uncertainty of the Brier score and Brier skill score. For probability forecasts, these are the most commonly used measures of accuracy and skill. In the following sections, we derive exact or approximate analytical expressions for the sampling uncertainty of these two measures, and then use Monte Carlo simulation for synthetic forecasting examples to evaluate the expressions and estimated confidence intervals.

2. Forecast verification framework

A probability forecast is one that assigns a probability to the occurrence of a discrete (dichotomous) event. In this case, there are only two possible outcomes: either the forecast event occurs or it does not. A familiar example is a probability-of-precipitation forecast. Probability forecasts can also be constructed from ensemble forecasts (Toth et al. 2003; Zhu 2005); an event is defined as the exceedance (or nonexceedance) of a particular threshold value (e.g., precipitation in excess of a given depth), then the probability forecast for an event occurrence can be made using the ensemble members.

Using a sample of probability forecasts and the resulting (binary) event outcomes, the Brier score and Brier skill score (Brier 1950) are often computed to characterize the accuracy and skill of the forecast. To evaluate the uncertainty of these verification measures from sampling theory, assumptions must be made about the statistical properties of forecasts and observations. The diagnostic framework for forecast verification introduced by Murphy and Winkler (1992), known as the distributions-oriented (DO) approach, explicitly defines a stochastic process commonly assumed for verification. Specifically, forecasts and observations are treated as random variables. Each forecast–observation pair is assumed to be independent of all other pairs, and identically distributed. The relationship between forecasts and observations is defined by their joint distribution.

Using the assumptions of the DO approach and the notation described by Murphy (1997), we will evaluate the sampling uncertainty of the Brier score and Brier skill score. In particular, 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 the 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 $x = 1$). The joint distribution $p(f, x)$ describes the relationship between the forecasts and observations.

a. Distribution moments

The moments of the random variables characterize aspects of the joint distribution. The first moment (mean) of the observations is

$$E[x] = \mu_x, \tag{1}$$

where μ_x is interpreted as the climatological probability of the event occurrence. Because x is a Bernoulli random variable, $x^m = x$ for all positive integers m, so the higher-order noncentral moments are all

$$E[x^m] = \mu_x. \tag{2}$$

Using (1) and (2), the variance of the observations σ_x^2 is

$$\sigma_x^2 = \mu_x(1 - \mu_x). \tag{3}$$

The forecast f may be either a discrete or continuous random variable. Let μ_f be the first moment (mean) of the forecasts. The higher-order noncentral moments are denoted as

$$E[f^m] = \mu'_{(m)f}. \tag{4}$$

One can also characterize the moments of f conditioned on the observation. For instance, the first conditional moments are denoted as

$$E[f|x = 0] = \mu_{f|x=0} \quad \text{and} \tag{5}$$

$$E[f|x = 1] = \mu_{f|x=1}. \tag{6}$$

Higher-order noncentral conditional moments can also be defined, and are denoted for the mth moment as $\mu'_{(m)f|x=0}$ and $\mu'_{(m)f|x=1}$.

The moments of the products of forecasts and observations may also be defined. For positive integers m and n,

$$E[f^{(m)}x^{(n)}]= \mu_x\mu'_{(m)f|x=1}. \tag{7}$$

Appendix A summarizes how to compute estimates of these moments using a verification data sample.

b. *Accuracy measure*

The *accuracy* is an attribute describing the closeness of the forecasts f to the observations x (Murphy 1997). A commonly used measure of the accuracy of the forecasts is the mean squared error:

$$\mathrm{MSE}(f, x) = E[(f - x)^2]. \tag{8}$$

For probability forecasts of a dichotomous event, MSE can be defined in terms of moments of f and x (Bradley et al. 2003). By expanding the terms in Eq. (8),

$$\begin{aligned}\mathrm{MSE}(f, x) &= E[f^2] - 2E[xf] + E[x^2]\\ &= \mu'_{(2)f} - 2\mu_x\mu_{f|x=1} + \mu_x\\ &= \mu'_{(2)f} + \mu_x(1 - 2\mu_{f|x=1}).\end{aligned} \tag{9}$$

c. *Skill measure*

The accuracy of the forecasts relative to a reference forecasting system is known as the *skill*. Let MSE(r, x) donate the accuracy for the reference forecasts r. The mean squared error skill score $\mathrm{SS_{MSE}}$ is then

$$\mathrm{SS_{MSE}}(r, f, x) = 1 - \frac{\mathrm{MSE}(f, x)}{\mathrm{MSE}(r, x)}. \tag{10}$$

Using a climatology forecast μ_x as the reference forecast,

$$\mathrm{MSE}(\mu_x, x) = E[(x - \mu_x)^2] = \sigma_x^2. \tag{11}$$

Therefore, $\mathrm{SS_{MSE}}$ using climatology as a reference forecast is

$$\mathrm{SS_{MSE}}(\mu_x, f, x) = 1 - \frac{\mathrm{MSE}(f, x)}{\sigma_x^2}. \tag{12}$$

3. Sampling uncertainty

Verification measures of accuracy and skill are computed for a forecasting system from a *sample* of forecasts and observations. Let x_i be the observation at time i. Let f_i be the probability forecast of the event at time i. Assume that the forecast–observation pairs $\{f_i, x_i, i = 1, \ldots, N\}$ are a random sample drawn from the joint distribution $p(f, x)$. The verification data sample will be used to *estimate* forecast accuracy and skill. In the following sections, we define sample estimators for the measures of forecast accuracy and skill, and derive expressions for their sampling uncertainty.

a. *Brier score*

The sample estimator for MSE is

$$\widehat{\mathrm{MSE}}\,(f, x) = \frac{1}{N}\sum_{i}^{N}(f_i - x_i)^2, \tag{13}$$

where the "hat" is used to indicate a sample estimate. For probability forecasts, this estimator is referred to as the Brier (or half-Brier) score (Brier 1950).

The expected value of $\widehat{\mathrm{MSE}}$ is

$$\begin{aligned}E[\widehat{\mathrm{MSE}}(f, x)] &= \frac{1}{N}\sum_{i}^{N}E[(f_i - x_i)^2]\\ &= \frac{1}{N}\sum_{i}^{N}E[(f_i^2 - 2x_if_i + x_i^2)].\end{aligned} \tag{14}$$

Because $x_i^2 = x_i$ for a Bernoulli random variable, Eq. (14) reduces to

$$\begin{aligned}E[\widehat{\mathrm{MSE}}\,(f, x)] &= \frac{1}{N}\sum_{i}^{N}(E[f_i^2] - 2E[x_if_i] + E[x_i])\\ &= \mu'_{(2)f} - 2\mu_x\mu_{f|x=1} + \mu_x\\ &= \mu'_{(2)f} + \mu_x(1 - 2\mu_{f|x=1}).\end{aligned} \tag{15}$$

Because $E[\widehat{\mathrm{MSE}}]$ is equal to MSE (see section 2b), the Brier score $\widehat{\mathrm{MSE}}$ is an unbiased estimator of MSE.

b. *Sampling variance of the Brier score*

The variance of a sample estimator is a measure of its sampling uncertainty. The variance of the Brier score $\widehat{\mathrm{MSE}}$ is

$$\begin{aligned}V[\widehat{\mathrm{MSE}}\,(f, x)] &= V\left[\frac{1}{N}\sum_{i}^{N}(f_i - x_i)^2\right]\\ &= \frac{1}{N^2}\sum_{i}^{N}V[(f_i - x_i)^2]\\ &= \frac{1}{N}V[(f - x)^2].\end{aligned} \tag{16}$$

The variance term on the right-hand side can be expanded as

$$\begin{aligned}V[(f - x)^2] &= E[(f - x)^4] - E[(f - x)^2]^2\\ &= E[(f - x)^4] - \mathrm{MSE}(f, x)^2.\end{aligned} \tag{17}$$

Using the binomial expansion, the first term is

$$E[(f-x)^4] = E[f^4] - 4E[f^3x] + 6E[f^2x^2] - 4E[fx^3] + E[x^4]$$
$$= \mu'_{(4)f} - \mu_x[4\mu'_{(3)f|x=1} - 6\mu'_{(2)f|x=1} + 4\mu_{f|x=1}] + \mu_x. \tag{18}$$

Combining all of the terms,

$$V[\widehat{\mathrm{MSE}}\,(f,x)] = \frac{1}{N}\{\mu'_{(4)f} + \mu_x[1 - 4\mu'_{(3)f|x=1} + 6\mu'_{(2)f|x=1} - 4\mu_{f|x=1}] - [\mathrm{MSE}(f,x)]^2\}. \tag{19}$$

The standard error of a sample estimator is also used to describe its sampling uncertainty. The standard error is equivalent to the standard deviation of the estimator. Therefore, the standard error for the Brier score $\widehat{\mathrm{MSE}}$ is defined as

$$\sigma_{\widehat{\mathrm{MSE}}} = \sqrt{V[\widehat{\mathrm{MSE}}\,(f,x)]}. \tag{20}$$

c. Brier skill score

The sample estimator for $\mathrm{SS}_{\mathrm{MSE}}$, where climatology is used as a reference forecast, is

$$\widehat{\mathrm{SS}}_{\mathrm{MSE}}\ \ (\mu_x, f, x) = 1 - \frac{\widehat{\mathrm{MSE}}\,(f,x)}{\hat{\sigma}_x^2}. \tag{21}$$

For probability forecasts, this estimator is known as the Brier skill score.

Because it involves a ratio of sample estimators, the Brier skill score is a biased estimator of $\mathrm{SS}_{\mathrm{MSE}}$ (Mason 2004). Using a Taylor series expansion, a second-order approximation of the bias $B(\widehat{\mathrm{SS}}_{\mathrm{MSE}})$ is (Benjamin and Cornell 1970)

$$\tilde{B}[\widehat{\mathrm{SS}}_{\mathrm{MSE}}(\mu_x, f, x)] = c_1V[\widehat{\mathrm{MSE}}\,(f,x)] + c_2V[\hat{\sigma}_x^2] + c_3\,\mathrm{cov}[\widehat{\mathrm{MSE}}\,(f,x), \hat{\sigma}_x^2], \tag{22}$$

where

$$c_1 = \frac{1}{2}\cdot\frac{\partial^2\widehat{\mathrm{SS}}_{\mathrm{MSE}}}{\partial\widehat{\mathrm{MSE}}^2}\bigg|_{\bar{\mu}} = 0, \tag{23}$$

$$c_2 = \frac{1}{2}\cdot\frac{\partial^2\widehat{\mathrm{SS}}_{\mathrm{MSE}}}{\partial(\hat{\sigma}_x^2)^2}\bigg|_{\bar{\mu}} = -\frac{E[\widehat{\mathrm{MSE}}\,(f,x)]}{E[\hat{\sigma}_x^2]^3}, \quad \text{and} \tag{24}$$

$$c_3 = \frac{\partial^2\widehat{\mathrm{SS}}_{\mathrm{MSE}}}{\partial\widehat{\mathrm{MSE}}\,\partial\hat{\sigma}_x^2}\bigg|_{\bar{\mu}} = \frac{1}{E[\hat{\sigma}_x^2]^2}. \tag{25}$$

As shown in section 3a, $\widehat{\mathrm{MSE}}$ is an unbiased estimator of MSE. However, for a Bernoulli random variable with parameter μ_x, the sample estimator of σ_x^2 is biased (see appendix B):

$$E[\hat{\sigma}_x^2] = \frac{N-1}{N}\sigma_x^2. \tag{26}$$

By substitution and simplification,

$$c_2 = -\frac{1 - \mathrm{SS}_{\mathrm{MSE}}(\mu_x, f, x)}{\sigma_x^4}\left(\frac{N}{N-1}\right)^3 \quad \text{and} \tag{27}$$

$$c_3 = \frac{1}{\sigma_x^4}\left(\frac{N}{N-1}\right)^2. \tag{28}$$

For a Bernoulli random variable, the variance of the sample estimator $\hat{\sigma}_x^2$ is (Kenney and Keeping 1951)

$$V[\hat{\sigma}_x^2] = \frac{(N-1)}{N^3}[(N-1) + \sigma_x^2(6-4N)]\sigma_x^2. \tag{29}$$

Using the result shown in appendix B, the covariance term is

$$\mathrm{cov}[\widehat{\mathrm{MSE}}\,(f,x), \hat{\sigma}_x^2] = \left(\frac{N-1}{N^2}\right)\sigma_x^2(1-2\mu_x)\cdot\{[\mu'_{(2)f|x=1} - \mu'_{(2)f|x=0}] + (1 - 2\mu_{f|x=1})\}. \tag{30}$$

d. Sampling variance of the Brier skill score

The variance of the Brier skill score $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$ is

$$V[\widehat{\mathrm{SS}}_{\mathrm{MSE}}(\mu_x, f, x)] = V\left[1 - \frac{\widehat{\mathrm{MSE}}\,(f,x)}{\hat{\sigma}_x^2}\right]. \tag{31}$$

Here, we employ a first-order approximation (Benjamin and Cornell 1970) to estimate the variance of $\mathrm{SS}_{\mathrm{MSE}}$:

$$\tilde{V}[\widehat{\mathrm{SS}}_{\mathrm{MSE}}(\mu_x, f, x)] = d_1V[\widehat{\mathrm{MSE}}\,(f,x)] + d_2V[\hat{\sigma}_x^2] + d_3\,\mathrm{cov}[\widehat{\mathrm{MSE}}\,(f,x), \hat{\sigma}_x^2], \tag{32}$$

where

$$d_1 = \frac{\partial\widehat{\mathrm{SS}}_{\mathrm{MSE}}}{\partial\widehat{\mathrm{MSE}}}\bigg|_{\bar{\mu}}^2 = \frac{1}{\sigma_x^4}\left(\frac{N}{N-1}\right)^2, \tag{33}$$

$$d_2 = \frac{\partial\widehat{\mathrm{SS}}_{\mathrm{MSE}}}{\partial\hat{\sigma}_x^2}\bigg|_{\bar{\mu}}^2 = \frac{[1 - \mathrm{SS}_{\mathrm{MSE}}(\mu, f, x)]^2}{\sigma_x^4}\left(\frac{N}{N-1}\right)^4, \tag{34}$$

and

$$d_3 = 2 \cdot \frac{\partial \widehat{\mathrm{SS}}_{\mathrm{MSE}}}{\partial \widehat{\mathrm{MSE}}}\bigg|_{\overline{\mu}} \cdot \frac{\partial \widehat{\mathrm{SS}}_{\mathrm{MSE}}}{\partial \hat{\sigma}_x^2}\bigg|_{\overline{\mu}} = -\frac{2\,[1 - \mathrm{SS}_{\mathrm{MSE}}(\mu_x, f, x)]}{\sigma_x^4}\left(\frac{N}{N-1}\right)^3, \tag{35}$$

The standard error for the Brier skill score $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$ is approximated by

$$\sigma_{\widehat{\mathrm{SS}}_{\mathrm{MSE}}} = \sqrt{\tilde{V}[\widehat{\mathrm{SS}}_{\mathrm{MSE}}(\mu_x, f, x)]}. \tag{36}$$

4. Forecasting examples

To evaluate the analytical expressions for the sampling uncertainties, we used Monte Carlo simulations to generate verification datasets for several synthetic forecasting examples. For each example, the joint distribution $p(f, x)$ and the true values of the accuracy and skill measures are known. Ten thousand verification data samples are generated for each example, for sample sizes ranging from 50 to 1000 pairs. For each verification data sample, sample estimates of the accuracy and skill (the Brier score and Brier skill score) are computed. To determine the "true" standard error of the measures for comparison with the analytical expressions, we compute the standard deviation of the measures for the 10 000 verification data samples. Note that with a Monte Carlo sample of this size, the uncertainty of the true value is less than about 1.5% (at a 95% confidence level).

A stochastic model based on the calibration–refinement factorization of the joint distribution $p(f, x)$ (Murphy 1997) was used to construct the synthetic forecasting examples for the Monte Carlo simulations. Forecast–observation pairs were created from the stochastic model as follows. First, the probability forecast f_i for the discrete event is randomly generated. The marginal distribution of the forecasts $s(f)$ is modeled by a beta distribution (Krzysztofowicz and Long 1991):

$$s(f) = \frac{(1-f)^{\upsilon-1} f^{\omega-1}}{B(\upsilon, \omega)}, \tag{37}$$

where υ and ω are parameters, and $B(\upsilon, \omega)$ is the beta function. Next, the conditional expected value of the observation $\mu_{x|f}$, given the forecast f, is evaluated. A linear model (Murphy and Wilks 1998) is used:

$$\mu_{x|f} = a + bf, \tag{38}$$

where a and b are parameters. Finally, a corresponding observation x_i is generated for the given forecast f_i. This is accomplished by generating a uniform random variable over the range from 0 to 1; if the value is less than $\mu_{x|f_i}$, then observation x_i is set to 1 (event occurrence). Otherwise, the observation is set to 0 (event nonoccurrence).

We use the results for selected examples to illustrate the properties of the uncertainty estimators. In particular, we show results for forecasts of a rarely occurring event ($\mu_x = 0.05$) and for a more commonly occurring event ($\mu_x = 0.25$). For both cases, we generated forecasts with low ($\mathrm{SS}_{\mathrm{MSE}} = 0.2$), medium ($\mathrm{SS}_{\mathrm{MSE}} = 0.4$), and high skill ($\mathrm{SS}_{\mathrm{MSE}} = 0.6$). In all the cases, the forecast is assumed to be unconditionally unbiased ($\mu_f = \mu_x$), which imposes the following constraint on the linear model parameters:

$$a = \mu_x(1 - b). \tag{39}$$

Note that if $b = 1$, the forecast is also conditionally unbiased; if $b \neq 1$, the forecast has conditional biases. We examined both of these cases. For the conditionally unbiased forecast case,

$$\mu_{x|f} = f, \tag{40}$$

which implies that the forecast is perfectly reliable. For cases with a conditionally biased forecast, we selected $b = 0.8$ and used Eq. (39) to find a.

Table 1 shows the model parameters for the cases presented in the following section. Also shown are relative measures of forecast quality (Murphy 1997) for each case. By design, the relative measures are the same for a given skill level within the unbiased or the conditionally biased forecast cases. By definition, the reliability (REL) of the forecast is 0 for a perfectly reliable forecast; the reliability is nonzero for the cases with conditional biases. Also, the resolution (RES), discrimination (DIS), and sharpness (SHP) of the forecast are all higher for a more skillful forecast. Finally, in the case of a conditionally biased forecast with $b = 0.8$, the highest achievable skill score is 0.6. For this high-skill case, the forecast must be perfectly sharp (the probability forecast f is either 0 or 1). Hence, the relative sharpness (SHP/σ_x^2) for the high-skill case with conditionally biased forecasts achieves its maximum value of 1. Note that for this special case, the forecast f is a discrete random variable and is generated from a Bernoulli distribution with parameter μ_f.

5. Results

a. Standard errors

First we evaluate the analytical expressions for the standard errors of the Brier score and Brier skill score.

Table 1. Relative measures of forecast quality for Monte Carlo simulations of forecast–observation pairs. The parameters of the stochastic model (defined in section 4) are υ, ω, and b. The forecast quality measures are skill (SS_{MSE}), resolution (RES), reliability (REL), discrimination (DIS), type-2 conditional bias (B_2), and sharpness (SHP) (Murphy 1997).

Case	υ	ω	b	SS_{MSE}	RES/σ_x^2	REL/σ_x^2	DIS/σ_x^2	B_2/σ_x^2	SHP/σ_x^2
Rare event (μ_x = 0.05): no bias									
Low	0.2	3.8	1	0.200	0.200	0.000	0.040	0.640	0.200
Medium	0.075	1.425	1	0.400	0.400	0.000	0.160	0.360	0.400
High	0.0333	0.6333	1	0.600	0.600	0.000	0.360	0.160	0.600
Common event (μ_x = 0.25)*:* no bias									
Low	1	3	1	0.200	0.200	0.000	0.040	0.640	0.200
Medium	0.375	1.125	1	0.400	0.400	0.000	0.160	0.360	0.400
High	0.1667	0.5	1	0.600	0.600	0.000	0.360	0.160	0.600
Rare event (μ_x = 0.05): conditional bias									
Low	0.1	1.9	0.8	0.200	0.213	0.013	0.071	0.538	0.333
Medium	0.025	0.475	0.8	0.400	0.427	0.027	0.284	0.218	0.667
High	—	—	0.8	0.600	0.640	0.040	0.640	0.040	1.000
Common event (μ_x = 0.25): conditional bias									
Low	0.5	1.5	0.8	0.200	0.213	0.013	0.071	0.538	0.333
Medium	0.125	0.375	0.8	0.400	0.427	0.027	0.284	0.218	0.667
High	—	—	0.8	0.600	0.640	0.040	0.640	0.040	1.000

True values for the sampling uncertainty are represented by the 10 000-sample estimates from the Monte Carlo simulation. These are compared to the analytical expression estimates made using the known (true) values for the moments.

Figure 1 shows the standard error estimates for the Brier score $\widehat{\mathrm{MSE}}$ for medium-skill ($SS_{MSE} = 0.4$) forecasts. Results are shown for an unbiased and a conditionally biased forecast, for both rare- ($\mu_x = 0.05$) and common-event ($\mu_x = 0.25$) occurrences. Note that the analytical expression is exact for the sampling variance; not surprisingly, the standard error estimates are virtually identical to the true values in all cases. For a forecast with conditional biases, the standard errors are slightly higher than those for a forecast with no conditional biases. Furthermore, the standard errors for $\widehat{\mathrm{MSE}}$ for forecasts of common-event occurrences are higher than those for rare-event occurrences. This occurs because the magnitude of *MSE* tends to increase as the inherent uncertainty σ_x^2 increases; for probability forecasts, σ_x^2 is small as μ_x approaches 0 (or 1), and reaches its maximum for $\mu_x = 0.5$ [see Eq. (3)]. Therefore, the absolute magnitudes of the standard error for $\widehat{\mathrm{MSE}}$ are not directly comparable for rare- and common-event cases, because they depend on the event occurrence frequency.

For the Brier skill score $\widehat{SS}_{MSE}$, the analytical expression for its sampling variance is a first-order approximation. The quality of the approximation for the standard error is illustrated in Fig. 2 for unbiased forecasts with low skill ($SS_{MSE} = 0.2$) and high skill ($SS_{MSE} = 0.6$). For large sample sizes, the standard error estimates are very good in all cases. However, significant underestimation of the standard error occurs for forecasts of the rare event for small sample sizes. Unlike MSE, SS_{MSE} is directly comparable for rare- and common-event cases because it is nondimensionalized by the variance of the observations. For the forecasts of the rare event, there are fewer occurrences of the forecast event, resulting in relatively high sampling variance. Still, the deviations of the standard error approximation for these cases are evident only with small sample sizes (200 or less), when there are very few

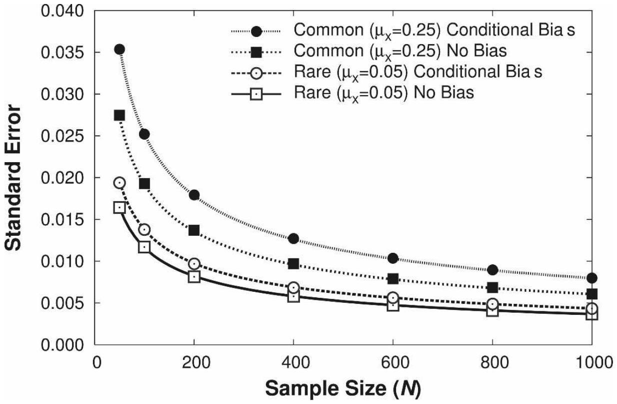

Fig. 1. Standard error of the Brier score for a medium-skill forecast ($SS_{MSE} = 0.4$). The symbols are the true standard errors derived from the Monte Carlo simulations. The curves are the estimates based on the analytical expression [Eq. (20)]. Results are shown for rare-event ($\mu_x = 0.05$) and common-event ($\mu_x = 0.25$) occurrences, for both an unbiased and a conditionally biased forecast.

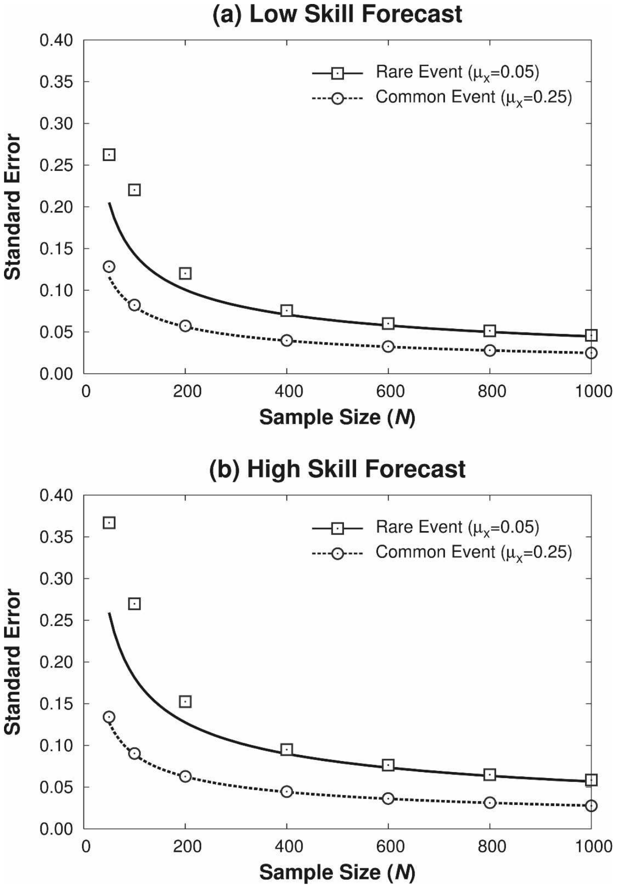

FIG. 2. Standard error of the Brier skill score for (a) a low-skill forecast ($SS_{MSE} = 0.2$) and (b) a high-skill forecast ($SS_{MSE} = 0.6$). The symbols are the true standard errors derived from the Monte Carlo simulations. The curves are the estimates based on the analytical expression [Eq. (36)]. Results are shown for rare-event ($\mu_x = 0.05$) and common-event ($\mu_x = 0.25$) occurrences, for an unbiased forecast.

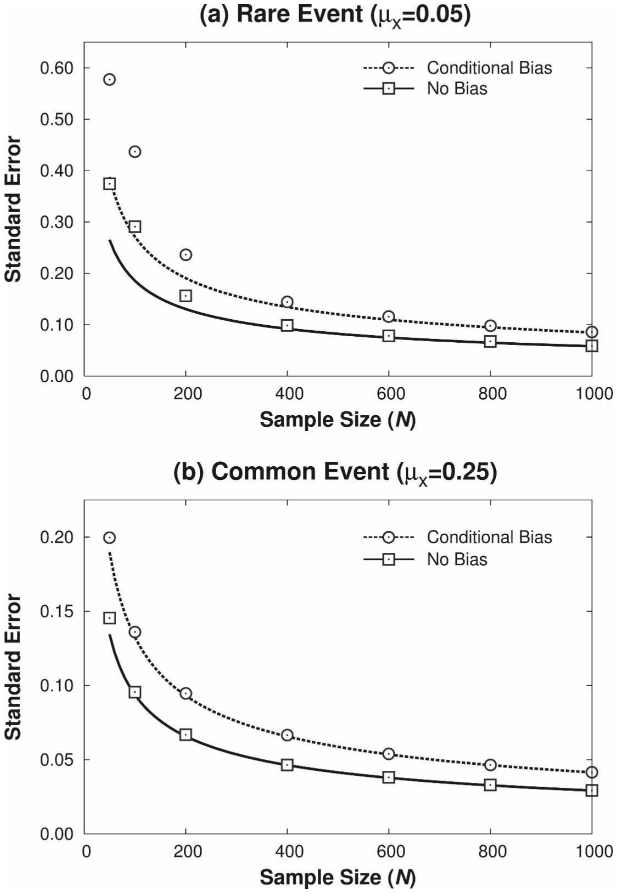

FIG. 3. Standard error of the Brier skill score for a medium-skill forecast ($SS_{MSE} = 0.4$) for (a) a rare event ($\mu_x = 0.05$) and (b) a common event ($\mu_x = 0.25$). The symbols are the true standard errors derived from the Monte Carlo simulations. The curves are the estimates based on the analytical expression [Eq. (36)]. Results are shown for both an unbiased and a conditionally biased forecast.

rare-event occurrences in the sample (e.g., 10 or less). Note too that the standard errors for both the rare- and common-event cases increase slightly with increasing skill (see Fig. 2), but the increase is proportionally smaller than the increase in the skill itself. For the high-skill forecast, SS_{MSE} is larger than the standard error for all sample sizes, whereas for the low-skill forecast, the standard error for the rare-event case exceeds SS_{MSE} for very small sample sizes.

The impact of conditional forecast biases is illustrated in Fig. 3 for the case of medium-skill (SS_{MSE} = 0.4) forecasts. As is the case for $\widehat{\mathrm{MSE}}$, the standard errors for $\widehat{SS}_{MSE}$ are higher when the forecast is conditionally biased. Approximate standard errors are nearly exact for common-event forecasts, even with conditional biases (see Fig. 3b). However, the same deviation in the approximation occurs for the rare-event case when conditional forecast biases are present (see Fig. 3a).

b. *Bias in the Brier skill score*

Although the sample estimator of MSE is unbiased, the sample estimator of SS_{MSE} is biased. Figure 4 shows the expected skill score $E[\widehat{SS}_{MSE}]$ for medium-skill forecasts based on the Monte Carlo simulation, and that evaluated with the second-order approximation of the bias shown in Eq. (22). Clearly, biases are largest for the forecasts of the rare event; the biases are only slightly larger when the forecast is conditionally biased. The magnitude of the bias can be quite large. For rare-event occurrences with a true SS_{MSE} of 0.4, the expected skill score is 0.34 or less for a sample of fewer than a hundred forecast–observation pairs. Although not shown, the magnitude of the bias does not vary

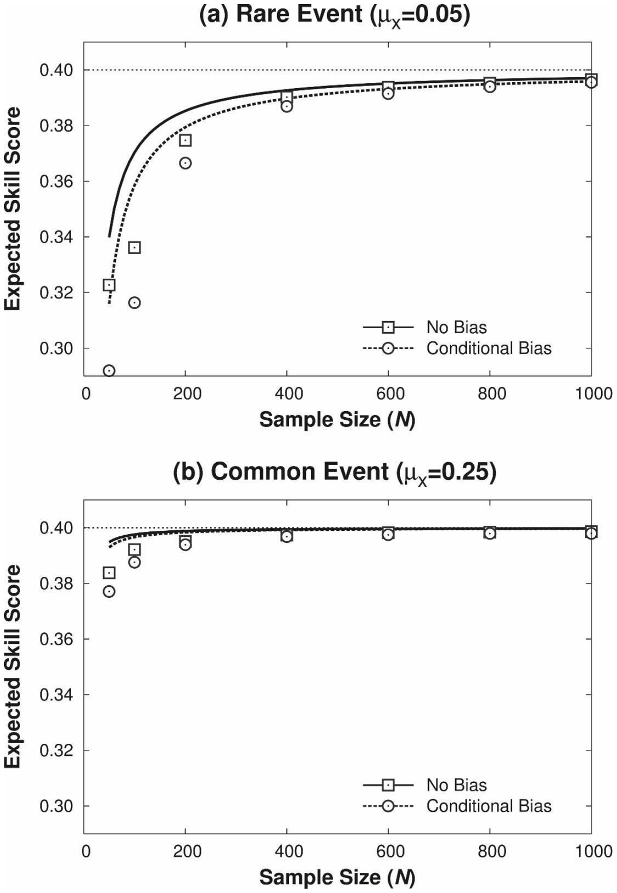

FIG. 4. Bias of the Brier skill score for a medium-skill forecast ($\mathrm{SS}_{\mathrm{MSE}} = 0.4$) for (a) a rare event ($\mu_x = 0.05$) and (b) a common event ($\mu_x = 0.25$). The symbols are the true expected values of the Brier skill score derived from the Monte Carlo simulations. The curves are the estimates made using the analytical expression for bias [Eq. (22)]. The expected values are shown for both an unbiased and a conditionally biased forecast.

greatly with the magnitude of the forecast skill in these forecasting examples.

The second-order approximation of the biases produces a reasonable prediction of the bias for rare-event occurrences (see Fig. 4a). The curves follow the overall trend found in the Monte Carlo experiments, but the magnitude of the bias is underestimated. When the actual bias is less, as is the case for common-event occurrences (see Fig. 4b), the approximation is poor. In particular, the magnitudes of the observed biases are much larger for small sample sizes than predicted for the common-event occurrences.

c. Approximate confidence intervals

In practice, the analytical expressions would be used to estimate the uncertainty of the Brier score and Brier skill score computed with a verification dataset. Uncertainty estimates are obtained by replacing the moments in the standard error and bias equations in section 3 with moment estimates from a sample (see appendix A for traditional sample estimators of the moments). Note that for the standard error of the Brier score, this approach is mathematically similar to the sample estimator proposed by Ferro (2007) for ensemble forecasts. Using the resulting sample estimates of uncertainty, one can construct error bars, perform hypothesis testing, or construct confidence intervals for individual sample estimates. To evaluate the performance of the uncertainty expressions when sample moments are used, we examine confidence interval estimates for both MSE and $\mathrm{SS}_{\mathrm{MSE}}$ for verification data samples.

In principle, confidence interval estimation requires knowledge of the sampling distributions for $\widehat{\mathrm{MSE}}$ and $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$. However, for sufficiently large sample sizes, the central limit theorem states that the sampling distribution of a sum of independent and identically distributed random variables will converge to a normal distribution. Therefore, a normal distribution approximation is often used to construct confidence intervals. For $\widehat{\mathrm{MSE}}$, which involves the sum of the forecast errors squared, it is reasonable to assume that its sampling distribution will converge to a normal distribution for large sample sizes. However, the situation is not as straightforward for $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$, which essentially involves the ratio of two sums.

Figure 5 shows the empirical sampling distribution for $\widehat{\mathrm{MSE}}$ from the Monte Carlo simulation for medium-skill forecasts with no bias. For rare-event occurrences (see Fig. 5a), the lower bound at 0 causes deviations from a normal distribution for small sample sizes. Otherwise, the normal distribution approximation is reasonable for the sampling distribution. For common-event occurrences (see Fig. 5b), the normal distribution approximation is reasonable, except perhaps in the tails of the distribution.

Figure 6 shows the sampling distribution for $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$ for the same medium-skill forecasts. For rare-event occurrences (see Fig. 6a), the sampling distributions are negatively skewed due to the upper bound at 1, and the skewness is significant for most sample sizes shown. Therefore, a normal distribution is a poor approximation for the observed sampling distribution for $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$. For common-event occurrences (see Fig. 6b), the skewness of the sampling distribution is not as severe. For sample sizes of 400 or greater, a normal distribution approximation is not unreasonable.

Despite the deviations observed, we examined the characteristics of confidence intervals constructed for

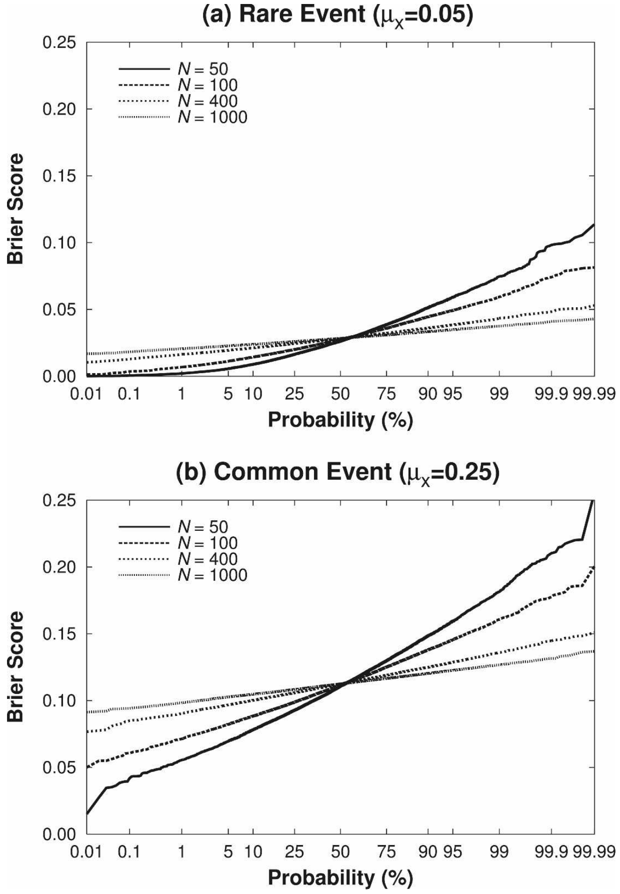

FIG. 5. Sampling distribution of the Brier score for an unbiased forecast with medium skill (SS_{MSE} = 0.4) for (a) a rare event (μ_x = 0.05) and (b) a common event (μ_x = 0.25). The empirical distribution is based on the 10 000-sample estimates from the Monte Carlo simulation. A normal distribution would plot as a straight line.

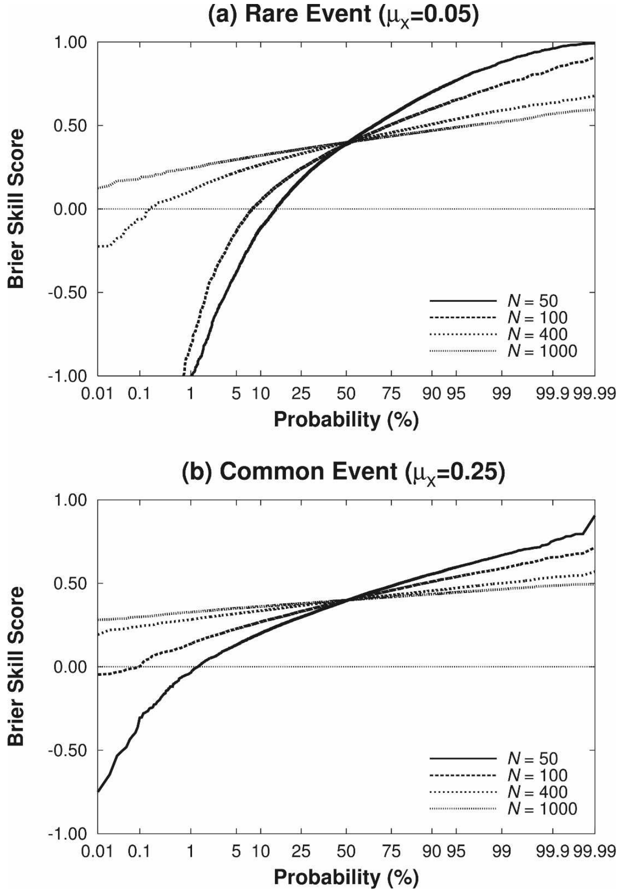

FIG. 6. Sampling distribution of the Brier skill score for an unbiased forecast with medium skill (SS_{MSE} = 0.4) for (a) a rare event (μ_x = 0.05) and (b) a common event (μ_x = 0.25). The empirical distribution is based on the 10 000-sample estimates from the Monte Carlo simulation. A normal distribution would plot as a straight line.

individual samples using a normal approximation. For the case where a sample statistic $\hat{Y}$ is normally distributed with unknown variance, the 100(1 − α)% confidence interval for the true parameter μ_Y is

$$\hat{Y} \pm t_{\alpha/2,N-1}\hat{\sigma}_Y, \tag{41}$$

where $t_{\alpha/2,N-1}$ is the critical value of the t distribution for a two-sided test with a significance level of α and $N - 1$ degrees of freedom, and $\hat{\sigma}_Y$ is the sample estimate of the standard error of the test statistic. For all forecast cases, we constructed confidence limits for individual verification datasets. In particular, we computed the sample moments (see appendix A) for each dataset, and we used the sample estimates in the standard error equations shown in section 3 to obtain standard error estimates for $\widehat{\mathrm{MSE}}$ and $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$ and confidence intervals. Then, the fractions of cases where the true values of MSE and SS_{MSE} fell within the confidence limits were computed. The results are shown for 95% confidence intervals in Table 2 for unbiased forecasts.

For the forecasts of the rare event, the 95% confidence intervals constructed using sample estimates contain the true values much less than 95% of the time for small sample sizes. Given the underestimation of the standard error for $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$ for the forecasts of the rare event, underestimation of the interval is not surprising. However, the variance estimator $\widehat{\mathrm{MSE}}$ is exact, and the underestimation is of a similar magnitude for a given skill level and sample size. Although deviations from the normal distribution approximation may contribute to the underestimation, the dominate cause is the use of sample (rather than the true) moments to estimate the sampling variances of $\widehat{\mathrm{MSE}}$ and $\widehat{\mathrm{SS}}_{\mathrm{MSE}}$. Note that the expressions involve higher-order moments (up to

TABLE 2. Percentage of samples where the true MSE or SS_{MSE} falls within the 95% confidence interval constructed for the sample. Results are shown for forecasting examples without conditional forecast biases.

Skill	Sample size 50	100	200	400	600	800	1000
			MSE rare events (μ_x = 0.05)				
Low	84.5	89.0	92.4	94.0	94.2	94.3	94.3
Medium	81.7	88.6	91.8	93.3	93.9	94.1	94.5
High	76.2	85.0	90.3	92.9	93.4	93.8	93.9
			SS_{MSE} rare events (μ_x = 0.05)				
Low	81.9	92.5	94.3	95.0	95.0	94.9	94.9
Medium	77.3	91.2	93.9	94.0	94.8	94.5	95.0
High	70.8	86.9	91.4	93.5	94.1	94.7	94.4
			MSE common events (μ_x = 0.25)				
Low	93.4	94.6	94.8	94.5	95.1	95.1	95.2
Medium	92.3	93.9	94.4	94.6	94.7	94.7	95.0
High	91.1	93.0	93.8	94.4	94.7	94.9	94.8
			SS_{MSE} common events (μ_x = 0.25)				
Low	95.8	95.7	95.2	95.3	95.2	95.1	95.3
Medium	94.1	94.8	94.8	95.0	94.9	94.8	95.0
High	92.8	94.1	94.5	94.9	95.0	95.1	95.1

fourth order), so sample estimates of these moments have large sampling uncertainty for a small sample size. For the forecasts of the common event, the 95% confidence intervals performed much better, even though slight deviations occur for small sample sizes. For all cases, the confidence interval results are best (closer to 95%) for low skill forecasts. The intervals for $\widehat{MSE}$ are better for the rare-event occurrences, but those for $\widehat{SS}_{MSE}$ are generally better for the common-event occurrences. So, despite the approximation used for the variance of $\widehat{SS}_{MSE}$, estimated confidence intervals for both $\widehat{MSE}$ and $\widehat{SS}_{MSE}$ perform similarly using the normal distribution approximation.

6. Discussion

As the results of the Monte Carlo experiments show, forecasters can reliably use the standard error and confidence limit estimators presented here to characterize uncertainty in many forecasting situations. Still, significant deviations can occur with forecasts made for rarely occurring events for small sample sizes (roughly less than a few hundred forecast–observation pairs). Unfortunately, the desire to characterize sampling uncertainty for rare (high impact) events is usually the greatest when the forecast–observation data sample is small. In these situations, standard error estimates for the Brier skill score, and confidence limits for both the Brier score and Brier skill score, underestimate the actual uncertainty.

On the other hand, the estimated standard errors and confidence limits, although deficient, do indicate the large uncertainties that exists for such cases. As an example, Fig. 7 shows the 95% confidence limits for SS_{MSE} as a function of the sample size for a low-skill forecast of a rare event ($\mu_x = 0.05$), computed using the approximate standard error estimators and the true model parameters. Note that for sample sizes up to about 300, one cannot reject the hypothesis that the true skill score is 0 for this case. For a sample size of 50, the estimated 95% confidence interval for SS_{MSE} covers a range of [−0.26, 0.57]. Even though these estimates understate the true uncertainty range, the interval is still so large that one cannot conclude that the

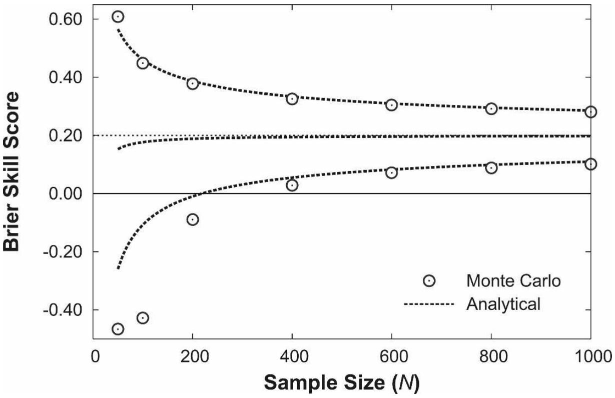

FIG. 7. The 95% confidence intervals and expected Brier skill score for an unbiased forecast with low skill (SS_{MSE} = 0.2) for a rare event (μ_x = 0.05). Lines show the results based on analytical expressions evaluated with the true parameters. Symbols show the empirical confidence intervals from the Monte Carlo simulations.

forecast of the rare event is skillful from such a small sample. Although it may be possible to devise corrections that improve uncertainty estimates for small sample sizes and rare events, the conclusions drawn in such cases would not be substantially different.

A second-order approximation of the bias for the SS_{MSE} estimator was also derived, which could be applied to correct biases for sample estimates. However, the utility of such a correction is debatable. As suggested by Fig. 7, the magnitude of the bias is quite small compared to the sampling variability, especially for large sample sizes. As shown in Fig. 4, the approximation performed best for rare-event occurrences for small sample sizes, where the sampling uncertainty is too large to make statistical inferences. For common-event occurrences, the bias is much smaller, and the approximation underestimates the bias to such an extent that its use for bias correction would not substantially change the results.

Another important consideration in the use of the uncertainty estimators is their limit of applicability. Recall that the assumptions of the DO approach to forecast verification (Murphy 1997) were used to derive the estimators. One assumption is that forecast–observation pairs are identically distributed; a verification dataset is a random sample drawn from this unchanging (or stationary) joint distribution. Yet in many applications, a verification dataset pools together forecasts issued at different times of the year, and for different locations. In this case, the assumption of identically distributed forecast–observation pairs may be violated, because the joint distribution of forecasts and observations may depend on the season and location.

Hamill and Juras (2006) provide some insightful examples of how nonstationarity in the form of temporal and spatial variations in climatology affects verification measures. In particular, if the joint distribution varies with time or location, verification measures will be unreliable (biased) and misleading. Whenever such variations are a concern, Hamill and Juras (2006) recommend computing verification measures with subsamples where stationarity of the joint distribution can be reasonably assumed. Of course, the drawback to subdividing the verification data sample into smaller subsamples (to avoid such biases) is the inherent increase in sampling variability.

If one adopts this approach when dealing with pooled samples, the uncertainty estimators derived here still have applications. Consider the case where a pooled sample of size N is drawn from M distinct groups, each with its own (stationary) joint distribution. Mathematically, the sample Brier score computed with the pooled sample $\widehat{\mathrm{MSE}}_p$ is equivalent to

$$\widehat{\mathrm{MSE}}_p = \sum_{i}^{M} \frac{N_i}{N} \widehat{\mathrm{MSE}}_i, \tag{42}$$

where MSE_i is the sample Brier score for the ith group and N_i is the sample size for the ith group. Using probability theory, and assuming independence of the subsamples, the variance of the pooled Brier score is

$$V[\widehat{\mathrm{MSE}}_p] = \sum_{i}^{M} \frac{N_i^2}{N^2} V[\widehat{\mathrm{MSE}}_i], \tag{43}$$

where $V[\widehat{\mathrm{MSE}}_i]$ is defined by Eq. (19) for each subsample.

In cases where a single skill measure is desired for a pooled sample, Hamill and Juras (2006) recommend using a weighted average of the skill score for (stationary) subsamples, rather than the skill score computed directly from the pooled sample. One suggested weighted-average skill measure is

$$\widehat{\mathrm{SS}}_p = \sum_{i}^{M} \frac{N_i}{N} \widehat{\mathrm{SS}}_i. \tag{44}$$

Again, assuming independence of the subsamples, the standard error of this weighted-average skill score is simply

$$V[\widehat{\mathrm{SS}}_p] = \sum_{i}^{M} \frac{N_i^2}{N^2} V[\widehat{\mathrm{SS}}_i], \tag{45}$$

where $V[\widehat{\mathrm{SS}}_i]$ is defined by Eq. (32) for the stationary subsample.

Another assumption of the DO approach is that each forecast–observation pair is independent of all others. The independence assumption is violated whenever forecasts or observations are significantly correlated in time or space. Temporal correlation for probability forecasts may be a concern for frequently updated event forecasts (e.g., daily probability of precipitation), but sample sizes also tend to be relatively large in this case. On the other hand, the spatial correlation for probability forecasts is often a concern for both meteorological and climate forecasts. If the forecasts or observations are stationary, but significantly correlated, the sample size is effectively reduced, and the estimators presented here would underestimate the true sampling uncertainty.

One clear advantage of the analytical estimators of sampling uncertainty is that they are easy to compute at the same time as verification measures. In contrast, with resampling approaches, a few hundred or thousand resampled datasets must be constructed, and verification measures computed for each case. Clearly, the added computational burden is of no consequence for any one verification dataset. However, the computational effort can be a significant obstacle when evaluating the suite of products for an ensemble forecasting system, where the ensemble forecasts are transformed into probability forecasts for multiple event thresholds (e.g., Bradley et al. 2004), and evaluated at many different locations and forecast times (e.g., Kruger et al. 2007). Still, if the underlying stochastic process for forecasts and observations assumed in the derivation of these estimators is not valid, resampling remains an attractive nonparametric alternative.

7. Summary and conclusions

The Brier score and Brier skill score are commonly used verification measures for probability forecasts. However, their sampling characteristics have not been explored in great detail. Using the joint distribution of forecasts and observations as defined in the distributions-oriented approach to forecast verification (Murphy and Winkler 1992), we derive analytical expressions for the sampling uncertainties of the Brier score and Brier skill score. The expressions depend only on the moments of the forecasts and observations, making sampling uncertainties easy to estimate at the same time as the scores themselves. The expressions can be used with verification datasets to estimate standard errors and biases, construct confidence intervals, or perform hypothesis testing (e.g., Hamill 1999) on the Brier score and Brier skill score.

Monte Carlo experiments using synthetic forecasting examples illustrate the performance of the expressions. In general, the estimates provide very reliable information on uncertainty. However, the quality of an estimate depends on both the sample size and the occurrence frequency of the forecast event. For the Brier skill score, the approximation underestimates the standard error at small sample sizes (a few hundred forecast–observation pairs or less) for infrequently occurring events. In contrast, the bias estimator for the Brier skill score only provides reasonable estimates when the bias is large, a situation that exists only for infrequently occurring events. Confidence interval estimates constructed for individual samples using a normal distribution approximation perform well except at small sample sizes, where the sampling distributions are skewed because of large sampling variances and the lower (upper) bound on the Brier score (Brier skill score).

Although this paper focuses on uncertainty estimation for measures of accuracy and skill, there are additional measures that describe other aspects of the forecast quality for probability forecasts (Murphy 1997; Wilks 2006). Bradley et al. (2003) showed that other distributions-oriented measures for probability forecasts can also be expressed as a function of the moments of the joint distribution. We are currently exploring the use of sampling theory to develop approximate uncertainty estimates for these measures as well.

Acknowledgments. This work was supported in part by National Oceanic and Atmospheric Administration (NOAA) Grant NA04NWS4620015, from the National Weather Service (NWS) Office of Hydrologic Development, and Grant NA16GP1569, from the Office of Global Programs as part of the GEWEX Americas Prediction Project (GAPP). We gratefully acknowledge this support. We also 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 estimator for the mean of the observations is

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

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

$$\hat{\sigma}_x^2 = \hat{\mu}_x(1 - \hat{\mu}_x). \tag{A2}$$

The sample estimators for the mean and higher-order noncentral moment estimators for the forecasts are

$$\hat{\mu}_f = \frac{1}{N}\sum_i^N f_i \quad \text{and} \tag{A3}$$

$$\hat{\mu}'_{(m)f} = \frac{1}{N}\sum_i^N f_i^m. \tag{A4}$$

To estimate the conditional mean and higher-order noncentral moments of the forecasts, 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 $\hat{\mu}_{f|x=0}$ and $\hat{\mu}_{f|x=1}$, and higher-order noncentral moments $\hat{\mu}'_{(m)f|x=0}$ and $\hat{\mu}'_{(m)f|x=1}$ are then estimated using the subsamples in a similar fashion as above.

Note that sample estimates of the accuracy and skill measures shown in sections 2b and 2c, or the bias and standard errors shown in section 3, can be obtained by replacing the moments with the sample moment estimates shown above.

APPENDIX B

Covariance Term

The covariance of the sample estimators of the mean squared error $\widehat{\mathrm{MSE}}$ (Brier score) and the variance of the observations $\hat{\sigma}_x^2$ is

$$\mathrm{cov}[\widehat{\mathrm{MSE}}, \hat{\sigma}_x^2] = E[\widehat{\mathrm{MSE}} \cdot \hat{\sigma}_x^2] - E[\widehat{\mathrm{MSE}}]E[\hat{\sigma}_x^2]. \tag{B1}$$

By substituting sample moments into the expressions for the MSE and σ_x^2 in Eqs. (10) and (3), the first term on the right-hand side is

$$E[\widehat{\mathrm{MSE}} \cdot \hat{\sigma}_x^2] = E[[\hat{\mu}'_{(2)f} + \hat{\mu}_x(1 - 2\hat{\mu}_{f|x=1})]\hat{\mu}_x(1 - \hat{\mu}_x)]. \tag{B2}$$

Expanding the terms,

$$\begin{aligned} E[\widehat{\mathrm{MSE}} \cdot \hat{\sigma}_x^2] = {} & E[\hat{\mu}_x\hat{\mu}'_{(2)f}] - E[\hat{\mu}_x^2\hat{\mu}'_{(2)f}] \\ & + E[\hat{\mu}_x^2(1 - 2\hat{\mu}_{f|x=1})] \\ & - E[\hat{\mu}_x^3(1 - 2\hat{\mu}_{f|x=1})]. \end{aligned} \tag{B3}$$

We will evaluate expectations by conditioning on the number of occurrences of $x = 1$, denoted N_1. For a known sample size N, N_1 is a random variable. Because x is a Bernoulli random variable, N_1 has a binomial distribution with the parameter μ_x. Note that the number of occurrences of $x = 0$, denoted N_0, is related to N_1 by the relationship $N = N_0 + N_1$. Conditioning on N_1, the first term is

$$E[\hat{\mu}_x\hat{\mu}'_{(2)f}|N_1] = E\left[\left(\frac{1}{N}\sum_i^N x_i\right)\left(\frac{1}{N}\sum_j^N f_j^2\right)\middle| N_1\right] = \frac{N_1}{N^2}\sum_j^N E[f_j^2|N_1]. \tag{B4}$$

Using the conditional expectations for the noncentral moments of f,

$$\begin{aligned} E[\hat{\mu}_x\hat{\mu}'_{(2)f}|N_1] &= \frac{N_1}{N^2}[N_0\mu'_{(2)f|x=0} + N_1\mu'_{(2)f|x=1}] = \frac{N_0N_1}{N^2}\mu'_{(2)f|x=0} + \frac{N_1^2}{N^2}\mu'_{(2)f|x=1} \\ &= \left(\frac{N_1}{N} - \frac{N_1^2}{N^2}\right)\mu'_{(2)f|x=0} + \frac{N_1^2}{N^2}\mu'_{(2)f|x=1}. \end{aligned} \tag{B5}$$

The unconditional expectation is

$$E[\hat{\mu}_x\hat{\mu}'_{(2)f}] = E[E[\hat{\mu}_x\hat{\mu}'_{(2)f}|N_1]] = \left(\frac{E[N_1]}{N} - \frac{E[N_1^2]}{N^2}\right)\mu'_{(2)f|x=0} + \frac{E[N_1^2]}{N^2}\mu'_{(2)f|x=1}. \tag{B6}$$

Because N_1 is a binomial random variable, the expected values for the noncentral moments are

$$E[N_1] = \mu_x N \quad \text{and} \tag{B7}$$

$$E[N_1^2] = \mu_x^2N^2 + \sigma_x^2N. \tag{B8}$$

By substitution,

$$\begin{aligned} E[\hat{\mu}_x\hat{\mu}'_{(2)f}] &= E[E[\hat{\mu}_x\hat{\mu}'_{(2)f}|N_1]] = \left(\mu_x - \mu_x^2 - \frac{\sigma_x^2}{N}\right)\mu'_{(2)f|x=0} + \left(\mu_x^2 + \frac{\sigma_x^2}{N}\right)\mu'_{(2)f|x=1} \\ &= \mu_x[(1 - \mu_x)\mu'_{(2)f|x=0} + \mu_x\mu'_{(2)f|x=1}] + \frac{\sigma_x^2}{N}[\mu'_{(2)f|x=1} - \mu'_{(2)f|x=0}]. \end{aligned} \tag{B9}$$

By definition,

$$\mu'_{(m)f} = (1 - \mu_x)\mu'_{(m)f|x=0} + \mu_x\mu'_{(m)f|x=1}. \tag{B10}$$

Therefore, the expectation simplifies to

$$E[\hat{\mu}_x\hat{\mu}'_{(2)f}] = \mu_x\mu'_{(2)f} + \frac{\sigma_x^2}{N}[\mu'_{(2)f|x=1} - \mu'_{(2)f|x=0}]. \tag{B11}$$

Using the same approach with other terms in Eq. (B3), the second term is

$$E[\hat{\mu}_x^2\hat{\mu}'_{(2)f}] = \mu_x^2\mu'_{(2)f} + \frac{\sigma_x^2}{N}[(1 - 3\mu_x)\mu'_{(2)f|x=0} + 3\mu_x\mu'_{(2)f|x=1}] + \frac{\sigma_x^2}{N^2}(1 - 2\mu_x)[\mu'_{(2)f|x=1} - \mu'_{(2)f|x=0}], \tag{B12}$$

the third term is

$$E[\hat{\mu}_x^2(1 - 2\hat{\mu}_{f|x=1})] = \mu_x^2(1 - 2\mu_{f|x=1}) + \frac{\sigma_x^2}{N}(1 - 2\mu_{f|x=1}), \tag{B13}$$

and the fourth term is

$$E[\hat{\mu}_x^3(1 - 2\hat{\mu}_{f|x=1})] = \mu_x^3(1 - 2\mu_{f|x=1}) + \frac{3\sigma_x^2}{N}\mu_x(1 - 2\mu_{f|x=1}) + \frac{\sigma_x^2}{N^2}(1 - 2\mu_x)(1 - 2\mu_{f|x=1}). \tag{B14}$$

Because the sample estimator of MSE is unbiased, the expected value of $\widehat{\text{MSE}}$ is

$$E[\widehat{\text{MSE}}] = \mu'_{(2)f} + \mu_x(1 - 2\mu_{f|x=1}). \tag{B15}$$

The expected value of $\hat{\sigma}_x^2$ is

$$E[\hat{\sigma}_x^2] = \frac{N-1}{N}\sigma_x^2, \tag{B16}$$

which implies that the sample estimator is biased.

Combining all of the terms, the covariance term simplifies to

$$\text{cov}(\widehat{\text{MSE}}, \hat{\sigma}_x^2) = \left(\frac{1}{N} - \frac{1}{N^2}\right)\sigma_x^2(1 - 2\mu_x)\cdot\{[\mu'_{(2)f|x=1} - \mu'_{(2)f|x=0}] + (1 - 2\mu_{f|x=1})\}. \tag{B17}$$

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting,* **18,** 918–932.

——, ——, ——, ——, and ——, 2005: Verification of precipitation forecasts from two limited-area models over Italy and comparison with ECMWF forecasts using a resampling technique. *Wea. Forecasting,* **20,** 276–300.

Benjamin, J., and C. Cornell, 1970: *Probability, Statistics, and Decision for Civil Engineers*. McGraw-Hill, 684 pp.

Bradley, A. A., T. Hashino, and S. S. Schwartz, 2003: Distributions-oriented verification of probability forecasts for small data samples. *Wea. Forecasting,* **18,** 903–917.

——, S. S. Schwartz, and T. Hashino, 2004: Distributions-oriented verification of ensemble streamflow predictions. *J. Hydrometeor.,* **5,** 532–545.

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

Briggs, W., 2005: A general method of incorporating forecast cost and loss in value scores. *Mon. Wea. Rev.,* **133,** 3393–3397.

——, and D. Ruppert, 2005: Assessing the skill of yes/no forecasts. *Biometrics,* **61,** 799–807.

Carpenter, T., and K. Georgakakos, 2001: Assessment of Folsom Lake response to historical and potential future climate scenarios: 1. forecasting. *J. Hydrol.,* **249,** 148–175.

Doblas-Reyes, F., V. Pavan, and D. Stephenson, 2003: The skill of multi-model seasonal forecasts of the wintertime North Atlantic oscillation. *Climate Dyn.,* **21,** 501–514.

Ebert, E. E., L. J. Wilson, B. G. Brown, P. Nurmi, H. E. Brooks, J. Bally, and M. Jaeneke, 2004: Verification of nowcasts from the WWRP Sydney 2000 Forecast Demonstration Project. *Wea. Forecasting,* **19,** 73–96.

Efron, B., 1981: Nonparametric estimates of standard error: The jacknife, the bootstrap and other methods. *Biometrika,* **68,** 589–599.

Ferro, C. A., 2007: Comparing probabilistic forecasting systems with the Brier score. *Wea. Forecasting,* **22,** 1076–1088.

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

——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.,* **132,** 2905–2923.

Jolliffe, I., 2007: Uncertainty and inference for verification measures. *Wea. Forecasting,* **22,** 637–650.

——, and D. Stephenson, 2003: Introduction. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. Jolliffe and D. Stephenson, Eds., John Wiley, 1–12.

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.

Kenney, J. F., and E. S. Keeping, 1951: *Mathematics of Statistics, Part 2.* 2nd ed. Van Nostrand, 429 pp.

Kruger, A., S. Khandelwal, and A. A. Bradley, 2007: Ahpsver: A Web-based system for hydrologic forecast verification. *Comput. Geosci.,* **33,** 739–748.

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

Mason, S. J., 2004: On using climatology as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **132,** 1891–1895.

——, and G. M. Mimmack, 1992: The use of bootstrap confidence intervals for the correlation coefficient in climatology. *Theor. Appl. Climatol.,* **45,** 229–233.

——, and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.,* **128,** 2145–2166.

Murphy, A. H., 1997: Forecast verification. *Economic Value of Weather and Climate Forecasts,* R. Katz and A. H. Murphy, Eds., Cambridge University Press, 19–74.

——, and R. L. Winkler, 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.

Schwartz, S. S., 1992: Verifying probabilistic water supply outlooks for the Potomac River basin. Preprints, *28th Conf. and Symp. on Managing Water Resources during Global Change,* Reno, NV, American Water Resources Association, 153–161.

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

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.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. Jolliffe and D. Stephenson, Eds., John Wiley, 137–163.

Welles, E., S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action and collaboration. *Bull. Amer. Meteor. Soc.,* **88,** 503–511.

Wilks, D. S., 1996: Statistical significance of long-range "optimal climate normal" temperature and precipitation forecasts. *J. Climate,* **9,** 827–839.

——, 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 91, Academic Press, 648 pp.

Zhang, H., and T. Casey, 2000: Verification of categorical probability forecasts. *Wea. Forecasting,* **15,** 80–89.

Zhu, Y. J., 2005: Ensemble forecast: A new approach to uncertainty and predictability. *Adv. Atmos. Sci.,* **22,** 781–788.