Summary Verification Measures and Their Interpretation for Ensemble Forecasts

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

(Manuscript received 4 December 2009, in final form 26 July 2010)

ABSTRACT

Ensemble prediction systems produce forecasts that represent the probability distribution of a continuous forecast variable. Most often, the verification problem is simplified by transforming the ensemble forecast into probability forecasts for discrete events, where the events are defined by one or more threshold values. Then, skill is evaluated using the mean-square error (MSE; i.e., Brier) skill score for binary events, or the ranked probability skill score (RPSS) for multicategory events. A framework is introduced that generalizes this approach, by describing the forecast quality of ensemble forecasts as a continuous function of the threshold value. Viewing ensemble forecast quality this way leads to the interpretation of the RPSS and the continuous ranked probability skill score (CRPSS) as measures of the weighted-average skill over the threshold values. It also motivates additional measures, derived to summarize other features of a continuous forecast quality function, which can be interpreted as descriptions of the function's geometric shape. The measures can be computed not only for skill, but also for skill score decompositions, which characterize the resolution, reliability, discrimination, and other aspects of forecast quality. Collectively, they provide convenient metrics for comparing the performance of an ensemble prediction system at different locations, lead times, or issuance times, or for comparing alternative forecasting systems.

1. Introduction

Ensemble prediction systems are widely used to make weather, climate, and hydrologic forecasts. Often forecasters need to make comparative assessments of ensemble forecast quality. For instance, they may wish to determine whether system enhancements improve the forecasts, or compare the performance of alternative systems. Even with a single forecasting system, forecasters may need to assess how the quality of ensemble forecasts depends on the time of year (or time of day) when the forecasts are issued, or how the quality varies from one location to another. To accomplish these tasks, forecasters need summary verification measures suitable for comparison. Suitable measures are ones where differences can be attributed to differences in the forecast system performance, and not differences in the nature of the forecast variable itself (e.g., differences in its climatology at different times or locations).

Finding suitable summary measures for ensemble prediction systems is challenging, as the nature of the forecast itself is complex. Transforming ensemble forecasts into single value (deterministic) forecasts (e.g., ensemble mean), or into probability forecasts for a discrete event (e.g., outcome above some threshold), are common approaches used in ensemble forecast verification (Hamill and Colucci 1997; Markus et al. 1997; Buizza and Palmer 1998; Hamill and Colucci 1998; Atger 1999; Carpenter and Georgakakos 2001; Ebert 2001; Hou et al. 2001; Kumar et al. 2001; Mullen and Buizza 2001; Grimit and Mass 2002; Cong et al. 2003; Franz et al. 2003; Kirtman 2003; Shamir et al. 2006, among others). To examine the performance of ensemble forecasts as probabilistic forecasts of a continuous variable, measures like the continuous ranked probability score (CRPS) are used (Matheson and Winkler 1976; Unger 1985). Decomposition of scores into measures of reliability, resolution, and discrimination (Murphy and Winkler 1992; Murphy 1997; Wilks 2000; Hersbach 2000;

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

DOI: 10.1175/2010MWR3305.1

Candille and Talagrand 2005, among others) is helpful for diagnosing how these aspects affect forecast accuracy.

In this paper, we examine summary verification measures for assessing the overall performance of ensemble forecasts as probabilistic forecasts. We begin by introducing a theoretical framework that generalizes common verification approaches, describing forecast quality as a continuous function of the forecast variable (or its climatological probability). This description naturally leads to a set of measures—some based on traditional measures and others introduced here—which summarizes aspects of ensemble forecast quality, and can be interpreted as measures of the ''geometric shape'' of forecast quality functions.

2. Forecast verification framework

Forecast verification is often viewed in terms of data samples and the sample statistics (verification measures) computed from them. But the verification process can also be described using basic probability theory. As an example, probability theory is central to the distribution-oriented (DO) approach to verification (Murphy and Winkler 1987; Murphy 1997). Forecasts and observations are treated as random variables. Verification measures are defined based on the joint distribution of the forecasts and observations. We will employ a similar approach throughout this paper to define aspects of forecast quality.

Consider an ensemble prediction system that produces a forecast of some continuous random variable Y. The forecast issued represents the probability distribution of Y, conditioned on the state of the system ξ (e.g., initial conditions) at the time of the forecast. This *probability distribution forecast* can be defined by a conditional cumulative distribution function $F(y|\xi)$ as

$$F(y|\xi) = P\{Y \le y|\xi\}, \tag{1}$$

where $P\{Y \le y|\xi\}$ is the probability that the forecast variable Y is less than or equal to some value y. Figure 1 illustrates such a probability distribution forecast from an ensemble prediction system.

Because of the complex nature of a probability distribution forecast, it is commonly transformed into a simpler *probability forecast* of a discrete event for verification. Consider the discrete event $\{Y \le y_p\}$, defined by the threshold value y_p. Then $F(y_p|\xi)$ is a random variable that represents the forecast probability that the event occurs. Figure 1 illustrates how $F(y_p|\xi)$ is defined by the ensemble forecast for the specific threshold value y_p.

Whether the event occurs or not depends on the observation of Y. Let $X(y_p)$ be a random variable that denotes the observation of the discrete event, defined as

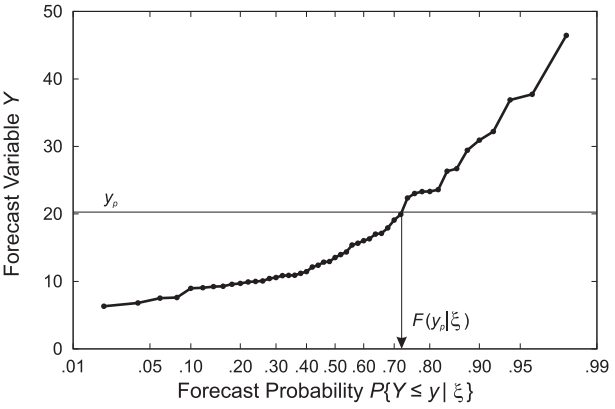

FIG. 1. An example ensemble forecast. The forecast is an empirical probability distribution of the forecast variable Y defined by the ensemble members. For a specific nonexceedence event $\{Y \le y_p\}$ defined by the threshold y_p, a probability forecast $F(y_p|\xi)$ of the event occurrence can be determined from the ensemble probability distribution. The ensemble forecast could also be used to determine a probability forecast for an exceedance event.

$$X(y_p) = \begin{cases} 1 & \text{if} \quad Y \le y_p \\ 0 & \text{if} \quad Y > y_p \end{cases}. \tag{2}$$

In other words, $X(y_p)$ is 1 if the event $\{Y \le y_p\}$ occurs and 0 if it does not.

We employ the notation y_p for the threshold because later it will become convenient to replace the threshold with the climatological probability of the event it defines. The climatological probability p is defined by the *unconditional probability* of the occurrence of the event $\{Y \le y_p\}$:

$$p = P\{Y \le y_p\}. \tag{3}$$

By definition, the climatological probability p is equivalent to the expected value of the binary observation $X(y_p)$:

$$\mu_x(y_p) = E[X(y_p)] = p. \tag{4}$$

Since $X(y_p)$ is a Bernoulli random variable, its variance is simply

$$\sigma_x^2(y_p) = p(1-p). \tag{5}$$

Although the probability forecasts and observations are defined above for a nonexceedance event, $F(y_p|\xi)$ and $X(y_p)$ could have been defined instead for an exceedance event $\{Y > y_p\}$. The climatological probability p would then represent the unconditional probability of the exceedance event.

a. Forecast quality measures

Traditional measures used in forecast verification can be employed to evaluate the quality of probability forecasts for a discrete event. For instance, a measure of the accuracy of the probability forecast is the mean square error (MSE; or Brier score; Brier 1950). MSE is defined for the threshold y_p as

$$\mathrm{MSE}(y_p) = E[\{F(y_p|\xi) - X(y_p)\}^2]. \quad (6)$$

A common measure of skill, or accuracy relative to a reference forecast, is the MSE (or Brier) skill score (SS). Using climatology as a reference forecast, SS for the threshold y_p is defined as

$$\mathrm{SS}(y_p) = 1 - \frac{\mathrm{MSE}(y_p)}{\sigma_x^2(y_p)}. \quad (7)$$

In practice, these and other verification measures are *estimated* based on a random sample of probability forecasts and observations. The appendix illustrates how a verification dataset of ensemble forecasts and continuous observations may be used to create a sample of probability forecasts and observations for a discrete event.

b. Generalization for probability distribution forecasts

From the definition of $F(y|\xi)$ in Eq. (1), it follows that a probability distribution forecast contains a probability forecast for the event $\{Y \le y\}$, for any value of the threshold y. A generalization of the above verification approach is to define and evaluate the forecast quality of probability distribution forecasts as a *continuous function* of threshold value. Let $Q(y_p)$ denote a forecast quality measure (e.g., skill) for a specific threshold y_p. Then the function $Q(y)$ characterizes the forecast quality measure as a continuous function of y. If instead we index the threshold by the climatological probability p of the event defined by y_p, then the function $Q(p)$ characterizes the forecast quality measure as a function of the event occurrence frequency p. Such an approach has been used to describe ensemble forecast quality by Bradley et al. (2004) and Gneiting et al. (2007).

Figure 2 illustrates the concept of a *forecast quality function* for a set of ensemble drought forecasts for the Des Moines River. The figure shows the skill functions SS(y) and SS(p) for the ensemble forecasts. This generalization provides an important insight on the nature of ensemble forecasts. A single value of skill cannot completely characterize the relative accuracy of ensemble forecasts; forecast skill varies depending on the threshold value. In this example, probability forecasts for low flow volume thresholds are quite skillful; however, for the same set of ensemble forecasts, the probability forecasts for high flow volume thresholds have virtually no skill.

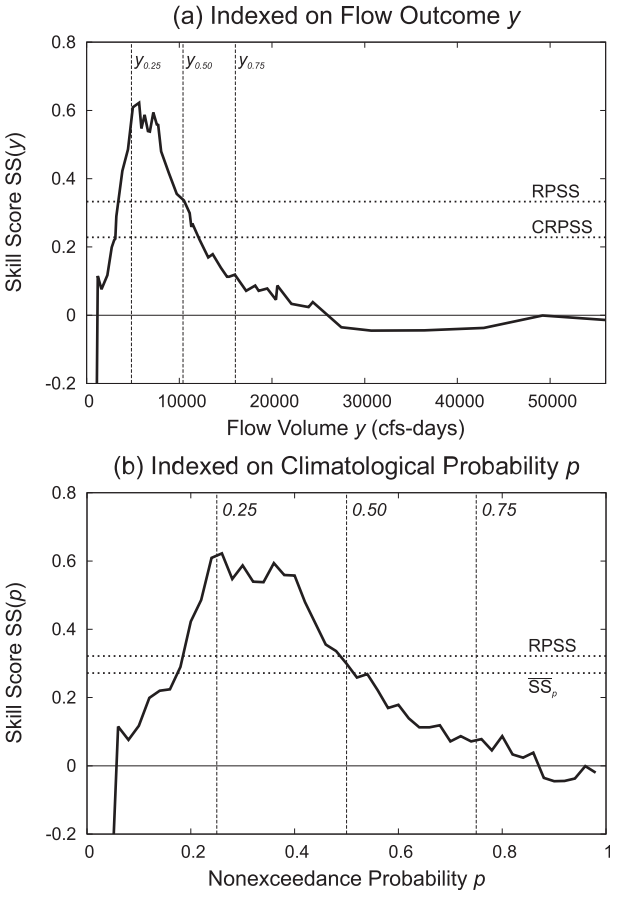

FIG. 2. Forecast quality functions for the MSE skill score (a) using y as the index variable and (b) using p as the index variable. The example is based on ensemble drought forecasts issued in April for the Des Moines River basin; the ensemble predicts the minimum 7-day flow volume over the next 90-day period. The RPSS is the average skill for multicategory probability forecasts, defined in this example by the flow quartiles [Eq. (14)]. CRPSS is the average skill over the entire range of continuous flow outcomes y [Eq. (17)]. $\overline{\mathrm{SS}}_p$ is the average skill over the entire range of probability thresholds p [Eq. (22)].

Even though Fig. 2 shows the forecast skill for nonexceedance events, the forecast skill for exceedance events is readily deduced. For a given threshold y_p, MSE and SS are the same for exceedance and nonexceedance events. Therefore, a plot of SS(y) for exceedance events is the same as shown in Fig. 2a. However, a plot of SS(p) for exceedance events is the mirror image of that shown in Fig. 2b, since the corresponding probability index for an exceedance event is $1 - p$.

Forecast quality functions can be defined that describe other aspects of the ensemble forecasts. As an example, the MSE skill score can be decomposed as (Murphy and Winkler 1992)

$$\mathrm{SS} = \rho_{fx}^2 - [\rho_{fx} - (\sigma_f/\sigma_x)]^2 - [(\mu_f - \mu_x)/\sigma_x]^2, \quad (8)$$

where ρ_{fx} is the correlation of the forecasts and observations; σ_f and σ_x are the standard deviation of the forecasts and the observations, respectively; and μ_f and μ_x are the mean of the forecasts and the observations, respectively. The first term of the decomposition (right-hand side) is known as the potential skill, and is a measure of the resolution of the forecasts (i.e., the skill if there were no biases). The second term is known as the slope reliability, and is a measure of the conditional bias. The third term is the standardized mean error, a measure of the unconditional bias.

Using this decomposition with the skill function $\mathrm{SS}(p)$, we can also create functions that describe each component of the decomposition:

$$\mathrm{SS}(p) = \mathrm{PS}(p) - \mathrm{CB}(p) - \mathrm{UB}(p), \quad (9)$$

where $\mathrm{PS}(p)$ represents the potential skill function [first term in Eq. (8)], $\mathrm{CB}(p)$ represents the conditional bias function (second term), and $\mathrm{UB}(p)$ represents the unconditional bias function (third term). Likewise, the decomposition of the skill function can also be carried out for $\mathrm{SS}(y)$. Although we will use this particular skill score decomposition as an example in the remainder of the paper, the same approach can be used with other decompositions (see Murphy 1997).

In principle, any probability forecast verification measure could be represented as a continuous function of the threshold y or its climatological probability p. However, some measures are not as well suited for graphical comparison (as in Fig. 2). For instance, the magnitude of MSE depends on the event occurrence frequency p. For the most part, a plot of the function $\mathrm{MSE}(y)$ or $\mathrm{MSE}(p)$ illustrates this dependency (see Bradley et al. 2004), and not the differences in forecast quality at different thresholds. A better way to assess the forecast quality as a continuous function is with a *relative measure*. Skill measures, and their decompositions, make better candidates for graphical comparison. We will focus exclusively on these relative measures for the remainder of this paper.

3. Summary measures of forecast quality functions

Forecast quality functions provide a very detailed description of the quality of probability distribution forecasts for a single forecast variable. But there is still a need for simpler measures that can summarize information contained in these functions, to allow forecasters to compare system performance for different locations or lead times, or even for different forecasting systems. In this section, we look at some common measures used in ensemble forecast verification to see what information they summarize, and derive additional measures that can summarize other properties of forecast quality functions.

a. Average forecast quality

The ranked probability score (RPS) is a common measure of the accuracy of probability forecasts for multicategory events (Epstein 1969; Murphy 1971; Wilks 2006). It is well known that the average RPS is equivalent to the MSE (or average Brier score) of probability forecasts for the thresholds defining the discrete categories (Toth et al. 2003). Using the notation derived above:

$$\overline{\mathrm{RPS}} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}(y_i), \quad (10)$$

where y_i is a threshold, and k is the number of thresholds defining the $k + 1$ categories. Note that $\overline{\mathrm{RPS}}$ for climatological forecasts is

$$\overline{\mathrm{RPS}}_{\mathrm{clim}} = \frac{1}{k}\sum_{i=1}^{k}\sigma_x^2(y_i). \quad (11)$$

Hence, the ranked probability skill score (RPSS), using climatology as the reference forecast, is

$$\begin{aligned}\mathrm{RPSS} &= 1 - \frac{\overline{\mathrm{RPS}}}{\overline{\mathrm{RPS}}_{\mathrm{clim}}} \\ &= 1 - \frac{\sum_{i=1}^{k}\mathrm{MSE}(y_i)}{\sum_{i=1}^{k}\sigma_x^2(y_i)}.\end{aligned} \quad (12)$$

To see what information the RPSS summarizes about the skill function $\mathrm{SS}(y)$, we solve for the MSE in Eq. (7), and by substitution into the expression above:

$$\mathrm{RPSS} = \frac{\sum_{i=1}^{k}\sigma_x^2(y_i)\mathrm{SS}(y_i)}{\sum_{i=1}^{k}\sigma_x^2(y_i)}.$$

Gathering the terms involving $\sigma_x^2(y_i)$, which do not depend on forecasts, we define a function w as

$$w(y_i) = \frac{\sigma_x^2(y_i)}{\sum_{i=1}^{k} \sigma_x^2(y_i)}. \quad (13)$$

The RPSS then simplifies to

$$\text{RPSS} = \sum_{i=1}^{k} w(y_i)\text{SS}(y_i). \quad (14)$$

The interpretation of the RPSS is very intuitive when viewed in this form; it is a measure of the *weighted-average skill score* for probability forecasts defined by the discrete thresholds y_i. The weight applied to each threshold is related to the variance $\sigma_x^2(y_i)$, which depends only on the climatological probability of the event defined by y_i [see Eq. (5)].

The RPS applies to probability forecasts for discrete categories; the CRPS extends the concept to continuous forecasts [as defined in Eq. (1)]. The average CRPS for a verification dataset is defined as (Hersbach 2000)

$$\overline{\text{CRPS}} = \int_{-\infty}^{\infty} \text{MSE}(y)\, dy. \quad (15)$$

The continuous ranked probability skill score (CRPSS) is defined as

$$\text{CRPSS} = 1 - \frac{\overline{\text{CRPS}}}{\overline{\text{CRPS}}_{\text{clim}}}, \quad (16)$$

with climatology as the reference forecast. To see what information the CRPSS summarizes about the skill function SS(y), we follow a derivation similar to that shown above for the RPSS. The result is that the CRPSS is mathematically equivalent to

$$\text{CRPSS} = \int_{-\infty}^{\infty} w(y)\text{SS}(y)\, dy, \quad (17)$$

where the weight function is

$$w(y) = \frac{\sigma_x^2(y)}{\int_{-\infty}^{\infty} \sigma_x^2(y)\, dy}. \quad (18)$$

Therefore, CRPSS is a summary measure representing the *weighted-average skill score* over the continuous range of outcomes y.

Figure 2a illustrates this interpretation of summary measures of ensemble forecast skill for drought forecasts for the Des Moines River. The dashed horizontal lines indicate the computed RPSS and CRPSS for the set of ensemble forecasts. In this example, the RPSS is based on four categories defined by the flow quartiles, which are shown as dashed vertical lines. Hence, the RPSS is simply the weighted-average skill function values at these three thresholds. In contrast, the CRPSS is the weighted-average skill over the entire skill function.

The decomposition of the skill score shown in Eq. (9) naturally leads to the definition of weighted-average measures for other forecast quality functions. Substituting this expression into Eq. (17) for the weighted-average skill yields

$$\begin{aligned}\text{CRPSS} &= \int_{-\infty}^{\infty} w(y)\text{SS}(y)\, dy \\ &= \int_{-\infty}^{\infty} w(y)\{\text{PS}(y) - \text{CB}(y) - \text{UB}(y)\}\, dy \\ &= \overline{\text{PS}}_y - (\overline{\text{CB}}_y + \overline{\text{UB}}_y), \end{aligned} \quad (19)$$

where $\overline{\text{PS}}_y$, $\overline{\text{CB}}_y$, and $\overline{\text{UB}}_y$ are weighted-average measures of the potential skill, conditional bias, and unconditional bias, respectively. In general then, if $Q(y)$ denotes the MSE skill function or a component of a skill score decomposition, then

$$\overline{Q}_y = \int_{-\infty}^{\infty} w(y)Q(y)\, dy. \quad (20)$$

b. Summary measures using probability thresholds

As seen in section 3a, the CRPSS can be interpreted as a summary measure of a skill function indexed using the threshold value y. Here we derive an analogous measure for a skill function indexed using probability p.

For MSE(p), an integrated measure analogous to $\overline{\text{CRPS}}$ [Eq. (15)] is

$$\overline{\text{MSE}} = \int_0^1 \text{MSE}(p)\, dp, \quad (21)$$

where the symbol $\overline{\text{MSE}}$ is used since the right-hand side of Eq. (21) is a mathematical definition of the expected value of MSE(Y).

Following the same derivation as outlined in section 3a, a summary skill measure based on $\overline{\text{MSE}}$, which we denote here as $\overline{\text{SS}}_p$, is

$$\overline{\text{SS}}_p = \int_0^1 w(p)\text{SS}(p)\, dp, \quad (22)$$

where the weight function is

$$w(p) = \frac{p(1-p)}{\int_0^1 p(1-p)\, dp}. \quad (23)$$

Therefore, $\overline{SS}_p$ is a summary measure, analogous to CRPSS, representing the *weighted-average skill score* over the continuous range of probability thresholds p. Figure 2b illustrates this interpretation for the drought forecasts for the Des Moines River. Note that $\overline{SS}_p$ is not equal to CRPSS shown in Fig. 2a. However, $\overline{SS}_p$ is closely related to the RPSS; in the limit as the number of categories goes to infinity, the RPSS converges to $\overline{SS}_p$.

In general then, if $Q(p)$ denotes the MSE skill function or a component from a skill score decomposition, then

$$\overline{Q}_p = \int_0^1 w(p)Q(p)\,dp. \tag{24}$$

When using the probability threshold p as an index, the denominator of $w(p)$ [in Eq. (23)] is a constant 1/6, so $w(p)$ simplifies for this case to

$$w(p) = 6p(1-p). \tag{25}$$

Figure 3 shows this weight function. Probability thresholds near 0.5 have the largest weight. The weight approaches 0 as the probability threshold approaches 0 or 1. The weight functions $w(y)$ in Eqs. (13) and (18) have similar properties; their numerator is equal to $p(1-p)$ for the threshold y_p [see Eq. (5)].

Clearly, this weighting of forecast quality functions is not arbitrary; it arises from the mathematical definitions of RPSS, CRPSS, and $\overline{SS}_p$. These summary skill measures are based on $\overline{RPS}$, $\overline{CRPS}$, and $\overline{MSE}$, which sum or integrate MSE over the range of thresholds. Note that the magnitude of MSE at each threshold scales as $p(1-p)$, so just like the weight function in Fig. 3, rare events (where p approaches 0 or 1) naturally contribute less to the sum or integral than commonly occurring events (where p is near to 0.5). Our formulation replaces MSE with SS—a relative measure that eliminates this scale dependency—so this weighting now appears explicitly in the form of the weight function.

c. Higher-order summary measures

The weighted-average skill summarizes an extremely important property of the skill functions shown in Fig. 2. But other properties of the functions are also of interest. For instance, the skill function $SS(p)$ is not constant; it exhibits systematic departures from the weighted-average skill. Probability forecasts for low flow volume thresholds have high skill, whereas those for high flow volume thresholds have very little skill. This information is significant in this drought forecasting example, as a forecaster would prefer high skill probability forecasts of unusually low flow conditions (thresholds corresponding to low flow volumes). Although $\overline{SS}_p$ provides summary information on the average skill, it does not indicate that the skill is concentrated at low flow volume thresholds.

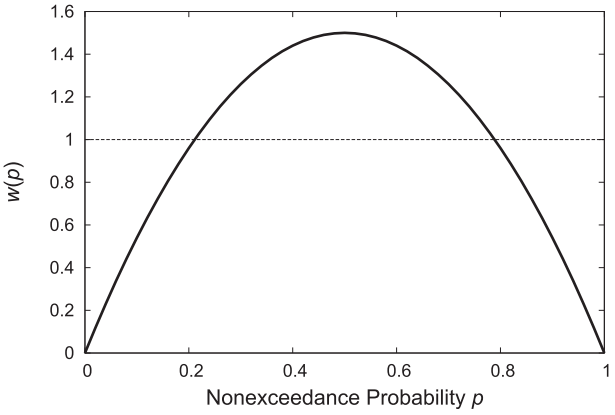

FIG. 3. Weight function $w(p)$ for summary measures of forecast quality functions indexed on the climatological nonexceedance probability p.

One way to measure this property of a forecast quality function $Q(p)$ is to use a geometric analogy. If one considers the weighted forecast quality $w(p)Q(p)$ to be like a "mass distribution," denoted here as $M(p)$, then the weighted-average forecast quality [Eq. (24)] is equivalent to

$$\overline{Q}_p = \int_0^1 w(p)Q(p)\,dp = \int_0^1 M(p)\,dp. \tag{26}$$

In other words, a *geometric interpretation* of $\overline{Q}_p$ is as the total "mass" of the weighted forecast quality function (or more simply, the area under the curve). In physics and engineering applications, the center of mass is defined as the balancing point for the mass distribution, a measure of where the mass is concentrated. The location of the center of mass for the weighted forecast quality function, denoted here as $\overline{p}_Q$, is

$$\overline{p}_Q = \frac{\int_0^1 pM(p)\,dp}{\int_0^1 M(p)\,dp} = \frac{1}{\overline{Q}_p}\int_0^1 pw(p)Q(p)\,dp. \tag{27}$$

Note that the combination of the center of mass $\overline{p}_Q$ and the average forecast quality $\overline{Q}_p$ has a geometric interpretation as the centroid of the weighted forecast quality function. This concept is illustrated graphically in Fig. 4.

We can further extend the geometric analogy to examine another property of the shape of the forecast quality function. The second moment about the center of mass, or the moment of inertia, is defined as

$$I_{Q_p} = \int_0^1 p^2 w(p)Q(p)\,dp - \overline{p}_Q^2\overline{Q}_p. \tag{28}$$

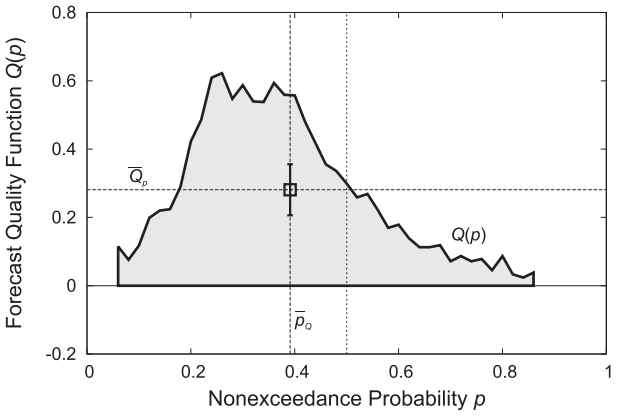

FIG. 4. Geometric interpretation of summary measures of a forecast quality function. The symbol shows the centroid of the forecast quality function (shaded area). The centroid is defined by the weighted-average $\overline{Q}_p$ and the center of mass $\overline{p}_Q$. The shape γ_{Q_p} is indicated by the bar. The length of the bar from the centroid is equal to $|\gamma_{Q_p}|(0.075)$, with the same scale and (probability) units as the x axis. The vertical bar indicates that the forecast quality function "mass" is more concentrated near the center of mass than a constant function ($\gamma_{Q_p} < 0$).

A related measure with the same units as $\overline{p}_Q$ is the radius of gyration k_{Q_p}, defined as

$$k_{Q_p} = \sqrt{\frac{I_{Q_p}}{\overline{Q}_p}}. \tag{29}$$

The moment of inertia and radius of gyration are small if the mass of the forecast quality function is concentrated near the center of mass, and are large if the mass is concentrated away from the center of mass.

A useful benchmark for comparison is the case where the forecast quality function is constant with an average of $\overline{Q}_p$. Then, $Q(p) = \overline{Q}_p$, and its center of mass $\overline{p}'_Q$ is

$$\begin{aligned}\overline{p}'_Q &= \int_0^1 pw(p)\,dp \\ &= 6\int_0^1 p^2(1-p)\,dp \\ &= \frac{1}{2}.\end{aligned} \tag{30}$$

In other words, for a constant forecast quality function, the center of mass is located at the median. If the forecast quality function is not constant, $\overline{p}_Q$ can shift from the median to reflect the concentration of mass (e.g., skill). This is depicted graphically by the centroid in Fig. 4.

For the benchmark case where the forecast quality function is constant, the moment of inertia I'_{Q_p} simplifies to

$$\begin{aligned}I'_{Q_p} &= 6\overline{Q}_p\int_0^1 p^3(1-p)\,dp - \overline{p}'^2_Q\overline{Q}_p \\ &= \overline{Q}_p\left[\frac{3}{10} - \left(\frac{1}{2}\right)^2\right] \\ &= \frac{\overline{Q}_p}{20}.\end{aligned} \tag{31}$$

A convenient measure of the distribution of mass of a forecast quality function using this benchmark case is

$$\begin{aligned}\gamma_{Q_p} &= k_{Q_p} - k'_{Q_p} \\ &= k_{Q_p} - \frac{1}{\sqrt{20}}.\end{aligned} \tag{32}$$

Here k'_{Q_p} is the radius of gyration for a constant forecast quality function. This measure is depicted by the bar at the centroid in Fig. 4. The length of the bar is equal to $|\gamma_{Q_p}|$, and its orientation (vertical or horizontal) indicates its sign. Since $\gamma_{Q_p} < 0$ in this example, a vertical bar is used to visually indicate that the mass is more concentrated near the centroid than the constant $\overline{Q}_p$ benchmark case. If $\gamma_{Q_p} > 0$, then a horizontal bar would be used to visually indicate that the mass is distributed farther away from the centroid than the benchmark case.

One important limitation with the geometric analogy arises when a forecast quality function contains negative values; although the geometric summary measures can still be evaluated, the geometric interpretation breaks down because mass cannot be negative. In the case of the skill score decomposition shown in Eq. (8), all the terms on the right-hand side are nonnegative. However, the skill score itself can take on values less than zero. One way to account for this is to define a nonnegative skill function $\mathrm{SS}_0(p)$ as

$$\mathrm{SS}_0(p) = \begin{cases}\mathrm{SS}(p) & \text{if} \quad \mathrm{SS}(p) \geq 0 \\ 0 & \text{if} \quad \mathrm{SS}(p) < 0\end{cases}. \tag{33}$$

In essence, $\mathrm{SS}_0(p)$ shows the quality of the ensemble forecasts only at thresholds where the probability forecasts are skillful (more accurate than climatology forecasts). One can then use the transformed skill function $\mathrm{SS}_0(p)$ as the function $Q(p)$ to find the average skill $\overline{\mathrm{SS}}_{0p}$, the center of mass $\overline{p}_{\mathrm{SS}_0}$, and the shape $\gamma_{\mathrm{SS}_{0p}}$, and the geometric analogy is exact. A drawback of using $\mathrm{SS}_0(p)$ is that the weighted average of MSE decomposition terms [like those shown in Eq. (19)] sum to $\overline{\mathrm{SS}}_p$ and not $\overline{\mathrm{SS}}_{0p}$.

4. Examples

The three derived measures—the weighted average $\overline{Q}_p$, the center of mass $\overline{p}_Q$, and the shape measure γ_{Q_p}—summarize geometric features of a forecast quality function. In this section, we illustrate some advantages

and applications of these summary verification measures. One is a hypothetical example that illustrates the added information using the geometric measures. The second is an application to an operational ensemble prediction system, which illustrates how the measures can be used to better understand the characteristics of forecasts at many different sites and forecast times.

a. Hypothetical forecasts with the same average skill

The first example compares the skill function for three hypothetical probabilistic forecasting systems. Figure 5 shows the skill functions for the three cases. By design, all three have the same average skill $\overline{SS}_p$, However, the quality of the forecasts is quite different, as demonstrated by the skill functions. One function has a constant skill for all probability thresholds; one is V shaped, with more skill concentrated at the low and high extremes; and one has an inverted-V shape, with more skill concentrated at intermediate thresholds.

Using the three geometric summary measures, we can produce graphical depictions of key features of three skill functions. The centroids (symbols) and shape (bars) show the differences in where skill is concentrated. For the constant skill function, the center of mass is at 0.5 without a bar ($\gamma_{SS_p} = 0$). For the V-shaped skill function, the centroid is shifted to the left since the mass is more concentrated at lower thresholds, and the bar is horizontal ($\gamma_{SS_p} > 0$) since the mass is greater at the extremes. For the inverted-V skill function, the mass peaks at higher thresholds and is low at the extremes, so the centroid shifts to the right and the bar is vertical ($\gamma_{SS_p} < 0$).

The conclusion is that even when alternative forecast systems are equivalent in terms of their average skill (e.g., CRPSS), the concentration of skill as shown by their skill functions can be very different. The additional summary measures provide a simple way of quantifying these differences in the ensemble forecasts. In this hypothetical example the measures are readily visualized as differences in a function's shape, since the shapes chosen are very simple. As will be seen in the next example, where function shapes are much more complex, the geometric meaning of the measures can still allow one to visualize (in an approximate way) the overall magnitude of a function, and the ways it differs from the benchmark case (a constant function).

b. Forecast system verification problem

The second example illustrates how a forecast system manager might utilize the summary measures for diagnostic verification of an ensemble prediction system. The example utilizes retrospective ensemble streamflow forecasts generated from the operational system at the National Weather Service (NWS) North Central River Forecast Center (NCRFC) for three locations along the Des Moines River: in the headwaters at Jackson (JCKM5), upstream of a flood control reservoir at Stratford (STRI4), and downstream of the reservoir at Des Moines (DESI4). The forecast variable is the minimum 7-day flow volume over a 90-day time horizon, used to anticipate the probability of drought (low flow volumes) during the upcoming season. Verification data samples are constructed using the 50 ensemble forecasts issued on the same calendar day (i.e., 1 forecast per year) over a 50-yr historical period (Kruger et al. 2007).

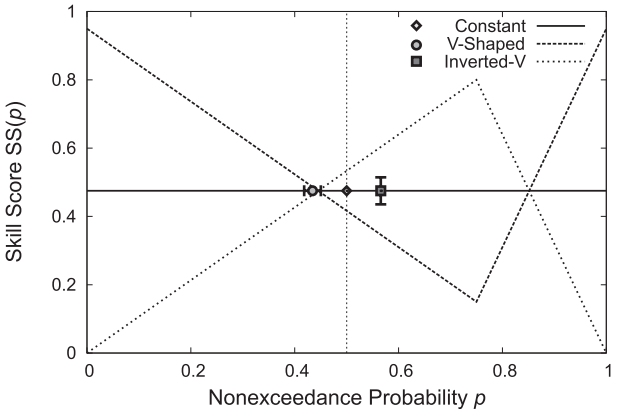

FIG. 5. Hypothetical forecast quality functions and a graphical depiction of their summary measures. The symbols are plotted at centroid of the skill functions. The centroid is defined by the weighted-average skill $\overline{SS}_p$ and the center of mass $\bar{p}_{SS}$. The shape γ_{SS_p} is indicated by the bars. The length of the bar from the centroid is equal to $|\gamma_{SS_p}|$, with the same scale and (probability) units as the x axis. A horizontal bar indicates that the skill function is higher in the extremes than a constant function ($\gamma_{SS_p} > 0$). A vertical bar indicates that the skill function is higher near the center of mass than a constant function ($\gamma_{SS_p} < 0$). The constant skill function is plotted without a bar ($\gamma_{SS_p} = 0$).

We begin by evaluating forecast quality functions for the ensemble forecasts at the upstream location (Jackson) issued at a single time of the year (April; see Fig. 6). The drought forecasts have skill [$SS(p) > 0$], except for extreme low and high threshold probabilities. However, the potential skill $PS(p)$ of the ensemble forecasts is much higher, indicating that forecast biases are reducing forecast accuracy. Still, the conditional bias $CB(p)$ is near zero, except at the extremes. Instead, the unconditional bias $UB(p)$ is the primary source of bias (especially at intermediate threshold levels), and the main reason why the potential skill is not realized.

Figure 6 also plots the geometric summary measures—the centroid (symbols) and shape (bars) for the forecast quality functions. Note that we summarize the function for $SS_0(p)$, as the skill function $SS(p)$ contains some negative values. Despite the complex shape of the functions, collectively the three summary measures indicate

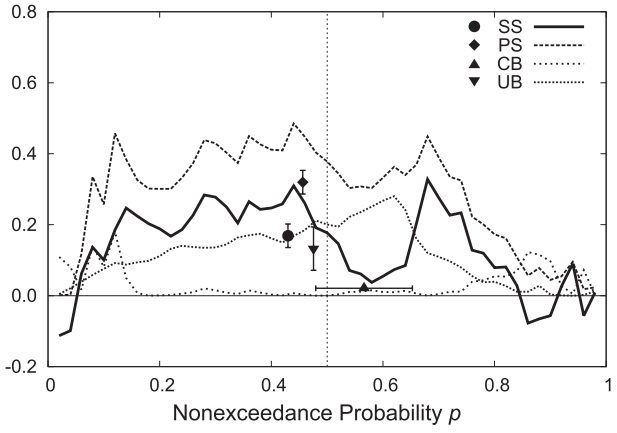

FIG. 6. Skill function and its decomposition for April ensemble drought forecasts for the Des Moines River at Jackson (JCKM5) and a graphical depiction of their summary measures. The symbols are plotted at the centroid of the functions and the bars indicate the shape of the functions (see definitions in Fig. 4). The drought forecast variable is the minimum 7-day flow volume forecast during the next 90 days. Verification is based on a set of 50 April ensemble forecasts (1 forecast per year from 1950–99). Each ensemble forecast contains 49 ensemble members.

how the functions differ from the benchmark (constant function) case. For both SS_0 and PS, the center of mass is less than 0.5 since $SS_0(p)$ and $PS(p)$ values are generally larger at lower probability thresholds; the vertical bars indicate that the mass is more concentrated near the centroid than a constant function. Considering the two bias terms, the weighted average of $CB(p)$ is near zero, but the long horizontal bars indicate that the largest $CB(p)$ values occur at the extremes. In contrast, $UB(p)$ is much larger, with its mass more concentrated near the centroid than the other functions (the longest vertical bar), and a center of mass nearer to 0.5.

To quickly determine whether the forecasts at downstream locations are better or worse, and how the quality of forecasts varies through the spring and early summer (April–July), we compare the summary measures of the forecast quality functions for all three forecast locations. All three sites have skillful ensemble forecasts in spring (see Fig. 7a), but their characteristics consistently change from April to May; skill decreases and the mass of the skill function shifts to higher flow thresholds (an unwelcome property for drought forecasting). Overall, the skill is higher downstream and lower upstream for the same forecast issuance date. By June and July, all three sites have little or no forecast skill.

The trends in the skill are explained by the summary measures for the skill decomposition. The potential skill PS is consistently high at all three sites for April and May forecasts, and a mass shift toward higher flow thresholds in May is apparent (see Fig. 7b). Potential skill drops precipitously in summer months. The conditional biases (CBs) are generally small in April, but by summer they are significantly larger and concentrated at higher flow thresholds (see Fig. 7c). It is the unconditional biases (UBs) that have the most significant affect on skill (see Fig. 7d). At all sites, UB increases from April through July.

Another way that summary measures can be compared is illustrated by the time series plots for the unconditional bias in Fig. 8. The bias $\overline{UB}_p$ significantly increases in the summer months, and is larger upstream and smaller downstream (see Fig. 8a). In these later months, biases at higher flow levels ($\overline{p}_{UB} > 0.5$) are the problem at Jackson (JCKM5) and Stratford (STRI4; see Fig. 8b). In contrast, larger biases occur at lower flow levels at Des Moines ($\overline{p}_{UB} < 0.5$) Biases are concentrated near the centroid ($\gamma_{UB} < 0$), except in May at Stratford and Des Moines, where the biases are relatively small (see Fig. 8c).

Overall, the conclusion is that even though the ensemble forecasts have lower potential skill in summer (June and July), large unconditional biases (or poor reliability) degrade whatever skill they might have. Implementing bias correction (Smith et al. 1992; Leung et al. 1999; Wood et al. 2002; Seo et al. 2006; Hashino et al. 2007) or forecast calibration (Eckel and Walters 1998; Atger 2003; Gneiting et al. 2005; Hamill and Whitaker 2007; Wilks and Hamill 2007; Hagedorn et al. 2008; Primo et al. 2009, among others) could eliminate unconditional biases and help the system realize the potential skill that exists.

Clearly, the utility of the summary measures for forecast system verification is that they permit a meaningful approximate comparison of the properties of forecast quality functions. Although the climatological distribution of the forecast variable is different at each site (and in each month), the nondimensional nature of the probability threshold summary measures facilitates a side-by-side comparison, and provides a consistent diagnostic framework for forecast system assessment.

5. Discussion

Although the previous example compares only three forecast points, it illustrates the utility of the summary measures for forecast system verification. In practice, verification is often done with a single metric like the RPSS or CRPSS—a measure of the weighted-average skill of the forecast skill function. Using the two additional summary measures can provide additional information on the shape of the skill function (and its decompositions) to characterize the performance of the forecast system. Still, it is important to recognize that the three summary measures are insufficient to completely define the shape of a forecast quality function. For example, it is possible to construct shapes that would have

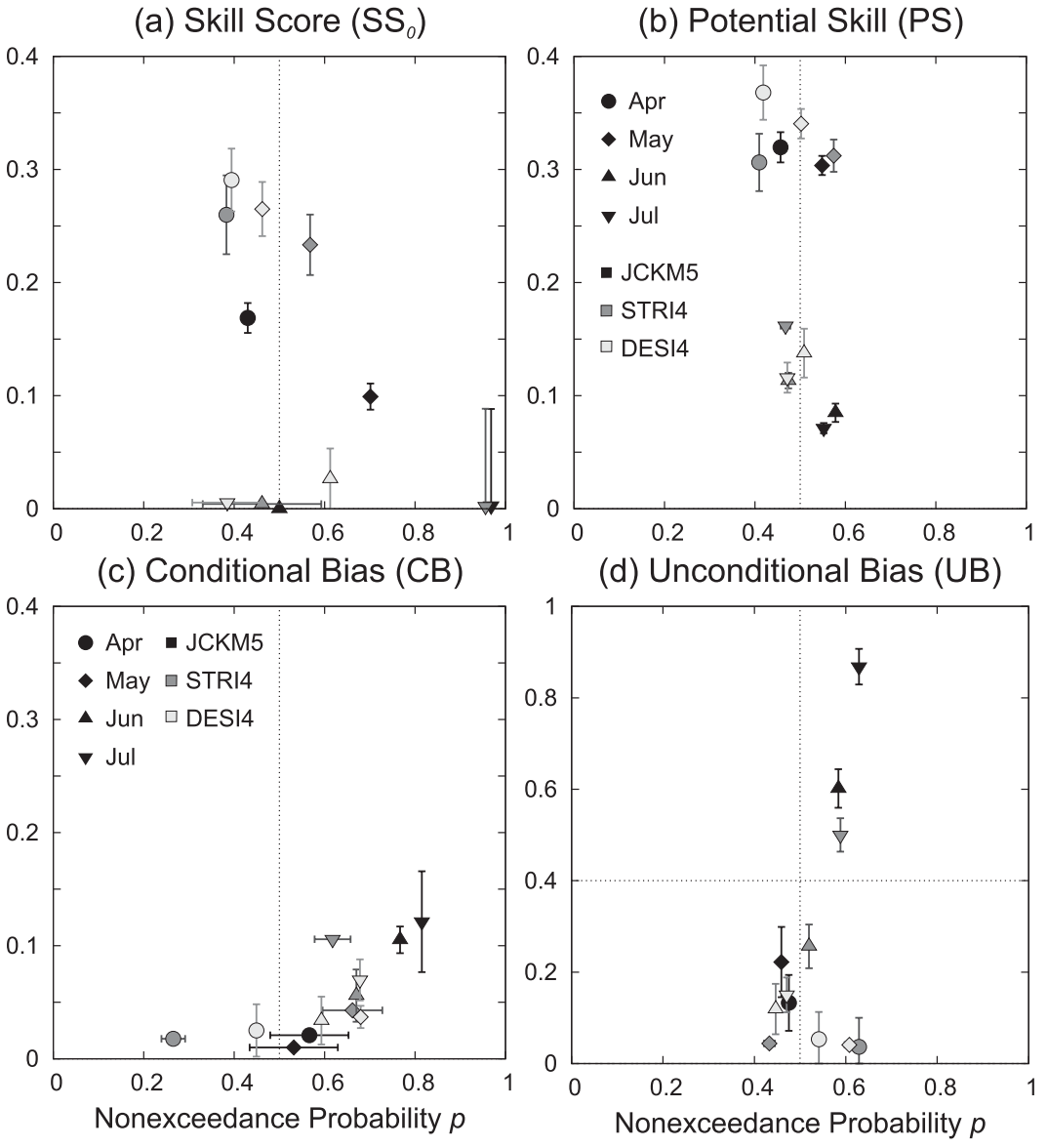

FIG. 7. Ensemble drought forecast summary measures for (a) SS_0, (b) PS, (c) CB, and (d) UB. The symbols are plotted at the centroid of the functions, and the bars indicate the shape of the functions (see definitions in Fig. 4). The drought forecast variable is the minimum 7-day flow volume forecast during the next 90 days. Results are shown for forecasts issued near the beginning of the months of April–July. Verification is based on a set of 50 ensemble forecasts (1 forecast per year from 1950–99) for each month. Each ensemble forecast contains 49 ensemble members. The forecast locations are on the Des Moines River at Jackson (JCKM5), at Stratford (STRI4), and at Des Moines (DESI4).

summary measures that are identical to those for a constant skill function. The constructed shapes must be perfectly symmetrical about a probability threshold of 0.5 (to produce a center of mass of 0.5), and any deviations from the constant function must be perfectly offset by deviations elsewhere of opposite sign (so that the radius of gyration is the same as for a constant function). Such shapes are unlikely in verification applications, but illustrate that there are certain limits in distinguishing between forecast quality function shapes with just three summary measures. In this way they are analogous to the moments of a probability distribution: the mean, variance, and skewness summarize information about the distribution, but cannot completely define its shape (unless it has a simple parametric form that is known a priori). But just as three moments say more

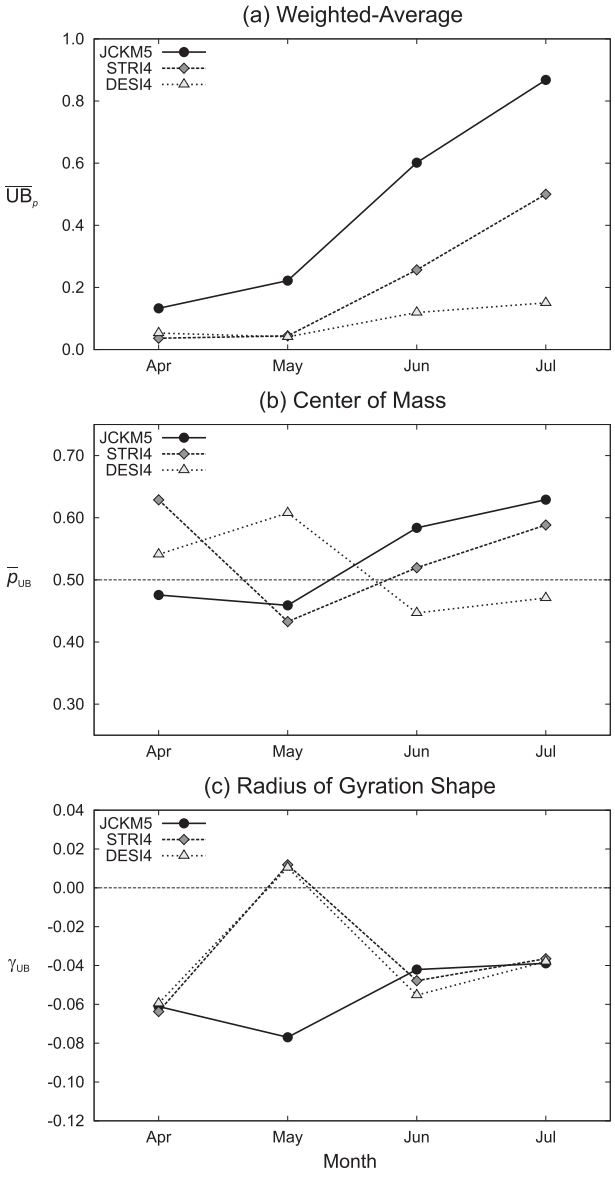

FIG. 8. Time series of ensemble drought forecast summary measures of UB: (a) weighted-average UB, (b) center of mass for UB, and (c) radius of gyration shape for UB. The plots show the summary measure information from Fig. 7d as time series. The drought forecast variable is the minimum 7-day flow volume forecast during the next 90 days. Results are shown for forecasts issued near the beginning of the months of April–July. Verification is based on a set of 50 ensemble forecasts (1 forecast per year from 1950–99) for each month. Each ensemble forecast contains 49 ensemble members. The forecast locations are on the Des Moines River at Jackson (JCKM5), at Stratford (STRI4), and at Des Moines (DESI4).

about a probability distribution than the mean alone, so, too, do the three summary measures say more about a forecast quality function than a weighted average (RPSS or CRPSS) alone.

Using either the forecast quality function indexed by the forecast variable $Q(y)$, or by its climatological nonexceedance probability $Q(p)$, the geometric interpretation of summary measures is valid. Although higher-order summary measures are derived only for $Q(p)$ in section 3, similar measures can be obtained for $Q(y)$. Still, there are compelling reasons for preferring summary measures based $Q(p)$ for ensemble forecast system comparison. First and foremost, the higher-order measures are essentially nondimensional (units of probability), allowing a meaningful comparison between forecast locations or forecasting systems. Also, their measures are readily interpretable. For example, a shift of $\overline{p}_Q$ away from 0.5 (the threshold corresponding to the median forecast variable) has the same meaning in every instance. In contrast, the center of mass and the radius of gyration for $Q(y)$ have dimensions of the forecast variable. For the benchmark case of a constant forecast quality function, the center of mass and radius of gyration depend on the climatological distribution of the forecast variable; they are not constant (as are $\overline{p}'_Q$ and k'_{Q_p}), or consistent in their meaning. One advantage of summary measures for $Q(y)$ is that the weighted-average skill is equal to the CRPSS, which has been used for verification (Candille et al. 2007; Hamill and Whitaker 2007; Ferro et al. 2008; Hagedorn et al. 2008; McCollor and Stull 2009; Candille 2009). Still, $\overline{SS}_p$ is closely analogous to the RPSS, and RPSS converges to $\overline{SS}_p$ as the number of categories grows large.

Since the new summary measures we derived in this paper are based on the RPSS and CRPSS, the forecast quality functions are all weighted by $w(p)$ or $w(y)$. As seen in section 3b, the weights are smaller for rare events, and larger for events near the median. Of course, it is possible to define a new class of summary measures, unrelated to RPSS or CRPSS, where equal weighting at all thresholds is used. One advantage of such an approach would be that the summary measures would describe the shape of the forecast quality functions $Q(p)$ or $Q(y)$, rather than their weighted functions $w(p)Q(p)$ or $w(y)Q(y)$. But in the case of forecast quality functions indexed by the threshold value y, which can be unbounded at the upper and/or lower end, finite values might not exist for equal-weighted measures (i.e., their integral values may be infinite). For functions indexed by p, which is bounded at 0 and 1, such a problem does not occur. Therefore, it would be worthwhile to explore such summary measures for $Q(p)$.

An important insight of the verification framework outlined in section 2 is the interpretation it offers on summary measures of ensemble forecasts. An ensemble probability distribution forecast of a continuous variable (a function) is more complex than a traditional

probability forecast of a discrete event (a single number). Visualizing ensemble forecast quality as a continuous function of the forecast variable is a natural extension of the discrete case. This also leads to the realization that scalar measures like the RPSS or CRPSS summarize features of the forecast quality function. For instance, when viewing a skill score function, the sensitivity of the RPSS to the selection of multicategory events is plain to see. But beyond the numerical result of a given summary measure, viewing forecast quality as varying by outcome also illustrates the fundamental nature of such forecasts. Whereas a single-value probability forecast is *either* skillful or not, an ensemble forecast can be *both* skillful and not, depending on the outcome considered. Using the proposed verification framework to evaluate forecast quality over the continuous range of outcomes can help in assessing the strengths, weaknesses, and potential applications of an ensemble prediction system.

6. Conclusions

Ensemble prediction systems produce forecasts that can represent the probability distribution of a continuous forecast variable. As such, ensemble forecasts are more complex than traditional probability forecasts of a discrete event. In essence, the ensemble forecast contains a probability forecast for any outcome. As a result, one can define and evaluate the forecast quality of a set of ensemble forecasts as a continuous function of the forecast variable (or its climatological nonexceedance probability). Rather than using a single value to describe some aspect of forecast quality, *a forecast quality function* is needed to completely describe that aspect over the range of continuous outcomes for ensemble forecasts. Indeed, the examples presented demonstrate that ensemble forecasts may contain probability forecast statements that are of high quality (skillful) for predicting some outcomes, but be of low quality (no skill) for predicting others.

Still, forecasters need summary measures of forecast quality to enable comparisons between ensemble forecasts made by different forecasting systems, or between forecasts made at multiple locations, issuance times, and lead times. We explore summary measures that describe properties of a forecast quality function. In particular, we show how traditional summary measures of ensemble forecast skill (the ranked probability skill score and the continuous ranked probability skill score) mathematically represent the weighted-average skill of a skill function. This concept can be extended to skill score decompositions to derive weighted-average measures of other aspects of forecast quality, like resolution and reliability (conditional and unconditional biases). Using a geometric analogy, we derive other summary measures that describe the center of mass of a forecast quality function, and the distribution of mass (related to the moment of inertia). Together, the three summary measures can be interpreted as descriptions of the geometric shape of the forecast quality function.

Several examples illustrate the advantages and applications of the summary measures. In particular, even when ensemble forecasts have the same average skill based on traditional summary measures, the differences in their distribution of skill over the range of outcomes has significant implications (to forecasters and forecast users). The additional summary measures showing where the skill is concentrated further differentiates the quality of the ensemble forecasts. Extending the geometric summary measures to skill score decompositions allows one to efficiently compare multiple sets of forecasts, characterize their differences, and diagnose attributes that contribute (or detract) from their skill. Although the geometric summary measures cannot completely replace the information contained in the forecast quality functions, for a forecast system manager evaluating the quality of hundreds of forecast elements, the geometric summary measures provide a way to concisely summarize important verification attributes for ensemble forecast system assessment.

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 NA09OAR4310196, from the Climate Program Office as part of the Climate Prediction Program for the Americas (CPPA). We gratefully acknowledge this support. We would also like to thank the two anonymous reviewers for their thoughtful comments and suggestions.

APPENDIX

Discrete Approximation of Functions and Summary Measures

This section describes how a verification dataset containing ensemble forecasts and outcomes (observations) is used to evaluate forecast quality functions and the geometric summary measures. Let $z_t(j)$, $j = 1, \ldots, M_t$ denote the ensemble forecast at time t, where $z_t(j)$ is the jth ensemble member, and M_t is the number of ensemble members in the forecast. Corresponding to each ensemble forecast, let y_t be the outcome (the observation of the forecast variable).

The first step is to select discrete thresholds where the forecast quality will be evaluated. To simplify the calculations, we define k discrete thresholds, selected at constant intervals in probability space. Let $\{p_i, i = 1, \ldots, k\}$ be the probability thresholds, defined as

$$p_i = \frac{i}{k+1}. \tag{A1}$$

In some applications, the climatology of the forecast variable is known (e.g., from a longer historical record). The threshold values corresponding to the probability thresholds, denoted y_{p_i}, are determined from the known climatology. In other applications, the climatology of the forecast variable is estimated from the verification sample. Let $\{y_{(i)}, i = 1, \ldots, N\}$ be the ranked observations for the N ensemble forecasts in the verification dataset. One could then use the midpoints between successive pairs $y_{(i)}$ and $y_{(i+1)}$ to define k discrete thresholds ($k = N - 1$):

$$y_{p_i} = \frac{y_{(i)} + y_{(i+1)}}{2}, \quad i = 1, \ldots, k. \tag{A2}$$

a. Forecast quality functions

For each threshold y_{p_i}, a set of probability forecasts $f_t(y_{p_i})$ and observations $x_t(y_{p_i})$ must be computed using the ensemble forecasts. Assuming that each ensemble outcome is equally likely, one way to estimate the forecast probability is by the fraction of ensemble members less than or equal to the threshold y_{p_i}:

$$f_t(y_{p_i}) = \frac{1}{M_t}\sum_{j=1}^{M_t} I[y_{p_i} - z_t(j)], \tag{A3}$$

where $I()$ is the indicator function defined as

$$I(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 & \text{if} \quad x \geq 0 \end{cases}. \tag{A4}$$

The observation is then

$$x_t(y_{p_i}) = I(y_{p_i} - y_t). \tag{A5}$$

Note that other approaches may be used to estimate probability forecasts. For example, as illustrated in Fig. 1, one could assign a nonexceedance probability to each ensemble member $z_t(j)$ using a plotting position formula, then interpolate to find the probability forecast $f_t(y_{p_i})$ corresponding to y_{p_i}. Also, if certain outcomes are more likely than others given the state of the system, unequal weighting of ensemble members may be used (e.g., see Smith et al. 1992; Werner et al. 2004).

The set of forecasts and observations for the threshold y_{p_i} are then used to calculate forecast quality measures (e.g., skill and its decomposition; Bradley et al. 2003). The process is repeated for all k thresholds values. The discrete approximation of the forecast quality function $Q(p)$ is $\{Q(p_i), i = 1, \ldots, k\}$.

b. Discrete approximation of summary measures

The summary measures $\overline{Q}_p$, $\overline{p}_Q$, and I_{Q_p} are defined as integrals over the range of outcomes indexed by the climatological probability p. These integrals are approximated by numerical integration. Since the probability interval $p_{i+1} - p_i$ is equal to $1/(k + 1)$ for each pair of thresholds [see Eq. (A1)], numerical integration using the trapezoidal rule reduces to the simple expressions shown below.

Using the discrete weight function $w(p_i)$ defined as

$$w(p_i) = \frac{p_i(1 - p_i)}{\sum_{i=1}^{k} p_i(1 - p_i)}, \tag{A6}$$

the approximation of the weighted-average forecast quality [Eq. (24)] is

$$\overline{Q}_p = \sum_{i=1}^{k} w(p_i)Q(p_i). \tag{A7}$$

The approximation of the center of mass [Eq. (27)] is

$$\overline{p}_Q = \frac{1}{\overline{Q}_p}\sum_{i=1}^{k} p_i w(p_i)Q(p_i). \tag{A8}$$

The approximation moment of inertia I_{Q_p} [Eq. (28)] is

$$I_{Q_p} = \sum_{i=1}^{k} p_i^2 w(p_i)Q(p_i) - \overline{p}_Q^2\overline{Q}_p. \tag{A9}$$

Note that in the case of the skill function $\mathrm{SS}(p)$, the approximation of $\overline{\mathrm{SS}}_p$ is mathematically equivalent to the RPSS for the $k + 1$ categories defined by the thresholds $\{y_{p_i}, i = 1, \ldots, k\}$.

REFERENCES

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.,* **127,** 1941–1953.

——, 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.,* **131,** 1509–1523.

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.

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.,* **126,** 2503–2518.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.,* **137,** 1655–1665.

——, and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.,* **131,** 2131–2150.

——, C. Cote, P. L. Houtekamer, and G. Pellerin, 2007: Verification of an ensemble prediction system against observations. *Mon. Wea. Rev.,* **135,** 2688–2699.

Carpenter, T. M., and K. P. Georgakakos, 2001: Assessment of Folsom Lake response to historical and potential future climate scenarios: 1. Forecasting. *J. Hydrol.,* **249** (1–4), 148–175.

Cong, S., J. C. Schaake, and E. Welles, 2003: Retrospective verification of ensemble stream predictions (ESP): A case study. Preprints, *17th Conf. on Hydrology,* Long Beach, CA, Amer. Meteor. Soc., JP3.8.

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

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting,* **13,** 1132–1147.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Ferro, C. A. T., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteor. Appl.,* **15,** 19–24.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales, 2003: Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin. *J. Hydrometeor.,* **4,** 1105–1118.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Series B-Stat. Method.,* **69,** 243–268.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting,* **17,** 192–205.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.,* **136,** 2608–2619.

Hamill, T. M., and S. J. Colucci, 1997: Verification of ETA-RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998: Evaluation of ETA-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, and J. S. Whitaker, 2007: Ensemble calibration of 500-hPa geopotential height and 850-hPa and 2-m temperatures using reforecasts. *Mon. Wea. Rev.,* **135,** 3273–3280.

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2007: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci.,* **11,** 939–950.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Hou, D. C., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX '98 ensemble forecasts. *Mon. Wea. Rev.,* **129,** 73–91.

Kirtman, B. P., 2003: The COLA anomaly coupled model: Ensemble ENSO prediction. *Mon. Wea. Rev.,* **131,** 2324–2341.

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

Kumar, A., A. G. Barnston, and M. P. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate,* **14,** 1671–1676.

Leung, L. R., A. F. Hamlet, D. P. Lettenmaier, and A. Kumar, 1999: Simulations of the ENSO hydroclimate signals in the Pacific Northwest Columbia River basin. *Bull. Amer. Meteor. Soc.,* **80,** 2313–2328.

Markus, M., E. Welles, and G. N. Day, 1997: A new method for ensemble hydrograph forecast verification. Preprints, *13th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology,* Long Beach, CA, Amer. Meteor. Soc., J106–J108.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.,* **22** (10), 1087–1096.

McCollor, D., and R. Stull, 2009: Evaluation of probabilistic medium-range temperature forecasts from the North American Ensemble Forecast System. *Wea. Forecasting,* **24,** 3–17.

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **129,** 638–663.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.,* **10,** 155–156.

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

——, and ——, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7** (4), 435–455.

Primo, C., C. A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson, 2009: Calibration of probabilistic forecasts of binary events. *Mon. Wea. Rev.,* **137,** 1142–1149.

Seo, D.-J., H. D. Herr, and J. C. Schaake, 2006: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci. Discuss.,* **3,** 1987–2035.

Shamir, E., T. M. Carpenter, P. Fickenscher, and K. P. Georgakakos, 2006: Evaluation of the National Weather Service operational hydrologic model and forecasts for the American River basin. *J. Hydrol. Eng.,* **11** (5), 392–407.

Smith, J. A., G. N. Day, and M. D. Kane, 1992: Nonparametric framework for long-range streamflow forecasting. *J. Water Resour. Plan. Manage.,* **118,** 82–91.

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

Unger, D. A., 1985: A method to estimate the continous ranked probability score. Preprints, *Ninth Conf. on Probability and Statistics in Atmospheric Sciences,* Virginia Beach, VA, Amer. Meteor. Soc., 206–213.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay, 2004: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts. *J. Hydrometeor.,* **5,** 1076–1090.

Wilks, D. S., 2000: Diagnostic verification of the Climate Prediction Center long-lead outlooks, 1995–98. *J. Climate,* **13,** 2389–2403.

——, 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 59, Academic Press, 629 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390.

Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.,* **107,** 4429, doi:10.1029/2001JD000659.