Distributions-Oriented Verification of Ensemble Streamflow Predictions

A. ALLEN BRADLEY

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

STUART S. SCHWARTZ

Center for Environmental Science, Technology and Policy, Cleveland State University, Cleveland, Ohio

TEMPEI HASHINO

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

(Manuscript received 29 July 2003, in final form 2 December 2003)

ABSTRACT

Ensemble streamflow prediction systems produce forecasts in the form of a conditional probability distribution for a continuous forecast variable. A distributions-oriented approach is presented for verification of these probability distribution forecasts. First, a flow threshold is used to transform the ensemble forecast into a probability forecast for a dichotomous event. The event is said to occur if the observed flow is less than or equal to the threshold; the probability forecast is the probability that the event occurs. The distributions-oriented approach, which has been developed for meteorological forecast verification, is then applied to estimate forecast quality measures for a verification dataset. The results are summarized for thresholds chosen to cover the range of possible flow outcomes. To aid in the comparison for different thresholds, relative measures are used to assess forecast quality. An application with experimental forecasts for the Des Moines River basin illustrates the approach. The application demonstrates the added insights on forecast quality gained through this approach, as compared to more traditional ensemble verification approaches. By examining aspects of forecast quality over the range of possible flow outcomes, the distributions-oriented approach facilitates a diagnostic evaluation of ensemble forecasting systems.

1. Introduction

Ensemble streamflow prediction systems make forecasts by using a hydrologic model to simulate streamflow time series (traces) for alternate input weather sequences. The weather inputs are often from historical records, although inputs from ensemble weather and climate forecasts might be used as well. Each hydrologic simulation is initialized to reflect the current moisture state of the watershed. The result is an ensemble of conditional streamflow traces. Increasingly, ensemble prediction techniques are being used for streamflow forecasting and water resources decision making (Day 1985; Smith et al. 1992; Croley 1993; Georgakakos et al. 1995; Georgakakos et al. 1998; Hamlet and Lettenmaier 1999; Carpenter and Georgakakos 2001; Duce 2001; Faber and Stedinger 2001; Krzysztofowicz 2001; Krzysztofowicz and Herr 2001; Yao and Georgakakos 2001; Hamlet et al. 2002; Hay et al. 2002; Wood et al. 2002). At the National Weather Service (NWS), ensemble prediction techniques are being implemented nationwide for long-range streamflow forecasting (out to 90 days), as a part of Advanced Hydrologic Prediction Services (AHPS) (Braatz et al. 1997; Connelly et al. 1999).

The nature of a forecast from an ensemble prediction system is very different from traditional hydrologic and meteorologic forecasts. For example, a short-range flood forecast is an example of a *deterministic* forecast (Jolliffe and Stephenson 2003). That is, the forecast is a single outcome, and no statement about its uncertainty is provided. In contrast, a probability of precipitation (PoP) forecast is an example of a *probabilistic* forecast for a discrete event (Jolliffe and Stephenson 2003). That is, the forecast is the probability that the event (i.e., rain) will occur. Ensemble streamflow predictions are usually used to make a probabilistic forecast. Rather than a forecast probability for a discrete event, however, the ensemble defines an empirical probability distribution for an outcome (e.g., streamflow volume) that is a continuous variable; for a discussion on interpreting ensembles as probabilistic forecasts see Krzysztofowicz (2001).

Corresponding author address: Dr. A. Allen Bradley, IIHR-Hydroscience and Engineering, 523A C. Maxwell Stanley Hydraulics Laboratory, University of Iowa, Iowa City, IA 52242-1585.
E-mail: allen-bradley@uiowa.edu

Verification of forecasts is an essential step for their operational use in decision making (Livezey 1990; Mjelde et al. 1993; Murphy 1993; Roebber and Bosart 1996; Doswell and Brooks 1998; Stern 2001; Wilks 2001; Hartmann et al. 2002; Palmer 2002; Pagano et al. 2002; among others). Forecast verification is carried out using a verification dataset, which contains a record of forecasts and subsequent observations. A comparison of the forecasts with the observations is then made to assess forecast quality. To generate a verification dataset, it may be necessary to reconstruct forecasts for the past, where a record of streamflow observations is available (a technique also known as hindcasting).

Numerous techniques have been used for verification of ensemble predictions for meteorologic (Anderson 1996; Hamill and Colucci 1997; Wilson et al. 1999; Hersbach 2000; Hou et al. 2001; among others) and hydrologic forecasts (Day et al. 1992; Schwartz 1992; Markus et al. 1997; Hamlet and Lettenmaier 1999; Carpenter and Georgakakos 2001; Duce 2001; Cong et al. 2003; among others). In this paper, we present a *distributions-oriented* approach for ensemble forecast verification. The distributions-oriented approach has been developed and used extensively for meteorologic forecast verification of discrete forecasts and observations (both probabilistic and deterministic) (Murphy and Winkler 1992; Brooks and Doswell 1996; Brooks et al. 1997; Murphy and Wilks 1998; Wilks 2000; among others). Here, we describe a framework for applying distributions-oriented concepts to probability distribution forecasts of continuous outcomes. The approach is presented through an application to ensemble streamflow predictions for the Des Moines River basin. The application illustrates how the distributions-oriented approach adds insight on the quality of forecasts over the information available with more traditional approaches.

2. Problem statement

Let Y_i be a continuous random variable representing a forecast variable at time *i*. Examples might include the discharge on a specified date, the cumulative discharge volume for some period, or the minimum discharge during the forecast interval. Assume that the forecasting system produces forecasts of the probability distribution of Y_i conditioned on the current state of the system. The *probability distribution forecast* $G_i(y)$ is defined as

$$G_i(y) = P\{Y_i \le y \,|\, \xi_i\}, \tag{1}$$

where $P\{Y_i \le y \,|\, \xi_i\}$ is the probability that the forecast variable Y_i is less than or equal to y, conditioned on the state of the hydroclimatic system ξ_i. Our goal for verification is to characterize the quality of probability distribution forecasts $G_i(y)$ for a specific forecasting system.

a. Verification methodology

Forecast verification requires comparison of the forecasts with observations. However, the form of the forecast $G_i(y)$ as a probability distribution function, and the outcome Y_i as a continuous variable, makes a comparison difficult for ensemble forecasts. In practice, probability distribution forecasts are usually reformatted (or transformed), and verification is carried out with a simpler form of the forecast.

One common approach is to compare the ensemble mean with the observations (Day et al. 1992; Hou et al. 2001; Duce 2001; Cong et al. 2003; among others). This approach essentially transforms $G_i(y)$ into a deterministic forecast for verification. A second approach is to define discrete outcomes, and transform $G_i(y)$ into a probabilistic forecast for each category (e.g., Carpenter and Georgakakos 2001; Cong et al. 2003). An example would be the use of flow terciles to define above-, below-, and near-normal categories. Although both approaches provide valuable information on aspects of forecast quality, neither completely evaluates the quality of forecasts over the continuous range of outcomes. Here we address this problem through a generalization of the second approach.

The meteorological literature defines a *probability forecast* as one that assigns a probability to the occurrence of a discrete event. From the definition of $G_i(y)$ in Eq. (1), it follows that the probability distribution forecast is a probability forecast for the event $\{Y_i \le y\}$, for any and all values of y. For a specific value y^*, let $f_i(y^*)$ be the probability forecast that the nonexceedance event $\{Y_i \le y^*\}$ occurs. By definition,

$$f_i(y^*) = G_i(y^*). \tag{2}$$

Let $x_i(y^*)$ be a discrete random variable defined as

$$x_i(y^*) = \begin{cases} 1 & \text{if } Y_i \le y^* \\ 0 & \text{if } Y_i > y^*. \end{cases} \tag{3}$$

In other words, $x_i(y^*)$ is 1 when the event $\{Y_i \le y^*\}$ occurs and 0 when it does not. Using a set of probability forecasts $f_i(y^*)$ and discrete observations $x_i(y^*)$, forecast verification methods can be applied to evaluate forecast quality for the event $\{Y_i \le y^*\}$.

Let $Q(y^*)$ denote a specific forecast quality measure (e.g., skill) for a specific threshold y^*. Then the function $Q(y)$ characterizes the forecast quality measure for the probability distribution forecast $G_i(y)$. In principle, $Q(y)$ must be determined for the continuous range of y. In practice, however, we will evaluate $Q(y)$ at a set of threshold values spanning the range of possible outcomes.

b. Distributions-oriented approach

Verification methods for probability forecasts of discrete variables are well established in the meteorological literature (Brier 1950; Wilks 1995; Murphy 1997; Zhang

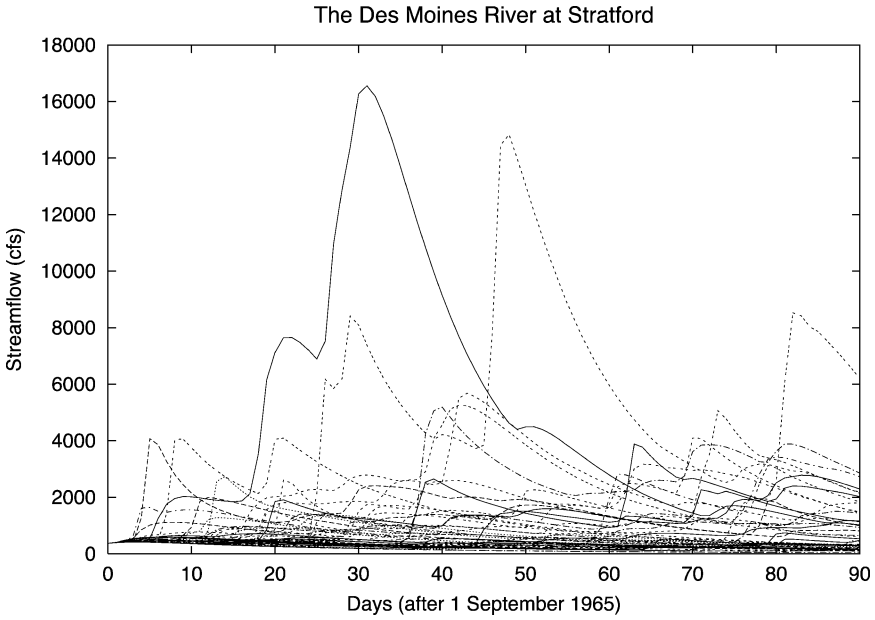

FIG. 1. Ensemble traces simulated from 1 Sep 1965 with historical weather sequences for the Des Moines River at Stratford.

and Casey 2000). However, recent applications have demonstrated the appeal of the distributions-oriented (DO) approach to forecast verification (Murphy and Winkler 1992; Brooks and Doswell 1996; Brooks et al. 1997; Murphy and Wilks 1998; Wilks 2000; among others). Forecast verification using a DO approach centers on the joint distribution of forecasts and observations $h(f, x)$ (Murphy and Winkler 1987; Murphy 1997). In the case of probability forecasts of a dichotomous event, f denotes the forecast probability, and x is the binary outcome (0 or 1). If forecast-observation pairs (f, x) at different times are independent and identically distributed, the relationship between f and x is completely described by $h(f, x)$. Hence, all aspects of the forecast quality can be determined directly from the joint distribution (Murphy 1997).

Murphy (1997) provides a comprehensive presentation of forecast verification based on the DO approach, and describes summary measures for aspects of forecast quality. In the approach proposed here for probability distribution forecasts, forecast quality measures derived from the joint distribution $h(f, x)$ are used for verification. These measures will be estimated for forecasts $f_i(y)$ and observations $x_i(y)$ for the nonexceedance event threshold y, and the results will be summarized over the entire range of possible flow thresholds.

3. Study area

An application of the DO approach for verification of probability distribution forecasts is presented for an experimental streamflow forecasting system for the Des Moines River basin. This forecast system makes ensemble streamflow predictions using a hydrological model driven with historical weather observations for the period from 1948 to 1997. The hydrological model used is the Hydrological Simulation Program-FORTRAN (HSPF), a lumped-parameter conceptual model developed for the U.S. Environmental Protection Agency (Donigian et al. 1995).

Verification is carried out for monthly flow volume forecasts for September to illustrate the proposed approach. Forecasts are made on 1 September of each year for the cumulative streamflow volume through the end of the month. The location selected for forecast verification is the Des Moines River at Stratford, Iowa (U.S. Geological Survey stream gauge 5481300). This location drains a 5452 mi^2 area, and is directly upstream of Saylorville Reservoir. Discharge volume at this location is of interest because it represents the majority of the inflow to the reservoir. Saylorville Reservoir is operated for flood control, drought management, and other uses by the U.S. Army Corps of Engineers.

A verification dataset, made up of forecasts and observations for many events, is needed to assess the quality of the forecasting system. In this application, September flow volume forecasts for 1949–96 ($N = 48$) were reconstructed for verification. Streamflow traces were generated using the initial conditions valid on the forecast date and the weather data for each year in the historical record (see Fig. 1). Therefore, for each forecast date, there are 48 traces available. To correct for biases, the simulated volumes were adjusted based on simulation results and observations from the historical record, using a technique similar to that used by Wood et al. (2002) for weather inputs. Probability distribution

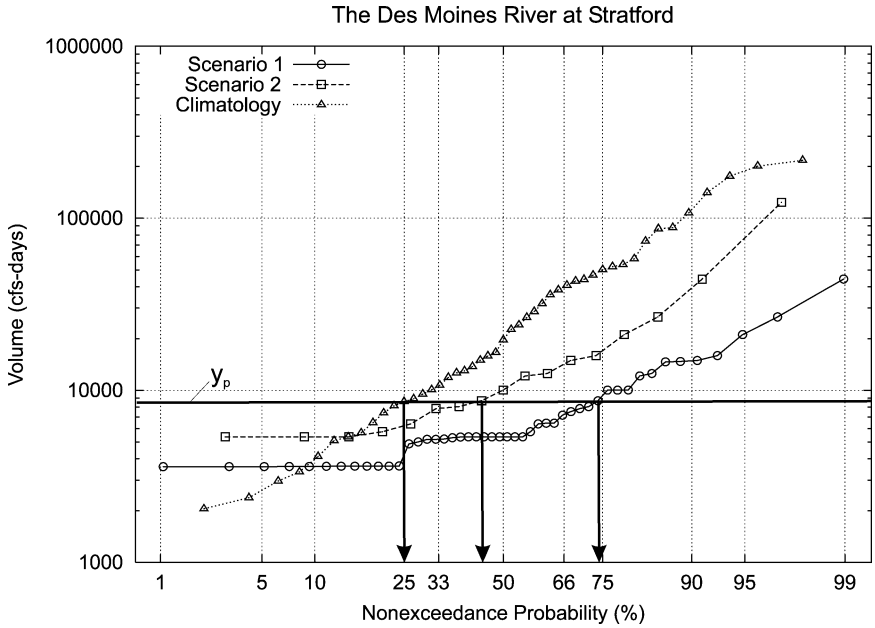

FIG. 2. Probability distribution forecasts of Sep streamflow volume for 1 Sep 1965 for the Des Moines River at Stratford. Forecasts are shown for scenario 1 (no climate forecast) and scenario 2 (skillful climate forecast). Also shown is the empirical distribution of observed flows over the 48-yr verification period (climatology).

forecasts of monthly flow volume were then determined using the extended (or ensemble) streamflow prediction (ESP) approach (Day 1985; Smith et al. 1992).

Two alternative ESP forecasts were generated from the traces for comparison. Scenario 1 assumes that no skillful climate forecast is available at the time the ESP forecast is made. Thus, all 48 traces are considered to be ''equally likely'' outcomes, and all are used to define the probability distribution forecast. Scenario 2 assumes that a skillful climate forecast is available at the time of the ESP forecast. Specifically, we assume the climate forecast predicts whether the September precipitation is above, below, or near normal, as defined by the terciles of its unconditional (climatological) distribution. Furthermore, we assume that this hypothetical climate forecast is ''perfect'' in that it always predicts the correct precipitation category for the month. The probability distribution forecasts for streamflow volume are then derived using only those traces generated with precipitation in the forecast precipitation category. Figure 2 shows an example ESP probability distribution forecast for 1 September 1965 for the two scenarios. For comparison, the empirical distribution of the observed September flow volume for the 48-yr historical record is also shown. Figure 3 shows a time series of ESP forecasts and observed volumes for the period from 1960 to 1970.

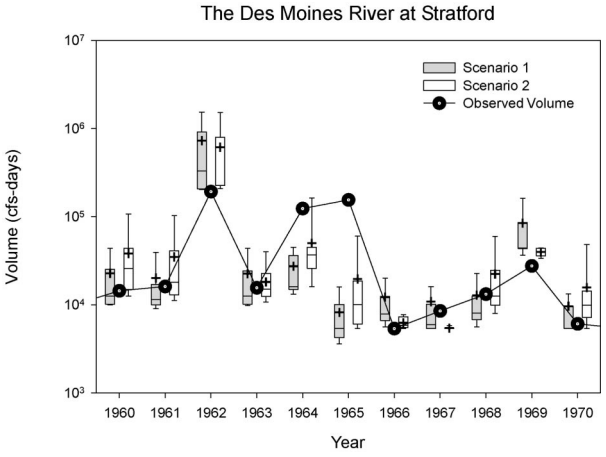

FIG. 3. Probability distribution forecasts and observed Sep streamflow volume from 1960 to 1970 for the Des Moines River at Stratford. Forecasts are shown for scenario 1 (no climate forecast) and scenario 2 (skillful climate forecast). The boxes show the 25% and 75% percentile forecast volumes; the ensemble median is indicated within the boxes. The whiskers indicate the 10% and 90% percentile forecast volumes. The ensemble mean is shown by the crosses.

Clearly, scenario 2 is not selected for its operational realism. Indeed, it is unrealistic to expect that perfect forecasting of future precipitation categories is achievable. Note too that the approach for making a probability distribution forecast based on the hypothetical climate forecasts is but one way (possibly not the best) to use this information for streamflow forecasting (e.g., see Croley 2000 and Perica 1998). Instead, this scenario is selected because one might reasonably expect that perfect insight into next month's precipitation category will result in higher-quality streamflow forecasts. The verification in the following sections will show whether this is so.

4. Application

The first step in the application is to select flow thresholds where we will apply the DO verification approach. Let $\{y_{(i)}, i = 1, \ldots, N\}$ be the ranked observed flow volumes for the N years of the verification period. We will use the midpoints between successive pairs $y_{(i)}$ and $y_{(i+1)}$ as thresholds. The thresholds will be indexed by their nonexceedance probability p. For the midpoint between $y_{(i)}$ and $y_{(i+1)}$, p is estimated by i/N. For this application ($N = 48$), this approach results in 47 separate thresholds, with nonexceedance probabilities p ranging from 1/48 to 47/48.

The next step is to construct the set of forecast-observation pairs for each flow threshold. Figure 2 illustrates the process for the September 1965 forecast. The horizontal line shows the $y_{0.25}$ quantile threshold in relation to the probability distribution forecasts for the two scenarios and the empirical distribution of observed flows. Note that the empirical distribution can be thought of as the climatological (or unconditional) forecast of September streamflow volume. By definition then, the climatological probability forecast f associated with the $y_{0.25}$ threshold is 0.25. The conditional forecasts for the two scenarios predict a much higher probability of flow below the threshold. The forecast probability is 0.737 for scenario 1, and 0.432 for scenario 2.

For each year in the verification dataset, the forecast probability f_i is determined for each scenario. Table 1 shows the results for the $y_{0.25}$ threshold. Also shown are the observed flow volumes y_i and the observations transformed into discrete outcomes x_i using Eq. (3). For the 1965 forecasts shown in Fig. 2, Table 1 shows that observed September flow was greater than $y_{0.25}$, and hence, x_i is 0. Therefore, at least for this threshold, the 1965 forecasts for the two scenarios were worse than the climatology forecast; that is, the forecasts indicated a greater chance of observing flows below $y_{0.25}$ than climatology, but the actual outcome was above the threshold. Viewed in isolation, this result might tempt some to conclude that the quality of the ensemble forecasts is poor. While the result is clearly undesirable, a larger verification dataset is needed to make a meaningful assessment of forecast quality. As will be seen, the quality of the forecasts for both scenarios is better than climatology forecasts when viewed in this proper context.

With the forecast-observation pairs $\{(f_i, x_i), i = 1, \ldots, N\}$ for each threshold, the last step is to estimate the DO measures of forecast quality. As the DO approach was originally presented (Murphy and Winkler 1987; Murphy 1997), the forecasts f must take on a set of discrete forecast values. However, forecast probabilities produced by the ensemble prediction system are essentially continuous variables. Recently, Bradley et al. (2003) developed techniques for estimating DO summary measures when probability forecasts f are continuous variables. The appendix summarizes the sample estimators of the forecast quality measures based on these techniques. In the following sections, we describe the forecast quality measures, and present sample results for the Des Moines example.

TABLE 1. Forecast verification dataset for the 0.25 quantile threshold of Sep flow volume. Probabilistic forecasts f_i are made for two scenarios. Scenario 1 assumes that no climate forecast is available. Scenario 2 used a skillful climate forecast to selectively choose ensemble traces for the forecast. The observed outcome x_i is 1 if the observed volume y_i is less or equal to the 0.25 quantile threshold (8603 cfs-days), and 0 if it is greater than the threshold.

Year	Observed volumes (cfs-days) y_i	Observed outcomes x_i	Scenario 1 f_i	Scenario 2 f_i
1949	4913	1	0.863	0.897
1950	17113	0	0.559	0.457
1951	88280	0	0	0
:	:	:	:	:
1962	192390	0	0	0
1963	15586	0	0	0.007
1964	123690	0	0	0
1965	155020	0	0.737	0.432
1966	5364	1	0.539	1
1967	8498	1	0.638	1
1968	13197	0	0.553	0.149
:	:	:	:	:
1987	29767	0	0	0
1988	3633	1	0.792	0.530
1989	8733	0	0.763	0.776
1990	22590	0	0	0
1991	22635	0	0	0
1992	48624	0	0	0
1993	226390	0	0	0
1994	42683	0	0	0
:	:	:	:	:

a. Bias

The first moments of the joint distribution $h(f, x)$ are used to characterize the *bias*. The first moments are the expected value of the forecasts μ_f and the expected value of the observations μ_x. Note that for the case where the forecast-observations pairs are defined by the quantile y_p, the expected value of the forecasts is

$$\mu_f = E[f(y_p)] = E[G_i(y_p)], \tag{4}$$

and the expected value of the observations is

$$\mu_x = E[x(y_p)] = P\{x(y_p) = 1\} = p. \tag{5}$$

Hence, the forecasts are unbiased if the mean forecast probability μ_f equals the unconditional nonexceedance probability p. Figure 4 compares sample estimates of μ_f and p for the range of quantile thresholds y_p for the September volume forecasts. A 1:1 line is shown to help assess whether the sample average is close to p. Both forecast scenarios produce probability distribution forecasts that closely follow the 1:1 line; the results for scenario 2 are only slightly closer to the 1:1 line. Hence, the probability forecasts have very low bias for both forecast scenarios.

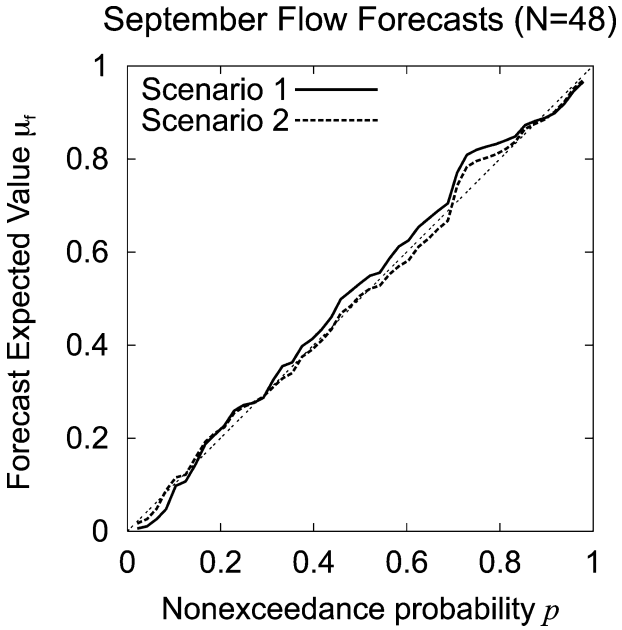

FIG. 4. Bias of probability forecasts for Sep flow volumes. For unbiased forecasts, the estimated forecast expected value μ_f equals the climatological nonexceedance probability p, as shown by the 1:1 line. Solid line shows results for scenario 1 (no climate forecast). Dashed line shows results for scenario 2 (skillful climate forecast).

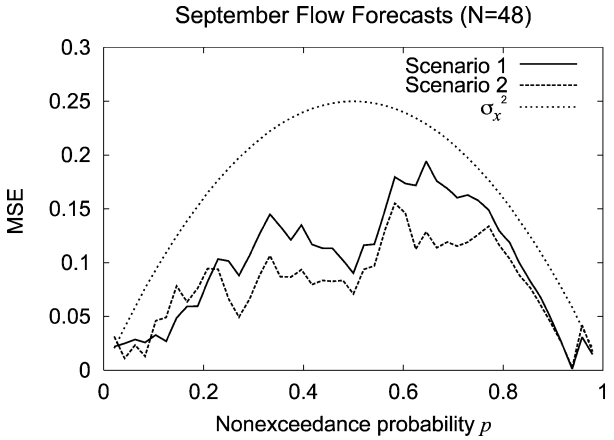

FIG. 5. The MSE of the probability forecasts for Sep flow volumes corresponding to the climatological nonexceedance probability p. Solid line shows results for scenario 1 (no climate forecast). Dashed line shows results for scenario 2 (skillful climate forecast). The dotted line shows σ_x^2, a measure of the inherent uncertainty of the observations.

b. Accuracy

A measure of the *accuracy* of the forecasts is the mean-square error (MSE), which is defined using the joint distribution $h(f, x)$ as

$$\mathrm{MSE}(f, x) = \sum_f \sum_x h(f, x)(f - x)^2. \qquad (6)$$

For probability forecasts, the MSE is also known as the Brier score. Figure 5 shows the MSE for the two forecasting scenarios plotted over the range of thresholds y_p. As was the case for the bias, we use the nonexceedance probability p as an index to the results for the y_p threshold. The shape of the plot is concave, with low MSE for probability forecasts for extreme thresholds (low and high). Also, the MSE is consistently lower for scenario 2, except perhaps at extreme thresholds. Hence, the skillful climate forecasts improve the accuracy of the probability forecasts for nonextreme flow thresholds, but do not improve the forecast accuracy for extreme events.

The MSE results shown in Fig. 5 can be somewhat misleading in that they seem to suggest that the forecasts for extreme thresholds are more accurate (because the MSE is lower). However, the magnitude of the MSE depends on the variability of the forecasts f and the observations x. For binary observations, the variance is simply

$$\sigma_x^2 = p(1 - p). \qquad (7)$$

Note the variance σ_x^2 is a measure of the inherent *uncertainty* of the observations, and is independent of the forecasts that are made (Murphy 1997). Figure 5 also shows this uncertainty measure, which has a symmetrical concave shape with a maximum at $p = 0.5$. Hence, even though the MSE is much higher for thresholds near the median, it is still much lower than its inherent uncertainty σ_x^2.

Because the magnitude of the MSE depends on p, a better way to assess forecast quality over the range of flow thresholds is with a relative measure. A nondimensional measure of the accuracy is the forecast *skill*. The skill of the forecast is the accuracy relative to a reference forecast methodology. The MSE skill score using climatology as a reference (i.e., the forecast probability f is always the climatological mean μ_x, or in this case, p) is

$$\mathrm{SS}_{\mathrm{MSE}}(f, \mu_x, x) = 1 - [\mathrm{MSE}(f, x)/\sigma_x^2]. \qquad (8)$$

For probability forecasts, the $\mathrm{SS}_{\mathrm{MSE}}$ is also known as the Brier skill score. Figure 6a shows the estimated MSE skill scores.

For both scenarios, the forecasts have skill (i.e., are better than climatology forecasts) for all but the most extreme flow thresholds. In particular, the 0.25 quantile threshold forecasts have considerable skill, despite the bad forecasts for September 1965. Note also that the use of a skillful climate forecast in scenario 2 produces only modest improvements in streamflow forecasting skill for the intermediate flow thresholds. Forecasts made without a climate forecast (scenario 1) still have considerable skill, indicating that knowledge of the initial moisture state contributes more to skill than knowledge of the future precipitation category. The relative measure of accuracy also reveals a slight asymmetry in the forecast accuracy; forecasts for lower-flow thresholds tend to have higher (relative) accuracy than for higher-flow thresholds.

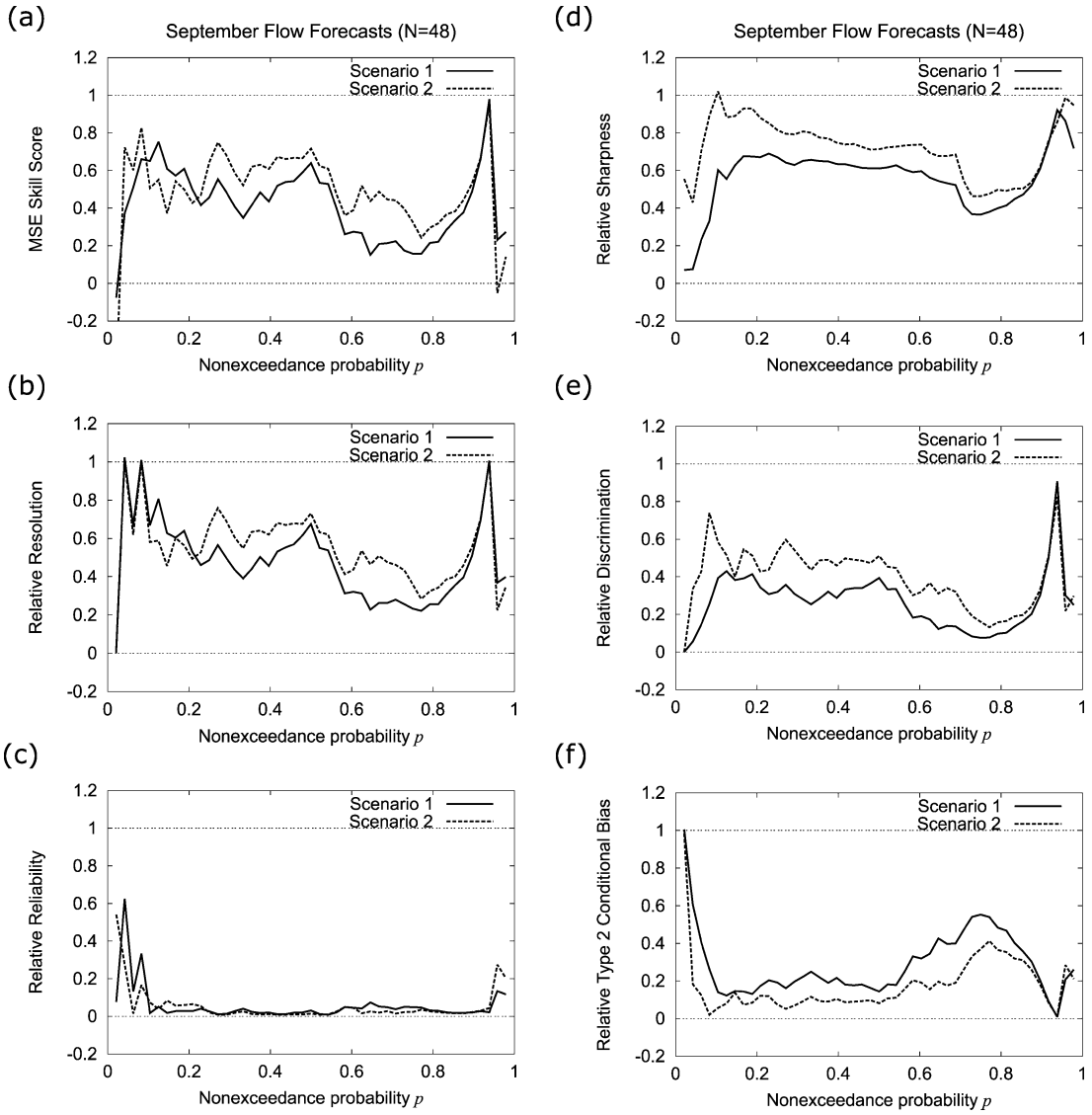

FIG. 6. The DO measures of forecast quality for the Sep flow volumes forecasts: (a) MSE skill score; (b) relative resolution; (c) relative reliability; (d) relative sharpness; (e) relative discrimination; (f) relative type-2 conditional bias. Solid line shows results for scenario 1 (no climate forecast). Dashed line shows results for scenario 2 (skillful climate forecast).

The obvious exception is the spike in the skill score at the 0.938 quantile. The rapid change in skill is a result of random sampling variability. At the spike, there are only 3 instances (out of 48) where the observed flow exceeds the threshold, and the probability forecasts for these 3 cases are remarkably good. Still, by examining the probability forecast skill over the range of thresholds, one can deduce that this result is an artifact of the particular verification data sample. This result emphasizes the fact that forecast quality measures are sample estimates based on a relatively small sample. The implications of sampling variability will be discussed in detail in a later section.

It is important to recognize that even for thresholds where the estimated skill for the two scenarios is similar, the probability forecasts and their attributes are still quite different. The remaining panels in Fig. 6 show attributes contributing to the forecast skill. In the following sections, we describe these attributes with an example for a single threshold, then return to see how these combine to produce forecast skill.

c. Calibration-refinement measures

Using the calibration-refinement (CR) factorization of the joint distribution $h(f, x)$, one can investigate the statistical properties of event occurrences (or nonoccurrences) for specific forecast probabilities. The CR factorization is

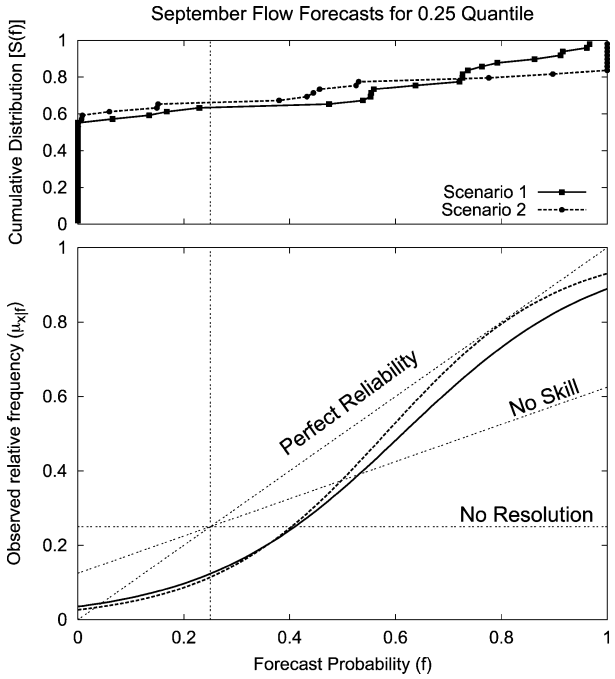

FIG. 7. Reliability diagram for probability forecasts for the 0.25 quantile threshold of Sep flow volume. Solid line shows results for scenario 1 (no climate forecast). Dashed line shows results for scenario 2 (skillful climate forecast). The relative frequency $\mu_{x|f}$ is estimated from the verification data sample using logistic regression (Bradley et al. 2002).

$$h(f, x) = q(x|f)s(f), \tag{9}$$

where $q(x|f)$ is the conditional distribution of the observations given the forecast, and $s(f)$ is the marginal distribution of the forecasts.

Figure 7 shows a reliability diagram, which graphically presents components of the CR decomposition (Wilks 1995). The figure shows the results for the 0.25 quantile threshold for the two forecasting scenarios. The top panel shows the marginal cumulative distribution of the forecasts $S(f)$. In most reliability diagrams, the marginal density $s(f)$ is usually plotted as a histogram. However, because f has a mixed distribution, with mass points at 0 and 1, the cumulative distribution provides a clearer graphical display.

The *sharpness* is defined as the degree to which probability forecasts approach 0 and 1. Ideally, forecasts will be perfectly sharp, consisting only of 0 and 1. The marginal distribution $S(f)$ in Fig. 7 shows that the forecasts are sharp for both scenarios; over half of the forecasts are 0. However, the forecasts for scenario 2 are sharper, as a forecast of 1 is also issued about 15% of the time. This result is not unexpected. Note that the hypothetical precipitation-category forecasts used in scenario 2 are perfectly sharp; utilizing only those traces corresponding to the forecast category results in slightly sharper streamflow forecasts.

The bottom panel of Figure 7 indicates the relative frequency of event occurrence $\mu_{x|f}$ for a given forecast probability f. The *reliability* describes the conditional bias of the forecasts. For conditionally unbiased forecasts (i.e., perfect reliability), the relative frequency $\mu_{x|f}$ equals f. For both scenarios, events (i.e., volumes less than the 0.25 quantile) occur less frequently than their forecast probability of occurrence. In other words, the scenarios tend to overforecast event occurrences, resulting in a small conditional bias. An integrated measure of the reliability is

$$\text{REL} = E_f(\mu_{x|f} - f)^2. \tag{10}$$

For perfectly reliability forecasts, REL is zero. For a climatology forecast, REL is equal to σ_x^2.

The *resolution* describes the degree to which the observations for a specific forecast f differ from the unconditional mean (or climatological probability). If the relative frequency of event occurrences $\mu_{x|f}$ are the same, regardless of the forecast f, the forecasts have no resolution (see line in Fig. 7). For both scenarios, the forecasts have good resolution. When probability forecasts f of the event $\{Y_i \le y_{0.25}\}$ are near 0, the event rarely occurs, and when forecasts are near 1, the event usually occurs. An integrated measure of the resolution is

$$\text{RES} = E_f(\mu_{x|f} - \mu_x)^2. \tag{11}$$

For forecasts with no resolution, RES is zero. The maximum resolution is obtained for perfectly reliability forecasts ($\mu_{x|f} = f$), where RES is σ_x^2.

The connection between the reliability and resolution of the forecasts, and the MSE, can be seen through a decomposition of the MSE. Conditioning on the forecast leads to the so-called CR decomposition (Murphy 1996):

$$\text{MSE}_{\text{CR}}(f, x) = \sigma_x^2 + \text{REL} - \text{RES}. \tag{12}$$

Substituting (12) into Eq. (8) for the skill score shows that

$$\text{SS}_{\text{CR}}(f, \mu_x, x) = \frac{\text{RES}}{\sigma_x^2} - \frac{\text{REL}}{\sigma_x^2}, \tag{13}$$

where the first term on the right-hand side is a relative measure of the resolution, and the second is a relative measure of the reliability. In other words, through the CR decomposition, skill can be thought of as the difference between the resolution of the forecasts and their conditional bias (the reliability).

The relative resolution and reliability of the September volumes forecasts are summarized over the entire range of flow quantile thresholds in Figs. 6b,c. Clearly, the spike in skill is related to the spike in the resolution of the forecasts. With the exception of the spike, the resolution of the forecasts tends to be asymmetric; forecasts for lower-flow thresholds have good resolution, but those for higher thresholds have poorer resolution. The asymmetry of the resolution (and skill) is not too surprising for streamflow forecasts. The persistence of dry conditions produces low-flow volumes, thus the

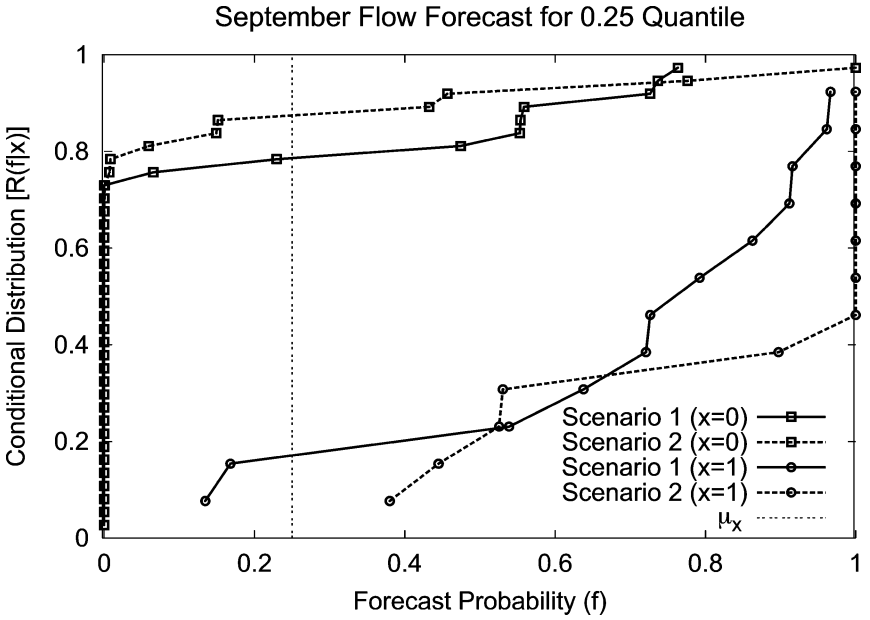

FIG. 8. Discrimination diagram for probability forecasts for the 0.25 quantile threshold of Sep flow volume. Solid line shows results for scenario 1 (no climate forecast). Dashed line shows results for scenario 2 (skillful climate forecast).

probable occurrence of low flow is highly predictable based on the moisture state at the time the forecast is made. The occurrence of high-flow volumes depends on the precipitation over the month. Hence, knowledge of the September precipitation category in scenario 2 leads to better resolution (at least for quantile thresholds between about 0.2 and 0.9).

Note, however, that the high resolution of the probability forecasts does not translate into significantly better skill for very low flows (see Figs. 6a–c for p of less than 0.1). Conditional biases (poor reliability) offset the high resolution. The reason for conditional biases is the poor simulation of low flows by the forecast model (e.g., see low flows in Fig. 2). Hence, this analysis reveals that there is room for improvement for low flows (and to a lesser extent, very high flows). For example, conditional biases could be reduced or eliminated by adjustment (calibration) of the probability forecasts (e.g., Stewart and Reagan-Cirincione 1991; Krzysztofowicz and Sigrest 1999). In contrast, the forecasts for nearly all other thresholds are already reliable, so improvements through calibration are not possible.

d. Likelihood-base-rate measures

Using the likelihood-base-rate (LBR) factorization of the joint distribution $h(f, x)$, one can evaluate the statistical properties of the forecasts that are made when events occur (or do not occur). The LBR factorization is

$$h(f, x) = r(f|x)t(x), \tag{14}$$

where $r(f|x)$ is the conditional distribution of the forecasts given the observation, and $t(x)$ is the marginal distribution of the observations.

Figure 8 shows a discrimination diagram, which graphically presents components of the LBR decomposition (Wilks 1995). The figure shows the results for the 0.25 quantile for the two forecasting scenarios. For each scenario, the conditional cumulative distributions $R(f|x = 0)$ and $R(f|x = 1)$ are shown. Although discrimination diagrams are usually plotted with the conditional densities $r(f|x)$, we again use the cumulative distributions to more clearly portray the mixed distribution for f.

The *discrimination* describes the degree to which the forecasts differ for a specific observation x. Note that the cumulative distributions for $x = 0$ and $x = 1$ are very different, suggesting the forecasts have good discrimination. When the event does not occur ($x = 0$), a forecast of 0 was issued over 70% of the time. When the event occurs ($x = 1$), the forecasts are much closer to 1. Scenario 2 forecasts have better discrimination. Note that for $x = 1$, a forecast of 1 was issued about 55% of the time for scenario 2, whereas forecast probabilities as high as 1 were not issued for scenario 1. An integrated measure of the discrimination is

$$\text{DIS} = E_x(\mu_{f|x} - \mu_f)^2, \tag{15}$$

where $\mu_{f|x}$ is the expected value of the forecasts conditioned on the observation. If forecasts have no discrimination, then $\mu_{f|x}$ is the same (and equal to μ_f) regardless of whether the event occurs or not, and DIS is zero.

Similar to the reliability for the CR factorization, the *type-2 conditional bias* describes the bias conditioned on the observation. An integrated measure of the type-2 conditional bias is

$$B_2 = E_x(\mu_{f|x} - x)^2. \quad (16)$$

Note that B_2 is zero (no conditional bias) only in the case of perfect forecasts.

The connection between the discrimination and type-2 conditional bias of the forecasts, and the MSE, can be seen through another decomposition of the MSE. Conditioning on the observations leads to the LBR decomposition (Murphy 1996):

$$\mathrm{MSE}_{\mathrm{LBR}}(f, x) = \sigma_f^2 + B_2 - \mathrm{DIS}, \quad (17)$$

where the first term in the decomposition is a measure of the sharpness of the forecasts. Substituting (17) into Eq. (8) for the skill score shows that

$$\mathrm{SS}_{\mathrm{LBR}}(f, \mu_x, x) = 1 - \frac{\sigma_f^2}{\sigma_x^2} + \frac{\mathrm{DIS}}{\sigma_x^2} - \frac{B_2}{\sigma_x^2}, \quad (18)$$

where the second term on the right-hand side is the relative sharpness, the third is the relative discrimination, and the fourth is the relative type-2 conditional bias.

There is a complex interrelation between sharpness, discrimination, and type-2 condition bias. As an example, for the case of a climatology forecast (i.e., $f = \mu_x$ in all cases), the sharpness of the forecasts is zero. However, the discrimination is also zero, and the type-2 conditional bias is maximized at σ_x^2 (resulting in a skill of zero). Hence, for forecasts to possess skill, they must be sharp and have discrimination.

The relative sharpness, discrimination, and type-2 conditional bias of the September volumes forecasts are summarized over the entire range of flow quantile thresholds in Figs. 6d–f. The forecasts for scenario 2 are generally sharper, have higher discrimination, and have lower type-2 conditional biases. This combination results in the higher skill for the intermediate-flow thresholds. Still, the results for nonexceedance probabilities between 0.1 and 0.2 are noteworthy. The forecasts for scenario 2 are much sharper than those for scenario 1. That is, the forecasts made based on a skillful climate forecast (scenario 2) go ''out on a limb'' more, with forecast probabilities tending to be closer to 0 or 1 for the event occurrence. However, these sharper forecasts do not have significantly better discrimination. Hence, for the given increase in sharpness, the improvement in discrimination is insufficient to improve the overall accuracy (skill) of the forecasts. This helps explain why the skill for scenario 1 is better in this range.

The results also highlight that skill is an incomplete measure of the quality of the forecasting system. Note that near $p = 0.22$, the skill for the two scenarios is similar (Fig. 6a). However, the nature of the forecasts are clearly very different; the forecasts made using a skillful climate forecast (scenario 2) are sharper with slightly more discrimination than those made without a climate forecast (scenario 1). Such differences can affect the value of the forecasts in decision making, even though the overall accuracy (skill) of the forecasts is similar.

TABLE 2. Traditional verification measures. The Rmse and MSE skill score are for the ensemble mean. The RPSS is for probabilistic forecasts of tercile categories, defined by the empirical of tercile categories, defined by the empirical distribution of Sep flow volumes.

Forecast scenario	Rmse (cfs-days)	MSE skill score	Tercile category RPSS
	Entire 48-yr sample		
Scenario 1	129 539	−4.67	0.28
Scenario 2	96 173	−2.13	0.50
	Sample without 1962 and 1993		
Scenario 1	44 216	−0.10	—
Scenario 2	29 394	0.51	—

5. Discussion

In this section, we discuss several issues related to the proposed methodology for ensemble forecast verification. First, we compare the DO approach results with those from traditional measures used in ensemble verification. We then discuss how the use of forecasts affects the choice of verification methodology. And finally, we examine the effect of sampling uncertainty in hydrologic forecast verification.

a. Traditional measures

Traditional approaches to forecast verification often involve calculation of various summary measures meant to assess specific attributes of the forecasts (Wilks 1995). To assess whether ensemble forecasts have accuracy (or skill), a very common approach is to compare the ensemble mean with the observation using the root-mean-square error (rmse), or its related MSE skill score. (Day et al. 1992; Hou et al. 2001; Duce 2001; among others). Table 2 summarizes these measures for the two scenarios. Although both measures are better for ensemble forecasts made using a skillful climate forecast (scenario 2), the MSE skill scores are much less than zero. This would seem to suggest that the ensemble forecasts are quite poor. However, a significant problem with these measures is that they heavily penalize large errors for individual forecasts. In this case, the errors for just two forecasts (1962 and 1993) account for almost 90% of the MSE. These errors are caused by a similar situation. Both years experienced very wet summers, followed by a drier than average September; with high moisture conditions at the start of September, most ensemble traces were significantly higher than the observed flow (see Fig. 3 for 1962). Table 2 shows that excluding these two years results in a much lower rmse;

the MSE skill score is still below zero for scenario 1, but the results for scenario 2 now suggest that the ensemble mean has significant skill.

To assess probabilistic forecasts from ensemble predictions for multiple-ordered categories, a common measure of skill is the ranked probability skill score (RPSS; Wilks 1995). One advantage of this measure is that forecast probabilities in categories farther away from the observed category are penalized greater. Note that the RPSS is mathematically related to the MSE in Eq. (6). If $\{y_j, j = 1, \ldots, M\}$ are thresholds defining the boundary between the $(M + 1)$ flow categories, then

$$\text{RPSS} = 1 - \frac{\sum_{j}^{M} \text{MSE}_j}{\sum_{j}^{M} p_j(1 - p_j)}, \tag{19}$$

where MSE_j is the MSE for the jth threshold, and p_j is the nonexceedance probability for the jth threshold. Therefore, the RPSS can be computed from the results shown in Fig. 5 for any set of categories. Table 2 summarizes the RPSS of probability forecast for three flow categories, defined by the terciles of the distribution of observed September flow volumes. The RPSS shows that forecasts for tercile categories are skillful for both scenarios. Still, the forecasts made with a skillful climate forecast (scenario 2) are clearly much better.

Although the conclusion by these traditional measures is that streamflow forecasts made using a skillful climate forecast (scenario 2) are better, it is important to contrast this result with the insights gained by using the DO approach. Specifically, the traditional measures do not reveal that forecast probabilities for extremes have almost no skill, or the forecast for lower-flow events are generally better than those for higher-flow events. They also do not reveal the reason why the forecasts for scenario 2 are more skillful. Perhaps more importantly, the traditional measures do not indicate that the improvement for scenario 2 only occurs for intermediate-flow thresholds; probability forecast for flow extremes are not improved by the skillful climate forecasts. Because extreme events are usually of great concern in water management, insights on the quality of forecasts for extreme events can be very valuable to water managers.

b. Verification methodologies

The verification methodology presented for ensemble streamflow predictions involves transforming the probability distribution forecasts into probability forecasts for discrete events, then evaluating the forecast quality. The approach implicitly assumes that one is interested in assessing the quality of the probability forecasts over the entire range of possible events. Another objective is to diagnose the attributes contributing to the forecast skill. Therefore, the verification results might be very useful to a forecaster, who wishes to evaluate the overall quality of the forecast system, and find ways to improve it.

However, from a forecast user's perspective, the verification approach should be tailored to the decision problem where the forecast will be used. Clearly, certain users will find different verification approaches more helpful. For example, for a water manager who considers only a single drought or flood threshold for decision making, the DO verification approach could be applied to the probability forecasts for this one event (threshold). Still, other verification measures, such as those based on the relative operating characteristics (ROC), can also provide valuable information for decision making in this situation (Mason 1982; Centor 1991; Harvey et al. 1992; Mason and Graham 1999, 2002). For some users, assessing the quality of probability forecasts for specific flow ranges is important (e.g., terciles, or critical flow ranges defined by system constraints or regulations). Here, the ensemble predictions would be used to assign forecast probabilities to each discrete category. A DO verification approach is still applicable in this situation, but other verification measures may also be useful.

c. Sampling uncertainty

The single most significant obstacle to verification of ensemble streamflow predictions is their inherently small sample sizes. In the example application for September volumes, there is only one forecast made each year (for a given lead time). Most flow records have about 50 or fewer years of data, so the size of the verification dataset is severely limited. In contrast, sample sizes for most meteorological forecasts are an order of magnitude or more larger. Because measures computed using the verification dataset are sample estimates of the forecast quality attributes, small sample sizes lead to high sampling variability (Schwartz 1992). Furthermore, the sampling variability strongly depends on the threshold. The uncertainty for intermediate thresholds, where exceedances (or nonexceedances) occur relatively frequently, is much less than that for extreme thresholds, where exceedances (or nonexceedances) rarely occur.

Using Monte Carlo simulation, Bradley et al. (2003) estimated standard errors of DO measures for the Des Moines example presented here, using a stochastic model developed to represent the September volume forecasts (Hashino et al. 2002). For a threshold with a nonexceedance probability of 0.25, the standard errors in estimates of MSE, REL, and RES, for a sample size of 50, ranged from $\pm 5\%$ to 20%. However, for a threshold with a nonexceedance probability of 0.05, the standard errors ranged from $\pm 60\%$ to 110%. Clearly, for the extreme thresholds shown in Fig. 6, the uncertainty of the forecast quality measures is quite large. One benefit of evaluating the forecast quality measures at so many thresholds is that sampling variability and artifacts can been seen; Fig. 6 shows higher variations at the extremes

(e.g., the spike), and smoother variations at intermediate thresholds. However, due to sampling uncertainty, quantitative inferences on forecast quality at the extremes should not be made. Instead, the results for the extremes should be interpreted only in a comparative sense; the results document differences in how well (or poorly) the two forecast scenarios performed for this particular verification sample.

In some cases, it may be possible to increase the sample size, and reduce sampling uncertainties, by pooling forecasts made at different times into a single verification dataset. However, this approach assumes that the joint distributions of the forecasts and observations are the same for each time period, and the forecasts issued at different times are (statistically) independent. If these assumptions do not hold, the reduced sampling variability from pooling the forecasts may be offset by biases in the forecast quality estimates.

Regardless, additional work is needed to assess the uncertainty of forecast quality estimates. Ideally, estimates of the uncertainty of each measure could be made, and the standard errors or confidence limits plotted along with the sample values. Exact or approximate analytical expressions for the standard errors of the forecast quality measures based on sampling theory (e.g., Schwartz 1992; Carpenter and Georgakakos 2001) are being explored for this purpose. Due to the small sample sizes, assessment of uncertainty in the forecast quality measures needs to be an integral part of ensemble forecast verification.

6. Summary and conclusions

Forecasts from ensemble prediction systems can be issued in several ways. For ensemble streamflow predictions, they are usually issued as probability distribution forecasts for a continuous flow variable. A distributions-oriented (DO) approach is presented to verify ensemble forecasts in this form. Flow thresholds are used to transform the ensemble forecasts into a probability forecast for a dichotomous event. The event is defined as the occurrence of flow below the threshold; the probability forecast is the probability of the event occurrence. The DO approach is then applied to a verification dataset constructed for the threshold, which is based on an archive of ensemble forecasts and observations. The process is repeated for thresholds covering the range of possible outcomes; relative measures are then used to examine aspects of forecast quality over this range.

The application of the DO approach illustrates the insights gained from this verification approach. For September volume forecasts for the Des Moines experimental system, the skill of ensemble forecasts was seen to vary over the range of flow outcomes. The skill was highest for intermediate flows; the forecasts had very little skill for low- and high-flow extremes. Forecasts were compared for two scenarios. One uses all the ensemble traces to make the forecast; the other uses a hypothetical skillful climate forecast to select traces for making the forecast. The verification results showed that a skillful climate forecast improved the skill of the streamflow forecasts for intermediate-flow thresholds. Significantly though, the forecasts for low- and high-flow extremes were not improved. This kind of information can be extremely important for water managers, because extreme events are often the most significant for decision making.

The DO approach also provides information to help diagnose the reason for the skill in probability distribution forecasts. For the Des Moines forecasts, the reason that ensemble forecasts made using the hypothetical climate forecast had more skill for intermediate flows is that they had better resolution. The forecasts were also sharper, even for low flows, and had better discrimination (but still not enough to improve the accuracy for low flows). More traditional approaches to verification also suggested that ensemble forecasts made using the hypothetical climate forecast had higher skill. However, such measures would not inform the user that the improvement is only for intermediate flows, and not flow extremes, or the reason for the improvement in skill.

The DO approach to ensemble streamflow forecast verification is designed to provide a diagnostic evaluation of the quality of probability forecasts over the range of possible flow outcomes. Although the methodology is presented for streamflow forecasts, the approach would also be useful for verification of meteorological ensemble forecasts. However, there are unique challenges for hydrologic applications. Verification datasets for ensemble streamflow predictions typically have small sample sizes. Because forecast quality measures are sample estimates, small sample sizes result in large sampling uncertainty. Additional work is needed to assess uncertainty of forecast quality measures as part of the verification process.

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

APPENDIX

Estimating Forecast Quality Measures

For probability forecasts of a dichotomous event, Bradley et al. (2003) derived simplified expressions for the distributions-oriented forecast quality measures. Most of the forecast quality measures can be estimated

analytically using sample moments of forecasts and observations from the verification data sample. Other measures require a statistical model for estimation. The sample estimators for these measures are briefly outlined below.

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 sample estimators used to evaluate bias are the traditional sample means $\hat{\mu}_f$ and $\hat{\mu}_x$. A sample estimator for the MSE is

$$\widehat{\mathrm{MSE}}(f, x) = \frac{1}{N}\sum_{i}^{N}(f_i - x_i)^2. \quad \text{(A1)}$$

The likelihood-base-rate measures depend on just the conditional and unconditional means. The sample estimator of the discrimination (DIS) is

$$\widehat{\mathrm{DIS}} = (1 - \hat{\mu}_x)(\hat{\mu}_{f|x=0} - \hat{\mu}_f)^2 + \hat{\mu}_x(\hat{\mu}_{f|x=1} - \hat{\mu}_f)^2; \quad \text{(A2)}$$

and the sample estimator for the type-2 conditional bias (B_2) is

$$\hat{B}_2 = (1 - \hat{\mu}_x)\hat{\mu}^2_{f|x=0} + \hat{\mu}_x(\hat{\mu}_{f|x=1} - 1)^2, \quad \text{(A3)}$$

where $\hat{\mu}_{f|x=0}$ and $\hat{\mu}_{f|x=1}$ are the sample conditional means. To estimate the conditional means, the verification data sample is partitioned into two sets, one for the case where the event does not occur ($x = 0$), and the other for the case where the event occurs ($x = 1$). The conditional means are then estimated by the subsample means.

The calibration-refinement measures depend on higher-order moments. The sample estimator for the reliability (REL) is

$$\widehat{\mathrm{REL}} = \hat{E}_f(\mu^2_{x|f}) - 2\hat{\mu}_x\hat{\mu}_{f|x=1} + \hat{\sigma}^2_f + \hat{\mu}^2_f, \quad \text{(A4)}$$

and the sample estimator for the resolution (RES) is

$$\widehat{\mathrm{RES}} = \hat{E}_f(\hat{\mu}^2_{x|f}) - 2\hat{\mu}_x\hat{\mu}_f + \hat{\mu}^2_f, \quad \text{(A5)}$$

where $\hat{\sigma}^2_f$ is the sample variance of the forecasts, and is estimated by

$$\hat{\sigma}^2_f = \frac{1}{N}\sum_{i}^{N}(f_i - \hat{\mu}_f)^2. \quad \text{(A6)}$$

The term $\hat{E}_f(\hat{\mu}^2_{x|f})$ cannot be simplified if the form of the marginal distribution for f is unknown. Bradley et al. (2003) show two approaches for estimating the term. One approach is to use logistic regression to estimate $\mu_{f|x}$ from the verification data sample $\{(f_i, x_i), i = 1, \ldots, N\}$. With the logistic regression, the sample estimate for the term is

$$\hat{E}_f(\hat{\mu}^2_{x|f}) = \frac{1}{N}\sum_{i}^{N}\hat{\mu}^2_{x|f_i}, \quad \text{(A7)}$$

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

Finally, the forecast quality measures are nondimensionalized using the uncertainty measure σ^2_x. Because x is a Bernoulli random variable, the sample estimator for the variance is simply

$$\hat{\sigma}^2_x = \hat{\mu}_x(1 - \hat{\mu}_x). \quad \text{(A8)}$$

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Braatz, D. T., J. B. Halquist, R. J. Warvin, J. Ingram, J. J. Feldt, and M. S. Longnecker, 1997: NWS hydrologic products and services: Moving from the traditional to the technically advanced. Preprints, *13th Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology,* Long Beach, CA, Amer. Meteor. Soc., J63–J69.

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.

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

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

——, A. Witt, and M. D. Eilts, 1997: Verification of public weather forecasts available via the media. *Bull. Amer. Meteor. Soc.,* **78,** 2167–2177.

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,** 148–175.

Centor, R. M., 1991: Signal detectability—The use of ROC curves and their analyses. *Med. Decis. Making,* **11,** 102–106.

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

Connelly, B. A., D. T. Braatz, J. B. Halquist, M. M. DeWeese, L. Larson, and J. J. Ingram, 1999: Advanced hydrologic prediction system. *J. Geophys. Res.,* **104** (D16), 19 655–19 660.

Croley, T. E., 1993: Probabilistic Great Lakes hydrology outlooks. *Water Resour. Bull.,* **29,** 741–753.

——, 2000: *Using Meteorology Probability Forecasts in Operational Hydrology.* American Society of Civil Engineers, 206 pp.

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

——, L. Brazil, C. S. McCarthy, and D. P. Laurine, 1992: Verification of the National Weather Service Extended Streamflow Prediction procedure. *Proc. 28th Conf. and Symp. on Managing Water Resources during Global Change,* Reno, NV, American Water Resources Association, 163–172.

Donigian, A. S., B. R. Bicknell, and J. C. Imhoff, 1995: Hydrological Simulation Program—Fortran (HSPF). *Computer Models of Watershed Hydrology,* V. P. Singh, Ed., Water Resources Publications, 395–442.

Doswell, C. A., and H. E. Brooks, 1998: Budget cutting and the value of weather services. *Wea. Forecasting,* **13,** 206–212.

Duce, D. J., 2001: Insights from a history of seasonal inflow forecasting with a conceptual hydrologic model. *J. Hydrol.,* **249,** 102–112.

Faber, B. A., and J. R. Stedinger, 2001: Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts. *J. Hydrol.,* **249,** 113–133.

Georgakakos, A. P., H. Yao, M. G. Mullusky, and K. P. Georgakakos, 1998: Impacts of climate variability on operational forecasts and management of the upper Des Moines River basin. *Water Resour. Res.,* **34,** 799–821.

Georgakakos, K. P., D.-H. Bae, M. G. Mullusky, and A. P. Georgakakos, 1995: Hydrologic variability in midwestern drainage basins: Diagnosis, prediction and control. *Preparing for Global Change: A Midwestern Perspective*, E. Folk, Ed., SPB Academic Publications, 61–90.

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

Hamlet, A. F., and D. P. Lettenmaier, 1999: Columbia River streamflow forecasting based on ENSO and PDO climate signals. *J. Water Resour. Plann. Manage.*, **125,** 333–341.

——, D. Huppert, and D. P. Lettenmaier, 2002: Economic value of long-lead streamflow forecasts for the Columbia River basin. *J. Water Resour. Plann. Manage.*, **128,** 91–101.

Hartmann, H. C., T. C. Pagano, S. Sorooshian, and R. Bales, 2002: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.*, **83,** 683–698.

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Moss, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120,** 863–883.

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2002: Verification of probabilistic streamflow forecasts. IIHR Rep. 427, IIHR-Hydroscience and Engineering, Iowa City, IA, 125 pp.

Hay, L. E., M. P. Clark, R. L. Wilby, W. J. Gutowski, G. H. Leavesley, Z. Pan, R. W. Arritt, and E. S. Takle, 2002: Use of regional climate model output for hydrologic simulations. *J. Hydrometeor.*, **3,** 571–590.

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

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

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Sciences*. John Wiley & Sons, 240 pp.

Krzysztofowicz, R., 2001: The case for probabilistic forecasting in hydrology. *J. Hydrol.*, **249,** 2–9.

——, and A. A. Sigrest, 1999: Calibration of probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **14,** 427–442.

——, and H. D. Herr, 2001: Hydrologic uncertainty processor for probabilistic river stage forecasting: Precipitation-dependent model. *J. Hydrol.*, **249,** 46–68.

Livezey, R. E., 1990: Variability of skill of long-range forecasts and implications for their use and value. *Bull. Amer. Meteor. Soc.*, **71,** 300–309.

Markus, M., E. Welles, and G. 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.

Mason, I. B., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.*, **30,** 291–303.

Mason, S. J., and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14,** 713–725.

——, and ——, 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.

Mjelde, J. W., D. S. Peel, S. T. Sonka, and P. J. Lamb, 1993: Characteristics of climate forecast quality—Implications for economic value to midwestern corn producers. *J. Climate*, **6,** 2175–2187.

Murphy, A. H., 1993: What is a good forecast: An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8,** 281–293.

——, 1996: General decompositions of MSE-based skill scores: Measures of some basic aspects of forecast quality. *Mon. Wea. Rev.*, **124,** 2353–2369.

——, 1997: Forecast verification. *Economic Value of Weather and Climate Forecasts*, R. W. 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,** 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.

Pagano, T. C., H. C. Hartmann, and S. Sorooshian, 2002: Factors affecting seasonal forecast use in Arizona water management: A case study of the 1997–98 El Niño. *Climate Res.*, **21,** 259–269.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128,** 747–774.

Perica, S., 1998: Integration of meteorological forecasts/climate outlooks into extended streamflow prediction (ESP) systems. Preprints, *14th Conf. on Probability and Statistics in the Atmospheric Sciences*, Phoenix, AZ, Amer. Meteor. Soc., 130–133.

Roebber, P. J., and L. F. Bosart, 1996: The complex relationship between forecast skill and forecast value: A real-word analysis. *Wea. Forecasting*, **11,** 544–559.

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

Smith, J. A., G. N. Day, and M. D. Kane, 1992: Nonparametric framework for long-range streamflow forecasting. *J. Water Resour. Plann. Manage.*, **118,** 82–92.

Stern, H., 2001: The application of weather derivatives to mitigate the financial risk of climate variability and extreme weather events. *Aust. Meteor. Mag.*, **50,** 171–182.

Stewart, T. R., and P. Regan-Cirincione, 1991: Coefficients for debiasing forecasts. *Mon. Wea. Rev.*, **119,** 2047–2051.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

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

——, 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.*, **8,** 209–219.

Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.*, **127,** 956–970.

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.

Yao, H., and K. P. Georgakakos, 2001: Assessment of Folsom Lake response to historical and potential future climate scenarios: 1. Reservoir management. *J. Hydrol.*, **249,** 176–196.

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