Bayesian Correlation Score: A Utilitarian Measure of Forecast Skill

ROMAN KRZYSZTOFOWICZ

Department of Systems Engineering, University of Virginia, Charlottesville, Virginia

(Manuscript received 21 September 1990, in final form 23 March 1991)

ABSTRACT

From the theory of sufficient comparisons of experiments, a measure of skill is derived for categorical forecasts of continuous predictands. Called Bayesian correlation score (BCS), the measure is specified in terms of three parameters of a normal-linear statistical model that combines information from two sources: a prior (climatological) record of the predictand and a verification record of forecasts. Three properties characterize the BCS: (i) It is meaningful for comparing alternative forecasts of the same predictand, as well as forecasts of different predictands, though in a limited sense; (ii) it is interpretable as correlation between the forecast and the predictand; and, most significantly, (iii) it orders alternative forecast systems consistently with their ex ante economic values to rational users (those who make decisions by maximizing the expected utility of outcomes under the posterior distribution of the predictand). Thus, by maximizing the BCS, forecasters can assure a utilitarian society of the maximum potential economic benefits of their forecasts.

1. Introduction

a. Skill measures

The performance of a forecast system is often characterized in terms of two attributes of forecasts: the lead time and the skill. The *lead time* of a forecast is the time interval elapsed from the instant up to which the data for preparing the forecast have been observed to the earliest instant at which the actual state of the predictand could be observed. The *skill* of a forecast lacks a unique definition. Instead, it is customarily defined in terms of some measure designed to capture one's intuitive notion of the forecast skill, quality, goodness, or informativeness—the attributes which we shall view here as synonyms.

Among frequently encountered skill measures for categorical forecasts of continuous predictands, one finds (Murphy and Daan 1985; Murphy and Epstein 1989): (i) a metric of distance between the forecasted and the actual state, such as the mean square error or the mean absolute error; (ii) a statistic of association, such as covariance or correlation; and (iii) a normalized metric of distance—a skill score. For a fixed lead time, a skill score enables one to compare the performance of different forecast systems, or to gauge the performance of a given system relative to two limiting cases: a perfect forecast, specifying the actual state; and a naive forecast, specifying a mean of the state (as in a climatological forecast) or an extrapolation of the latest state observation (as in a persistence forecast).

Corresponding author address: Professor Roman Krzysztofowicz, Department of Systems Engineering, University of Virginia, Thornton Hall, Charlottesville, VA 22901.

Popular skill measures have two snags. First, different skill measures may imply different preference orderings of alternative forecast systems, and there are no normative principles for rationally reconciling such incoherences. Second, these skill measures have no assured relevance to the actual or potential uses of forecasts in decision making: planning, management, and operation of weather-sensitive activities. That is to say, a modification of the forecast system perceived as an improvement by forecasters on the basis of a skill measure does not necessarily imply an improvement to the users. The converse is also true: a deterioration of the skill measure does not necessarily imply inferior forecasts to their users.

b. Utilitarian criterion

In a utilitarian society, the ultimate measure for evaluating and comparing systems producing forecasts for public use, at a fixed cost, is the ex ante economic value of forecasts. This value is a function of both the lead time and the skill of forecasts. When comparing alternative forecast systems, or alternative improvements of a given system at a specified cost, one should choose the alternative assuring society of the highest economic value. This value is equal to the sum of the economic values accrued by all forecast users. Inasmuch as each user has his own utility function for outcomes of his activity, the utilitarian approach necessitates performing a decision analysis for each individual user. This is an enormous undertaking. In practice, the decision analysis is performed for a few representative users, and the results are extrapolated to a population (Krzysztofowicz and Davis 1983); for agricul-

tural users, the decision analysis is performed per acre of crops, so the results can be extrapolated to a region (Katz et al. 1982). But even such approximate analyses may be too costly for routine evaluations of forecast systems.

Recent developments in Bayesian decision theory offer a promise. They have led to operationalizations of the binary relation of sufficiency that enables one to order alternative forecast systems according to the utilitarian criterion, yet without the necessity of estimating the economic values themselves. A practical procedure for establishing such a preference order is the subject of this article.

c. *Overview*

The procedure is developed for forecasts which are categorical, in the sense that they do not convey the degree of uncertainty, but specify only point estimates of continuous predictands such as temperature or precipitation amount. Two measures of skill, sufficient for ordering alternative forecasts consistently with their ex ante economic values to the users, are derived in progression: the sufficiency characteristic (SC), for ordering forecasts of the same predictand; and the standardized sufficiency characteristic (SSC), for ordering forecasts of different predictands. The SSC implies the SC. Next, a relationship is derived between the SSC and a Bayesian estimator of correlation between the forecast and the predictand. Taking advantage of this relationship, we propose the Bayesian correlation score (BCS) as a utilitarian measure of forecast skill. A plot of the BCS versus the lead time of the forecast summarizes the performance tradeoffs available to the users.

The procedure is illustrated with numerical examples for forecasts of runoff volumes during the snowmelt season prepared by the Soil Conservation Service and the National Weather Service. The snowmelt season covers several months, depending on the geographic location: from January to May in Arizona, from April to September in Montana. Forecasts are issued at the beginning of each month from January through May for 533 river gauging stations in 11 western states. The first forecast is thus prepared with the lead time of 5–9 months, depending upon the location. Each subsequent forecast is a revision of the earlier forecast. An example referred to throughout the article is based on the verification record for a gauging station on the Weiser River near Weiser, Idaho. The snowmelt season covers four months from April through July; the unit in which the runoff is expressed, both forecasted and actual, is the percentage of the 25-yr (1961–1985) mean seasonal runoff volume (414 100 acre-feet).

2. Models of predictand and forecasts

We begin with a description of models that provide statistics necessary for defining the skill measures. There are two models: the model of a predictand, called the prior distribution; and the model of forecasts, called the likelihood function.

a. *Prior distribution*

Let ω denote a continuous predictand. The uncertainty about ω that exists before the preparation of any forecast is described by a prior probability density function g. Suppose this density is normal with the mean and variance:

$$E(\omega) = M,$$
$$\text{var}(\omega) = S^2. \tag{1}$$

The prior parameters M and S may be estimated from a climatological record. In principle, this should be the longest record available that satisfies the hypothesis that the process generating ω has been stationary.

b. *Likelihood function*

A categorical forecast specifies a point estimate x of ω. Let $f(x|\omega)$ denote the relationship between state ω and its forecast x. For a fixed ω, the function $f(\cdot|\omega)$ represents the probability density function of the forecast x. For a fixed x, the function $f(x|\cdot)$ represents the likelihood function of the state ω. The likelihood functions characterize the predictive capabilities of the forecaster from the viewpoint of a user.

We shall concentrate on a particular form of f, arising when the relationship between ω and x is modeled in terms of a linear equation

$$x = a\omega + b + \theta, \tag{2}$$

where a and b are fixed parameters, and θ is a random variable, stochastically independent of ω, and having a normal density k with moments

$$E(\theta) = 0,$$
$$\text{var}(\theta) = \sigma^2. \tag{3}$$

Consequently, the likelihood function is specified by the relation: $f(x|\omega) = k(x - a\omega - b)$. It follows that $f(\cdot|\omega)$ is a normal density with moments

$$E(x|\omega) = a\omega + b,$$
$$\text{var}(x|\omega) = \sigma^2. \tag{4}$$

The likelihood parameters a, b, and σ may be estimated via the least-squares method applied to a historical or simulated record of forecasts and actual states. An illustration of the procedure in Fig. 1 shows the sample points, a plot of the conditional mean $E(x|\omega)$ versus ω (the regression line), and estimates of the parameters.

The posterior distribution of the state ω, conditional on forecast x, is not needed for our development. Nevertheless, it is presented in appendix A for the sake of completeness. Forecasters may employ it as a cali-

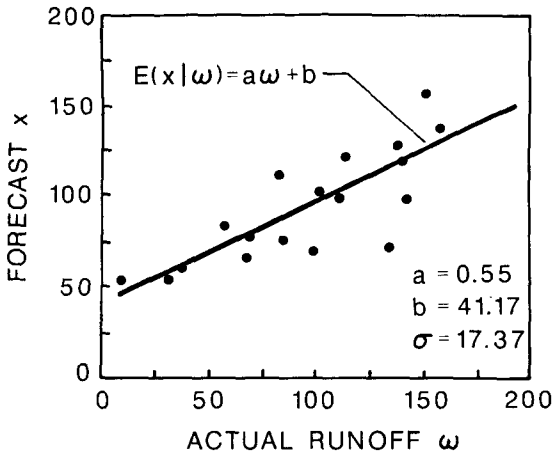

FIG. 1. Estimation of the likelihood function for a seasonal runoff volume forecast. (Weiser River, January forecast, historical record 1971–1988.)

brating filter, while users may input it into their decision procedures.

c. *Likelihood parameters*

The normal-linear model of forecasts plays an important practical role: it enables us to summarize the entire verification record in just three parameters that, from a decision-theoretic point of view, completely characterize the forecaster's predictive capabilities. If forecasts were perfect, then we would have $a = 1$, $b = 0$, and $\sigma = 0$. If forecasts were randomly generated from an arbitrary distribution having mean N and standard deviation T, then we would obtain $a = 0$, $b = N$, and $\sigma = T$; such forecasts would be worthless, of course. These limiting cases suggest an interpretation of the likelihood parameters: the slope a measures forecast information (or "signal" carried by the forecast), while the standard deviation σ measures forecast uncertainty (or "noise" in the forecast). Intuitively, one may anticipate that as the signal increases and the noise decreases, forecasts become more valuable. We shall revisit this interpretation after deriving the supporting mathematical expressions.

Per traditional definition, forecasts are said to be unbiased if $E(x) = E(\omega)$. Accordingly, the intercept b could be interpreted as a conveyor of the forecast bias. A bias is present whenever $b \neq (1 - a)M$, the condition that arises as follows. Under model (1)–(3), the mean of the predictand is $E(\omega) = M$, while the mean of the forecast, derived in appendix A, is $E(x) = aM + b$. Hence, when forecasts are unbiased, we find $b = (1 - a)M$; and since the intercept b is a function of the slope a, just two parameters, a and σ, summarily characterize the forecaster.

As a property of categorical forecasts, the unbiasedness, in the sense of $E(x) = E(\omega)$, is inconsequential to rational users, while it may be either desirable or detrimental for other users, depending upon the nature of their decision rules and shapes of their utility functions. These facts explain why the parameter b will not appear in the skill measures derived from the utilitarian point of view. (Other users are those who employ suboptimal decision rules, of which there are many; for instance, a common suboptimal rule is to ignore forecast uncertainty and prescribe a decision as if the categorical forecast were perfect.)

d. *Forecast error*

To further interpret the normal-linear model of forecasts, it is helpful to consider the forecast error: $\epsilon = x - \omega$. A simple transformation of (2) gives

$$\epsilon = (a - 1)\omega + b + \theta. \tag{5}$$

Two cases may be distinguished. If $a \neq 1$, then the error ϵ depends linearly on the actual state ω. If $a = 1$, then ϵ is stochastically independent of ω. The density of ϵ, conditional on ω, is normal with moments:

$$E(\epsilon|\omega) = (a - 1)\omega + b,$$
$$\text{var}(\epsilon|\omega) = \sigma^2. \tag{6}$$

The marginal density of ϵ is again normal with moments:

$$E(\epsilon) = (a - 1)M + b,$$
$$\text{var}(\epsilon) = (a - 1)^2S^2 + \sigma^2. \tag{7}$$

A popular measure of forecast skill is the mean square error

$$E(\epsilon^2) = \text{var}(\epsilon) + E^2(\epsilon), \tag{8}$$

or its transformation, the quadratic score

$$\text{QS} = 1 - \frac{E(\epsilon^2)}{\text{var}(\omega)}, \tag{9}$$

where $\text{var}(\omega) = E[(M - \omega)^2] = S^2$ represents the mean square error of a forecast that always specifies the climatological mean $E(\omega) = M$. Relations (7)–(9) reveal the structure of both measures in terms of the parameters of the prior distribution and the likelihood function. We shall demonstrate later that, unless the forecast and the posterior mean coincide, $x = E(\omega|x)$, the mean square error and the quadratic score are not rational measures of forecast skill from the utilitarian point of view.

3. Decision-theoretic framework

a. *Dominance order of forecasts*

Suppose each user makes a decision based on forecast x by following the Bayesian principles of rationality (Harsanyi 1978). Accordingly, having assessed his prior density g and estimated the likelihood functions f, the user obtains the posterior density of the state ω, con-

ditional on forecast x, and then finds the optimal decision that maximizes his expected utility of outcomes under the posterior density. The ex ante economic value of the forecast depends thus upon three elements of the user's decision model: the prior density g, the likelihood functions f, and the utility function u (which encodes the user's preferences for outcomes of his decisions).

Let us turn now to the problem of comparing two forecast systems, say i and j, producing, with identical lead times, forecasts x_i and x_j of the same predictand ω. In general, the ordering of forecasts in terms of their economic values is user dependent. We are interested, though, in a special situation wherein each user indicates the same preference order between x_i and x_j. For if such a unanimous order exists, it is the preferred order for society as a whole. Formally, this situation is described as follows:

Definition. Forecast x_i dominates forecast x_j if for every prior density g and utility function u, forecast x_i has an economic value at least as high as forecast x_j.

The question now arises whether it might be possible to infer the dominance order between forecasts x_i and x_j solely from their likelihood functions f_i and f_j. This avenue is investigated next.

b. *Relation of sufficiency between forecasts*

A preference order between forecasts, consistent with their dominance order, may be established via a binary relation of sufficiency defined in terms of the likelihood functions. Introduced originally by Blackwell (1951, 1953) for the purpose of comparing statistical experiments, the concept of sufficiency has been recently revived and applied in comparisons of forecasts by Alexandridis and Krzysztofowicz (1982), DeGroot and Fienberg (1982, 1986), Ehrendorfer and Murphy (1988), Krzysztofowicz and Long (1990), and others.

Definition. Forecast x_i is sufficient for forecast x_j if there exists a stochastic transformation—a family of conditional densities ψ, such that for every ω and x_j,

$$f_j(x_j|\omega) = \int \psi(x_j|x_i) f_i(x_i|\omega) dx_i . \qquad (10)$$

Insight into the sufficiency relation may be obtained by considering the task of simulation of forecasts (DeGroot 1970, p. 434; Alexandridis and Krzysztofowicz 1982). For a fixed state ω, forecast x_i can be generated from the density $f_i(\cdot|\omega)$. Forecast x_j can be generated either in one step from the density $f_j(\cdot|\omega)$ or, according to (10), in two steps. First, x_i is generated from $f_i(\cdot|\omega)$; next, given this x_i, x_j is generated from $\psi(\cdot|x_i)$. Thus, in comparison with the generator of x_i, the two-step generator of x_j involves an auxiliary randomization. One may expect, therefore, that this "additional randomness" of forecast x_j, in comparison with forecast x_i, will translate into a consistent difference in the evaluation of performance. This is indeed the case, as Blackwell's theorem attests: if forecast x_i is sufficient for forecast x_j, then x_i dominates x_j.

We shall next harness this general theoretical framework for comparing forecasts in rationalizing a new measure of the forecast skill. This measure evolves through three forms, and we derive them progressively.

4. Sufficiency characteristic

Suppose a forecast is characterized in terms of the likelihood function (2)–(4). The *sufficiency characteristic* (SC) of such a forecast is defined as the ratio of the conditional standard deviation of the forecast error (a measure of "noise" in the forecast) to the absolute value of the slope coefficient of the regression line between the forecasted and actual states (a measure of "signal" carried by the forecast):

$$\mathrm{SC} = \frac{\sigma}{|a|} . \qquad (11)$$

The units of the SC are the same as the units of the forecast x. For the perfect forecast, SC = 0. For the forecast produced by guessing, or a random number generator, SC = ∞.

In order to rationalize the SC as a measure of the forecast skill, let us suppose that the predictand ω is forecasted by two systems, say i and j. System n ($n = i, j$) issues forecast x_n having likelihood parameters (a_n, b_n, σ_n) and the sufficiency characteristic $\mathrm{SC}_n = \sigma_n/|a_n|$. The comparison of these forecasts is governed by the following theorem (Krzysztofowicz 1987): Forecast x_i is sufficient for forecast x_j if, and only if,

$$\mathrm{SC}_i < \mathrm{SC}_j . \qquad (12)$$

Now recall that if forecast x_i is sufficient for forecast x_j, then x_i dominates x_j. Consequently, the ordering of alternative forecasts in terms of their SCs (from the lowest to the highest) coincides with the ordering of forecasts in terms of their ex ante economic values (from the highest to the lowest) for each user, and thereby for society as a whole.

In summary, a forecaster who evaluates his performance in terms of the SC, and who chooses a new forecast system or an improvement of the existing system by minimizing the SC, acts as if he were maximizing the ex ante economic benefit of forecasts to each user. In that sense, the SC constitutes a sufficient measure of the forecast skill from the utilitarian point of view.

5. Standardized sufficiency characteristic

a. *Motivation and definition*

In some situations, it may be desirable to compare the performance of forecasts of different predictands. Evaluations of seasonal snowmelt runoff forecasts call

for such comparisons. For instance, the Yellowstone River near Billings, Montana, receives in January a forecast of the runoff volume during April–September, and in May a forecast of the runoff volume during May–September. Thus, not only is the lead time of these forecasts different, but also the predictand. In other situations, one may wish to compare the performance of forecasts of the same quantity, say daily maximum wind speed, for different stations. Still in other situations, it may be important to compare the performance of forecasts of different quantities, such as the mean seasonal temperature and the mean seasonal precipitation, in order to prioritize the needs for improving various forecasting services.

Suppose a predictand has the prior density specified by (1) and its forecast is characterized in terms of the likelihood function (2)–(4). The *standardized sufficiency characteristic* (SSC) of such a forecast is defined as the ratio of the SC to the prior standard deviation of the predictand:

$$\mathrm{SSC} = \frac{\sigma}{|a|S}. \tag{13}$$

The SSC is dimensionless. For the perfect forecast, SSC = 0. For the forecast produced by guessing, or a random number generator, SSC = ∞.

b. Interpretation

Let us now consider two different predictands and their forecasts, both indexed by $n(n = i, j)$. Predictand ω_n has prior parameters (M_n, S_n), and its forecast x_n has likelihood parameters (a_n, b_n, σ_n) and the standardized sufficiency characteristic $\mathrm{SSC}_n = \sigma_n/|a_n|S_n$. In appendix B, we justify the following statement: forecast x_i of predictand ω_i is sufficient for forecast x_j of predictand ω_j if, and only if,

$$\mathrm{SSC}_i < \mathrm{SSC}_j. \tag{14}$$

The interpretation of this sufficiency relation in terms of economic values of forecasts is not straightforward. We are comparing not only different forecasts, but also different predictands, which may have very different uses in decision making. We must, therefore, resort to an abstract concept of a standardized decision problem. Imagine a user whose utility of outcomes depends upon his decision and a standard normal variate $v \sim N(0, 1)$. This variate may be obtained from either predictand through the usual standardization:

$$v = \frac{\omega_n - M_n}{S_n}. \tag{15}$$

Likewise, the corresponding forecast of v may be obtained:

$$z_n = \frac{x_n - M_n}{S_n}. \tag{16}$$

Furthermore, imagine that the user is free to choose the predictand, ω_i or ω_j, that will generate v and consequently the outcome of his decision. Since both predictands give the same prior distribution of $v \sim N(0, 1)$, they offer no basis for choice. If, however, the user could obtain a forecast of either predictand, then he could choose between z_i and z_j. In appendix B we prove that condition (14) is equivalent to the statement: z_i is sufficient for z_j. Next, Blackwell's theorem may be invoked to conclude that z_i dominates z_j. Therefore, the user would prefer z_i over z_j. In other words, the user would choose predictand ω_i as the outcome-generating variable and employ its forecast x_i in decision making.

This interpretation of the sufficiency condition (14) could be summarized like this: if forecast x_i of predictand ω_i is sufficient for forecast x_j of predictand ω_j, then, in every standardized decision problem, the predictand–forecast pair (ω_i, x_i) has the ex ante economic value at least as high as the pair (ω_j, x_j) does.

c. Distinction between SSC and SC

Our final discussion is devoted to the distinction between sufficient comparisons of (i) forecasts of the same predictand and (ii) forecasts of different predictands. The first type of comparisons is solely between forecasts and involves their likelihood functions. The second type of comparisons is between predictand–forecast pairs and involves both the prior distributions and the likelihood functions. To illuminate these distinctions, let us consider, for instance, forecasts of daily maximum wind speed for two locations, say i and j. Suppose that from climatological records one estimated $S_i > S_j$, implying a larger natural variability in the first location. Furthermore, suppose that from forecast verification records (which need not overlap with the climatological records) one estimated (a_i, σ_i) and (a_j, σ_j) and found $\mathrm{SC}_i = \mathrm{SC}_j$. Hence, $\mathrm{SSC}_i < \mathrm{SSC}_j$, demonstrating that when the climatological variability of the predictands is taken as a benchmark, the forecast for location i exhibits a higher skill than the forecast for location j. This preference order may also be interpreted in terms of economic values of forecasts, though in a restricted sense, applying only to users confronted with a standardized decision problem.

6. Bayesian correlation score

The SC and SSC have been presented in their parsimonious forms. Inasmuch as both characteristics are ordinal scales, they may always be transformed monotonically, without affecting the resultant preference order of forecasts. While the admissible transformations are endless, there is one that we wish to explore because it establishes an insightful relationship between the SSC and a Bayesian estimator of correlation between the forecast and the predictand.

a. Correlation between forecast and predictand

The marginal density of forecast x (predictive—in the Bayesian sense) results from the prior density g and the likelihood function f via the total probability law. It is a normal density with moments

$$E(x) = aM + b,$$
$$\text{var}(x) = a^2S^2 + \sigma^2. \quad (17)$$

Under the assumptions of the linear model (2)–(3), the covariance of x and ω is: $\text{cov}(x, \omega) = a\ \text{var}(\omega)$. Consequently, the correlation between x and ω takes the form:

$$\rho = a\left[\frac{\text{var}(\omega)}{\text{var}(x)}\right]^{1/2} = \frac{aS}{(a^2S^2 + \sigma^2)^{1/2}}, \quad (18)$$

which may be rearranged into

$$\rho = (\text{sign of } a)\left(\frac{\sigma^2}{a^2S^2} + 1\right)^{-1/2}. \quad (19)$$

This expression prompts three observations. First, the correlation estimator is Bayesian in nature, since information from different sources may be used to estimate its component parameters, S and (a, σ). Hence, ρ need not be equal to the empirical correlation estimated from a verification record alone. Second, ρ could also be interpreted as a Bayesian estimator of the anomaly correlation—that is, the correlation between the forecasted departure $(x - M)$ and actual departure $(\omega - M)$ from the climatological mean $M = E(\omega)$; this correlation is exactly $\text{cor}[(x - M), (\omega - M)] = \text{cor}(x; \omega) = \rho$. Third, the first term in the bracket is the square of the already familiar SSC. Thus, it turns out that the Bayesian estimator of the correlation between the forecast and predictand is a function of the standardized sufficiency characteristic. The monotonicity of this function depends upon the sign of a, but for the purpose of sufficient comparisons of forecasts, the correlation sign is irrelevant. (This is so because a Bayesian decision procedure would recognize the negative slope coefficient a and automatically invert the direction of the dependence between x and ω in the posterior distribution, as shown in appendix A.)

b. The score and its interpretation

We have thus arrived at the definition of the *Bayesian correlation score:* $\text{BCS} = |\rho|$, or more explicitly

$$\text{BCS} = \left(\frac{\sigma^2}{a^2S^2} + 1\right)^{-1/2}. \quad (20)$$

The score is bounded, $0 \leq \text{BCS} \leq 1$, with $\text{BCS} = 1$ for the perfect forecast and $\text{BCS} = 0$ for the forecast produced by guessing, or a random number generator.

The score establishes a complete order between forecasts as follows: Forecast x_i of predictand ω_i is sufficient for forecast x_j of predictand ω_j if, and only if,

$$\text{BCS}_i > \text{BCS}_j. \quad (21)$$

When predictands ω_i and ω_j are different, the interpretation of the sufficiency condition (21) follows the interpretation of condition (14) in terms of the SSC. When predictands ω_i and ω_j are identical, so that $\omega_i = \omega_j = \omega$, the interpretation of the sufficiency condition (21) parallels the interpretation of condition (12) in terms of the SC. In either case, the BCS is interpretable as the correlation between the forecast and the predictand. Inasmuch as the correlation is an absolute scale, the BCS constitutes a universally comparable measure of the forecast skill. This measure assumes that the climatological distribution of the predictand constitutes prior (benchmark) information, and orders alternative forecasts consistently with the utilitarian criterion of choice.

c. Practical verification procedure

A practical implementation of the verification procedure entails four steps: (i) From a stationary climatological record of the predictand, estimate its standard deviation S. (ii) From a record of the forecasted and actual states, which may be historical or simulated, estimate the parameters a and σ of the linear regression. (iii) Verify that the assumptions of a normal prior distribution and a normal-linear likelihood function are acceptable. (iv) Compute the BCS. In appendix C, we expand upon several questions that might be raised concerning the interpretation and special cases of this verification procedure.

Persistence-type forecasts, often generated in order to establish a benchmark of skill for other types of forecasts, may be verified via the proposed procedure as well. In such a case, two BCSs would be reported, one for the persistence-type forecast and the other for the forecast of primary interest.

When the data cry out against the assumption of a normal prior distribution, the general approach is to eliminate the nonnormality via a suitable transformation of the predictand (Box and Tiao 1973, Chapter 10). Violations of the assumptions underlying the normal-linear likelihood function may manifest themselves in two ways: (i) a nonlinearity of the regression $E(x|\omega)$—this may often be eliminated via a suitable transformation of the forecast x; (ii) a dependence of the conditional variance $\text{var}(x|\omega)$ upon the state ω—this may be impossible to eliminate, and therein may lie a limit on the universality of the BCS.

Transformations of the predictand ω and the forecast x need not be identical, but each must be strictly monotone (increasing or decreasing). The admissibility of such transformations rests on the fact that they do not alter the binary relation of sufficiency between

forecasts. Hence, the verification procedure described above may be applied to records of the transformed variables. Transformations that guarantee the normality of the prior distribution, and that are also necessary, though not sufficient, for the linearity of the regression, are probability integral transformations. The mechanics and effectiveness of such transformations are illustrated in appendix D.

7. Examples and discussion

a. Sufficient comparison versus mean square error

Shown in Table 1 is an exemplary comparison of three systems ($n = 1, 2, 3$) producing forecasts of the same predictand. Hence, the ordering of forecasts according to their ex ante economic values may be established via the SC_n, whose computation does not involve the prior variance S^2. This ordering, implied by increasing SC_n, is 3, 2, 1. The example also illustrates the impact of the prior variance S^2 on the Bayesian correlation score BCS_n and the mean square error $E(\epsilon_n^2)$. Since we set $M = 0$ and $b_n = 0$, $E(\epsilon_n^2)$ is equal to the unconditional variance of the forecast error $\text{var}(\epsilon_n)$.

The example demonstrates that neither the conditional variance of the forecast error $\text{var}(\epsilon_n|\omega) = \sigma_n^2$, nor the mean square error $E(\epsilon_n^2)$, is a proper surrogate measure for forecast value. In terms of increasing σ_n^2, the forecasts are ranked 1, 2, 3. In terms of increasing $E(\epsilon_n^2)$, the forecasts are ranked 1, 2, 3 when the prior variance is $S^2 = 10^2$, and 2, 3, 1 when the prior variance is $S^2 = 50^2$. Not only, then, does the ranking of forecasts in terms of $E(\epsilon_n^2)$ depend upon the natural variability of the predictand, but it is also inconsistent with the ordering of forecasts in terms of their economic values.

This ordering, implied by decreasing BCS_n, is 3, 2, 1, regardless of the prior variance S^2. Naturally, the magnitude of the BCS_n depends upon S^2: the score is uniformly larger for the higher variance, reflecting the fundamental notion that the economic value of a given forecast is relative to the natural variability of the predictand.

The quadratic score QS_n, which in both cases implies the same ordering of forecasts as the mean square error $E(\epsilon_n^2)$, takes on the negative sign when the prior variance is $S^2 = 10^2$. As per traditional interpretation, these forecasts would be said to have negative skill. This is a dangerous form of conclusion: it may mislead the recipients and harm the forecasters and users alike, because despite the negative QS_n, each of the three forecasts has a nonnegative economic value for every rational user. Moreover, forecast $n = 3$, with the most negative QS_n, has the highest BCS_n and, therefore, the highest economic value.

While interpreting the Bayesian correlation score, we should keep in mind that it is an ordinal measure of economic value. To wit, the magnitude of the BCS does not say whether the economic value is sizeable or infinitesimal—this depends upon a decision problem. Likewise, the equality $BCS_i = \alpha BCS_j$, with $\alpha > 0$, does not imply that forecast i is α times as valuable as forecast j. Ergo, it would be inappropriate to arbitrarily set a threshold on the BCS scale in order to dichotomize all forecasts into "skillful" and "unskillful," or "valuable" and "worthless." Only by performing a full-fledged decision analysis could one establish a correspondence between the magnitude of the BCS and the magnitude of the economic value for a particular user. Most likely, this correspondence takes a different form for different users. It is possible, though, that numerous case studies performed for a variety of users could offer guidelines by delineating a lower range of the BCS where forecasts of a given predictand have no significant economic value to most users.

b. Performance of seasonal runoff forecasts

Continuing the example introduced earlier, we report results of an evaluation of seasonal snowmelt runoff forecasts for the Weiser River near Weiser, Idaho. All five forecasts are for the same predictand, the runoff volume during April–July, but each is issued with a different lead time: the first forecast, issued in January, has the lead time of seven months, and each subsequent forecast has the lead time shorter by one month. The prior parameters M and S were estimated from a record of 38 observations (1951–1988) of the predictand; the likelihood parameters a, b, and σ were estimated from a record of 18 realizations (1971–1988) of the forecast and predictand.

TABLE 1. Sensitivity of the ordering of forecasts in terms of the Bayesian correlation score BCS_n, the mean square error $E(\epsilon_n^2)$, and the quadratic score QS_n, to the prior variance S^2.

				$S = 10$			$S = 50$		
n	a_n	σ_n	SC_n	BCS_n	$[E(\epsilon_n^2)]^{1/2}$	QS_n	BCS_n	$[E(\epsilon_n^2)]^{1/2}$	QS_n
1	0.30	15	50.00	0.196	16.55	−1.74	0.707	38.08	0.42
2	0.50	24	48.00	0.204	24.52	−5.01	0.721	34.66	0.52
3	0.70	33	47.14	0.208	33.14	−9.98	0.728	36.25	0.47

Since $M = 0$ and $b_n = 0$, $E(\epsilon_n^2) = \text{var}(\epsilon_n)$.

TABLE 2. Evaluation of performance of seasonal runoff volume forecasts for Weiser River near Weiser, Idaho.

Forecast month n	Regression parameters: Slope a_n	Regression parameters: Intercept b_n	Conditional standard deviation of forecast error σ_n	Sufficiency characteristic SC_n	Bayesian correlation score BCS_n
1	0.55	41.17	17.37	31.58	0.79
2	0.68	26.27	19.60	28.82	0.81
3	0.70	27.44	14.38	20.54	0.89
4	0.79	18.60	14.39	18.22	0.91
5	0.81	15.80	12.40	15.31	0.94

Prior mean $M = 99.86$; prior standard deviation $S = 40.40$. Runoff volume is for the period April–July.

Table 2 reports estimates of all parameters as well as the SC_n and BCS_n for $n = 1, \cdots, 5$. As one would expect, BCS_n increases with n, implying that each forecast is sufficient for every forecast issued earlier. If all forecasts were available to the users at the same epoch, then the ordering of forecasts in terms of their economic values would be from x_5 to x_1.

c. *Trade-offs between skill and lead time*

To elucidate the kinds of analyses that the BCS allows, the seasonal snowmelt runoff forecasts have been evaluated for three stations: Yellowstone River near Billings, Montana; Boise River near Twin Springs, Idaho; and Salt River near Roosevelt, Arizona. Inasmuch as the forecasts are for different stations and different runoff seasons, the predictands are not identical random variables. Hence, the SSC or BCS are the appropriate measures for comparing the skill of these forecasts.

For each station, Table 3 lists the runoff seasons, and Fig. 2 shows a plot of the BCS_n as a function of the forecast month n. Several observations can be made. (i) For every station, the forecast skill generally improves as the lead time becomes shorter. Most of these improvements take place between the January and March forecast months. (ii) Trade-offs between the skill and lead time of forecasts are distinct for each river. Forecasts for the Boise River exhibit the highest skill, even though they do not have the shortest lead times. Forecasts for the Yellowstone River have the longest lead times, yet their skill is not uniformly the lowest. (iii) The April forecast for the Yellowstone River (with the lead time of six months) has about the same skill as the January forecast for the Boise River (with the lead time of seven months).

TABLE 3. Runoff seasons for which forecasts are prepared.

Station	Forecast month	Runoff season
Yellowstone	1, 2, 3, 4	April–September
	5	May–September
Boise	1, 2, 3, 4, 5	April–July
Salt	1	January–May
	2	February–May
	3	March–May
	4	April–May

8. Closure

These analyses exemplify the kinds of cross-comparisons of the forecast skill that are admissible. Moreover, comparisons of the BCSs are meaningful not only in the statistical sense, as comparisons of correlations, but also in the economic sense: a higher BCS implies a higher economic value of forecasts, either (i) to all rational users if one compares forecasts of the same predictand, or (ii) to a class of rational users facing the standardized decision problem if one compares forecasts of different predictands. This interpretive limitation notwithstanding, the BCS is a measure that ensures coherence, in the broadest sense allowed by the theory of sufficient comparisons, between a statistical verification analysis and a full-fledged economic anal-

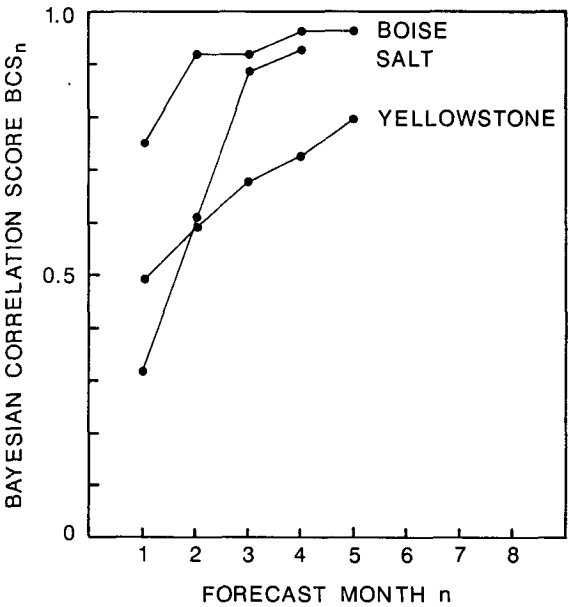

FIG. 2. Bayesian correlation scores of seasonal runoff volume forecasts for three stations.

ysis of all uses of forecasts. For these reasons, the BCS might be worth considering as a standard measure for reporting the skill of categorical forecasts of continuous predictands, from the viewpoint of utilitarian users.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Award ECE-8352536, "Presidential Young Investigator Award," and by the Soil Conservation Service of the United States Department of Agriculture which provided matching funds. Valuable comments by two reviewers on an earlier version of the manuscript are gratefully acknowledged.

APPENDIX A

Posterior Distribution and Forecast Calibration

a. *Posterior distribution*

The predictive density of the forecast is

$$\xi(x) = \int f(x|\omega)g(\omega)d\omega. \tag{A1}$$

Given the models of g and f specified in section 2, ξ is a normal density with moments

$$E(x) = aM + b,$$
$$\text{var}(x) = a^2S^2 + \sigma^2. \tag{A2}$$

The posterior density of the predictand ω, conditional upon the forecast x, results from Bayes' theorem:

$$\eta(\omega|x) = \frac{f(x|\omega)g(\omega)}{\xi(x)}. \tag{A3}$$

Given the models of g and f specified in section 2, $\eta(\cdot|x)$ is a normal density with moments

$$E(\omega|x) = Ax + B,$$
$$\text{var}(\omega|x) = s^2, \tag{A4}$$

where

$$A = \frac{aS^2}{a^2S^2 + \sigma^2}, \quad B = \frac{M\sigma^2 - abS^2}{a^2S^2 + \sigma^2}, \tag{A5}$$

$$s^2 = \frac{S^2\sigma^2}{a^2S^2 + \sigma^2}. \tag{A6}$$

Relations (A4)–(A6) reveal the manner in which the climatological information about the predictand, encoded in M and S, is combined with information about the forecaster, encoded in a, b, and σ. The posterior mean $E(\omega|x)$ is a linear function of the forecast x, while the posterior variance $\text{var}(\omega|x)$ is constant. Thus, for any forecast x, the posterior density $\eta(\cdot|x)$ is obtained by translating the fixed density $N(0, s^2)$ to the location $Ax + B$. The posterior variance s^2 is never greater than the prior variance S^2, no matter the magnitude of a and σ. Thus, the Bayesian processor (A3) automatically guards the user against notoriously poor forecasts.

b. *Calibration of categorical forecasts*

The forecaster may employ the posterior distribution $H(\cdot|x)$, corresponding to the density $\eta(\cdot|x)$, as a calibrating filter in the following sense. Suppose a point estimate y of ω is desired that ensures a specified posterior probability p of the event $\{\omega \leq y\}$. Having produced x, the forecaster can obtain the desired estimate y as the p-probability fractile of the posterior distribution, such that $H(y|x) = p$. In general, a categorical forecast may be calibrated in this way to any posterior probability, $0 < p < 1$, of the event $\{\omega \leq y\}$. In particular, when $p = \frac{1}{2}$, the calibrated estimate y equals the posterior median of ω which, given the normal distribution $H(\cdot|x)$, coincides with the posterior mean $E(\omega|x)$; hence the calibrating filter is given by a linear expression: $y = Ax + B$. (It may be of interest to note that this expression is a form of the Kalman filter, an estimator popular among electrical engineers.)

From the viewpoint of rational users, nothing is gained by recalibrating estimate x into estimate y, whatever the probability p, because an optimal decision rule itself utilizes the entire posterior distribution $H(\cdot|x)$. From the viewpoint of other users, who employ suboptimal decision rules, recalibration may be desirable, provided the forecaster knows the posterior probability p that produces an optimal estimate y for a given class of users. The common belief that the posterior mean is a universally optimal estimator is unfounded. For instance, $y = E(\omega|x)$ is a nonoptimal estimator for a user who makes a decision as if y were a perfect forecast of ω and whose utility is an asymmetric function of the forecast error $(y - \omega)$. These facts explain why the lack of calibration, in the sense of $x \neq E(\omega|x)$, does not bear on the measures of forecast skill derived from the utilitarian point of view.

APPENDIX B

Derivation of SSC

Given the original predictand–forecast pair (ω, x), modeled in terms of the prior density (1) and the likelihood function (2)–(4), define a transformed pair (ν, z), wherein

$$\nu = \frac{\omega - M}{S} \quad \text{and} \quad z = \frac{x - M}{S}. \tag{B1}$$

Consequently, $\nu \sim N(0, 1)$ and

$$\frac{x - M}{S} = \frac{a\omega + b + \theta - M}{S}$$
$$= a\frac{\omega - M}{S} + \frac{(a-1)M + b}{S} + \frac{\theta}{S}. \tag{B2}$$

After denoting

$$b' = \frac{(a-1)M + b}{S} \quad \text{and} \quad \theta' = \frac{\theta}{S}, \tag{B3}$$

we obtain

$$z = av + b' + \theta', \tag{B4}$$

which establishes a relationship between the transformed predictand v and its forecast z. Since $\theta' \sim N(0, \sigma^2/S^2)$, the sufficiency characteristic of forecast z, renamed into the standardized sufficiency characteristic, takes the following form:

$$\mathrm{SSC} = \left[\frac{\mathrm{var}(\theta')}{a^2}\right]^{1/2} = \frac{\sigma}{|a|S}. \tag{B5}$$

Given a task of comparing different predictand–forecast pairs, we transform each original pair (ω_n, x_n) into (v_n, z_n) and find $\mathrm{SSC}_n = \sigma_n/|a_n|S_n$, for $n = i, j$. Since v_i and v_j have identical prior distributions, we let $v = v_i = v_j$ and compare (v, z_i) with (v, z_j). Thus, the task has been reduced to a comparison of two forecasts z_i and z_j of predictand v. By applying Krzysztofowicz's (1987) theorem, we may now assert that z_i is sufficient for z_j if, and only if, $\mathrm{SSC}_i < \mathrm{SSC}_j$. Because the transformation between (v, z_n) and (ω_n, x_n) is one to one, the following definition may be introduced: forecast x_i of predictand ω_i is sufficient for forecast x_j of predictand ω_j if, and only if, z_i is sufficient for z_j.

APPENDIX C

Interpretations and Special Cases

a. The viewpoint of an observer

Two questions, probing the coherence of our Bayesian approach to forecast verification, were brought to our attention by a reviewer. The first question: Is it valid to estimate the prior distribution from a climatological record of the predictand, when the forecaster also utilizes the climatological record in his procedure? The answer is yes. The argumentation rests on the premise that models for forecast verification should be constructed from the viewpoint of a user—an outside observer of the forecast system, not a forecaster. The purpose of the prior distribution, then, is to describe the uncertainty about state ω, given all information the user would have for making decisions in the absence of forecasts. Since this information could be a climatological record, it is valid to estimate the prior distribution from such a record.

The second question: Should not the likelihood function take into account some features of the forecasting procedure? For example, when a forecaster himself applies the Bayesian procedure of appendix A, and calibrates his categorical forecast to a specified posterior probability, does the Bayesian approach to verification remain valid? Again, this dilemma may be resolved by consistently maintaining the viewpoint of a user. The purpose of the likelihood function is to describe the statistical relationship between state ω and its forecast x as seen by a user—an outside observer of the forecast system. This relationship is the sole characterization of forecasts that the user needs to know in order to make decisions rationally; knowing the bowels of the forecast system is irrelevant.

b. Posterior mean as a forecast

When a categorical forecast x coincides with the posterior mean $E(\omega|x)$, with respect to a specified prior distribution of the predictand ω (which may be the climatological distribution), the Bayesian verification procedure automatically detects this property of forecasts. For if $x = E(\omega|x)$, then with the aid of expressions (A4)–(A6) presented in appendix A, we find that the likelihood parameters must satisfy the following equalities:

$$b = (1-a)M,$$
$$\sigma^2 = a(1-a)S^2, \tag{C1}$$

with the condition $0 \leq a \leq 1$. The first equality implies that forecasts are unbiased in the sense of $E(x) = E(\omega)$. Furthermore, it implies that the regression line defined in (4) takes the form:

$$E(x|\omega) = \omega + (1-a)(M-\omega). \tag{C2}$$

Thus, a categorical forecast coinciding with the posterior mean, $x = E(\omega|x)$, exhibits a conditional bias since $E(x|\omega) \neq \omega$. Such a forecast will, on the average, overestimate the states below the prior mean, $\omega < M$, and underestimate the states above the prior mean, $\omega > M$.

Upon inserting σ^2 from (C1) into (20), the Bayesian correlation score becomes

$$\mathrm{BCS} = a^{1/2}. \tag{C3}$$

We have reached the simplest possible case of verification: when $x = E(\omega|x)$, just one parameter, the slope a of the regression line between forecast x and actual state ω, summarily characterizes the forecaster. In particular, $a = 1$ implies a perfect forecast, and for any imperfect forecast, $a < 1$.

Last, we shall examine implications of the equality $x = E(\omega|x)$ on the forecast error: $\epsilon = x - \omega$. Taking advantage of relations (C1), we insert them into (7) and obtain moments:

$$E(\epsilon) = 0,$$
$$\mathrm{var}(\epsilon) = (1-a)S^2; \tag{C4}$$

we note in passing that $\mathrm{var}(\epsilon) = \mathrm{var}(\omega|x)$, the posterior variance of the predictand. The mean square error, defined in (8), now becomes

$$E(\epsilon^2) = (1-a)S^2, \tag{C5}$$

and the quadratic score, defined in (9), takes the form

$$QS = a. \tag{C6}$$

By comparing (C6) with (C3), we find that

$$BCS = (QS)^{1/2}. \tag{C7}$$

In conclusion, if a categorical forecast x of a continuous predictand ω coincides with the posterior mean of the predictand, $E(\omega|x)$, with respect to a specified prior distribution of ω, then the quadratic score becomes a rational measure of forecast skill from the utilitarian point of view.

APPENDIX D

Probability Integral Transformations

Let v denote the original predictand and y denote the original forecast. The predictand–forecast pair (v, y) is processed into a pair (ω, x) through transformations that are necessary and sufficient for the normality of the predictand ω, and necessary for the linearity of the regression $E(x|\omega)$. The procedure is illustrated with a numerical example displayed in Table D1, wherein the original records of v and y are transformed into records of ω and x. The record of v was purposely made longer than the record of y to mimic the case found frequently in practice. The procedure consists of four steps.

Step 1. Given m observations of v, arranged in increasing order, $v_1 \leqslant v_2 \leqslant \cdots \leqslant v_m$, the empirical distribution R of v is constructed:

$$R(v_i) = P(v \leqslant v_i) = \frac{i}{m+1}, \quad i = 1, \cdots, m. \tag{D1}$$

The construction is detailed in Table D2 and the resultant distribution is plotted in Fig. D1. With Q^{-1} denoting the inverse of the standard normal distribution, define a variate

$$\omega = Q^{-1}[R(v)]. \tag{D2}$$

Each observation v_i of v can thus be transformed into an observation ω_i of ω, as illustrated in Table D2.

TABLE D1. Original and transformed records of the predictand and the forecast.

Original records		Transformed records	
v	y	ω	x
110		1.150	
20		−0.675	
10	20	−1.150	−0.966
40	30	0.000	−0.430
70	60	0.675	0.430
30	40	−0.319	0.000
50	70	0.319	0.966

TABLE D2. Empirical distributions and probability integral transformations of the predictand and the forecast.

Predictand				Forecast			
i	v_i	$R(v_i)$	ω_i	i	y_i	$T(y_i)$	x_i
1	10	0.125	−1.150	1	20	0.167	−0.966
2	20	0.250	−0.675	2	30	0.333	−0.430
3	30	0.375	−0.319	3	40	0.500	0.000
4	40	0.500	0.000	4	60	0.667	0.430
5	50	0.625	0.319	5	70	0.833	0.966
6	70	0.750	0.675				
7	110	0.875	1.150				

Whatever the distribution R of v, the distribution G of ω is normal with mean $M = 0$ and variance $S^2 = 1$. The upper part of Fig. D2 shows G obtained in our example.

Step 2. Given n observations of y, arranged in increasing order, $y_1 \leqslant y_2 \leqslant \cdots \leqslant y_n$, the empirical distribution T of y is constructed:

$$T(y_i) = P(y \leqslant y_i) = \frac{1}{n+1}, \quad i = 1, \cdots, n. \tag{D3}$$

Define a variate

$$x = Q^{-1}[T(y)]. \tag{D4}$$

Each observation y_i of y can thus be transformed into an observation x_i of x, as illustrated in Table D2.

Step 3. The transformed record of the predictand-forecast pair (ω, x) is now used to estimate the parameters a, b, and σ of the linear regression defined in (4). Table D1 shows the transformed record, and the lower part of Figure D2 depicts the sample points, the regression line, and estimates of the parameters. What remains to be verified is (i) whether the linearity of the regression $E(x|\omega)$ is an acceptable assumption and (ii) whether the conditional variance $\text{var}(x|\omega)$ can be assumed independent of the predictand ω.

Step 4. The estimates $a = 0.883$, $\sigma = 0.492$, and $S = 1$ are input into (20) that gives BCS = 0.873.

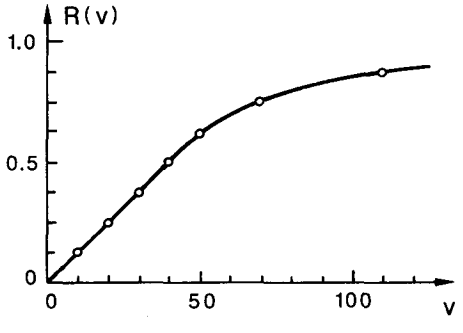

FIG. D1. Empirical distribution of the original predictand.

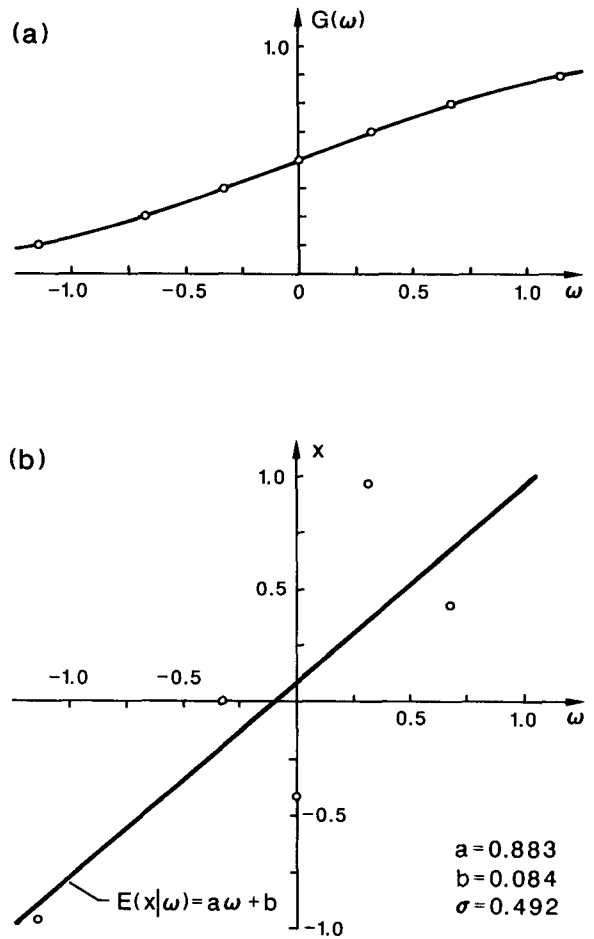

FIG. D2. Distribution of (a) the transformed predictand and (b) estimation of the likelihood function from the transformed records of the predictand and forecast.

REFERENCES

Alexandridis, M. G., and R. Krzysztofowicz, 1982: Value of categorical and probabilistic temperature forecasts for scheduling of power generation. Rep. No. 282, Ralph M. Parsons Laboratory, Massachusetts Institute of Technology, Cambridge.

Blackwell, D., 1951: Comparison of experiments. *Proc. of the Second Berkeley Symp. on Mathematical Statistics and Probability,* J. Neyman, Ed., University of California Press, 93–102.

——, 1953: Equivalent comparisons of experiments. *Ann. Math. Stat.,* **24,** 265–272.

Box, G. E. P., and G. C. Tiao, 1973: *Bayesian Inference in Statistical Analysis.* Addison–Wesley.

DeGroot, M. H., 1970: *Optimal Statistical Decisions.* McGraw-Hill.

——, and S. E. Fienberg, 1982: Assessing probability assessors: Calibration and refinement. *Statistical Decision Theory and Related Topics III,* Vol. 1. S. S. Gupta and J. O. Berger, Eds., Academic Press, 291–314.

——, and ——, 1986: Comparing probability forecasters: Basic binary concepts and multivariate extensions. *Bayesian Inference and Decision Techniques.* P. K. Goel and A. Zellner, Eds., Elsevier Science Publishers, 247–264.

Ehrendorfer, M., and A. H. Murphy, 1988: Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. *Mon. Wea. Rev.,* **116,** 1757–1770.

Harsanyi, J. C., 1978: Bayesian decision theory and utilitarian ethics. *The American Economic Review,* **68**(2), 223–228.

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

Krzysztofowicz, R., 1987: Markovian forecast processes. *J. Amer. Stat. Assoc.,* **82**(397), 31–37.

——, and D. R. Davis, 1983: A Methodology for evaluation of flood forecast-response systems. Part 1: Analyses and concepts, Part 2: Theory, Part 3: Case studies. *Water Resour. Res.,* **19**(6), 1423–1454.

——, and D. Long, 1990: Fusion of detection probabilities and comparison of multisensor systems. *IEEE Trans. Systems, Man, Cybernetics,* **20**(3), 665–677.

Murphy, A. H., and H. Daan, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences.* A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

——, and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.