EVALUATION OF FORECASTS

ALLAN H. MURPHY[1] and MARTIN EHRENDORFER[2]

[1] Professor Emeritus, College of Oceanic and Atmospheric Sciences,
Oregon State University, Prediction and Evaluation Systems
Corvallis, Oregon 97330, USA

[2] National Center for Atmospheric Research
Boulder, Colorado 80307, USA
On leave from the Institute for Meteorology and Geophysics,
University of Vienna, Hohe Warte 38, A-1190 Vienna, Austria

Abstract

Evaluation of forecasts encompasses the processes of assessing both forecast quality and forecast value. These processes necessarily play key roles in any effort to improve forecasting performance or to enhance the usefulness of forecasts.

A framework for forecast verification (the process of assessing forecast quality) based on the joint distribution of forecasts and observations - and on the conditional and marginal distributions derived from factorizations of this joint distribution - is described. The joint, conditional, and marginal distributions relate directly to basic aspects of forecast quality, and evaluation methods based on these distributions - and associated statistics and measures - provide a coherent, diagnostic approach to forecast verification. This approach - and its attendant methodology - is illustrated using a sample of probabilistic long-range weather forecasts.

A decision-analytic approach to the problem of assessing the value of forecasts is outlined, and this approach is illustrated by considering the so-called fallowing-planting problem. In addition to providing estimates of the value of state-of-the-art and hypothetically improved long-range weather forecasts, the results of this case study illustrate some of the fundamental properties of quality/value relationships. These properties include the inherent nonlinearity of such relationships and the existence of quality thresholds below which the forecasts are of no value.

The sufficiency relation is used to explore quality/value relationships; this relation embodies the conditions that must exist between the joint distributions of two forecasting systems to ensure that one system's forecasts are better in all respects (i.e., in terms of quality and value) than the other system's forecasts. The applicability of the sufficiency relation is illustrated by comparing forecasting systems that produce prototypical long-range weather forecasts. This application also demonstrates that quality/value reversals can occur when the multifaceted nature of forecast quality is not respected.

Some outstanding problems in forecast evaluation are identified and briefly discussed. Recommendations are made regarding improvements in evaluation methods and practices.

1. Introduction

In the meteorological community, both individuals who formulate numerical and/or statistical forecasting models (modelers) and individuals who actually make forecasts on

J. Grasman et al. (eds.), *Predictability and Nonlinear Modelling in Natural Sciences and Economics*
© Kluwer Academic Publishers 1994

an operational basis (forecasters) are concerned with the quality of their forecasts. Identification of the characteristics of forecasting performance is an essential first step in the processes of model refinement and forecast improvement. Since the underlying rationale for developing forecasting systems in the first place is to provide information that can enhance the welfare (economic or otherwise) of potential users of the forecasts, modelers and forecasters are - or should be - concerned as well with the value of their forecasts. Forecast evaluation, which encompasses the processes of assessing both forecast quality and forecast value, is thus a problem of considerable importance.

Notwithstanding the importance of this problem, traditional methods of assessing forecast quality and forecast value are seriously deficient and potentially misleading. With regard to quality, forecast verification (the name usually given to the component of the evaluation process concerned with forecast quality) generally consists of the calculation of one or two measures of the overall correspondence between forecasts and observations (e.g., a skill score, a correlation coefficient). Consideration of the dimensionality of most verification problems suggests that this measures-oriented approach is inadequate as a means of characterizing the fundamental strengths and weaknesses in forecasting performance. Forecast value is seldom assessed in anything but the most informal and ad hoc manner. It is often simply assumed that improvements in forecast quality, as reflected by an increase in a skill score or a correlation coefficient, necessarily are accompanied by increases in forecast value. It is relatively easy to show that this assumption generally is unwarranted (see section 4).

Coherent approaches to the problems of assessing forecast quality and forecast value are described in this paper. In the case of forecast quality, considered in section 2, the approach is based on the concept that the joint distribution of forecasts and observations contains all of the (nontime-dependent) information relevant to forecast verification. Section 3 outlines a decision-analytic approach to the assessment of forecast value. Both sections 2 and 3 contain brief descriptions of applications of evaluation methods associated with the respective approaches. The relationship between forecast quality and forecast value is discussed in section 4, with emphasis on the role and usefulness of the sufficiency relation in comparative evaluation. Section 5 identifies some of the outstanding methodological and practical problems in this area and includes some recommendations regarding changes in current practices.

2. Forecast Quality

2.1 Approach and Methodology

The quality of forecasts produced by a model or a forecaster (denoted by F) can be defined as the totality of the statistical characteristics embodied in the joint distribution of forecasts (f) and observations (x), $p(f,x)$ (Murphy and Winkler 1987). Under the assumption that the joint probabilities that constitute $p(f,x)$ are stationary parameters, they can be estimated from the joint relative frequencies obtained from a single realization of a verification process that extends over a reasonably long but finite time period. The information contained in $p(f,x)$ becomes more accessible when this distribution is factored into conditional and marginal distributions. Two such factorizations are possible: (1) $p(f,x) = p(x|f)\, p(f)$ and (2) $p(f,x) = p(f|x)p(x)$. These expressions are referred to as the calibration-refinement and likelihood-base rate factorizations of $p(f,x)$, respectively. In effect, these factorizations provide two different but complementary - and obviously related -

descriptions of forecast quality in the context of absolute verification problems (i.e., problems involving the quality of a single forecasting system).

If the dimensionality (*d*) of an absolute verification problem is defined as the number of parameters (e.g., joint probabilities) that must be specified to determine $p(f,x)$, then it follows that $d = 3$ when both f and x are binary, and $d = 21$ when f can assume any of 11 equally-spaced probability values and x is binary (Murphy 1991). (The inadequacy of the traditional measures-oriented approach to verification problems, described briefly in section 1, should now be clear.) The multidimensional nature of forecast quality also can be appreciated by recognizing that quality possesses several aspects, including bias, accuracy, skill, reliability, resolution, sharpness, and discrimination (Murphy and Winkler 1987; Murphy 1993). These aspects of quality can be related directly to the aforementioned joint, conditional, and marginal distributions (see section 2b).

To accommodate comparative verification, which is concerned with the relative quality of two sets of forecasts (produced by systems, models, or forecasters F and G), the framework for absolute verification must be extended to include two joint distributions, $p(f,x)$ and $p(g,x)$. (For simplicity, we assume throughout this paper that F and G produce forecasts for the same events at the same location on the same occasions.) In this context, it is necessary to compare the conditional and marginal distributions associated with the respective factorizations of these joint distributions. Clearly, comparative verification is considerably more complex than absolute verification, and space limitations preclude detailed treatment of this problem here (however, see section 4 for some relevant discussion).

Within the framework of this distributions-oriented approach to forecast verification, three general classes of verification methods can be identified: (1) the joint, conditional, and marginal distributions themselves; (2) statistics of these distributions (e.g., means, medians, standard deviations, interquartile ranges); and (3) performance measures and terms in decompositions of performance measures. These methods can be used to evaluate - both qualitatively and quantitatively - various aspects of forecast quality, and some of these methods are illustrated in section 2b.

In describing the distributions-oriented approach to verification problems, it has been assumed implicitly that both f and x are scalar quantities, in the sense that they denote forecasts and observations at a specific location. However, this approach can be extended readily to situations in which the arguments of the joint distribution are vectors; for example, these vectors might define two-dimensional arrays of the forecast and observed (or analyzed) values of a meteorological variable at a particular time. Verification of these fields could proceed by applying distributions-oriented methods to the data set obtained by pooling the forecast-observation pairs from all locations. More general methods that take into account the spatial characteristics of the arrays of forecasts and observations would require a further extension of the approach described in this paper.

2.2 Applications

To illustrate the distributions-oriented approach to forecast verification, we describe briefly an application of this methodology to long-range weather forecasts formulated by forecasters at the Climate Analysis Center of the U.S. National Weather Service (Murphy and Huang 1993). The results presented here involve probabilistic forecasts of monthly mean temperatures (MT) and monthly precipitation amounts (MP), which were produced twice each month for approximately 100 locations during the period 1982-1991. These forecasts relate to below-normal, near-normal, and above-normal categories of MT or MP, which are defined at each location such that their historical probabilities of occurrence are equal to 0.30, 0.40, and 0.30, respectively. These probabilities are referred to here as the climatological probabilities. Due to space limitations we will focus primarily on the quality of the forecast probabilities assigned to the below-normal category. These probabilities are denoted here by f. Since the probability assigned to the near-normal category always is fixed at 0.40, the range of values of f is $0 \leq f \leq 0.60$. Since the below-normal category either occurs or does not occur on each occasion (i.e., each month-location combination), the corresponding observation x equals 1 or 0.

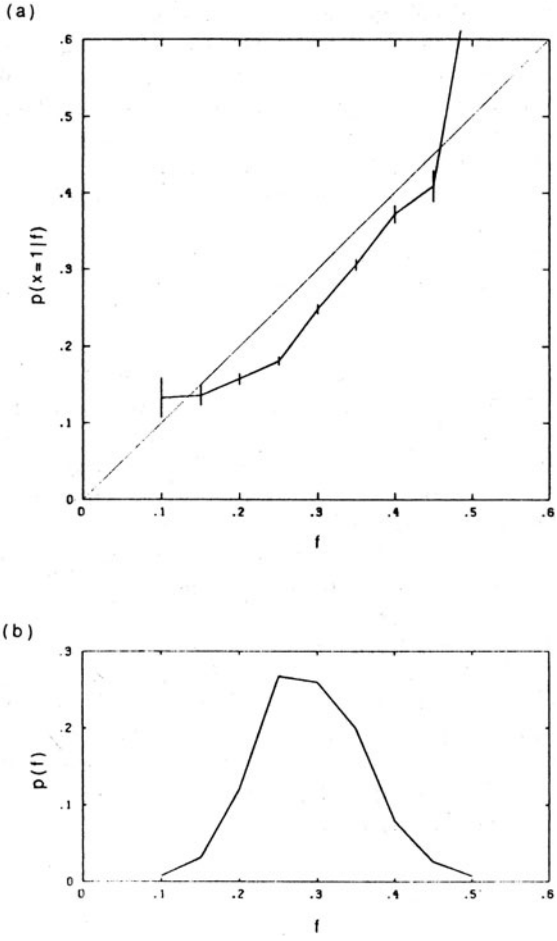

Figure 1. (a) Reliability diagram and (b) sharpness diagram, for monthly temperature forecasts.

Figures 1 and 2 contain reliability diagrams for the below-normal MT and MP forecasts. Probabilistic forecasts are completely reliable (or well-calibrated) if, for all distinct forecasts, the conditional relative frequency of occurrence of the event of interest given a particular forecast is equal to that probability [i.e., $p(x=1\,|\,f) = f$ for all f]. The reliability diagram for the MT forecasts (Fig. 1a) indicates that these forecasts exhibit modest but consistent overforecasting [$p(x=1\,|\,f) < f$] for $0.20 \leq f \leq 0.45$ and possibly some underforecasting [$p(x=1\,|\,f) > f$] for $f = 0.50$. The MP forecasts (Fig. 2a) exhibit quite good reliability over a limited range of probability values.

To complete the description of forecast quality based on the calibration-refinement factorization of $p(f,x)$, sharpness (refinement) diagrams also are included in Figures 1 and 2. Probabilistic forecasts are sharp if low and high probabilities are used relatively frequently and intermediate probabilities are used relatively infrequently [i.e., if $p(f)$ possesses a u-shaped distribution]. It is evident that the MT and MP forecasts (see Figs. 1b and 2b) are not very sharp, although the former are somewhat sharper than the latter. Most forecast probabilities fall in the range between 0.20 and 0.40; that is, within 0.10 of the climatological value of 0.30.

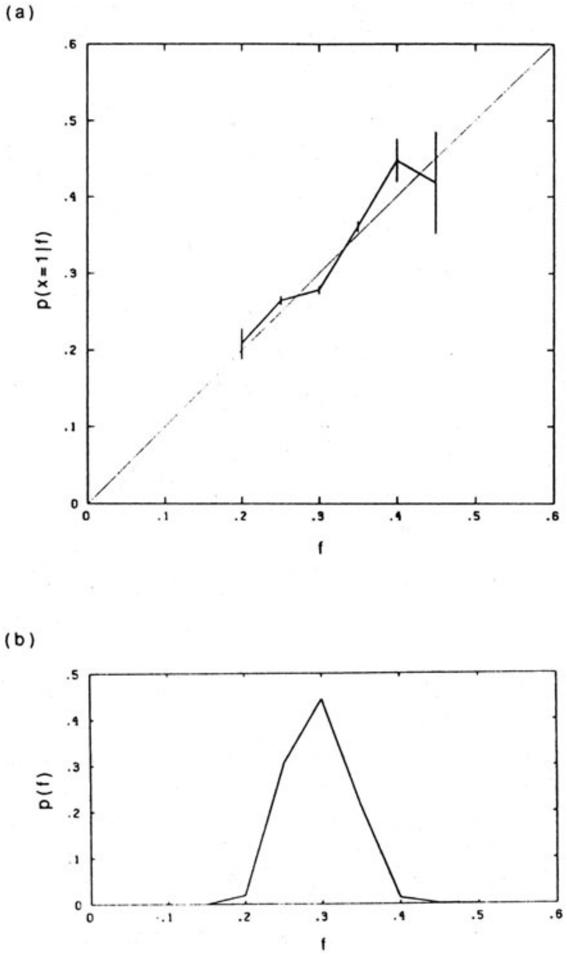

Figure 2. Same as Figure 1, except for monthly precipitation forecasts.

Discrimination (likelihood) diagrams for the below-normal MT and MP forecasts are included in Figures 3 and 4. These diagrams depict the conditional distributions of the forecasts given that the event in question does and does not occur; that is, $p(f|x=1)$ and $p(f|x=0)$, respectively. Perfect discrimination would be represented by conditional distributions that do not overlap. The greater the overlap between $p(f|x=1)$ and $p(f|x=0)$, the weaker the discrimination. Figures 3 and 4 reveal only modest discrimination in the case of the MT forecasts and little if any discrimination in the case of the MP forecasts. Evidently, it is difficult to achieve strong discrimination for these long lead-time forecasts. To complete the description of forecast quality based on the likelihood-base rate factorization of $p(f,x)$, the sample climatological probabilities (i.e., base rates), $p(x=0)$ and $p(x=1)$, are included as inserts in the discrimination diagrams.

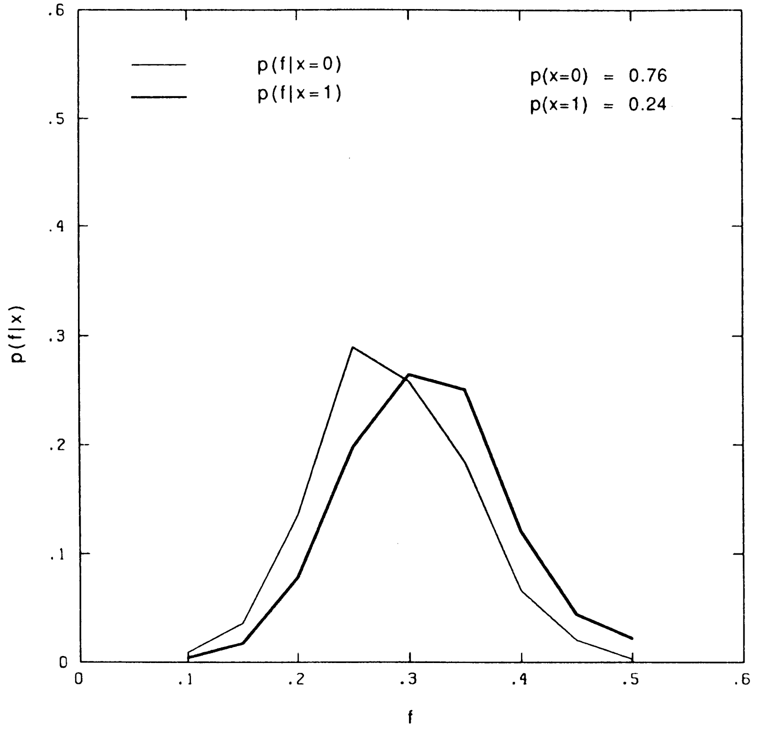

Figure 3. Discrimination diagram for monthly temperature forecasts.

Some statistics related to the joint, conditional, and marginal distributions are included in Table 1. These statistics (Table 1a) involve only the forecast probabilities assigned to the below-normal category. Comparison of $\bar{f}$ and $\bar{x}$ reveals that the MT forecasts exhibit some overall bias (i.e., overforecasting), whereas the MP forecasts exhibit no systematic bias at all. The standard deviations of both types of forecasts are considerably smaller than the respective standard deviations of the observations, reflecting in part a basic difference between these two quantities (i.e., the f's are probabilities whereas the x's are binary numbers). Note also that s_f(MT) is almost twice as large as s_f(MP). The correlation between f and x is relatively modest for both the MT and MP forecasts, with the magnitude of r_{fx}(MT) being more than twice that of r_{fx}(MP).

Conditional means of the distributions $p(f|x=1)$ and $p(f|x=0)$ also are included in Table 1a. The difference between these means is an overall one-dimensional measure of discrimination. In the case of both types of forecasts this difference is quite small, indicating relatively little discrimination (cf. Figs. 3 and 4). Since the conditional means of the distributions $p(x|f)$ represent the points that define the empirical curves in the reliability diagrams (see Figs. 1a and 2a), these means are not reproduced here.

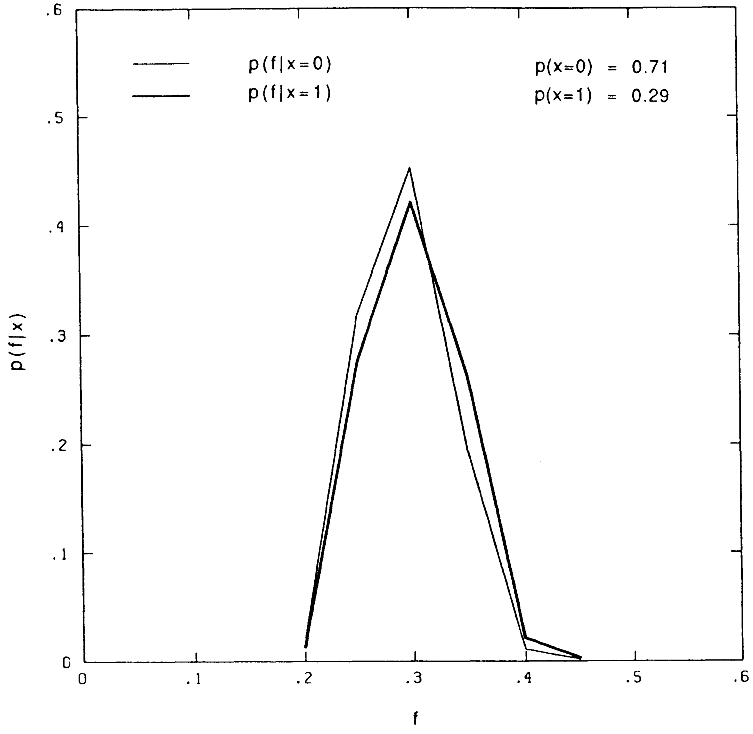

Figure 4. Same as Figure 3, except for monthly precipitation forecasts.

Table 1b contains overall measures of accuracy and skill (these measures involve the forecasts and observations for all three categories). The measure of accuracy is the ranked probability score (RPS) (Epstein 1969; Murphy 1971), a mean-square-error measure defined as the average squared difference between the forecasts and observations after they have been translated into empirical cumulative distribution functions. The skill score (SS) represents the fractional improvement in the RPS of the forecasts over the RPS of constant forecasts based solely on historical climatological probabilities. The latter indicates a 4% (1%) improvement in the case of the MT (MP) forecasts.

Finally, the terms in two decompositions of the RPS (see Murphy and Winkler 1987) are presented in Table 1c. The first decomposition (RPS = s_x^2 + REL - RES) is related to the calibration-refinement factorization of $p(f,x)$. This decomposition indicates that the failure of the MT and MP forecasts to achieve perfect reliability (see Figs. 1a and 2a) contributes very little to the magnitude of this error measure (i.e., REL is small). On

Table 1. Statistics of basic distributions, performance measures, and decompositions of RPS for long-range monthly temperature (MT) and monthly precipitation (MP) forecasts. See section 2b for additional details.

(a) Statistics of joint, conditional, and marginal distributions

Type of forecast	$\bar{f}$	$\bar{x}$	s_f	s_x	r_{fx}	$\bar{f}\|x = 1$	$\bar{f}\|x = 0$	n
MT	0.29	0.24	0.07	0.43	0.19	0.31	0.28	22374
MP	0.29	0.29	0.04	0.46	0.09	0.30	0.29	22262

(b) Performance measures

Type of forecast	RPS	SS
MT	0.42	0.04
MP	0.43	0.01

(c) Decompositions of RPS

Type of forecast	RPS	=	s_x^2 (s_f^2)	+	REL(DIS1)	−	RES(DIS2)
MT	0.42		0.42 (0.01)		0.01 (0.41)		0.02 (0.00)
MP	0.43		0.43 (0.00)		0.00 (0.43)		0.00 (0.00

the other hand, both types of forecasts exhibit very little resolution (i.e., RES is small). (For perfectly reliable forecasts, resolution is equivalent to sharpness.) Moreover, the fact that RES is small and only slightly larger than REL indicates that forecast skill, although positive, is low.

The second decomposition (RPS = s_f^2 + DIS1 - DIS2) is related to the likelihood-base rate factorization of $p(f,x)$. This decomposition indicates that the magnitude of the overall error measure is due to the fact that the conditional mean forecasts (i.e., $\bar{f}|x$=1 and $\bar{f}|x$=0) differ very little from each other and from the unconditional mean forecast (i.e., $\bar{f}$). As a result, these conditional means are quite distant from the observations x = 1 and x = 0 (i.e., DIS2 is small and DIS1 is large; cf. Figs. 3 and 4). The variability of the forecasts (i.e., s_f^2) contributes very little to the magnitude of the RPS.

Space limitations have precluded a more detailed examination and interpretation of the quality of these long-range weather forecasts. For an in-depth diagnostic verification of these forecasts, see Murphy and Huang (1993). Recent applications of these distributions-oriented methods to short-range weather forecasts include Murphy et al. (1989) and Murphy and Winkler (1992).

3. Forecast Value

3.1 Approach and Methodology

Forecasts possess no intrinsic value. Instead, they acquire value through their use by individuals whose decisions are influenced by the information contained in the forecasts. Here we adopt a decision-analytic approach to the problem of assessing the value of forecasts (Raiffa 1968; Bunn 1984; Clemen 1991). In this approach the basic determinants of forecast value are: (1) the courses of action available to the decision maker; (2) the payoff structure associated with the decision-making problem; (3) the quality of the information on which decisions are based in the absence of the forecasts; and (4) the quality of the forecasts themselves (Hilton 1981). Changes in any of these determinants (e.g., the addition or deletion of an action, the reevaluation of a payoff) can lead to changes in forecast value. This prescriptive approach to decision-making problems assumes that decision makers behave in a coherent manner and choose the actions that maximize their expected payoffs (expected payoffs are the probability-weighted averages of the payoffs associated with the outcomes, where the relevant probabilities are derived from the information on which the decisions are based).

In the context of this paper, it is important to note that forecast value depends on both the quality of the forecasts and the quality of the information on which decisions are based in the absence of the forecasts. In particular, if the quality of the forecasts is such that the decision maker makes the same decisions with and without the forecasts, then the forecasts are of no value. To simplify this discussion we will assume that payoffs expressed in monetary terms reflect the true worth of the outcomes to the decision maker. Under this assumption (of a linear utility function), the value of the forecasts is simply the difference between the expressions for expected payoff when the individual's decisions are made with and without the forecasts. The expression for expected payoff with the forecasts involves both the conditional distributions of the observations given the forecasts, $p(x|f)$, and the marginal distribution of the forecasts, $p(f)$. Thus, forecast value generally depends on forecast quality in its full dimensionality.

3.2 Applications

As an example of an application of the decision-analytic approach to forecast-value assessment, we present some results of the so-called fallowing-planting problem (Brown et al. 1986; Katz et al. 1987). This study involved farmers in the drier areas of the northern Great Plains region of the U.S. who must decide each year whether or not to plant a spring wheat crop. These farmers routinely grow a crop every other year, thereby allowing the land to lie fallow in alternate years. The primary reason for this practice is to ensure the availability of sufficient soil moisture at planting time the following year to grow an economically viable crop. However, information in the form of a forecast of growing-season precipitation amount might influence the farmer's decision to let the land lie fallow, especially on those occasions on which the forecast indicated a relatively high probability of above-normal precipitation.

A dynamic decision-analytic model of the fallowing-planting problem was formulated. In this model the farmer was assumed to make the decision to plant or fallow each year for a period of 50 years (the lifetime of the farmer) and to act in such a way as to maximize total expected economic return over this period. The model is dynamic in the sense that next year's soil moisture at planting time is assumed to depend on this year's soil moisture at planting time, the farmer's decision as to whether or not to plant a crop this year, and the precipitation amount in this year's growing season. Soil moisture at the time a crop is planted is the so-called state variable in this dynamic decision-making model (precipitation amount augments soil moisture). The expected yield when a crop is planted is based on a linear regression model in which the independent variables are soil moisture and growing-season precipitation amount. These yields were translated into economic returns to the farmer using 1983 estimates of crop prices, and a discount factor of 0.90 was applied to all future returns (equivalent to an interest rate of about 11%). The method of stochastic-dynamic programming was used to determine the farmer's optimal strategies and to obtain forecast-value estimates.

The forecasts of interest here are growing-season forecasts of precipitation amount. Forecasts of three types were considered: (1) climatological forecasts; (2) imperfect forecasts; and (3) perfect forecasts. Climatological forecasts are forecasts based solely on historical climatological probabilities and represent the zero point on the scale of forecast quality. In the absence of (imperfect) forecasts, the farmers are assumed to base their decisions on climatological forecasts. (In the dynamic decision-analytic model, decisions based on climatological forecasts lead to fallowing and planting in alternate years; see Katz et al. 1987.) Perfect forecasts, although unattainable, provide useful upper bounds on the quality and value of all forecasts. The imperfect forecasts considered here are defined in terms of a simple model of the performance of the seasonal forecasts of precipitation amount produced by the U.S. National Weather Service (these forecasts are similar in format to the monthly forecasts described in section 2b). Specifically, these forecasts are assumed to be completely reliable and to characterize the current level of forecasting performance. Given these assumptions, the variance of the forecasts represents a reasonable one-dimensional measure of forecast quality (a larger variance indicates higher quality).

The decision-analytic model (including the model of growing-season precipitation amount forecasts) was applied to representative farmers in Havre, Montana, and Williston, North Dakota. These two locations were considered to capture some of the natural variability in climatological precipitation amounts over the region. Both soil moisture and climatological precipitation amount are greater at Williston than at Havre. The overall results indicate that the value of imperfect forecasts of current quality is $4 per acre at Havre, whereas these forecasts are of no value at Williston. Perfect forecasts, on the other hand, would be worth $79 per acre at Havre and $47 per acre at Williston. Thus, the value of current forecasts at Havre is about 5% of the value of perfect forecasts.

These results, as well as results related to the value of hypothetical improvements in current forecast quality, are depicted in relative terms in Figure 5. This figure also reveals some general features of the relationships between forecast quality and forecast value (see also section 4). Specifically, quality/value relationships are inherently nonlinear and frequently possess quality thresholds, below which forecasts are of no value. In the fallowing-planting context, considerable improvement in forecast quality would be required before seasonal forecasts are of positive value at Williston. On the other hand, substantial increases in forecast value could be realized by relatively modest improvements in forecast quality at Havre. The "kink" in the quality/value curves corresponding to pseudoperfect forecasts (i.e., forecasts in which a probability of 0.60 is assigned to either the below-normal or above-normal precipitation amount category and the respective category subsequently occurs) may be an artifact of the definition of these forecasts and/or the way in which the forecasts were improved.

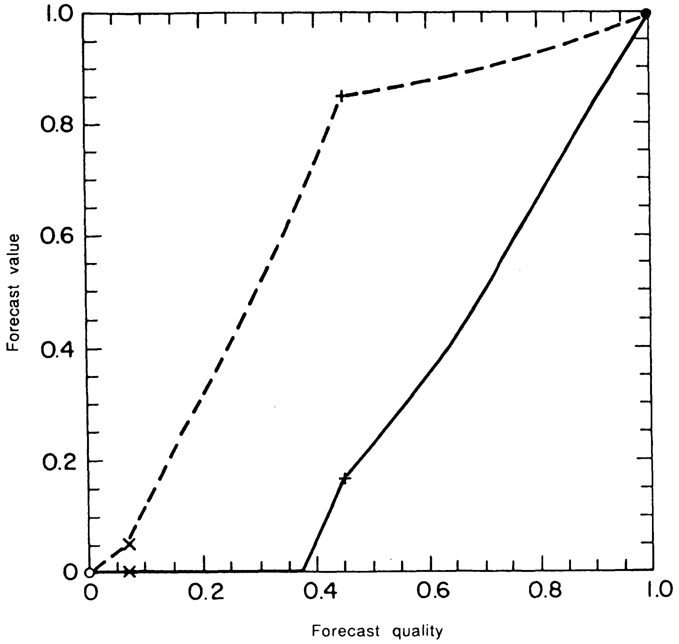

Figure 5. Relationship between forecast quality and forecast value, measured relative to the quality and value of perfect forecasts, in the fallowing-planting problem. Dashed (solid) curve represents Havre (Williston). Point (1,1) denotes perfect forecasts, point (0,0) denotes climatological forecasts, crosses (x) denote forecasts of current quality (in 1983), and pluses (+) denote pseudoperfect forecasts. See section 3b for additional details.

In addition to the fallowing-planting study, several other decision-analytic studies of the value of weather forecasts have been conducted in recent years. These studies include analyses of both prototypical (i.e., idealized) decision-making problems (e.g., Epstein and Murphy 1988; Katz 1993; Katz and Murphy 1990; Murphy 1985; Murphy et al. 1985; Wilks 1991; Winkler et al. 1983) and real-world decision-making problems (e.g., Katz et al. 1982; Mjelde et al. 1988; Sonka et al. 1987; Wilks and Murphy 1986; Wilks et al. 1993). For a more complete list of such studies, see Ehrendorfer and Murphy (1992a).

4. Quality/Value Relationships

4.1 Approach and Methods

This discussion of relationships between forecast quality and forecast value focuses on the following question: What conditions must exist between the joint distributions of forecasting systems F and G, $p(f,x)$ and $p(g,x)$, to ensure that F's forecasts can be judged unambiguously to be better in all respects than G's forecasts (or vice versa)? These conditions are embodied in the sufficiency relation, originally developed by Blackwell

(1953) to compare statistical experiments, further refined by Marschak (1971) in the context of information systems, and introduced into the forecasting literature by DeGroot and Fienberg (1982). According to this relation, *F*'s forecasts are sufficient for *G*'s forecasts when *G*'s likelihoods, $p(g \mid x)$, can be obtained from *F*'s likelihoods, $p(f \mid x)$, by a stochastic transformation. When such a stochastic transformation exists, it consists of a set of conditional probabilities defined over all possible combinations of *F*'s and *G*'s forecasts, and these conditional probabilities are used as weights to transform *F*'s likelihoods into *G*'s likelihoods (e.g., see Ehrendorfer and Murphy 1992b). The existence of such a stochastic transformation possesses two important consequences: (1) *F*'s forecasts are of higher quality than *G*'s forecasts in all relevant aspects and (2) *F*'s forecasts are of greater value than *G*'s forecasts to all decision makers regardless of their payoff structures. These powerful consequences make the sufficiency relation an attractive framework within which to perform comparative evaluation studies and to investigate quality/value relationships. However, the stringent conditions imposed by the sufficiency relation raise questions about its applicability in real-world situations.

4.2 Applications

Relationships between forecast quality and forecast value, in the context of comparative evaluation, are illustrated here by applying the sufficiency relation to prototypical forecasting systems that produce forecasts similar in format to the forecasts considered in sections 2b and 3b. Each forecast specifies the probabilities of three events, which represent below-normal, near-normal, and above-normal weather conditions, respectively. Moreover, the forecasting systems are constrained to use only three distinct forecasts: (1) a climatological forecast, with probabilities equal to the historical climatological probabilities (namely, 0.30, 0.40, and 0.30); (2) a forecast with a higher (lower) probability of below-normal (above-normal) conditions; and (3) a forecast with lower (higher) probability of below-normal (above-normal) conditions. In the cases of these latter two non-climatological forecasts, the probability assigned to near-normal conditions remains fixed at its climatological value (0.40). The quality of the forecasts produced by these prototypical forecasting systems can be described completely by two parameters, denoted here by δ and π. The parameter $\delta(-0.30 \le \delta \le 0.30)$ represents the magnitude of the deviation of the non-climatological forecast from the climatological forecast, whereas the parameter $\pi(0 \le \pi \le 0.50)$ specifies the relative frequency with which each non-climatological forecast is used (for a more detailed description of these forecasting systems, see Ehrendorfer and Murphy 1992b). The implications of the sufficiency relation for the comparative evaluation of these prototypical forecasting systems can be investigated within the framework of a sufficiency diagram. For these forecasting systems, this diagram is two-dimensional with coordinates δ and π. An example of such a sufficiency diagram is presented in Figure 6; in this case, the reference system F is defined by $\delta = -0.10$ and $\pi = 0.15$. Given *F*, the two-dimensional sufficiency diagram is divided into three regions: (1) a region *S* containing the systems *G* for which *F* is sufficient (denoted by diamonds); (2) regions *S′* containing the systems *G* that are sufficient for *F* (denoted by crosses); and (3) regions *I* containing the systems *G* that are insufficient for *F* (blank). Systems *G* represented by regions *I* are neither sufficient for *F*, nor is *F* sufficient for these systems.

Inspection of this diagram reveals that systems *G* that use more extreme non-climatological forecasts more frequently than the reference system *F* are sufficient for *F*. However, it also can be seen that systems *G* that use the non-climatological forecasts less frequently than the reference system F can still be sufficient for F, provided that the non-climatological forecast used by G is extreme enough. On the other hand, a system G that makes more frequent use of a less extreme non-climatological forecast can never be sufficient for F. Since sufficiency is determined by the values of the parameters that define the respective forecasting systems, it provides a coherent approach to the problem of comparative evaluation. Unfortunately, as this example illustrates, the stringent conditions imposed by the sufficiency relation may lead to situations in which the forecasting systems of interest are not comparable, in the sense that they are insufficient for each other. To investigate other features of quality/value relationships in this context, we introduce specific measures of both forecast quality and forecast value. As a one-dimensional measure of quality (in particular that aspect of quality called accuracy), we use the expected ranked probability score (ERPS) (see Ehrendorfer and Murphy 1992b).

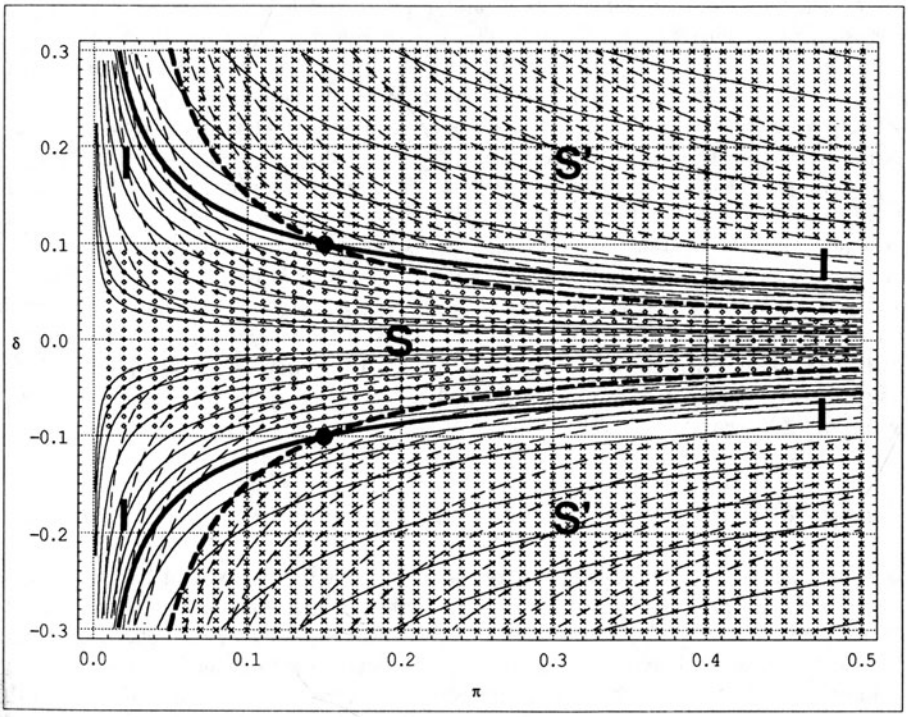

Figure 6. Example of a sufficiency diagram. The large dot represents the reference system $F(\delta = -0.10$ and $\pi = 0.15)$. The regions denoted by S (diamonds), S' (crosses), and I (blank) represent those alternative systems G for which F is sufficient, which are sufficient for F, and which are insufficient for F, respectively. Broken (solid) lines represent VF (ERPS) - contoured at unequal spacing - with decreasing (increasing) numerical values from the lower-right and upper-right corners towards the middle of the diagram. See section 4b for additional details.

To measure the value of these forecasts, it is necessary to consider a specific decision-making problem. Here we consider an extension of the so-called cost-loss ratio problem (see Murphy 1985), in which a decision maker must choose among three actions involving different levels of protection against adverse weather conditions that can occur on three different levels of severity. The payoff structure for this simple problem is described by a single parameter, the cost-loss ratio. The expected value of the forecasts (VF) is calculated as the difference between the expected expense when decisions are based on climatological information [i.e., on $p(x)$] and the expected expense when decisions are based on the forecasts [i.e., on $p(x|f)$ and $p(f)$] (see section 3). Numerical values of ERPS and VF are displayed in Figure 6 in the form of isopleths, where the solid lines are ERPS-contours and the'broken lines are VF-contours (the value of the cost-loss ratio was taken to be 0.3). Examination of these isopleths leads to two important conclusions. First, a one-dimensional measure of quality such as the ERPS is inadequate in this situation because it cannot discriminate between sufficiency and insufficiency. Specifically, isopleths of ERPS pass through regions in which F is sufficient for G (or G is sufficient for F) and regions in which F and G are insufficient for each other. Second, forecast quality, as measured by a one-dimensional measure of accuracy, and forecast value do not possess a one-to-one relationship. A given isopleth of ERPS intersects more than one isopleth of VF (and vice versa), which implies that an increase in forecast accuracy (as measured by ERPS) can be associated with a decrease in forecast value. The existence of such accuracy/value reversals raises serious questions about the use of one-dimensional measures of aspects of forecast quality as surrogates for measures of forecast value. To establish whether system F is better in all respects than system G it is necessary to consider forecast quality in its full dimensionality, as embodied in the sufficiency relation or some other equivalent representation of the relationships between the respective forecasts and observations.

Other applications of the sufficiency relation as a means of investigating (inter alia) various aspects of the relationship between forecast quality and forecast value include Ehrendorfer and Murphy (1988), Krzysztofowicz (1992), Krzysztofowicz and Long (1991a, 1991b), and Murphy and Ye (1990). Many of the forecast-value studies referenced at the end of section 3b also contain results and discussions related to quality/value relationships for weather forecasts.

5. Conclusion

This paper has addressed the problem of forecast evaluation, which includes the assessment of both forecast quality (forecast verification) and forecast value. A general frame work for forecast verification based on the joint distribution of forecasts and observations - and on conditional and marginal distributions associated with factorizations of this joint distribution - has been described. This framework is consistent with the multidimensional nature of verification problems and the multifaceted nature of forecast quality. From the perspective of this framework, forecast verification is seen as a means of obtaining a coherent and complete assessment of forecast quality, identifying basic strengths and weaknesses in forecasting performance, and guiding efforts to improve the systems and models (numerical, statistical, and conceptual) that are used to produce the forecasts. Some results of an application of verification methods consistent with this framework to long-range weather forecasts were presented to illustrate this distributions-oriented approach to verification problems. Since most forecasts in other fields (e.g., business, economics) possess formats similar to those of weather forecasts, these distributions-oriented methods also could be used to verify such forecasts.

A decision-analytic approach to forecast-value assessment was outlined. An application of this approach to a problem involving fallowing-planting decisions and long-range weather forecasts was described and forecast-value estimates were reported. Since these estimates are derived from an approach based on an optimization process (i.e., the decision maker is assumed to choose courses of action that maximize expected payoffs or minimize expected expenses), it follows that the relationship between forecast quality and forecast value is nonlinear. Moreover, this and other decision-analytic case studies have demonstrated that many forecast-sensitive decision-making problems involve quality thresholds, below which the forecasts are of no value.

Relationships between forecast quality and forecast value in the context of comparative evaluation were described with the aid of the sufficiency relation. This relation, which defines a stochastic transformation between the likelihoods associated with the two forecasting systems of interest [i.e., between $p(f \mid x)$ and $p(g \mid x)$; see section 4a], embodies the conditions under which one system can be judged to be better in all respects - that is, in terms of both quality and value - than the other system. In an application involving prototypical long-range weather forecasting systems, it was shown that the stringent conditions imposed by the sufficiency relation imply that it is not always possible to show that one system is sufficient for another system. Moreover, this application was used to demonstrate the possibility of so-called quality/value reversals (i.e., decreases in value associated with increases in an aspect of quality), which can occur when the multifaceted nature of forecast quality is not respected.

Although the evaluation methods described in this paper appear to provide a reasonable framework for assessing both forecast quality and forecast value, as well as for investigating quality/value relationships, a variety of methodological issues and practical problems exist that warrant further attention. With regard to forecast quality (forecast verification), for example, it is not entirely clear what constitutes an adequate or complete assessment of forecast quality. In the context of absolute verification, is it always necessary to examine the conditional and marginal distributions associated with *both* factorizations of the joint distribution of forecasts and observations (in the sense that the respective factorizations contain different information)? This question assumes even greater significance in the context of comparative verification, in view of the substantially greater complexity and dimensionality of these problems. In this regard, it would be useful to explore ways of reducing the dimensionality of verification problems. One possibility would be to fit statistical models to the joint distribution of forecasts and observations, or to the conditional and marginal distributions associated with factorizations of this joint distribution. Given that acceptable fits can be obtained, forecast verification could be based on the parameters of the model(s), thereby reducing dimensionality. Use of statistical models (instead of empirical relative frequencies) in the verification process also should reduce the impact of sampling variability on the results of such assessments.

With regard to forecast value and quality/value relationships, it is clear that additional case studies of forecastsensitive decision-making problems are needed. In addition to yielding forecast-value estimates, such studies would provide further insight into the general and specific characteristics of quality/value relationships as well as potentially useful information regarding the structure of decision makers' payoff functions. Even partial knowledge of the structure and behavior of payoff functions in particular decision-making situations could be used as a basis for developing "tailored" versions of the sufficiency relation (the basic version of the sufficiency relation assumes that nothing is known about the decision makers' payoff functions). Tailored versions of the sufficiency relation generally would impose less stringent conditions on the joint distributions associated with the respective forecasting systems, thereby increasing the likelihood that one system could be judged unambiguously superior (or inferior) to another system in the situations of concern. In view of the strong conclusions that can be drawn when the conditions for sufficiency are met, efforts to enhance the practical applicability of this relation appear to be especially worthwhile.

From a practical point of view, the methods commonly used to evaluate forecasts need to be improved (they are incomplete and potentially misleading; see section 1). In the case of forecast quality, for example, the distributions-oriented approach described here provides the basis for a more coherent and insightful body of verification methods than is generally used in practice. Moreover, in view of recent technological developments, organizations involved in routine forecasting operations should develop :on-line" forecast verification systems that can provide a wide variety of outputs in real time (to satisfy the needs of managers, modelers, and forecasters). Such systems, if properly designed and implemented, would be a valuable source of information for those individuals who are responsible for formulating forecasts as well as for those individuals who are concerned primarily with improving forecasting methods and models.

Finally, the discussion of forecast verification in this paper has focused on the use of the verification process as a means of obtaining insight into the basic strengths and weaknesses in forecasting performance. Clearly, this process is an essential part of any effort to develop and implement state-of-the-art forecasting systems and necessarily plays a vital role in efforts to improve such systems. Notwithstanding the importance of these particular uses of the verification process and its output, it should be noted here that this process also can produce information related to the predictability of the events of concern and the uncertainty inherent in forecasts of these events. Statistical studies of the deterioration of forecast quality (or aspects of quality) as lead time increases can provide empirical estimates of predictability. Moreover, studies of the joint distribution of forecasts and observations, and the associated conditional and marginal distributions, can produce quantitative information regarding the overall uncertainty inherent in the forecasts. In the future it may prove useful to compare empirical estimates of predictability and uncertainty derived from the verification process with theoretical estimates of these quantities obtained from modeling studies.

Acknowledgment

This work was supported in part by the National Science Foundation under Grant SES-9106440.

References

Blackwell, D., (1953), 'Equivalent comparisons of experiments'. *Annals of Mathematical Statistics*, **24**, 265-272.

Brown, B.G., R.W. Katz, and A.H. Murphy, (1986), 'On the economic value of seasonal-precipitation forecasts: the fallowing-planting problem'. *Bulletin of the American Meteorological Society*, **67**, 833-841.

Bunn, D., (1984), *Applied Decision Analysis*. New York, McGraw-Hill, 251 pp.

Clemen, R.T., (1991), *Making Hard Decisions: An Introduction to Decision Analysis*. 4 Boston, PWS-Kent, 557 pp.

DeGroot, M.H., and S.E. Fienberg, (1982), 'Assessing probability assessors: calibration and refinement'. *Statistical Decision Theory and Related Topics III,* **1** (S.S. Gupta and J.O. Burger, Editors). New York, Academic Press, 291-314.

Ehrendorfer, M., and A.H. Murphy, (1988), 'Comparative evaluation of weather forecasting systems: sufficiency, quality, and accuracy'. *Monthly Weather Review,* **116**, 1757-1770.

Ehrendorfer, M., and A.H. Murphy, (1992a), 'On the relationship between the quality and value of weather and climate forecasting systems'. *Idojaras,* **96**, 187-206.

Ehrendorfer, M., and A.H. Murphy, (1992b), 'Evaluation of prototypical climate forecasts: the sufficiency relation'. *Journal of Climate,* **5**, 876-887.

Epstein, E.S., (1969), 'A scoring system for probabilities of ranked categories'. *Journal of Applied Meteorology,* **8**, 985-987.

Epstein, E.S., and A.H. Murphy, (1988), 'Use and value of multiple-period forecasts in a dynamic model of the cost-loss ratio situation'. *Monthly Weather Review* **116**, 746-761.

Hilton, R.W., (1981), 'The determinants of information value: synthesizing some general results'. *Management Science,* **27**, 57-64.

Katz, R.W., (1993), 'Dynamic cost-loss ratio decision-making model with an autocorrelated climate variable'. *Journal of Climate,* **6**, 151-160.

Katz, R.W., B.G. Brown, and A.H. Murphy, (1987), 'Decision-analytic assessment of the economic value of weather forecasts: the fallowing/planting problem'. *Journal of Forecasting,* **6**, 77-89.

Katz, R.W., and A.H. Murphy, (1990), 'Quality/value relationships for imperfect weath\-er forecasts in a prototype multistage decision-making model'. *Journal of Forecasting,* **9**, 75-86.

Katz, R.W., A.H. Murphy, and R.L. Winkler, (1982), 'Assessing the value of frost forecasts to orchardists: a dynamic decision-making approach'. *Journal of Applied Meteorology,* **21**, 518-531.

Krzysztofowicz, R., (1992), 'Bayesian correlation score: a utilitarian measure of forecast skill'. *Monthly Weather Review,* **120**, 208-219.

Krzysztofowicz, R., and D. Long, (1991a), 'Forecast sufficiency characteristic: construction and application'. *International Journal of Forecasting,* **7**, 39-45.

Krzysztofowicz, R., and D. Long, (1991b), 'Beta likelihood models of probabilistic forecasts'. *International Journal of Forecasting,* **7**, 47-55.

Marschak, J., (1971), 'Economics of information systems'. *Journal of the American Statistical Association,* **66**, 191-219.

Mjelde, J.W., S.T. Sonka, B.L. Dixon, and P.J. Lamb, (1988), 'Valuing forecast characteristics in a dynamic production system'. *American Journal of Agricultural Economics,***70**, 674-684.

Murphy, A.H., (1971), 'A note on the ranked probability score'. *Journal of Applied Meteorology,* **10**, 155-156.

Murphy, A.H., (1985), 'Decision making and the value of forecasts in a generalized model of the cost-loss ratio situation'. *Monthly Weather Review,* **113**, 362-369.

Murphy, A.H., (1991), 'Forecast verification: its complexity and dimensionality'. *Monthly Weather Review,* **119**, 1590-1601.

Murphy, A.H., (1993), 'What is a good forecast? An essay on the nature of goodness in weather forecasting'. *Weather and Forecasting,* **8**, 281-293.

Murphy, A.H., B.G. Brown, and Y.-S. Chen, (1989), 'Diagnostic verification of temperature forecasts'. *Weather and Forecasting,* **4**, 485-501.

Murphy, A.H., and J. Huang, (1993), 'Diagnostic verification of the Climate Analysis Center's probabilistic monthly and seasonal forecasts'. In preparation.

Murphy, A.H., R.W. Katz, R.L. Winkler, and W.-R. Hsu, (1985), 'Repetitive decision-making and the value of forecasts in the cost-loss ratio situation: a dynamic model'. *Monthly Weather Review*, **113**, 801-813.

Murphy, A.H., and R.L. Winkler, (1987), 'A general framework for forecast verification'. *Monthly Weather Review*, **115**, 1330-1338.

Murphy, A.H., and R.L. Winkler, (1992), 'Diagnostic verification of probability forecasts'. *International Journal of Forecasting*, **7**, 435-455.

Murphy, A.H., and Q. Ye, (1990), 'Comparison of objective and subjective precipitation probability forecasts: the sufficiency relation'. *Monthly Weather Review*, **118**, 1783-1792.

Raiffa, H., (1968), *Decision Analysis*. Reading, MA, Addison Wesley, 309 pp.

Sonka, S.T., J.W. Mjelde, P.J. Lamb, S.E. Hollinger, and B.L. Dixon, (1987), 'Valuing climate forecast information'. *Journal of Climate and Applied Meteorology*, **26**, 1080-1091.

Wilks, D.S., (1991), 'Representing serial correlation of meteorological events and forecasts in dynamic decision-analytic models'. *Monthly Weather Review*, **119**, 1640-1662.

Wilks, D.S., and A.H. Murphy, (1986), 'A decision-analytic study of the joint value of seasonal precipitation and temperature forecasts in a choice-of crop problem'. *Atmosphere-Ocean*, **24**, 353-368.

Wilks, D.S., R.E. Pitt, and G.W. Fick, (1993), 'Modeling optimal alfalfa harvest scheduling using short-range weather forecasts'. *Agricultural Systems*, **42**, 277-305.

Winkler, R.L., A.H. Murphy, and R.W. Katz, (1983), 'The value of climate information: a decision-analytic approach'. *Journal of Climatology*, **3**, 187-197.