A General Framework for Forecast Verification

ALLAN H. MURPHY

Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331

ROBERT L. WINKLER

Fuqua School of Business, Duke University, Durham, NC 27706

(Manuscript received 28 September 1986, in final form 18 December 1986)

ABSTRACT

A general framework for forecast verification based on the joint distribution of forecasts and observations is described. For further elaboration of the framework, two factorizations of the joint distribution are investigated: 1) the calibration-refinement factorization, which involves the conditional distributions of observations given forecasts and the marginal distribution of forecasts, and 2) the likelihood-base rate factorization, which involves the conditional distributions of forecasts given observations and the marginal distribution of observations. The names given to the factorizations reflect the fact that they relate to different attributes of the forecasts and/or observations. Several examples are used to illustrate the interpretation of these factorizations in the context of verification and to describe the relationship between the respective factorizations.

Some insight into the potential utility of the framework is provided by demonstrating that basic elements and summary measures of the joint, conditional, and marginal distributions play key roles in current verification methods. The need for further investigation of the implications of this framework for verification theory and practice is emphasized, and some possible directions for future research in this area are identified.

1. Introduction

Forecast verification is the process and practice of determining the quality of forecasts, and it represents an essential component of any scientific forecasting system. As such, forecast verification serves many important purposes. These purposes include assessing the state of the art of forecasting and recent trends in forecast quality, improving forecasting procedures and ultimately the forecasts themselves, and providing users with information needed to make effective use of the forecasts.

Meteorologists have devoted considerable attention to forecast verification, in terms of both the development of verification methods and the application of these methods in operational and experimental contexts (e.g., see Daan, 1984; Murphy and Daan, 1985). Verification measures have been formulated with a variety of purposes in mind and for a multitude of different situations. Here, the term "situations" relates to considerations such as the nature of the underlying variable (continuous/discrete, ordered/unordered, bounded/unbounded), the climatological likelihood of occurrence of relevant events (frequent/infrequent), and the type of forecast (categorical/probabilistic). As a result, verification measures have tended to proliferate, with relatively little effort being made to develop general concepts and principles, to investigate the relationships between measures, or to examine their relative strengths and weaknesses. This state of affairs has impeded the development of a science of forecast verification and undermined the utility of many verification concepts and methods.

A need exists for a general framework for forecast verification. To be useful, such a framework should (*inter alia*) (i) unify and impose some structure on the overall body of verification methodology, (ii) provide insight into the relationships among verification measures, and (iii) create a sound scientific basis for developing and/or choosing particular verification measures in specific contexts. Moreover, such a framework should minimize the number of distinct situations that must be considered. The primary purpose of this paper is to describe a framework that appears to meet many of these goals.

The basis for the framework for verification described here is the joint distribution of forecasts and observations, which contains all of the relevant information. This basic distribution is defined in section 2 and specific examples involving discrete and continuous variables are presented. Section 2 also contains a discussion of the relationships between the properties of the joint distribution and the quality of forecasts. In section 3, we describe two factorizations of the joint distribution, each of which leads to a distinct approach to forecast verification. The calibration-refinement factorization involves the conditional distributions of the observations given the forecasts and the marginal

distribution of the forecasts, whereas the likelihood-base rate factorization involves the conditional distributions of the forecasts given the observations and the marginal distribution of the observations. These factors are defined and examples of the conditional and marginal distributions are presented in section 3. This section also includes an interpretation of these factors in the context of verification and a discussion of the relationship between the alternative factorizations. Section 4 discusses current verification methodology from the perspective provided by the joint distribution and its respective factorizations. Section 5 contains a brief summary and some concluding remarks, including a short discussion of possible directions for future work in this area.

2. Joint distribution of forecasts and observations

The joint distribution of forecasts and observations provides the basis for our unified framework for forecast verification. Denoting the forecast by f and the observation (the observed event or the observed value of the variable of interest) by x, we let $p(f, x)$ represent the joint distribution of f and x. This distribution contains information about the forecast, about the observation, and about the relationship between the forecast and the observation. Concern with this relationship is in the spirit of DeGroot and Fienberg (1982, 1983), who note that the basic data in this context consist of pairs of forecasts and observations.

This framework is intended to be quite general. An observation could be a simple observed event in a dichotomous situation (e.g., precipitation/no precipitation) or a situation with multiple categories (e.g., clear/scattered/broken/overcast); the observed value of a single variable (e.g., maximum temperature, amount of precipitation); or even a multivariate observation (e.g., a joint observation of wind speed and visibility, a joint observation of temperature and type of precipitation). The forecast could be a categorical forecast (e.g., "rain tomorrow," "high temperature tomorrow 74°F"), a forecast with a qualitative expression of uncertainty (e.g., "chance of rain tomorrow"), or a probability forecast (e.g., "60% chance of rain tomorrow"). A probability forecast is the most informative of these three types of information, and a categorical forecast can be thought of as a special case of a probability forecast with probability one assigned to a particular event or value of the variable of interest.

In theory, $p(f, x)$ represents the joint probability distribution of f and x. In this sense, it is an ex ante distribution that is useful in comparing forecasters or forecast systems and in assisting decision makers who must choose a forecaster or system to consult for future forecasts. The theoretical joint probability distribution contains all information that is relevant to the evaluation of a forecaster or forecast system.

For decision-making purposes, the ex ante approach is appropriate, and a decision maker can assess a distribution $p(f, x)$, using any available information. Verification, however, usually involves ex post evaluation based on empirical analysis. Since our concern in this paper is with forecast verification, we will interpret $p(f, x)$ as an empirical relative frequency distribution based on a sample of past forecasts and observations $\{(f_i, x_i); i = 1, \cdots, n\}$. The relative frequency distribution summarizes the sample of data. In fact, if the time order of the forecast–observation pairs is not of interest, then the relative frequency distribution captures all relevant information in the sample. We can think of the relative frequency distribution as being an estimate of the theoretical probability distribution that we would ideally like to know.

Example A. The simplest verification situation involves categorical forecasts of a dichotomous event such as precipitation/no precipitation. Here we define

$$f = \begin{cases} 1, & \text{if precipitation is forecast,} \\ 0, & \text{if no precipitation is forecast,} \end{cases}$$

and

$$x = \begin{cases} 1, & \text{if precipitation occurs,} \\ 0, & \text{if precipitation does not occur.} \end{cases}$$

The joint distribution of f and x can be displayed in terms of a 2×2 contingency table containing the relative frequencies with which (f, x) equals (1, 1), (1, 0), (0, 1), and (0, 0).

Example B. Probability of precipitation (PoP) forecasts represent a generalization of Example A to the case where f takes on, say, 13 possible values (0.00, 0.02, 0.05, 0.10, 0.20, . . . , 0.90, 1.00). Therefore, the contingency table in this situation is 13×2 and contains relative frequencies such as the joint relative frequency with which $f = 0.20$ *and* $x = 1$.

Example C. Temperature forecasting provides an example with an underlying variable (temperature) that is essentially continuous. If x represents maximum temperature and f represents a categorical (point) forecast of maximum temperature, then the (f, x)-pairs can be displayed graphically as a scatter diagram of points in two-dimensional space. When temperature is forecast and observed to the nearest degree (Fahrenheit or Celsius) or in intervals of values, $p(f, x)$ can be depicted in terms of a contingency table.

If the joint distribution of forecasts and observations provides the basis for our general framework for forecast verification, we might ask which joint distributions indicate "good" forecasting and which indicate "bad" forecasting. On the good side, consider the extreme case of perfect forecasts. In Examples A and B, perfect forecasting implies that all of the relative frequencies are zero except for $p(1, 1)$ and $p(0, 0)$. In Example C, perfect forecasting implies that all of the (f, x)-pairs are on the line $f = x$ or on the principal diagonal of

the contingency table. Perfect forecasts are an unattainable ideal, but the closer a forecaster or forecast system can approach that ideal, the better.

The bad side is a bit more complicated, since many ways exist in which forecasts can be bad. Moreover, the perspective taken by the user can influence the definition of a bad forecast. For example, if the user takes the forecasts at face value, then the worst distributions in Examples A and B would consist of all of the relative frequencies being equal to zero except for $p(1, 0)$ and $p(0, 1)$, implying that the forecaster is always "dead wrong." However, to a user who realizes that the forecasts are always dead wrong, such forecasts are very helpful because the user can convert them appropriately ("if the forecast says rain, I am sure that it will not rain," etc.). Perhaps the worst, least useful forecasts are those yielding a joint distribution such that the column of relative frequencies $p(f, 1)$ is a constant multiple of the column of relative frequencies $p(f, 0)$, which implies that the forecasts and observations are independent in a statistical sense.

Keeping in mind our focus on forecast verification, which takes the forecasts at face value, the general statement can be made that we prefer joint distributions that assign high relative frequencies to (f, x)-pairs with f equal to or close to x and low relative frequencies to (f, x)-pairs with f not close to x. To learn more about specific characteristics of a forecaster or forecast system, we need to factor the joint distribution of f and x into a conditional distribution and a marginal distribution. This process can be accomplished in two ways, and the two factorizations are described and discussed in section 3.

3. Factorization of the joint distribution

Although the joint distribution of forecasts and observations contains all information relevant to verification, the information is more accessible when we factor the distribution. Any joint distribution can be factored into a conditional distribution and a marginal distribution in two ways. Thus, in considering both of these factorizations, we obtain two types of conditional distributions and two marginal distributions, each of which relates to particular aspects of verification. The two factorizations are presented in sections 3a and 3b, and some relationships between the factorizations are discussed in section 3c.

a. Calibration-refinement factorization

The first factorization we consider involves the conditional distributions of the observations given the forecasts and the marginal distribution of the forecasts:

$$p(f, x) = p(x|f)p(f). \quad (1)$$

Here we have a conditional distribution $p(x|f)$ for each possible forecast value f and a marginal distribution $p(f)$.

The conditional distribution $p(x|f)$ indicates how often different observations have occurred when a particular forecast f was given. In the case of categorical forecasts of precipitation/no precipitation (Example A), we have two conditional distributions, $p(x|1)$ and $p(x|0)$. The first, $p(x|1)$, tells us the proportion of occasions with precipitation and the proportion with no precipitation among all of the occasions on which a forecast of precipitation was given. The other conditional distribution, $p(x|0)$, tells us the proportion of occasions with and without precipitation given a forecast of no precipitation. Naturally, we would like $p(x = 1|f = 1)$ to be as large as possible and $p(x = 1|f = 0)$ to be as small as possible.

With PoP forecasts (Example B), we have 13 conditional distributions $p(x|f)$ corresponding to the 13 possible values of f. For instance, $f(x = 1|0.60)$ indicates the relative frequency of days with precipitation among the days with a precipitation probability forecast of 0.60. Ideally, we would like to have

$$p(x = 1|f) = f. \quad (2)$$

If (2) is satisfied for all f, then the forecaster or forecast system is said to be perfectly calibrated or perfectly reliable (e.g., Murphy and Winkler, 1977). Thus, the conditional distributions of the observations given the forecasts relate to the calibration or reliability of the forecasts. The values of $p(x = 1|f)$ are often plotted against f in a reliability diagram to depict the calibration of a set of probability forecasts.

When more than two values of x exist, as in temperature forecasting (Example C), the conditional distribution $p(x|f)$ consists of several relative frequencies and cannot be represented by just one of these frequencies. In this case, we can say that the set of forecasts is perfectly calibrated or reliable if

$$E(x|f) = f, \quad (3)$$

where $E(x|f)$ is the expected value of x given the forecast f. In the temperature example, a forecast system is perfectly calibrated if for any forecast f, the average observed temperature is equal to f.

The marginal distribution $p(f)$ indicates how often different forecast values are used. Consider PoP forecasts, for example. If the same forecast is always given, then the forecasts are said not to be refined, or sharp. A forecast system that always simply reported the climatological probability would never have any variation in forecasts and would not be able to distinguish between days with precipitation and days without precipitation. At the other extreme of perfect sharpness would be the forecast system that only gives PoP forecasts of zero and one.

Note that calibration and refinement are two separate concepts and both are of interest for verification purposes. We could have a forecaster who is perfectly

calibrated but shows no refinement (e.g., a forecaster who always takes climatology as the forecast). At the other extreme, we could have a forecaster who appears refined by giving only categorical forecasts of precipitation but who is not at all calibrated. For example, suppose that $p(f = 1) = 0.4$ and $p(f = 0) = 0.6$ but that $p(x = 1|f = 1) = p(x = 1|f = 0) = 0.4$. Neither of these extremes is very helpful. We would prefer a forecaster who is both well calibrated and quite refined. Another perspective is provided by recognizing that we would like the forecaster to be as refined as possible without sacrificing calibration. Maintaining calibration means that we can use the forecasts at face value, and better refinement means that the forecasts distinguish effectively among situations leading to different observations (values of x).

It may also be instructive to reconsider briefly the issue of the worst, least useful forecasts from the perspective of the conditional and marginal distributions, $p(x|f)$ and $p(f)$, introduced in this section. If independence of forecasts and observations is taken as the condition for the least useful forecasts, then this condition implies that $p(x|f) = p(x)$. In this case, the probability of occurrence of the events is independent of the forecasts and equals the respective sample climatological probability. Since this condition leads to a situation in which the forecasts are uninformative with respect to the observations, it seems quite reasonable to describe such forecasts as "least useful."

b. *Likelihood-base rate factorization*

The second factorization involves the conditional distributions of the forecasts given the observations and the marginal distribution of the observations:

$$p(f,x) = p(f|x)p(x). \tag{4}$$

In this factorization we have a conditional distribution $p(f|x)$ for each possible observation x and a marginal distribution $p(x)$.

The conditional distribution $p(f|x)$ indicates how often different forecasts are given before a particular value of x is observed. With categorical forecasts of precipitation/no precipitation (Example A), two conditional distributions exist, $p(f|1)$ and $p(f|0)$. The first, $p(f|1)$, gives the proportion of occasions with forecasts of precipitation and forecasts of no precipitation among those occasions on which precipitation eventually occurs. The second, $p(f|0)$, provides the same information for those occasions on which precipitation does not occur. With PoP forecasts (Example B), we also have two conditional distributions that indicate how often each of the 13 possible forecasts are given preceding days with precipitation and how often they are given preceding days with no precipitation. We would hope that high values of f are given more often when x turns out to be one and that low values of f are given more often when x turns out to be zero.

For a given forecast f, the conditional probabilities $p(f|1)$ and $p(f|0)$ are called the likelihoods associated with that forecast. This terminology, which comes from the field of statistics (e.g., see Winkler, 1972), reflects the fact that these probabilities tell us how likely the forecast f is given that we will observe $x = 1$ and given that we will observe $x = 0$. These likelihoods indicate how well the forecast f discriminates between days with $x = 1$ and days with $x = 0$. When more than two values of x are possible, as in the case of temperature forecasts (Example C), as many likelihoods exist as values of x, and the likelihoods indicate how well the forecast f discriminates among the various values of x. For a forecast f, if $p(f|x)$ is very similar for different x, the forecast is not very discriminatory. In the extreme, when $p(f|x)$ is the same for all x, the forecast is not at all discriminatory and provides us with no useful information about x. When the likelihoods are very different for different x, the forecast is much more discriminatory and hence very informative about x. The forecast f is perfectly discriminatory if $p(f|x)$ equals zero for all values of x except one. Then we can be certain, upon seeing this forecast f, that the value of x corresponding to the nonzero likelihood will occur.

The marginal distribution $p(x)$ indicates how often different values of x occur. With precipitation/no precipitation, it tells us the relative frequency of precipitation and the relative frequency of no precipitation. With observations falling into several categories, it gives the relative frequencies of all of the categories (e.g., clear/scattered/broken/overcast). In the case of a continuous variable such as maximum temperature, $p(x)$ provides an entire distribution for the values of x.

Note that $p(x)$ is the only element of either factorization that does not involve f in any way. It is a characteristic of the forecasting situation, not of the forecaster or forecast system. This characteristic is often referred to as the base rate (e.g., see Lichtenstein et al., 1982). In weather forecasting, the base rate is generally called sample climatology.

If independence of the forecasts and observations is once again taken as the condition characterizing the worst, least useful forecasts, then this condition implies that $p(f|x) = p(f)$. For such forecasts, the conditional distribution of the forecasts given an observation is the same for all observations and is therefore identical to the marginal distribution. In this situation, the observations are uninformative with respect to the forecasts.

The likelihood-base rate factorization gives us a separation of two types of information that may be helpful in predicting x. The base rate reflects historical observations (since it can be viewed as an estimate of the long-term climatology) and indicates how we should predict x in the absence of a forecast. The likelihoods then reflect the new information contained in the forecast and indicate how helpful this forecast is above and beyond the base rate of sample climatology.

c. Relationship between factorizations

The four elements in the two factorizations measure different characteristics of the forecaster or forecast system and of the forecast situation. However, some relationships among these elements must exist, since we know that

$$p(x|f)p(f) = p(f|x)p(x). \tag{5}$$

We can also rewrite (5) in a form known as Bayes' theorem:

$$p(x|f) = \frac{p(x)p(f|x)}{p(f)}. \tag{6}$$

Here the base rate $p(x)$ and the likelihoods $p(f|x)$ are multiplied to combine the two types of information that they represent. This product is then divided by $p(f)$ to normalize and yield $p(x|f)$, which reflects both the base rate information and the information contained in the forecast f.

To illustrate (5) and (6), we consider some data in the context of Example B. The joint distribution presented in Table 1 is based on 2820 PoP forecasts (and the corresponding observations) formulated by a particular National Weather Service (NWS) forecaster at Chicago, Illinois, during the period July 1972 through June 1976. The marginal distributions of f and x are given in the margins of this table, and the conditional distributions $p(x|f)$ and $p(f|x)$ are presented in Table 2. For any pair (f, x), (5) and (6) will be satisfied. For example, if we take $f = 0.40$ and $x = 1$ and use Tables 1 and 2,

$$p(x=1|f=0.40)p(f=0.40) = 0.3662(0.0609) = 0.0223$$

and

$$p(f=0.40|x=1)p(x=1) = 0.0895(0.2493) = 0.0223$$

are equal, as indicated by (5). Also, from (6),

$$p(x=1|f=0.40) = \frac{p(x=1)p(f=0.40|x=1)}{p(f=0.40)} = \frac{0.2493(0.0895)}{0.0609} = 0.3664$$

agrees with $p(x = 1|f = 0.40)$ as given in Table 2 except for a slight discrepancy at the fourth decimal place caused by roundoff error.

To some degree, then, we must have consistency between the factorizations. For example, if a forecast system is perfectly calibrated and perfectly refined, then the likelihoods must be perfectly discriminatory. However, the converse does not hold. The likelihood function can be perfectly discriminatory even though the forecaster is not well calibrated or refined. With PoP forecasts (Example B), suppose that $p(f = 0.40, x = 0) = 0.7$, $p(f = 0.80, x = 1) = 0.3$, and $p(f, x) = 0$ for all other (f, x)-pairs. These forecasts are perfectly discriminatory; f is either 0.40 or 0.80, with precipitation always following forecasts of 0.80 and never following forecasts of 0.40. However, the forecaster is not well-calibrated, since $p(x = 1|f = 0.40) = 0$ ($\neq 0.40$) and $p(x = 1|f = 0.80) = 1$ ($\neq 0.80$). Neither is the forecaster perfectly refined; forecasts of zero and one are never given. In this numerical example the forecaster is really capable of being a perfect forecaster but apparently does not realize it! Of course, perfect forecasting is not generally attainable, and considering extreme examples such as this case may not be particularly informative. However, it does provide some insight into differences between the two factorizations. The likelihood-base rate factorization focuses on how discriminatory a forecast is regardless of the "label" f. If f-values of 0.80 and 0.40 are perfectly discriminatory, then for the purposes of the likelihoods they are just as valuable as if they were one and zero. In this sense, the likelihood-base rate factorization might be thought of as being more oriented toward properties that indicate the *potential* value of the information contained in the forecasts if they are used appropriately. The calibration-refinement factorization is oriented more toward the labels assigned to the forecasts and toward the actual properties of the forecasts when they are taken at face value.

The difference between the factorizations in terms of impact upon the users of the forecasts depends on their degree of sophistication. If the users are sophisticated and are aware of the base rate and the likelihoods, then poor calibration and refinement are irrelevant for practical purposes. The base rate represents a user's information before seeing the forecast, and the likelihoods enable the user to revise this information after seeing the forecast. Note, however, that this process does require some effort and knowledge of the characteristics of the forecaster or forecast system. For less sophisticated users, who take the forecasts at face

TABLE 1. Joint and marginal distributions of PoP forecasts and observations for NWS forecaster at Chicago, Illinois.

		f													
		0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	$p(x)$
x	1	0.0014	0.0018	0.0028	0.0138	0.0316	0.0255	0.0223	0.0383	0.0309	0.0426	0.0216	0.0135	0.0032	0.2493
	0	0.0557	0.0500	0.0972	0.1901	0.1773	0.0656	0.0386	0.0337	0.0213	0.0138	0.0074	0.0000	0.0000	0.7507
$p(f)$		0.0571	0.0518	0.1000	0.2039	0.2089	0.0911	0.0609	0.0720	0.0522	0.0564	0.0290	0.0135	0.0032	

TABLE 2. Conditional distributions $p(x|f)$ and $p(f|x)$ for NWS forecaster at Chicago, Illinois.

	f												
	0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$p(x=1\|f)$	0.0245	0.0347	0.0280	0.0677	0.1513	0.2799	0.3662	0.5319	0.5920	0.7553	0.7448	1.0000	1.0000
$p(x=0\|f)$	0.9755	0.9653	0.9720	0.9323	0.8487	0.7201	0.6338	0.4681	0.4080	0.2447	0.2552	0.0000	0.0000
	f												
	0.00	0.02	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$p(f\|x=1)$	0.0056	0.0072	0.0112	0.0554	0.1268	0.1023	0.0895	0.1536	0.1239	0.1709	0.0866	0.0542	0.0128
$p(f\|x=0)$	0.0742	0.0666	0.1295	0.2532	0.2362	0.0874	0.0514	0.0449	0.0284	0.0184	0.0099	0.0000	0.0000

value, calibration and refinement are important. A poorly-calibrated forecasting system would lead such users astray. Since we cannot assume that consumers of weather forecasts are sophisticated, properties such as calibration and refinement are important in practice although they might be dismissed as relatively unimportant in theory in a perfect world with completely knowledgeable and rational users of forecasts.

4. The framework and current verification methods

The joint distribution of forecasts and observations described in section 2, together with the factorizations of this distribution discussed in section 3, appear to constitute a potentially useful, general framework for forecast verification. It now seems appropriate to ask to what extent this framework meets the conditions for usefulness set forth in section 1. For example, does the framework unify and impose some structure on the body of verification methodology? Does it provide insight into the relationships among verification measures? In this section, we will attempt to obtain some tentative answers to such questions, primarily by examining current verification methods from perspectives provided by this framework. In-depth evaluations of the framework and its ability to produce a sound scientific basis for developing and choosing verification methods are beyond the scope of this paper and will necessarily require a considerably more detailed and comprehensive assessment than that presented here.

First, we believe that the framework embodied by the joint distribution of forecasts and observations reveals the inherent unity of forecast verification. Recall that the joint distribution contains *all* of the relevant information. Therefore, whether the variable of concern is discrete or continuous or the forecasts are categorical or probabilistic, this distribution possesses the basic ingredients required for verification purposes. As a result, it should not be necessary to approach the verification problem from different perspectives in different situations; all situations should yield to a common approach.

Notwithstanding the overall unity provided by the joint distribution, the existence of two factorizations of this distribution appears to suggest that (at least) two distinct approaches to forecast verification are available. An approach based on the calibration-refinement factorization would naturally focus on the calibration (reliability) and the refinement (sharpness) of the forecasts. Alternatively, an approach based on the likelihood-base rate factorization would focus on the ability of the forecasts to discriminate among the observations and on the base rates (sample climatological probabilities) of these observations. In this regard, DeGroot and Fienberg (1982, 1983) employ the calibration-refinement factorization, although they also mention the likelihood-base rate factorization. The latter is the basis of the work by Lindley (1982) on improving probability forecasts. We believe that these factorizations constitute *complementary* rather than alternative ways to approach the verification problem. After all, as noted in section 3, the two factorizations are concerned with different attributes of the forecasts and/or observations. Thus, a complete verification study would necessarily involve the evaluation of factors associated with both factorizations.

We believe that the framework also imposes some useful structure on the body of verification methodology. For example, since the joint, conditional, and marginal distributions possess the basic ingredients needed for verification purposes, the distributions themselves—or elements thereof—can be considered to represent verification measures. Although recognition of this fact is implicit in many verification procedures and practices, it has seldom been explicitly acknowledged and its implications have not been fully explored. Moreover, the fundamental role played by these distributions suggests that common summary measures (means, variances, etc.) of such distributions should be useful verification measures in many situations. To what extent do current verification measures reflect these perspectives?

To answer this question in part, we consider here the conditional distributions of observations given forecasts and of forecasts given observations [$p(x|f)$ and $p(f|x)$, respectively] and briefly examine current verification procedures and practices in situations rep-

resented by Examples A, B, and C. In the case of categorical forecasts of precipitation occurrence (Example A), attention has generally focused on measures based on elements of the joint and/or marginal distributions, such as the fraction correct and various skill scores (e.g., see Daan, 1984). However, recognition of some deficiencies in these measures has led to the consideration of conditional probabilities such as $p(x = 1 | f = 1)$ and $p(f = 1 | x = 1)$ in certain situations. In the early verification literature, these conditional probabilities were referred to as "prefigurance" and "post agreement," but more recently the terms "hit rate" and "probability of detection" have gained favor. [In reality, it is $p(x = 0 | f = 1)$ rather than $p(x = 1 | f = 1)$ that is usually computed, and the former is generally referred to as the "false alarm rate."] These conditional probabilities are particularly useful measures of forecast quality in situations in which the base rates (i.e., climatological probabilities) of the two events are quite dissimilar.

The conditional distributions of observations given forecasts—or at least the expectations of these distributions—have been used to evaluate precipitation probability forecasts (Example B) for many years. As noted in section 3a, comparison of $p(x|f)$ and f in this context provides a means of assessing the reliability (calibration) of such forecasts. Unlike $p(x|f)$, the conditional distributions of the forecasts given the observations, $p(f|x)$, have generally not been considered in evaluating precipitation probability forecasts. Recent exceptions to this statement can be found in a paper by Mason (1982) describing verification methods based on signal detection theory and in an example discussed by Lindley (1982).

In the case of categorical forecasts of temperature (Example C), the conditional distributions $p(x|f)$ and $p(f|x)$ have seldom been reported in published verification studies. Interest here has centered almost exclusively on overall measures of quality such as the mean absolute error or mean square error. It should be noted that verification based solely on such measures assumes, in effect, that all errors of the same size are equally important and that the conditional distributions are similar for all values of f or x.

This admittedly superficial examination of the current use of the conditional distributions $p(x|f)$ and $p(f|x)$ in verification studies suggests that these fundamental and important distributions are frequently ignored or underutilized. Moreover, even when they are considered, full and consistent use of the information contained in these distributions is seldom achieved. We believe that careful and reasoned evaluation of $p(x|f)$ and $p(f|x)$—and of the marginal and joint distributions—could greatly enhance the insights provided by the verification process.

Although some use is made of the basic distributions—and elements thereof—in current verification studies, such studies traditionally rely primarily on so-called verification measures. These mathematical functions generally represent measures of particular attributes of the forecasts and/or observations. Since forecast verification is concerned to a considerable extent with the *correspondence* between forecasts and observations on either an individual or a collective basis, many verification measures are simple functions of the differences between (or ratios of) these quantities. In this context, then, it seems appropriate to ask the following question: What, if any, are the relationships between such measures and the general framework for verification described in this paper? As an illustration of some issues involved in attempting to answer this question, we consider a familiar measure of the accuracy of forecasts—the mean square error (MSE)—in the context of the situations represented by Examples A, B and C.

In terms of the notation employed in this paper, MSE can be expressed as follows:

$$\mathrm{MSE} = E[(f-x)^2] = \sum_f \sum_x (f-x)^2 p(f,x). \quad (7)$$

As its name implies, MSE is the average square difference (or distance) between the individual pairs of forecasts and observations. This measure is frequently used to evaluate categorical forecasts of temperature (Example C). Thus, it is of some interest to note that MSE, as defined in (7), is also identical to the Brier score (Brier, 1950), the most commonly used measure of the accuracy of precipitation probability forecasts (Example B). Moreover, in the case of categorical forecasts of precipitation occurrence (Example A), MSE $= p(1, 0) + p(0, 1)$, which implies that this measure is one minus the fraction correct.

It is also of interest to note that MSE itself can be decomposed into measures of other attributes of the forecasts and observations. For example, it is well known that MSE can be written as follows:

$$\mathrm{MSE} = \mathrm{Var}(f-x) + [E(f) - E(x)]^2. \quad (8)$$

The first term on the right-hand side (rhs) of (8) is the variance of the forecast errors, whereas the second term on the rhs of (8) is the square of the difference between the average forecast and average observation (this latter term is a measure of bias).

Moreover, since

$$\mathrm{Var}(f-x) = \mathrm{Var}(f) + \mathrm{Var}(x) - 2\,\mathrm{Cov}(f,x), \quad (9)$$

MSE in (8) can also be written as

$$\mathrm{MSE} = \mathrm{Var}(f) + \mathrm{Var}(x) - 2\,\mathrm{Cov}(f,x) + [E(f) - E(x)]^2. \quad (10)$$

Thus, MSE can be expressed in terms of means and variances of the marginal distributions, $p(f)$ and $p(x)$, and the covariance of the joint distribution, $p(f, x)$. This decomposition of MSE can be applied in any situation, including the situations represented by Ex-

amples A, B, and C. [The decomposition of MSE in (10) is similar to a recent decomposition of the Brier score described by Yates (1982).]

When x is a binary variable, as in the cases of Examples A and B, it is possible to obtain other useful decompositions of MSE by conditioning either on the forecasts or on the observations. In the case of the former, MSE can be expressed as follows (Murphy, 1973):

$$\text{MSE} = \text{Var}(x) + E_f[f - E(x|f)]^2 - E_f[E(x|f) - E(x)]^2, \quad (11)$$

in which $E_f(\cdot)$ represents the expectation with respect to the marginal distribution $p(f)$. The terms in this decomposition involve simple summary measures of the marginal distributions [$p(f)$ and $p(x)$] and of the conditional distributions of the observations given the forecasts [$p(x|f)$]. Thus, this decomposition is related to the calibration-refinement factorization of the joint distribution. Specifically, the second term on the rhs of (11) is obviously a measure of the reliability (calibration) of the forecasts. On the other hand, the third term on the rhs of (11) relates to the differences between the conditional expectations $E(x|f)$ and the marginal (unconditional) expectation $E(x)$ and, in view of the sign of this term, larger differences are preferred to smaller differences. This attribute of the forecasts has been referred to as "resolution" (Murphy and Daan, 1985).

In the case of conditioning on the observations, MSE can be decomposed as follows:

$$\text{MSE} = \text{Var}(f) + E_x[x - E(f|x)]^2 - E_x[E(f|x) - E(f)]^2, \quad (12)$$

in which $E_x(\cdot)$ represents the expectation with respect to the marginal distribution $p(x)$. (We are indebted to E. S. Epstein for identifying this particular decomposition of MSE. See also Murphy, 1986.) The decomposition in (12) is related to the likelihood-base rate factorization of the joint distribution, and the terms in this decomposition involve summary measures of the conditional distribution $p(f|x)$ and the marginal distributions $p(f)$ and $p(x)$. The second term on the rhs of (12) is the weighted squared difference between the observation (zero or one) and the average forecast associated with that observation (or event), where the weights are the base rates of the events. This term can be viewed as the weighted average of the errors in the (conditional) *average* forecasts. The third term on the rhs of (12) is the weighted squared difference between the average forecast associated with each observation and the overall (unconditional) average forecast. In view of the minus sign, larger differences are preferred here to smaller differences. This term measures the extent to which the average forecasts associated with the events differ from one another.

In this section, we have briefly examined some common verification methods and measures from the perspectives of the general framework. This discussion has provided some useful insights into the fundamental role that the basic distributions constituting this framework play in current verification methodology. Some other relationships between scoring rules (such as the Brier score) and the distribution of forecasts and observations are discussed in Winkler (1986). In the next section, we discuss the need for further investigation of the relationships between this framework (and its components) and verification methods.

5. Discussion and conclusion

In this paper we have described a general framework for forecast verification. The framework is based on the joint distribution of forecasts and observations which contains information about the forecasts, about the observations, and about the relationship between the forecasts and observations. The information embodied in the joint distribution is more accessible when this distribution is factored into conditional and marginal distributions. Two such factorizations exist, each of which relates to particular aspects of verification. The so-called calibration-refinement factorization involves the conditional distributions of the observations given the forecasts and the marginal distribution of the forecasts, and it contains information regarding the reliability and sharpness of the forecasts. On the other hand, the so-called likelihood-base rate factorization involves the conditional distributions of the forecasts given the observations and the marginal distribution of the observations, and it provides information regarding the ability of the forecasts to discriminate among the observations and regarding the forecasting situation itself (in terms of the relative frequencies of the observations). A preliminary investigation of some common verification measures in current use indicates that some of these measures are either elements of the basic distributions or are related to summary measures (e.g., mean, variance) of these distributions.

At this point, two questions suggest themselves. What are the implications of the general framework for verification theory and practice? What directions might future studies of verification take in light of this framework? With regard to the former, we believe that recognition of the fact that the joint distribution contains all of the information relevant to verification suggests that this distribution—or, equivalently, the factors involved in the factorizations of the joint distribution—should play a central role in verification studies. In recent years, attention has tended to focus on verification measures rather than on more basic indicators of performance such as the joint distribution (and its factorizations). Common verification measures are quite useful in situations in which the primary objective is simply to compare forecasting procedures or fore-

casters in some overall sense, but they are not particularly helpful when it comes to obtaining a more detailed understanding of the strengths and weaknesses in forecasts or to identifying ways in which the forecasts might be improved. A "diagnostic" approach to verification based at least in part on the general framework described here might provide the basis for a more effective verification system from the viewpoints of both the producers and users of the forecasts.

In section 4, we reported the results of a preliminary investigation of relationships between the general framework and some common verification measures. This investigation revealed that many such measures are directly related either to the basic distributions themselves or to summary measures of these distributions. We believe that more comprehensive and detailed studies of this type are warranted, since they may clarify the relationships between various verification methods and measures and may provide additional insight into the interpretation of verification data. In a related vein, studies of the empirical joint, conditional, and marginal distributions for various types of forecasts, including efforts to model these distributions, should also prove quite valuable. Recent examples of related modeling studies include Clemen and Winkler (1987), Krzysztofowicz (1986), and Mason (1982). The existence of reasonably faithful models of such distributions would have important implications for verification methods and practices.

As noted in section 1, the field of verification is currently characterized by a wide variety of apparently unrelated methods and measures. Perhaps the most immediate benefit of the general framework described in this paper is that it provides a new and unified way of looking at the verification problem. The perspective provided by this framework should prove to be quite useful, both to those concerned with verification methodology and those concerned with verification practice. Additional theoretical and practical work will be required before it will be possible to render a more precise assessment of the framework's ultimate utility.

Acknowledgments. Robert T. Clemen, Edward S. Epstein, and Ian B. Mason provided valuable comments on an earlier version of this paper. This research was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8507495.

REFERENCES

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

Clemen, R. T., and R. L. Winkler, 1987: Calibrating and combining precipitation probability forecasts, *Probability and Bayesian Statistics,* R. Viertl, Ed., Plenum (in press).

Daan, H., 1984: Scoring rules in forecast verification. World Meteorological Organization, Geneva, 60 pp.

DeGroot, M. H., 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 ——, 1983: The comparison and evaluation of forecasters. *The Statistician,* **32,** 14–22.

Krzysztofowicz, R., 1986: Stochastic model of seasonal runoff forecasts. *Water Resour. Res.*, **22,** 196–202.

Lichtenstein, S., B. Fischhoff and L. D. Phillips, 1982: Calibration of probabilities: the state of the art to 1980, *Judgment under Uncertainty: Heuristics and Biases,* D. Kahneman, P. Slovic, and A. Tversky, Eds., Cambridge University Press, 306–334.

Lindley, D. V., 1982: The improvement of probability judgments. *J. Roy. Stat. Soc.,* **A145,** 117–126.

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

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

——, 1986: A new decomposition of the Brier score: formulation and interpretation. *Mon. Wea. Rev.,* **114,** 2671–2673.

——, 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 R. L. Winkler, 1977: Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Stat.,* **26,** 41–47.

Winkler, R. L., 1972: *Introduction to Bayesian Inference and Decision,* Holt, Rinehart and Winston, 563 pp.

——, 1986: On "good probability appraisers." *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti,* P. Goel and A. Zellner, Eds., Elsevier, 265–278.

Yates, J. F., 1982: External correspondence: decompositions of the mean probability score. *Organizational Behavior and Human Performance,* **30,** 132–156.