



ELSEVIER

International Journal of Forecasting 11 (1995) 133–146

*international journal
of forecasting*

Screening probability forecasts: contrasts between choosing and combining

Robert T. Clemen^{a,*}, Allan H. Murphy^b, Robert L. Winkler^c

^aCollege of Business Administration, University of Oregon, Eugene, OR 97403-1208, USA

^bPrediction and Evaluation Systems, Corvallis, OR 97330-1139, USA

^cFuqua School of Business, Duke University, Durham, NC 27708-0120, USA

Abstract

In many forecasting situations, forecasts can be produced by several different methods. The ultimate objective of considering multiple methods may be to select a single method (the choosing scenario) or to aggregate the multiple forecasts into a single forecast (the combining scenario). Procedures for screening candidate forecasts—*sufficiency* in the choosing scenario and *extraneousness* in the combining scenario—are described here. Screening can identify forecasting methods that are dominated in the sense that their forecasts are clearly inferior to those of other methods or do not add any information to the combination of forecasts. These evaluation procedures are illustrated and contrasted by considering prototypical examples and an application involving precipitation probability forecasts. The value of screening is that it can reduce the set of candidate forecasting methods to a manageable number, which can then be evaluated in greater detail.

Keywords: Probability forecasts; Combining; Sufficiency; Screening; Extraneousness

1. Introduction

In many forecasting situations, it is possible to produce forecasts of interest by two or more different methods. The set of available methods may include various extrapolation methods, other statistical forecasting procedures, substantive models involving the questions of interest (e.g. economic models for economic forecasts or meteorological models for weather forecasts), subjective forecasts from different experts, or

some mix of these approaches. In this day and age of fast computers and rapid communication, the set of feasible alternatives to consider when deciding how to produce a forecast is often quite large.

Traditionally, the decision about how to produce a forecast has been viewed as a problem of identifying a single method. This decision involves some sort of evaluation of each individual method under consideration and a subsequent comparison of their performance, followed by the choice of a single method on the basis of this evaluation and comparison. We call

* Corresponding author.

this problem the *choosing scenario*. Some might frame this problem as the choice of a “best” forecasting method, or the choice of a method that produces the “best” forecasts. Our caveat with respect to this framing of the problem is that “best” is in the eye of the beholder and means best as viewed by the decision maker for the particular situation at hand.

When two or more forecasting methods are available it is also possible to combine the multiple forecasts into a single combined forecast. Methodological and practical issues related to combining have been studied extensively in recent years (Clemen, 1989), and combined forecasts have generally performed quite well in practice. When some sort of aggregation of forecasts is considered, the problem is typically viewed as a decision regarding which forecasts to combine and how to combine them. We call this the *combining scenario*. Of course, any combination of forecasts yields a single forecast. As a result, a particular combination of a given set of forecasts can itself be thought of as a forecasting method which could compete in the choosing scenario. In that sense, a broad view of the overall forecasting situation would simply include the combining scenario within the more general choosing scenario. Nonetheless, choosing and combining are often thought of separately, and it is convenient for the purposes of this paper to maintain this separation and to contrast the two.

Whether the basic problem is viewed as choosing or combining, it necessarily involves some sort of forecast evaluation methodology. In practice, an evaluation often involves one or two summary measures. We focus in this paper on probability forecasts, for which scoring rules (Winkler, 1967, 1986) can be used as summary measures of quality. Forecast quality, however, is multidimensional in nature, and one or two summary measures generally cannot completely describe the quality of a forecasting method or a set of forecasts. A more complete approach involves descriptions of relationships among forecasts and observations based on their joint distribution as well as multiple measures that characterize different aspects of forecasting performance (Murphy and Winkler, 1992).

In this paper we discuss procedures for screening probability forecasts. Screening can identify forecasting methods that are dominated in the sense that their forecasts are clearly inferior to those of other methods being considered in the choosing scenario or do not add any information to the combination in the combining scenario. These determinations can be made on the basis of the evidence available to the decision maker (e.g. sets of past forecasts and observations). Two basic screening methods are (a) sufficiency (DeGroot and Fienberg, 1982, 1983, 1986; DeGroot and Eriksson, 1985) and (b) extraneousness (Clemen, 1985; Clemen and Guerard, 1989). Sufficiency identifies conditions under which one set of forecasts can be unambiguously judged to be better than another set of forecasts, whereas extraneousness indicates conditions under which one set of forecasts contains all of the information (regarding the events of interest) possessed by another set of forecasts. Although sufficiency and extraneousness are similar, they are not equivalent; some distinctions between the two concepts will be illustrated and clarified in Sections 3–5.

The roles of sufficiency and extraneousness in screening forecasts are analogous to the role of stochastic dominance in screening options in decision making under uncertainty. The screening process typically will not completely solve the problem at hand; that is, it generally will not screen out all but one candidate method in the choosing scenario or reduce the set of methods to be combined as much as might be desirable in the combining scenario. The value of screening is that it can reduce the set of candidate methods to a manageable number, which can then be evaluated in more detail.

The objectives of this paper are to investigate the use of sufficiency and extraneousness as screening methods and to contrast the choosing and combining scenarios in terms of screening. Forecast evaluation in the two scenarios is briefly discussed in Section 2, and the screening methods are defined and interpreted in Section 3. Prototypical examples involving probability forecasts are considered in Section 4 to illustrate some differences between screening in the choosing scenario and screening in the combining

scenario. A real-world example involving operational probabilistic weather forecasts is considered in Section 5, with sufficiency and extra-neousness examined and compared. Section 6 contains a discussion of the implications of these results and some concluding remarks.

2. Forecast evaluation in the choosing and combining scenarios

2.1. Choosing scenario

The choosing scenario has the objective of selecting a single forecasting method. The role of forecast evaluation, therefore, is to evaluate the methods individually and then to compare their performance. For each method, the basis of the evaluation is the relationship between the probability forecast and the subsequent observation, as described by the joint distribution of forecasts and observations (Murphy and Winkler, 1992). Moreover, conditional and marginal distributions associated with factorizations of this joint distribution characterize specific aspects of forecast quality. The existence of these different aspects demonstrates the multidimensional nature of forecast quality. A diagnostic evaluation of a set of probability forecasts from a given forecasting method, along with the corresponding observations, may reveal strong performance on one dimension and weak performance on another dimension. For instance, the probabilities could be poorly calibrated yet very discriminating between occasions on which the event of interest did and did not occur.

When the relative quality of two or more forecasting methods is of concern, comparison of the values of one or two overall performance measures is inadequate. It is necessary to consider the basic characteristics of forecast quality embodied in the respective joint distributions of forecasts and observations. To evaluate n forecasting methods in the choosing scenario, then, it is necessary to look at n bivariate distributions of forecasts and observations. Ultimately, choosing among methods may involve tradeoffs among these characteristics. For instance, a decision maker may have to decide whether giving up

some discriminatory ability for better calibration is a worthwhile tradeoff.

Before the decision maker reaches the stage in which tradeoffs are considered, screening may help to reduce the set of methods being considered. Sufficiency provides a formal means of screening alternative forecasting methods that is consistent with the multidimensional nature of forecast quality. Moreover, it maintains a monotonic relationship between forecast quality and forecast value. That is, when method A is sufficient for method B, it follows that A's forecasts are of higher quality and greater value to all users than B's forecasts. In this sense, choosing method A over method B is not a matter of tradeoffs; method A simply dominates method B, so the choice is obvious.

2.2. Combining scenario

The combining scenario has the objective of aggregating forecasts from two or more methods. As a result, it is not adequate to evaluate the methods individually. The relationships among the methods are important in combining, and separate evaluations will not capture these relationships. For example, two methods may have identical characteristics when evaluated individually; for the purposes of combining, it will make a big difference if their probabilities always tend to be very similar or tend to be quite different. In the former case, the methods may be highly redundant, whereas in the latter case the degree of redundancy might be much less.

As described above, it is necessary in the choosing scenario to examine n joint (two-dimensional) distributions of forecasts and observations in order to evaluate and compare n forecasting methods. These n distributions, however, provide no information about the independent information content of the alternative forecasting methods. In the combining scenario with n forecasting methods, we must consider the $(n + 1)$ -dimensional distribution of the n forecasts and the corresponding observations. Thus, the fundamental difference between forecast evaluation in the choosing and combining scenarios is that the choosing scenario involves n bivariate distributions while the combining

scenario involves one multivariate, $(n + 1)$ -dimensional distribution.

Screening in the combining scenario can be used to reduce the number of candidate methods being considered for inclusion in the combination. Extraneousness provides a formal means of screening forecasting methods to eliminate methods that provide no additional information to the combined forecast beyond the information given in forecasts from the remaining methods. That is, when considering methods A, B, and C, if method A is extraneous, then a combination of forecasts from A, B, and C is no more informative than a combination of forecasts from just B and C.

3. Screening methods

3.1. Sufficiency

In the choosing scenario, the forecasting methods are evaluated individually, and sufficiency can be investigated by comparing pairs of methods from the n available methods. The discussion of sufficiency is therefore presented in terms of two methods. Consider two forecasting methods (e.g. models and/or forecasters) 1 and 2 and suppose that these methods produce probabilistic forecasts f_1 and f_2 for an event, where $f_1, f_2 \in F$, a set of permissible probabilities. Thus $f_1 = P_1(x = 1)$ and $f_2 = P_2(x = 1)$, where $x = 1$ if the event occurs and $x = 0$ otherwise. We will assume that these forecasts are well-calibrated in the sense that $P(x = 1 | f_1) = f_1$ and $P(x = 1 | f_2) = f_2$. Further, let $v_1(f) = P(f_1 = f)$ and $v_2(g) = P(f_2 = g)$, where $f, g \in F$, with $0 \leq f, g \leq 1$. The functions v_1 and v_2 characterize the refinement of 1's and 2's forecasts (the degree to which the forecasts are near the extreme probabilities of zero or one). Given these definitions, method 1 is *sufficient* for method 2 if a stochastic transformation $h(g | f)$ exists such that

$$\sum_f h(g | f)v_1(f) = v_2(g) \quad \text{for all } g \in F \quad (1)$$

and

$$\sum_f h(g | f)fv_1(f) = gv_2(g) \quad \text{for all } g \in F \quad (2)$$

(DeGroot and Fienberg, 1986). The function $h(g | f)$ satisfies the conditions for a stochastic transformation if $0 \leq h(g | f) \leq 1$ and $\sum_g h(g | f) = 1$ for all $f \in F$. Eq. (1) embodies the main result; Eq. (2) ensures that if 1's forecasts are well-calibrated, then the transformed forecasts will also be well-calibrated.

An interpretation of Eq. (1) can be obtained by observing that the stochastic transformation represents an auxiliary randomization that introduces noise into 1's forecasts. The new forecasts provided by this process are distributionally equivalent to 2's forecasts. Thus, 2's forecasts can be interpreted as containing greater uncertainty than 1's forecasts.

The significance of the sufficiency relation in the choosing scenario resides in its ability to order some pairs of forecasting methods in terms of their relative quality and value. If method 1 can be shown to be sufficient for method 2, then 1's forecasts are necessarily of higher quality in all respects—and of greater value to all users—than 2's forecasts. However, 1's forecasts may be sufficient for 2's forecasts, 2's forecasts may be sufficient for 1's forecasts, or 1's and 2's forecasts may be insufficient for each other. If neither method is sufficient for the other, it may be that 1's forecasts are more valuable than 2's forecasts for some users whereas 2's forecasts are more valuable than 1's forecasts for other users.

From the point of view of the implementation of the sufficiency relation, several different approaches are available. First, a straightforward approach can be taken involving the search for auxiliary randomizations that satisfy the conditions for a stochastic transformation. This approach has been taken by Ehrendorfer and Murphy (1988, 1992) in their studies of primitive weather and climate forecasts.

Second, Krzysztofowicz and Long (1991) have formulated an approach involving the so-called forecast sufficiency characteristic (FSC), based on a theorem presented by Blackwell and Girshick (1954). This approach requires the construction and comparison of the respective FSCs, which are derived from the likelihoods (the

conditional probabilities of the forecasts given the events). Krzysztofowicz and Long (1991) and Murphy and Ye (1990) have used this approach in comparative analyses of different types of precipitation probability forecasts.

Third, Krzysztofowicz (1992) has derived a measure of skill for non-probabilistic forecasts of continuous predictands—the so-called Bayesian correlation score (BCS)—based on the theory of sufficient comparisons of experiments (Blackwell, 1951, 1953). The BCS is specified in terms of the parameters of a normal linear model, which combines information regarding the climatology of the predictand with characteristics of forecast quality. Under these modeling assumptions, the BCS can be used to compare alternative forecasting methods involving the same predictand or (in a limited sense) different predictands, and it orders the methods in terms of the quality and value of their forecasts (in a manner similar to the sufficiency relation).

Fourth, DeGroot and Eriksson (1985) have demonstrated the equivalence between the sufficiency relation and second-order stochastic dominance. In this context, second-order stochastic dominance refers to the relationship between (cumulative) distribution functions of the forecasts produced by the respective forecasting methods. A comparative evaluation of alternative forecasting systems employing the concept of stochastic dominance is illustrated in Section 5.

3.2. Extraneousness

In a combining scenario involving n forecasting methods, the $(n + 1)$ -dimensional distribution $P(f_1, f_2, \dots, f_n, x)$ represents the appropriate framework within which to evaluate the methods. This distribution contains information about the forecasts produced (individually) by the n methods as well as information about the relationships among the n forecasts and the observations x . This distribution can be factored into univariate conditional distributions and a marginal distribution as follows:

$$P(f_1, \dots, f_n, x) = P(x | f_1, \dots, f_n) P(f_1, \dots, f_n) \quad (3)$$

where $P(x | f_1, \dots, f_n)$ represents conditional distributions of the observations given the n forecasts and $P(f_1, \dots, f_n)$ represents the joint distribution of the n forecasts. Since the conditional distributions $P(x | f_1, \dots, f_n)$ describe the relationship between the forecasts and the observations, it seems natural to focus attention on these distributions in this context.

Recall that we are concerned here with the question of whether or not forecasting methods (e.g. models and/or forecasters) provide independent information regarding future occurrences of the events of interest. If a method does not provide any such information, then it is referred to as *extraneous* (Clemen, 1985). With regard to the n methods, method i is extraneous if

$$P(x | f_1, \dots, f_n) = P(x | f_1, \dots, f_{i-1}, f_{i+1}, \dots, f_n) \quad (4)$$

In Eq. (4), x is conditionally independent of f_i given $f_1, \dots, f_{i-1}, f_{i+1}, \dots, f_n$. Thus, given the other $n - 1$ forecasts, forecast f_i contains no independent information regarding future values of x . Combinations of forecasts from all n methods will be of higher quality than combinations of forecasts from methods $1, \dots, i - 1, i + 1, \dots, n$ only if method i is not extraneous.

Extraneousness is always defined relevant to a given set of n methods, as indicated by the phrase “with regard to the n methods” that qualifies the condition given in (4) for the extraneousness of method i . The fact that deleting methods from the set could cause method i to no longer be extraneous is not very surprising. What is somewhat counterintuitive at first glance is that adding methods to the set might also mean that method i is no longer extraneous (e.g. see Clemen, 1987). When methods are added, method i might help us sort out relationships between the other original $n-1$ methods and the new methods. As a result, extraneousness should be re-examined whenever there is a change in the set of methods being considered. Sufficiency,

on the other hand, is a pairwise relationship; if method j is sufficient for method i , then the presence or absence of other methods in the set under consideration is irrelevant.

The concept of extraneousness has been used to study relationships between subjective (local) and "objective" (guidance) weather forecasts (Clemen and Murphy, 1986a; Murphy et al., 1988). Specifically, these studies have investigated whether either forecast is extraneous, in the sense that it provides no independent information regarding the weather events of interest. In the case of precipitation probability forecasts, Clemen and Murphy (1986a) have found that neither forecast is extraneous; a decision maker must in general consult both forecasts in order to obtain the maximum amount of information, although the independent information content in the local forecasts appears to exceed that in the guidance forecasts. This result suggests that a forecast produced by combining the two types of forecasts might be better than either forecast alone. Clemen and Murphy (1986b), in a companion paper, have investigated this possibility and found that modest improvements in forecasting performance could be achieved by averaging the two forecasts.

4. Some prototypical examples

In this section we present simple prototypical examples to illustrate some differences, and connections, between sufficiency and extraneousness as screening methods in the choosing and combining scenarios. For simplicity and ease of exposition the examples consider the case in which only two forecasting methods are available.

4.1. Example 1

Suppose that a decision maker wants a probability forecast for the occurrence of precipitation. Consider two forecasting methods A and B, each of which is well-calibrated and gives only precipitation probabilities of 0.1 and 0.9. Sup-

pose that their refinement functions are $v_A(0.1) = v_A(0.9) = v_B(0.1) = v_B(0.9) = 0.5$. Thus, A and B are exchangeable. Moreover, A and B are sufficient for each other. The stochastic transformation is $h(g|f) = 1$ if $g = f$ and $h(g|f) = 0$ otherwise. A decision maker faced with choosing between these two methods would be indifferent.

However, suppose that the forecasts from these methods possess the joint distribution $p(f_A, f_B)$ described in Fig. 1(a). That is, $p(f_A, f_B)$ is such that methods A and B make the same forecast 80% of the time (40% of the time they both say 0.1 and 40% of the time they both say 0.9). Their forecasts differ only 20% of the time (10% of the time A says 0.1 and B says 0.9, and 10% of the time A says 0.9 and B says 0.1).

Moreover, suppose that the conditional distribution of the event $x = 1$ (i.e. the occurrence of precipitation) given the two forecasts, $p(x =$

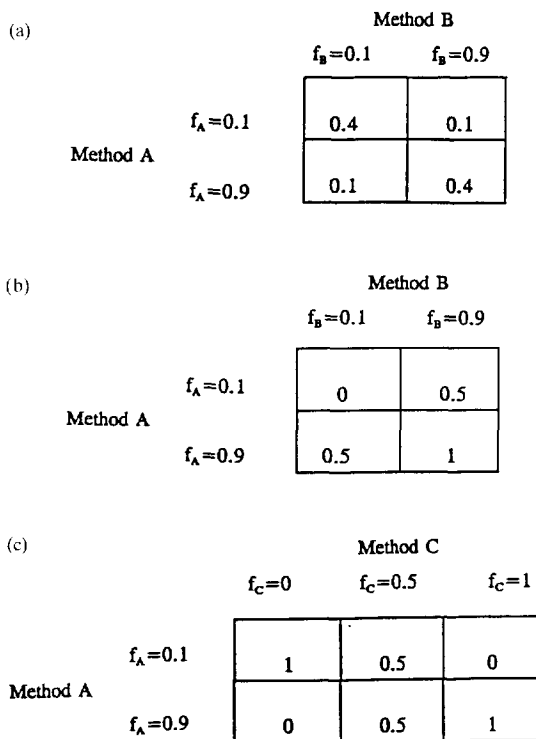


Fig. 1. For Example 1: (a) joint distribution $P(f_A, f_B)$; (b) conditional distribution $P(x = 1 | f_A, f_B)$; (c) stochastic transformation $h(f_A | f_C)$.

$1 | f_A, f_B$), is as described in Fig. 1(b). This conditional distribution indicates that if both A and B say 0.9, then the decision maker can be sure that precipitation will occur. Analogously, if both methods say 0.1, the decision maker can be sure that precipitation will not occur. When A and B disagree, the decision maker's posterior probability of precipitation is 0.5.

If A's and B's forecasts are combined using this information, the combined forecasts (denoted by C) would be well-calibrated and would possess the following refinement function: $v_C(0) = v_C(1) = 0.40$ and $v_C(0.5) = 0.20$. That is, the combined forecast would indicate that the probability of precipitation is 0 for 40% of the time, 1 for 40% of the time, and 0.50 for 20% of the time.

It is quite easy to show that the combined forecast C is sufficient for either A or B. The stochastic transformation is shown in Fig. 1(c), and it can be interpreted as follows: If $f_C = 0$, set $f_A = 0.1$, and if $f_C = 1$, set $f_A = 0.9$. If $f_C = 0.5$, toss a fair coin. If heads occurs, set $f_A = 0.1$, and if tails occurs, set $f_A = 0.9$. Identical expressions and instructions can be written for the stochastic transformation relating f_B and f_C . However, neither B nor A is sufficient for C, indicating that neither A nor B is extraneous.

This example illustrates a situation in which the trivariate distribution contains information that is useful in combining forecasts but does not help in choosing between the two forecasters. Moreover, even though methods A and B are exchangeable in the sense that it doesn't matter to the decision maker which method is used, neither method is extraneous.

4.2. Example 2

Consider the same methods as in Example 1. However, now suppose that they possess the joint distribution $p(f_A, f_B)$ described in Fig. 2. In order for the combined forecast (C) to remain well-calibrated, if both methods say 0.1, C must also say 0.1. Analogously, C must say 0.9 when both A and B say 0.9. As expected, these two methods are not only exchangeable, they are fully equivalent since they just parrot each other.

		Method B	
		$f_B=0.1$	$f_B=0.9$
Method A	$f_A=0.1$	0.5	0
	$f_A=0.9$	0	0.5

Fig. 2. Joint distribution $P(f_A, f_B)$ for Example 2.

In fact, it is easy to see that A (or B) is sufficient for C, indicating that B (or A) is extraneous.

In contrast to Example 1, this example demonstrates extraneous methods. Once the decision maker has a forecast from either A or B, then the other is extraneous. Moreover, it is easy to see that the combined forecast C must be equivalent to both A and B. The differences between Examples 1 and 2 are particularly intriguing because the two joint distributions are quite similar (cf. Figs.1(a) and 2).

4.3. Example 3

Now consider a situation in which the two methods are not exchangeable. In this example method A gives probabilities of 0.4 and 0.6, whereas method B gives probabilities of 0.1 and 0.9 (the same as in the two previous examples). Once again, it is assumed that both methods are well-calibrated.

Suppose that their respective refinement functions are as follows: $v_A(0.4) = v_A(0.6) = v_B(0.1) = v_B(0.9) = 0.5$. It is relatively easy to show that method B is sufficient for method A. The stochastic transformation is as follows: $h(f_A = 0.4 | f_B = 0.1) = 5/8$, $h(f_A = 0.6 | f_B = 0.1) = 3/8$, $h(f_A = 0.4 | f_B = 0.9) = 3/8$, and $h(f_A = 0.6 | f_B = 0.9) = 5/8$. That is, when B says 0.1, then with probability 5/8 the transformed forecast (A) is 0.4 and with probability 3/8 the transformed forecast (A) is 0.6. (A fair eight-sided die, with five red sides and three green sides, could be used to determine A's forecast). The probabilities 5/8 and 3/8 are reversed when B says 0.9.

Now suppose that the forecasts possess the joint distribution $p(f_A, f_B)$ described in Fig. 3(a). These methods are independent, at least in terms

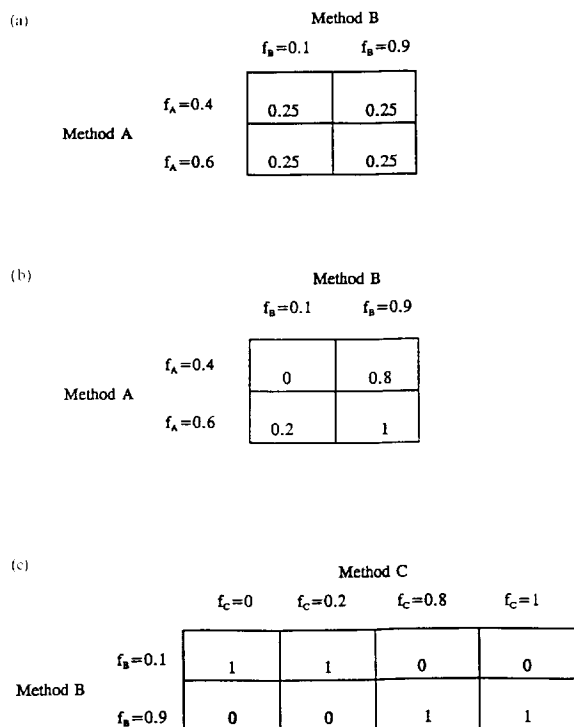


Fig. 3. For Example 3: (a) joint distribution $P(f_A, f_B)$; (b) conditional distribution $P(x=1|f_A, f_B)$; (c) stochastic transformation $h(f_B|f_C)$.

of this joint distribution. That is, $P(f_A=0.4, f_B=0.1) = P(f_A=0.4)P(f_B=0.1)$, etc.

Further, suppose that the conditional probability of precipitation given the two forecasts, $P(x=1|f_A, f_B)$, is as described in Fig. 3(b). Thus, if the methods “agree” in the sense that $f_A=0.4$ and $f_B=0.1$ or $f_A=0.6$ and $f_B=0.9$, the decision maker is sure about the outcome (as in Example 1). However, if the methods “disagree” ($f_A=0.4$ and $f_B=0.9$ or $f_A=0.6$ and $f_B=0.1$), the decision maker adopts a slightly modified version of B’s forecast.

As a result, the combined forecast (C), which is well-calibrated, has the following refinement function: $v_C(0) = v_C(0.2) = v_C(0.8) = v_C(1) = 0.25$. It is easy to show that C is sufficient for B (and hence is sufficient for A). The stochastic transformation is described in Fig. 3(c). That is, if C gives 0 or 0.2, then set the transformed forecast (B) to 0.1. Otherwise (i.e. $C = 0.8$ or 1),

set the transformed forecast (B) to 0.9. However, it can be shown that B is not sufficient for C, indicating that A is not extraneous.

This example is interesting because method A is dominated by method B in the choosing scenario. However, from the perspective of the combining scenario, A’s forecasts contain independent information (i.e. information not contained in B’s forecasts). Hence, the combined forecast C is sufficient for (and thus better than) B. In summary, B is sufficient for A but A is not extraneous.

5. Precipitation probability forecasts: an application

An area with extensive probability-forecast data is weather forecasting. Since 1965 the National Weather Service (NWS) of the United States has formulated and issued probability of precipitation (PoP) forecasts on a nationwide basis. These forecasts indicate a probability of measurable precipitation at a particular location during a specified time period. Although precipitation can take on different forms (rain, snow, etc.), for convenience we will refer to the occurrence and nonoccurrence of precipitation as “rain” ($x=1$) and “no rain” ($x=0$).

For any given forecast area and forecast period, two PoP forecasts are prepared. The forecast actually issued to the public is a forecast made by a weather forecaster in the local NWS office. In addition, the NWS prepares PoP forecasts based on a numerical-statistical forecasting model. We call these forecasts guidance forecasts because they are supplied to the forecasters for use in preparing the official local forecasts. Meteorologists have studied the relative performance of local and guidance forecasts; a review of this literature and a more complete review of the forecasting process are given in Murphy and Winkler (1984) and Murphy (1985).

The data analyzed in this paper consist of local and guidance PoP forecasts for the NWS office in Boston from April 1972 through September 1983. Local and guidance forecasts are made twice daily, in the morning and evening. On each

occasion, forecasts are formulated for three consecutive 12-h periods, or lead times: approximately 12-24 h, 24-36 h, and 36-48 h after the guidance forecast is made. These data have been analyzed previously in Clemen and Winkler (1990), where additional details regarding the data set can be found.

In evaluating PoP forecasts, one concern is whether the climatological probability of rain (overall relative frequency) is stable throughout the year. Meteorologists typically divide the year into cool (October-March) and warm (April-September) seasons to achieve approximate stability. In Boston, the climatological probabilities of rain for the warm and cool seasons are 0.233 and 0.215, respectively. On the basis of these calculations, we conclude that there is not a material difference in the two proportions and pool the data from both warm and cool seasons for our analysis. The pooled climatological probability of rain is 0.224.

For each (f_L, f_G) , where f_L represents the local PoP forecast and f_G represents the guidance

PoP forecast, we look at all occasions with forecast values (f_L, f_G) and find the relative frequency of occurrence of precipitation over those occasions. These relative frequencies, along with the number of occasions on which each is based, are given in Table 1. Because we need to estimate the probability of rain in each cell to obtain a combined forecast of rain given (f_L, f_G) , we ignore cells with less than ten observations. After excluding these cells, the data set contains a total of 12 729 pairs of local and guidance forecasts.

Table 1 fully characterizes both individual forecasts as well as the combined forecast. The right (lower) margin of the table gives information about the local (guidance) forecast. For example, the local forecast is 0.10 about 2766/12,729 = 21.7% of the time, and the relative frequency of precipitation in this case is 0.073. Thus, $v_L(0.10) = 0.217$, and $P(\text{rain} | f_L = 0.10) = 0.073$. The cells in the interior of the table give information about particular pairs of values of f_L and f_G . For example, the pair $(f_L = 0.30, f_G =$

Table 1
Relative frequency of precipitation and sample sizes

Local:	Guidance:														
	0	0.02	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1		
0	0.009	0.014	0.014	0.027	0.061	0.035	0.267								0.018
	1595	580	586	565	163	57	15								3541
0.1	0.028	0.088	0.063	0.063	0.103	0.161	0.143	0.167							0.073
	463	251	474	865	504	155	42	12							2766
0.2	0.027	0.052	0.08	0.125	0.129	0.235	0.268	0.271	0.226	0.571					0.164
	111	77	162	385	611	371	194	107	53	21					2092
0.3	0.135	0.294	0.136	0.169	0.233	0.274	0.301	0.275	0.290						0.244
	37	17	44	130	219	237	123	51	31						889
0.4	0.385		0.238	0.213	0.330	0.351	0.377	0.348	0.339	0.414					0.339
	13		21	61	109	148	167	92	62	29					702
0.5				0.361	0.350	0.355	0.410	0.457	0.395	0.565	0.618				0.417
				36	80	121	139	140	86	46	34				682
0.6				0.348	0.314	0.483	0.563	0.495	0.573	0.645	0.719	0.500			0.521
				23	51	58	103	107	124	62	32	10			570
0.7					0.471	0.613	0.500	0.589	0.658	0.693	0.833	0.700			0.654
					17	31	32	56	73	101	60	20			380
0.8					0.615	0.429	0.595	0.655	0.741	0.816	0.840	0.791	0.800		0.749
					13	21	42	58	81	114	125	67	10		531
0.9								0.500	0.750	0.765	0.933	0.893	0.963		0.849
								14	20	34	45	75	27		215
1									0.857	0.885	0.904	0.944	0.979		0.935
									21	26	73	90	141		351
	0.019	0.039	0.05	0.085	0.164	0.279	0.376	0.435	0.520	0.691	0.813	0.836	0.938		0.224
	2219	905	1287	2065	1767	1199	857	637	551	433	369	262	178		12729

0.40) was used 123 times, and precipitation occurred on 30.1% of these occasions. Thus, $v(f_L = 0.30, f_G = 0.40) = 23/12,729 = 0.0097$, and $P(\text{rain} | f_L = 0.30, f_G = 0.40) = 0.301$.

With the information given in Table 1, it is possible to investigate the local and guidance forecasts for both sufficiency and extraneousness. As indicated in Section 3, there are a number of operational ways to determine whether one forecast is sufficient for another. In this case, the most straightforward approach is to use the equivalence between second-order stochastic dominance and sufficiency (DeGroot and Eriksson, 1985). The analysis is done on the basis of the calibrated forecasts, $P(\text{rain} | f_i)$, which we refer to as the calibration function. The strategy is to assemble a complete list of possible values from the calibration functions of both forecasts under consideration. This list is sorted into ascending order and associated with

the relative frequencies (v_i functions) of the forecasts, including zeros where necessary when the two calibration functions do not overlap. Table 2 demonstrates this procedure for the local and guidance forecasts.

The condition for sufficiency in comparing two discrete, well-calibrated forecasters is given by theorem 2 of DeGroot and Eriksson (1985): Forecaster A is sufficient for Forecaster B if and only if

$$\int_0^s V_A(t) dt - \int_0^s V_B(t) dt \geq 0 \text{ for all } s, \quad 0 \leq s \leq 1, \tag{5}$$

where $V(t)$ equals the cumulative distribution function $\sum_{f \leq t} v(f)$. As with stochastic dominance, the relationship is called strict sufficiency if the inequality is strict for some value of s . The

Table 2
Sufficiency calculations for f_L and f_G

Calibrated Forecast	$v(f_L)$ Local	$v(f_G)$ Guidance	$V(f_L)$ Local	$V(f_G)$ Guidance	$\int V(f_L)$ Local	$\int V(f_G)$ Guidance	Difference
0.000			0.000	0.000	0.000	0.000	0.000
0.018	0.278		0.278	0.000	0.000	0.000	0.000
0.019		0.174	0.278	0.174	0.000	0.000	0.000
0.039		0.071	0.278	0.245	0.006	0.003	0.002
0.050		0.101	0.278	0.347	0.009	0.006	0.003
0.073	0.217		0.495	0.347	0.015	0.014	0.001
0.085		0.162	0.495	0.509	0.021	0.018	0.003
0.164	0.164	0.139	0.660	0.648	0.060	0.059	0.002
0.164			0.660	0.648	0.060	0.059	0.002
0.244	0.070		0.730	0.648	0.113	0.110	0.003
0.279		0.094	0.730	0.742	0.139	0.133	0.006
0.339	0.055		0.785	0.742	0.182	0.177	0.005
0.376		0.067	0.785	0.809	0.212	0.205	0.007
0.417	0.054		0.838	0.809	0.244	0.238	0.006
0.435		0.050	0.838	0.859	0.259	0.253	0.006
0.520		0.043	0.838	0.902	0.330	0.326	0.004
0.521	0.045		0.883	0.902	0.331	0.327	0.004
0.654	0.031		0.914	0.902	0.448	0.447	0.002
0.691		0.034	0.914	0.936	0.482	0.480	0.002
0.749	0.042		0.956	0.936	0.535	0.534	0.001
0.813		0.029	0.956	0.965	0.596	0.594	0.002
0.836		0.021	0.956	0.986	0.618	0.616	0.002
0.848	0.017		0.972	0.986	0.630	0.628	0.001
0.935	0.028		1.000	0.986	0.714	0.714	0.000
0.938		0.014	1.000	1.000	0.717	0.717	0.000
1.000			1.000	1.000	0.779	0.779	0.000

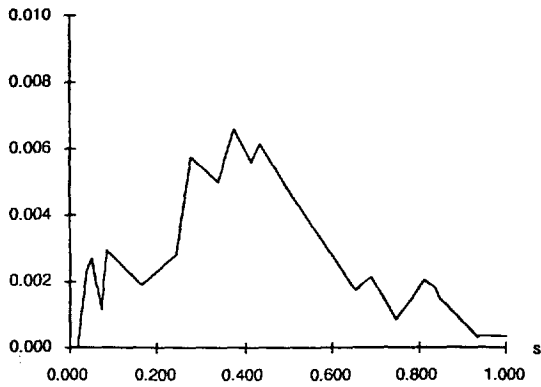


Fig. 4. Graph of $\int_0^s V_L(t) dt - \int_0^s V_G(t) dt$ for $0 \leq s \leq 1$.

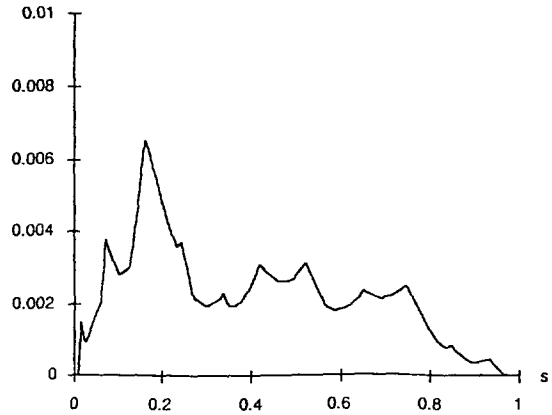


Fig. 5. Graph of $\int_0^s V_C(t) dt - \int_0^s V_L(t) dt$ for $0 \leq s \leq 1$.

calculations to compare f_L and f_G via this procedure are included in Table 2, and $\int_0^s V_L(t) dt - \int_0^s V_G(t) dt$ is graphed in Fig. 4. The fact that the curve in Fig. 4 is everywhere positive indicates that the local forecast is strictly sufficient for the guidance forecast. Thus, in a choosing scenario any decision maker would choose the local forecast. This result is consistent with intuition; the guidance forecast is actually used in preparing the local forecast, so it is no surprise that the local forecast performs better than the guidance forecast.

To check for extraneousness, we compare the combined forecast $f_C = P(\text{rain} | f_L, f_G)$ with the individual forecasts. In this case, we need only check f_L , because f_L has already been shown to be strictly sufficient for f_G . A procedure similar to the one described above can be performed to compare f_L and f_C , although the calculations are somewhat more tedious because f_C can potentially take on any of 88 different values (the number of non-empty cells in Fig. 1). Fig. 5 shows a plot of $\int_0^s V_C(t) dt - \int_0^s V_L(t) dt$; the fact that the curve is everywhere positive indicates that f_C is strictly sufficient for f_L . Thus, f_G is not extraneous and contributes information to the combined forecast.

Finally, we also show the forecast sufficiency characteristic (FSC) curves for f_G , f_L , and f_C in Fig. 6. These FSCs have been calculated according to the algorithm in Krzysztofowicz and Long (1991). Although the curves appear to fall nearly on top of each other, careful inspection shows

that $FSC_C \geq FSC_L \geq FSC_G$, confirming the results obtained above.

This application shows that real-world forecasts can have characteristics like those of Section 4.3. In a choosing scenario, one would always pick the local forecast because it is sufficient for the guidance forecast. However, this sufficiency does not mean that the guidance

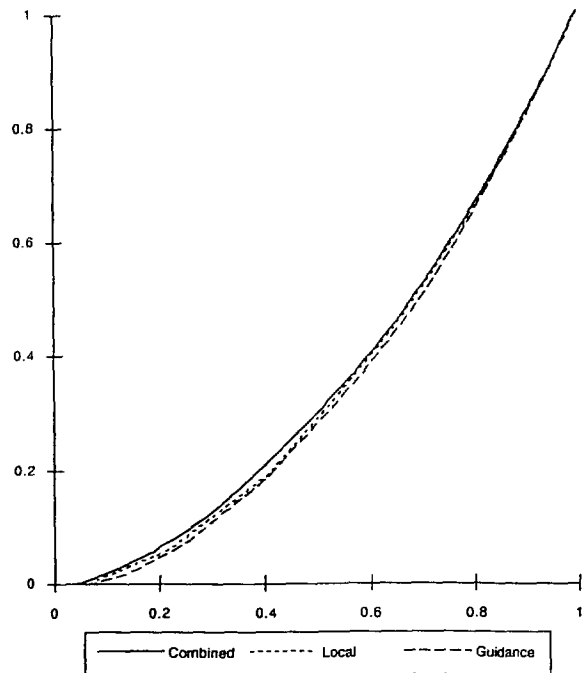


Fig. 6. Forecast sufficiency (FSC) curves for f_G , f_L , and f_C .

forecast is extraneous. In a combining scenario, the local forecast can be combined with the guidance forecast to advantage; the resulting combined forecast would be preferred by any decision maker because it improves on the local forecast in the sense of providing more information. Our result here is consistent with the more extensive study in Clemen and Murphy (1986a), who used a statistical procedure to test for extraneousness of the guidance and local forecasts.

6. Summary and discussion

In a given forecasting situation, we might consider a variety of forecasting methods because we want to choose a single method, and casting a wide net will help us wind up with “good” forecasts. Alternatively, looking at multiple forecasting methods opens the possibility of generating forecasts from two or more methods and then combining these forecasts. The appropriate forecast-evaluation methodology will depend on whether one must choose a single forecast or may combine multiple forecasts. In the choosing scenario, we can evaluate each method individually and then compare methods; in the combining scenario, on the other hand, we must evaluate the methods simultaneously in order to consider interrelationships among the methods as well as their individual performance.

In decision making, procedures for screening a set of alternatives can be useful in reducing the set to a more manageable size before a more detailed evaluation is conducted. Similarly, screening a set of forecast methods can reduce the number of methods under consideration. In this paper we have focused on procedures for screening probability forecasts, using the notions of sufficiency in the choosing scenario and extraneousness in the combining scenario. Examples have demonstrated the application of sufficiency and extraneousness and have illustrated some differences between these two concepts. Although the development here has been in terms of probability forecasts for single events, the same screening methods can be applied to prob-

ability forecasts for non-dichotomous variables and for point forecasts as well. Of course, as the forecasts become more complex (e.g. probability forecasts for continuous variables as opposed to single probabilities for events), the screening procedures may be more difficult to apply.

Screening imposes stringent conditions in the sense that sufficiency identifies methods that are clearly dominated by others and extraneousness identifies methods that provide no useful information beyond that contained in the remaining methods. After the screening process, a full evaluation of the remaining methods (or combinations of these methods) involves the appropriate distributions of forecasts and outcomes (Murphy and Winkler, 1992). For choosing a single method the bivariate distributions of forecasts and outcomes for the methods still under consideration are relevant. For combining, the full joint distribution of forecasts from all of these methods and the outcomes is relevant. In either case, these distributions, like the screening methods discussed in this paper, take full account of all information pertaining to the choosing or combining problem at hand. Furthermore, various aspects of forecast quality, such as calibration and refinement in the case of probability forecasts, can be isolated and studied on the basis of the distributions. In contrast, commonly used single-dimensional evaluation measures (e.g. scoring rules for probability forecasts, measures such as MSE for point forecasts) provide some information regarding the overall accuracy of the forecasts but do not provide a breakdown of forecast quality into separate aspects. Appropriate evaluation measures should reflect the full set of relevant information, thereby enabling a decision-maker to consider and make tradeoffs among different aspects of forecast quality.

References

- Blackwell, D., 1951, Comparison of experiments, in: J. Neyman, ed., *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, pp. 93–102.
- Blackwell, D., 1953, Equivalent comparisons of experiments, *Annals of Mathematical Statistics*, 24, 265–272.

- Blackwell, D. and M.A. Girshick, 1954. *Theory of Games and Statistical Decisions*. Wiley, New York.
- Clemen, R.T., 1985, Extraneous expert information, *Journal of Forecasting*, 4, 329–348.
- Clemen, R.T., 1987, Combining overlapping information, *Management Science*, 33, 373–380.
- Clemen, R.T., 1989, Combining forecasts: A review and annotated bibliography, *International Journal of Forecasting*, 5, 559–583.
- Clemen, R.T. and J. Guerard, 1989, Econometric GNP forecasts: Incremental information relative to naive extrapolation, *International Journal of Forecasting*, 5, 417–426.
- Clemen, R.T. and A.H. Murphy, 1986a, Objective and subjective precipitation probability forecasts: Statistical analysis of some interrelationships, *Weather and Forecasting*, 1, 58–65.
- Clemen, R.T. and A.H. Murphy, 1986b, Objective and subjective precipitation probability forecasts: Some methods for improving forecast quality, *Weather and Forecasting*, 1, 213–218.
- Clemen, R.T. and R.L. Winkler, 1990, Unanimity and compromise among probability forecasters, *Management Science*, 36, 767–779.
- DeGroot, M.H. and E.A. Eriksson, 1985, Probability forecasting, stochastic dominance, and the Lorenz curve, in: J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith, eds., *Bayesian Statistics 2*, North-Holland, Amsterdam, pp. 99–118.
- DeGroot, M.H. and S.E. Fienberg, 1982, Assessing probability assessors: Calibration and refinement, in: S.S. Gupta and J.O. Berger, eds., *Statistical Decision Theory and Related Topics III, Vol. 1*, Academic Press, New York, pp. 291–314.
- DeGroot, M.H. and S.E. Fienberg, 1983, The comparison and evaluation of forecasters, *The Statistician*, 32, 12–22.
- DeGroot, M.H. and S.E. Fienberg, 1986, Comparing probability forecasters: Basic binary concepts and multivariate extensions, in: P. Goel and A. Zellner, eds., *Bayesian Inference and Decision Techniques*, North-Holland, Amsterdam, pp. 247–264.
- 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, 1992, Evaluation of prototypical climate forecasts: The sufficiency relation, *Journal of Climate*, 5, 876–887.
- Krzysztofowicz, R., 1992, Bayesian correlation score: A utilitarian measure of forecast skill, *Monthly Weather Review*, 120, 208–219.
- Krzysztofowicz, R. and D. Long, 1991, Forecast sufficiency characteristic: Construction and application, *International Journal of Forecasting*, 7, 39–45.
- Murphy, A.H., 1985, Probabilistic weather forecasting, in: A. Murphy and R. Katz, eds., *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, Westview Press, Boulder, CO, pp. 337–377.
- Murphy, A.H. and R.L. Winkler, 1984, Probability forecasting in meteorology, *Journal of the American Statistical Association*, 79, 489–500.
- Murphy, A.H. and R.L. Winkler, 1992, Diagnostic verification of probability forecasts, *International Journal of Forecasting*, 8, 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.
- Murphy, A.H., Y.-S. Chen and R.T. Clemen, 1988, Statistical analysis of interrelationships between objective and subjective temperature forecasts, *Monthly Weather Review*, 116, 2121–2131.
- Winkler, R.L., 1967, The quantification of judgment: Some methodological suggestions, *Journal of the American Statistical Association*, 62, 1105–1120.
- Winkler, R.L., 1986, On 'good probability appraisers', in: P. Goel and A. Zellner, eds., *Bayesian Inference and Decision Techniques*, North-Holland, Amsterdam, pp. 265–278.

Biographies: Robert T. CLEMEN holds a Ph.D. in Business from Indiana University and is Associate Professor in the Lundquist College of Business at the University of Oregon. His research interests include decision analysis, decision theory, and the use and aggregation of expert information. His articles have appeared in a variety of scholarly journals, including *Management Science*, *International Journal of Forecasting*, and *Operations Research*. He is the author of the widely used text *Making Hard Decisions: An Introduction to Decision Analysis* (Duxbury, Belmont, CA, 1991).

Allan H. MURPHY is Professor Emeritus at Oregon State University and Principal of Prediction and Evaluation Systems of Corvallis, Oregon. In 1990–1991 he was University Corporation for Atmospheric Research senior visiting scientist at the National Meteorological Center (US National Weather Service). He received a B.S. from M.I.T. and M.S., M.A., and Ph.D. degrees from the University of Michigan. Dr. Murphy has held visiting positions at the International Institute for Applied Systems Analysis, University of Colorado, University of Vienna, and meteorological institutes in Finland, the Netherlands, Sweden and the United Kingdom. His primary research interests include probability forecasting, forecast verification, forecast use and value, decision analysis, and Bayesian statistics.

Robert L. WINKLER is James B. Duke Professor in the Fuqua School of Business and the Institute of Statistics and Decision Sciences at Duke University, Durham, NC 27706, where he is also serving as Senior Associate Dean for Faculty and Research in the Fuqua School. He received a B.S. from the University of Illinois and a Ph.D. from the University of Chicago, was at Indiana University prior to moving to Duke, and has held visiting positions at the University of Washington, the International Institute for Applied Systems Analysis, Stanford University, the National Center for Atmospheric Research, and INSEAD. His primary research interests include Bayesian inference, decision analysis, probability forecasting, combining forecasts, and risk assessment.