Comparative Evaluation of Weather Forecasting Systems: Sufficiency, Quality, and Accuracy

MARTIN EHRENDORFER* AND ALLAN H. MURPHY

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

(Manuscript received 22 June 1987, in final form 26 February 1988)

ABSTRACT

The concept of sufficiency, originally introduced in the context of the comparison of statistical experiments, has recently been shown to provide a coherent basis for comparative evaluation of forecasting systems. Specifically, forecasting system A is said to be sufficient for forecasting system B if B's forecasts can be obtained from A's forecasts by a stochastic transformation. The sufficiency of A's forecasts for B's forecasts implies that the former are of higher quality than the latter and that all users will find A's forecasts of greater value than B's forecasts. However, it is not always possible to establish that system A is sufficient for system B or vice versa. This paper examines the concept of sufficiency in the context of comparative evaluation of simple probabilistic weather forecasting systems and investigates its interpretations and implications from perspectives provided by a recently developed general framework for forecast verification.

It is shown here that if system A is sufficient for system B, then the basic performance characteristics of the two systems are related via sets of inequalities and A's forecasts are necessarily more accurate than B's forecasts. Conversely, knowledge of a complete set of performance characteristics makes it possible to infer whether A is sufficient for B, B is sufficient for A, or the two systems are insufficient for each other. In general, however, information regarding only relative accuracy, as measured by a performance measure such as the mean square error, will *not* be adequate to determine the presence or absence of sufficiency, except in situations in which the accuracy of the system of interest exceeds some relatively high critical value. These results, illustrated by means of numerical examples, suggest that comparative evaluation of weather forecasting systems should be based on fundamental performance characteristics rather than on overall performance measures.

Possible extensions of these results to situations involving more general forecasting systems, as well as some implications of the results for verification procedures and practices in meteorology, are briefly discussed.

1. Introduction

Comparative evaluation is concerned with comparing the performance of two (or more) forecasting systems or forecasters. In the context of weather forecasting, this comparison has traditionally been accomplished by computing an overall measure of performance (e.g., the mean square error or a skill score) for each system and then comparing the numerical values of these measures. It is implicitly assumed in this process that the measure of performance (a measure of accuracy or skill) completely characterizes the quality of the forecasts and that the forecasts judged to be of greater accuracy or skill are also of greater value to actual and potential users.

Recent studies related to quality/value relationships for weather forecasts in the cost–loss ratio problem (e.g., Chen et al. 1987; Murphy and Ehrendorfer 1987) reveal that overall measures of performance, by themselves, generally do not completely determine forecast quality. Specifically, these studies demonstrate that the relationships between performance measures and measures of forecast value are usually described by multivalued functions (i.e., envelopes). Such relationships are multivalued because accuracy (or skill) and value are different multivalued functions of basic characteristics of performance (i.e., the relationships are inherently multidimensional and therefore cannot be described by single-valued functions). The existence of such multivalued relationships implies that forecasts of greater accuracy (or skill) can actually be less valuable to users. Thus, the conditions under which one forecasting system can be judged to be unambiguously superior to another forecasting system are evidently not well defined, at least in the existing meteorological literature.

The fundamental work on the comparison of statistical experiments by Blackwell (1951, 1953) provides the basis for a coherent approach to comparative eval-

* Current affiliation: Institute for Meteorology and Geophysics, University of Vienna, Vienna, Austria.

Corresponding author address: Professor Allan H. Murphy, Department of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331-2209.

uation. This work identified the conditions under which one experiment could be considered to be more informative than—or sufficient for—another experiment. In a recent series of papers, DeGroot and Fienberg (1982, 1983, 1986) introduced the concept of sufficiency into a forecasting context, and they presented important results related to the comparative evaluation of forecasting systems (or forecasters). Specifically, they described the conditions under which one forecasting system can be considered to be sufficient for another forecasting system. The fact that system A is sufficient for system B implies that (i) A's forecasts are of higher quality than B's forecasts and (ii) A's forecasts are of greater value than B's forecasts for all users. These two implications (of sufficiency) provide a strong justification for adopting the concept of sufficiency as the basis for comparative evaluation of weather forecasting systems.

The purposes of this paper are (i) to discuss the concept of sufficiency and illustrate its interpretations and implications in the context of comparative evaluation of weather forecasting systems and (ii) to investigate the relationships among sufficiency, forecast quality (as determined by basic characteristics of forecasting systems), and forecast accuracy. In pursuit of these objectives, we will make extensive use of a recently developed general framework for forecast verification (Murphy and Winkler 1987). This framework, which is based on the joint distribution of forecasts and observations, provides three complementary perspectives on the issues of concern here: a perspective based on the joint distribution itself and two perspectives associated with factorizations of the joint distribution into conditional and marginal distributions. For simplicity, the present paper focuses on primitive probabilistic forecasting systems that produce forecasts for binary-event (e.g., precipitation/no precipitation) situations. The accuracy of these forecasting systems will be measured by the Brier score (Brier 1950), the mean square error of the probabilistic forecasts.

A definition of sufficiency for binary-event situations is presented in section 2. Four primitive probabilistic weather forecasting systems are also introduced in this section. These systems provide the basis for numerical examples and sample calculations that appear in subsequent sections of the paper. Section 3 contains both interpretations of sufficiency within the three major perspectives provided by the general framework for forecast verification and numerical examples to illustrate the three alternative means of determining sufficiency. Section 4 explores the implications of sufficiency for basic characteristics of forecast quality and measures of accuracy. Conversely, the extent to which characteristics of forecast quality and performance measures provide a basis for drawing inferences regarding sufficiency is investigated in section 5. Section 6 consists of a discussion and conclusion, including some remarks concerning the extension of these concepts and methods to more general weather forecasting systems and the implications of these results for procedures and practices in forecast evaluation.

2. The concept of sufficiency

a. *Definition of sufficiency*

The forecasting systems considered here are assumed to produce forecasts for a binary variable X, where $X = 1$ if the event of concern (e.g., precipitation) occurs and $X = 0$ otherwise. Unconditional or marginal probabilities of the two events are denoted by $p_1 = \Pr(X = 1)$ and $p_0 = \Pr(X = 0) = 1 - p_1$. In a meteorological context, these probabilities are the (sample) climatological probabilities. Extensions of the concepts, interpretations, and methods presented in this paper to situations involving more than two events will be discussed briefly in section 6.

A forecast (F) produced by such a system is assumed here to be a member of a finite set T of k numerical values, where $T = \{f_1, f_2, \cdots, f_k\}$. The members of this set specify various levels of the probability of occurrence of the event of concern; thus, the forecasts are probabilistic forecasts. For example, in the case of precipitation probability forecasts, the set T might contain the eleven equally spaced values 0.0(0.1)1.0. Given the climatological probabilities, these forecasting systems are completely characterized by the two conditional probabilities $p(f \mid 1)$ and $p(f \mid 0)$, where

$$p(f \mid 1) = \Pr(F = f \mid X = 1), \tag{1}$$

$$p(f \mid 0) = \Pr(F = f \mid X = 0) \tag{2}$$

[it is implicitly assumed here that $p(f \mid 1)$ and $p(f \mid 0)$ are stationary; that is, they are assumed to be time invariant]. These conditional probabilities are generally referred to as likelihoods, since they indicate the likelihoods of the forecasts given the respective events. For $f = 0.30$, for example, $p(0.30 \mid 1)$ and $p(0.30 \mid 0)$ are the likelihoods of a probability forecast of 0.30 given precipitation and no precipitation, respectively.

The probability of use of the forecast $F = f$ is denoted here by π_f; that is, $\pi_f = \Pr(F = f)$, where $\pi_f \geq 0$ and $\sum_{f \in T} \pi_f = 1$. These unconditional or marginal probabilities are generally referred to as predictive probabilities. For example, $\pi_{0.30}$ is the probability that a precipitation probability forecast of 0.30 is used. Estimates of probabilities such as π_f, as well as $p(f \mid x)$ and p_1, can be obtained from the empirical joint distribution of forecasts and corresponding events.

It is now possible to present a definition of the concept of sufficiency for binary events:

Definition of sufficiency: Let A and B denote two binary-event probabilistic forecasting systems. Then *system A is sufficient for system B* if a stochastic transformation $h(f^B \mid f^A)$ exists such that

$$\sum_{f^A \in T} h(f^B|f^A)p^A(f^A|1) = p^B(f^B|1) \quad \text{for each } f^B \in T, \quad (3)$$

$$\sum_{f^A \in T} h(f^B|f^A)p^A(f^A|0) = p^B(f^B|0) \quad \text{for each } f^B \in T, \quad (4)$$

where the superscripts A and B denote the respective forecasting systems.

The function $h(f^B|f^A)$ in (3) and (4) qualifies as a stochastic transformation if $0 \leq h(f^B|f^A) \leq 1$ and $\sum_{f^B \in T} h(f^B|f^A) = 1$ for each $f^A \in T$. It is assumed here that if systems A and B do not use the same set of permissible forecast values, then T denotes the union of the two sets of values.

Insight into the interpretation of the definition of sufficiency—and its implications—can be obtained by recognizing that the likelihoods in (1) and (2) describe the ability of the forecasting system to *discriminate* between different events (see Murphy and Winkler 1987). Thus, the existence of a stochastic transformation such as that defined in (3) and (4), tends to "average out" the discriminatory ability of the likelihoods; therefore, system B is necessarily less discriminatory than system A. This situation implies that forecasting system B contains greater uncertainty than forecasting system A concerning the likelihood of occurrence of the events. Therefore, the most important implication of the existence of such a stochastic transformation is that *all* users will prefer system A to system B. However, a "solution" $h(f^B|f^A)$ that qualifies as a stochastic transformation may not exist. As a result, it may not be possible to establish that system A is sufficient for system B, or vice versa. In such cases, the two systems will be said here to be *insufficient* for each other.

As defined above, the concept of sufficiency involves forecasts from a finite set T of k distinct values. For simplicity, it will be assumed here that the forecasts of concern employ only two probability values, f_1 and f_0. Two interpretations can be given to such *primitive* probabilistic forecasts: (i) the forecasts are calibrated versions of previously produced categorical forecasts (in which case, $f_1 = 1$ and $f_0 = 0$) or (ii) the forecasts are inherently probabilistic but involve the use of only two probability values. For convenience, the labels "1" and "0" will be used throughout this paper to represent f_1 and f_0, respectively. However, it should be kept in mind that these labels *do not necessarily* represent categorical forecasts (i.e., probabilities of one and zero) and that the labels "1" and "0" may have different meanings for the two forecasting systems. Thus, the set of permissible forecast values is now T^*, where $T^* = \{f_1^A, f_0^A, f_1^B, f_0^B\}$.

For the primitive probabilistic forecasts defined in the previous paragraph, $p(f|x)$ is zero except for f = "1" and f = "0" [i.e., only $p(\text{"1"}|1)$, $p(\text{"1"}|0)$, $p(\text{"0"}|1)$, and $p(\text{"0"}|0)$ are nonzero]. Thus, the definition of sufficiency [see (3) and (4)] for such forecasts reduces to the following set of equations (omitting the quotation marks around the labels):

$$h(f^B = 1|f^A = 1)p^A(f^A = 1|1) + h(f^B = 1|f^A = 0)p^A(f^A = 0|1) = p^B(f^B = 1|1), \quad (5)$$

$$h(f^B = 0|f^A = 1)p^A(f^A = 1|1) + h(f^B = 0|f^A = 0)p^A(f^A = 0|1) = p^B(f^B = 0|1), \quad (6)$$

$$h(f^B = 1|f^A = 1)p^A(f^A = 1|0) + h(f^B = 1|f^A = 0)p^A(f^A = 0|0) = p^B(f^B = 1|0), \quad (7)$$

$$h(f^B = 0|f^A = 1)p^A(f^A = 1|0) + h(f^B = 0|f^A = 0)p^A(f^A = 0|0) = p^B(f^B = 0|0). \quad (8)$$

Moreover, (6) and (8) are redundant with respect to (5) and (7), respectively. Therefore, using the simplified notation $p_{11}^A = p^A(f^A = 1|1)$, $p_{10}^A = p^A(f^A = 1|0)$, etc., this system of equations can be reduced to the following two linear independent equations:

$$h(f^B = 1|f^A = 1)p_{11}^A + h(f^B = 1|f^A = 0)(1 - p_{11}^A) = p_{11}^B, \quad (9)$$

$$h(f^B = 1|f^A = 1)p_{10}^A + h(f^B = 1|f^A = 0)(1 - p_{10}^A) = p_{10}^B. \quad (10)$$

Perhaps it should be noted at this point that the probabilities f_1 and f_0, used by systems A and B (see T^*), do not enter into the sufficiency relationship between the two forecasting systems in any way. They simply represent labels, and only their likelihoods are relevant to the determination of sufficiency.

b. *Hypothetical weather forecasting systems*

In order to illustrate various interpretations of the concept of sufficiency and to facilitate the discussion of some implications of sufficiency for forecast quality and forecast accuracy (and vice versa), we will consider four hypothetical binary-event primitive probabilistic weather forecasting systems in this paper. These systems are described in terms of samples of forecasts and the corresponding observations in Tables 1–4. In each case, the description begins with the joint frequency of $n = 10{,}000$ forecasts and observations and includes the joint, conditional, and marginal probabilities associated with these frequencies (the conditional probabilities ρ_{11} and ρ_{10} will be defined in section 3) as well as the expected half Brier score (see section 4b). The role of the reference system in these interpretations and discussions of implications will be played by forecasting system A (see Table 1). Forecasting systems B1, B2, and B3 (Tables 2, 3, and 4, respectively) represent alternative forecasting systems currently under considera-

TABLE 1. Description of reference forecasting system A.

Joint frequencies (n = 10 000):

		X = 1	X = 0
F	f_1^A	2400	1800
	f_0^A	1600	4200

Joint and marginal distributions ($c^A = p_1{}^A - a^A$, $d^A = p_0{}^A - b^A$):

		X = 1	X = 0	
F	f_1^A	$a^A = 0.24$	$b^A = 0.18$	$\pi_1{}^A = 0.42$
	f_0^A	$c^A = 0.16$	$d^A = 0.42$	$\pi_0{}^A = 0.58$
		$p_1 = 0.40$	$p_0 = 0.60$	1.00

Conditional distributions of forecasts given observations ($p_{01}^A = 1 - p_{11}^A$, $p_{00}^A = 1 - p_{10}^A$):

		X = 1	X = 0
F	f_1^A	$p_{11}^A = 0.6$	$p_{10}^A = 0.3$
	f_0^A	$p_{01}^A = 0.4$	$p_{00}^A = 0.7$
		1.0	1.0

Conditional distributions of observations given forecasts ($\rho_{01}^A = 1 - \rho_{11}^A$, $\rho_{00}^A = 1 - \rho_{10}^A$):

		X = 1	X = 0	
F	f_1^A	$\rho_{11}^A = 0.571$	$\rho_{01}^A = 0.429$	1.000
	f_0^A	$\rho_{10}^A = 0.276$	$\rho_{00}^A = 0.724$	1.000

Expected half Brier Score: $BS^A = 0.2187$

tion, and they will be compared with forecasting system A. Examination of these tables reveals that all four forecasting systems possess the same climatological probabilities; that is, they are *matched* forecasting systems in the sense that $p_1{}^A = p_1{}^B = p_1 = 0.4$ and $p_0{}^A = p_0{}^B = p_0 = 0.6$. Otherwise, these systems exhibit different characteristics as reflected by their joint, conditional, and predictive probabilities.

3. Interpretation and determination of sufficiency

a. LBR, CR, and JD interpretations

In this section we present three interpretations of the concept of sufficiency. These interpretations, which are tailored to the binary-event primitive probabilistic forecasting systems introduced in section 2a, are based on perspectives provided by a general framework for forecast verification (Murphy and Winkler 1987). In effect, this framework reveals that all of the (non-time-dependent) information relevant to forecast verification is contained in the joint distribution (JD) of forecasts and observations or, alternatively, in the calibration-refinement (CR) or likelihood-base rate (LBR) factorizations of the joint distribution into conditional and marginal distributions. The CR factorization involves the conditional distributions of observations given forecasts and the marginal distribution of the forecasts, whereas the LBR factorization involves the conditional distributions of forecasts given observations and the marginal distribution of the observations. Although it is necessary to utilize only a single interpretation in order to investigate sufficiency, we believe that it is both instructive and useful to consider all three interpretations in this introduction to—and expository treatment of—the sufficiency concept and its application.

The concept of sufficiency, as defined in section 2a, involves the likelihoods of forecasting systems A and B. Specifically, this concept is described in terms of p_{11}^A, p_{10}^A, p_{11}^B, and p_{10}^B. Thus, it is immediately possible, based on (9) and (10), to interpret the concept of sufficiency in terms of the LBR factorization as follows:

LBR interpretation of sufficiency. Consider two binary-event forecasting systems, A and B say, each of which produces primitive probabilistic forecasts. Then system A is sufficient for system B if a stochastic transformation characterized by u and v exists such that

TABLE 2. Description of forecasting system B1.

Joint frequencies (n = 10 000):

		X = 1	X = 0
F	f_1^{B1}	1667	1667
	f_0^{B1}	2333	4333

Joint and marginal distributions ($c^{B1} = p_1 - a^{B1}$, $d^{B1} = p_0 - b^{B1}$):

		X = 1	X = 0	
F	f_1^{B1}	$a^{B1} = 0.1667$	$b^{B1} = 0.1667$	$\pi_2{}^{B1} = 0.3333$
	f_0^{B1}	$c^{B1} = 0.2333$	$d^{B1} = 0.4333$	$\pi_0{}^{B1} = 0.6667$
		$p_1 = 0.4000$	$p_0 = 0.6000$	1.000

Conditional distributions of forecasts given observations ($p_{01}^{B1} = 1 - p_{11}^{B1}$, $p_{00}^{B1} = 1 - p_{10}^{B1}$):

		X = 1	X = 0
F	f_1^{B1}	$p_{11}^{B1} = 0.4167$	$p_{10}^{B1} = 0.2778$
	f_0^{B1}	$p_{01}^{B1} = 0.5833$	$p_{00}^{B1} = 0.7222$
		1.0000	1.0000

Conditional distributions of observations given forecasts ($\rho_{01}^{B1} = 1 - \rho_{11}^{B1}$, $\rho_{00}^{B1} = 1 - \rho_{10}^{B1}$):

		X = 1	X = 0	
F	f_1^{B1}	$\rho_{11}^{B1} = 0.50$	$\rho_{01}^{B1} = 0.50$	1.00
	f_0^{B1}	$\rho_{10}^{B1} = 0.35$	$\rho_{00}^{B1} = 0.65$	1.00

Expected half Brier score: $BS^{B1} = 0.2350$

TABLE 3. Description of forecasting system B2.

Joint frequencies (n = 10 000):

		X = 1	X = 0
F	f_1^{B2}	250	2250
	f_0^{B2}	3750	3750

Joint and marginal distributions ($c^{B2} = p_1 - a^{B2}$, $d^{B2} = p_0 - b^{B2}$):

		X = 1	X = 0	
F	f_1^{B2}	$a^{B2} = 0.0250$	$b^{B2} = 0.2250$	$\pi_1^{B2} = 0.2500$
	f_0^{B2}	$c^{B2} = 0.3750$	$d^{B2} = 0.3750$	$\pi_0^{B2} = 0.7500$
		$p_1 = 0.4000$	$p_0 = 0.6000$	1.0000

Conditional distributions of forecasts given observations ($p_{01}^{B2} = 1 - p_{11}^{B2}$, $p_{00}^{B2} = 1 - p_{10}^{B2}$):

		X = 1	X = 0
F	f_1^{B2}	$p_{11}^{B2} = 0.0625$	$p_{10}^{B2} = 0.3750$
	f_0^{B2}	$p_{01}^{B2} = 0.9375$	$p_{00}^{B2} = 0.6250$
		1.0000	1.0000

Conditional distributions of observations given forecasts ($\rho_{01}^{B2} = 1 - \rho_{11}^{B2}$, $\rho_{00}^{B2} = 1 - \rho_{10}^{B2}$):

		X = 1	X = 0	
F	f_1^{B2}	$\rho_{11}^{B2} = 0.1$	$\rho_{01}^{B2} = 0.9$	1.0
	f_0^{B2}	$\rho_{10}^{B2} = 0.5$	$\rho_{00}^{B2} = 0.5$	1.0

Expected half Brier Score: $BS^{B2} = 0.2100$

$$up_{11}^{A} + v(1 - p_{11}^{A}) = p_{11}^{B}, \quad (11)$$

$$up_{10}^{A} + v(1 - p_{10}^{A}) = p_{10}^{B}, \quad (12)$$

where $u = h(f^B = 1 | f^A = 1)$ and $v = h(f^B = 1 | f^A = 0)$.

In view of the fact that the LBR factorization involves the likelihoods and the climatological probabilities (i.e., base rates), it is interesting to note that the LBR interpretation of sufficiency does not depend explicitly on the climatological probabilities associated with the two forecasting systems.

For the primitive probabilistic forecasts of concern here, the relationships among the joint, conditional, and marginal probabilities can be expressed as follows (as elementary consequences of Bayes' theorem):

$$\Pr(F = 1, X = 1) = p_{11}p_1 = \rho_{11}\pi_1, \quad (13)$$

$$\Pr(F = 1, X = 0) = p_{10}p_0 = (1 - \rho_{11})\pi_1, \quad (14)$$

$$\Pr(F = 0, X = 1) = (1 - p_{11})p_1 = \rho_{10}\pi_0, \quad (15)$$

$$\Pr(F = 0, X = 0) = (1 - p_{10})p_0 = (1 - \rho_{10})\pi_0, \quad (16)$$

where

$$\rho_{11} = \Pr(X = 1 | F = 1), \quad (17)$$

$$\rho_{10} = \Pr(X = 1 | F = 0). \quad (18)$$

Note that (13) and (15), when added together, yield $\pi_1 = (p_1 - \rho_{10})/(\rho_{11} - \rho_{10})$.

The relationships among the conditional and predictive probabilities in (13)–(16), when considered in conjunction with the LBR interpretation of sufficiency [see (11) and (12)], can be used to derive the following CR interpretation of sufficiency:

CR interpretation of sufficiency. For binary-event forecasting systems A and B, both of which produce primitive probabilistic forecasts, system A is sufficient for system B if a stochastic transformation characterized by u and v exists such that

$$u\rho_{11}^{A}\pi_1^{A}/p_1^{A} + v\rho_{10}^{A}\pi_0^{A}/p_1^{A} = \rho_{11}^{B}\pi_1^{B}/p_1^{B}, \quad (19)$$

$$u(1 - \rho_{11}^{A})\pi_1^{A}/p_0^{A} + v(1 - \rho_{10}^{A})\pi_0^{A}/p_0^{A} = (1 - \rho_{11}^{B})\pi_1^{B}/p_0^{B}. \quad (20)$$

Moreover, if systems A and B are matched forecasting systems, then (19) and (20) become

TABLE 4. Description of forecasting system B3.

Joint frequencies (n = 10 000):

		X = 1	X = 0
F	f_1^{B3}	3429	857
	f_0^{B3}	571	5143

Joint and marginal distributions ($c^{B3} = p_1 - a^{B3}$, $d^{B3} = p_0 - b^{B3}$):

		X = 1	X = 0	
F	f_1^{B3}	$a^{B3} = 0.3429$	$b^{B3} = 0.0857$	$\pi_1^{B3} = 0.4286$
	f_0^{B3}	$c^{B3} = 0.0571$	$d^{B3} = 0.5143$	$\pi_0^{B3} = 0.5714$
		$p_1 = 0.4000$	$p_0 = 0.6000$	

Conditional distributions of forecasts given observations ($p_{01}^{B3} = 1 - p_{11}^{B3}$, $p_{00}^{B3} = 1 - p_{10}^{B3}$):

		X = 1	X = 0
F	f_1^{B3}	$p_{11}^{B3} = 0.8571$	$p_{10}^{B3} = 0.1429$
	f_0^{B3}	$p_{01}^{B3} = 0.1429$	$p_{00}^{B3} = 0.8571$
		1.0000	1.0000

Conditional distributions of observations given forecasts ($\rho_{01}^{B3} = 1 - \rho_{11}^{B3}$, $\rho_{00}^{B3} = 1 - \rho_{10}^{B3}$):

		X = 1	X = 0	
F	f_1^{B3}	$\rho_{11}^{B3} = 0.8$	$\rho_{01}^{B3} = 0.2$	1.0
	f_0^{B3}	$\rho_{10}^{B3} = 0.1$	$\rho_{00}^{B3} = 0.9$	1.0

Expected half Brier Score: $BS^{B3} = 0.1200$

$$u\rho_{11}^{A}\pi_1^{A} + v\rho_{10}^{A}\pi_0^{A} = \rho_{11}^{B}\pi_1^{B}, \quad (21)$$

$$u(1 - \rho_{11}^{A})\pi_1^{A} + v(1 - \rho_{10}^{A})\pi_0^{A} = (1 - \rho_{11}^{B})\pi_1^{B}, \quad (22)$$

respectively.

It is interesting to note that the sum of (21) and (22) yields

$$u\pi_1^{A} + v\pi_0^{A} = \pi_1^{B} \quad (23)$$

[see Eq. (2.8) of Theorem 2 in DeGroot and Fienberg 1986], which represents a randomization of A's predictive probabilities to obtain B's predictive probabilities. This relationship is consistent with the notion that B's forecasts are more uncertain than A's forecasts.

The relationships in (13)–(16), when considered in conjunction with the LBR interpretation of sufficiency [(11) and (12)], can also be used to obtain a JD interpretation of sufficiency. For convenience, let $a = \Pr(F = 1, X = 1)$ and $b = \Pr(F = 1, X = 0)$. Then, this interpretation can be expressed as follows:

JD interpretation of sufficiency. For binary-event forecasting systems A and B, both of which produce primitive probabilistic forecasts, system A is sufficient for system B if a stochastic transformation characterized by u and v exists such that

$$ua^{A}/p_1^{A} + v(p_1^{A} - a^{A})/p_1^{A} = a^{B}/p_1^{B}, \quad (24)$$

$$ub^{A}/p_0^{A} + v(p_0^{A} - b^{A})/p_0^{A} = b^{B}/p_0^{B}. \quad (25)$$

Moreover, if systems A and B are matched forecasting systems, then (24) and (25) become

$$ua^{A} + v(p_1 - a^{A}) = a^{B}, \quad (26)$$

$$ub^{A} + v(p_0 - b^{A}) = b^{B}, \quad (27)$$

respectively [see Eq. (2.9) of Theorem 2 in DeGroot and Fienberg 1986].

The three sets of equations (11) and (12), (21) and (22), and (26) and (27) define the conditions under which system A is sufficient for system B according to the LBR, CR, and JD interpretations, respectively (in the case of matched forecasting systems). These conditions are expressed in terms of the quantities u and v, and sufficiency is assured if $0 \leq u \leq 1$ and $0 \leq v \leq 1$. The pair of equations associated with any of the three interpretations can be used to determine the sufficiency (or insufficiency) of one forecasting system for another forecasting system.

Another possible interpretation that can be given to the conditions under which system A is sufficient for system B is provided by Theorem 2 in DeGroot and Fienberg (1986). In our notation this theorem states that, for binary-event matched forecasting systems A and B both of which produce primitive probabilistic forecasts, system A is sufficient for system B if and only if a stochastic transformation characterized by u and v exists such that (23) and (26) hold.

b. Determination of sufficiency: some examples

We will use the equations associated with the various interpretations of sufficiency to compare forecasting systems B1, B2, and B3 with the reference forecasting system A (see Tables 1–4). As noted in section 2b, these four systems are assumed to be matched forecasting systems, in the sense that they possess the same climatological probabilities ($p_1 = 0.4$ and $p_0 = 0.6$ in each case). This assumption is equivalent to assuming that each system has prepared forecasts for the same set of forecasting situations.

Consider the problem of investigating the sufficiency of system A for system B1, making use of the equations associated with the LBR interpretation [see (11) and (12)]. Using the appropriate numerical values for these systems (see Tables 1 and 2), we obtain the following equations:

$$0.6u^{B1} + 0.4v^{B1} = 0.4167, \quad (28)$$

$$0.3u^{B1} + 0.7v^{B1} = 0.2778. \quad (29)$$

Solving these equations for u^{B1} and v^{B1} yields $u^{B1} = 0.6019$ and $v^{B1} = 0.1389$. Since $0 \leq u^{B1} \leq 1$ and $0 \leq v^{B1} \leq 1$, the conditions for a stochastic transformation are satisfied and it is evident that system A is sufficient for system B. Using these same equations [or, alternatively, Eqs. (21) and (22) or Eqs. (26) and (27)] to investigate, in turn, the sufficiency of system A for systems B2 and B3, we obtain the following results: $u^{B2} = -0.3542$ and $v^{B2} = 0.6875$ and $u^{B3} = 1.8095$ and $v^{B3} = -0.5714$. Since one or both of these values lie outside of the unit interval in each case, no stochastic transformation exists and system A is not sufficient for either system B2 or system B3.

Although system A is not sufficient for systems B2 or B3, it is still possible for one or both of these latter systems to be sufficient for system A. To investigate these possibilities, it is necessary simply to interchange the labels A and B (actually, B2 or B3) on the terms in equations such as (11) and (12). In the case of B2, we would obtain the following equations:

$$0.0625u^{B2} + 0.9375v^{B2} = 0.6, \quad (30)$$

$$0.3750u^{B2} + 0.6250v^{B2} = 0.3. \quad (31)$$

Solving these equations yields $u^{B2} = -0.30$ and $v^{B2} = 0.66$, indicating that B2 is *not* sufficient for A. Since B2 is not sufficient for A and A is not sufficient for B2, systems A and B2 are insufficient for each other. Using the analogous equations for system B3, we obtain the solution $u^{B3} = 0.66$ and $v^{B3} = 0.24$. Since both values lie in the unit interval, it is evident that system B3 *is* sufficient for system A.

Thus, it is possible to use the equations defining the respective interpretations of sufficiency to determine whether one forecasting system is sufficient (or insufficient) for another forecasting system. To implement this procedure, it is necessary simply to be able to de-

scribe the respective forecasting systems in terms of their basic *performance characteristics* (i.e., p_{11} and p_{10} in the LBR interpretation, ρ_{11} and ρ_{10} in the CR interpretation, and a and b in the JD interpretation; see section 4). Further discussion of the procedures and conditions for determining sufficiency (or insufficiency) between two forecasting systems will be described in sections 4 and 5, in conjunction with the examination of the implications of sufficiency for forecast quality and forecast accuracy (and vice versa).

4. Implications of sufficiency for quality, accuracy, and value

Prior to the consideration of the issues of primary concern in this section, it is necessary to define some terminology. First, the term *performance measure* will be used here to describe verification measures that provide some—but generally incomplete—information about the performance of a forecasting system. An example of such a measure, in the context of probability forecasting, is the Brier score. As noted in section 1, it is the mean square error of probabilistic forecasts and, as such, is a measure of forecast accuracy. Second, the term *performance characteristics* will be used to describe basic components of forecast quality. Given a complete set of performance characteristics, it is possible to recover the entire relationship (i.e., joint distribution) between forecasts and observations. For example, from the perspective of the LBR factorization of the joint distribution, p_{11} and p_{10} represent a complete set of performance characteristics (assuming that p_1 is known). In general, performance measures are relatively complex, nonlinear functions of one or more performance characteristics (e.g., Chen et al. 1987).

Introduction of the concept of sufficiency and discussion of some of its interpretations in sections 2 and 3 raises the following question that will be addressed here: Given that forecasting system A is sufficient for forecasting system B, what are the implications of this condition for relationships between the respective (i) performance characteristics and (ii) performance measures associated with the two systems? For convenience, it will be assumed that the forecasts of interest are primitive probabilistic forecasts (as introduced in section 2a) and that the (sample) climatological probability (p_1) is the same for both systems and is known.

a. *Implications for forecast quality (performance characteristics)*

In the *LBR framework,* the condition that system A is sufficient for system B implies that the solution for u and v in (11) and (12) lies between zero and one (since a stochastic transformation exists). As a result, the following four inequalities on the performance characteristics p_{11}^B and p_{10}^B are obtained which must be satisfied simultaneously for A to be sufficient for B:

$$\mathrm{LR}_0(\mathrm{A}) \leq p_{11}^B/p_{10}^B \leq \mathrm{LR}_1(\mathrm{A}), \tag{32}$$

$$\mathrm{LR}_0(\mathrm{A}) \leq (1 - p_{11}^B)/(1 - p_{10}^B) \leq \mathrm{LR}_1(\mathrm{A}), \tag{33}$$

where $\mathrm{LR}_0(\mathrm{A}) = (1 - p_{11}^A)/(1 - p_{10}^A)$ and $\mathrm{LR}_1(\mathrm{A}) = p_{11}^A/p_{10}^A$ are the *likelihood ratios* of system A. In deriving (32) and (33) it has been assumed that $\mathrm{LR}_1(\mathrm{A}) > 1$, otherwise the inequalities must be reversed. From these restrictions on the likelihood ratios of system B, it is possible to identify those forecasting systems for which a specified reference system A is sufficient, and this procedure is illustrated by considering the hypothetical forecasting systems introduced in section 2b.

The four inequalities in (32) and (33) define a convex region in (p_{10}, p_{11})-parameter space. This region, denoted by S, is depicted in Fig. 1, in which p_{11}^B is plotted against p_{10}^B for the situation in which system A, characterized by $\mathrm{LR}_0(\mathrm{A}) = 4/7$ and $\mathrm{LR}_1(\mathrm{A}) = 2$ (see Table 1), is chosen as the reference system. The region S defines the set of all forecasting systems B for which system A is sufficient in this example [i.e., those systems B satisfying the inequalities in (32) and (33)].

Points representing forecasting systems A, B1, B2, and B3 are also depicted in Fig. 1. As expected, since system A is sufficient for system B1 (illustrated in section 3b), the point corresponding to system B1 falls in region S. The other regions and curves in this diagram will be discussed below (see section 5).

The boundaries of S are two pairs of parallel lines that are obtained from (32) and (33) when p_{11}^B is expressed as a (linear) function of p_{10}^B and the inequalities are replaced by equalities. Consideration of the respective nonparallel lines—for example,

$$p_{11}^B = \mathrm{LR}_1(\mathrm{A})p_{10}^B \quad \text{and} \quad p_{11}^B = 1 + \mathrm{LR}_0(\mathrm{A})(p_{10}^B - 1) \tag{34}$$

in the case of the right inequality in (32) and the left inequality in (33)—yields the two points of intersection in Fig. 1 with the coordinates $(p_{10}^B, p_{11}^B) = (p_{10}^A, p_{11}^A) = (0.3, 0.6)$ and $(p_{10}^B, p_{11}^B) = (1 - p_{10}^A, 1 - p_{11}^A) = (0.7, 0.4)$, respectively. The former is, of course, the point A (in Fig. 1) defining the reference forecasting system A in (p_{10}, p_{11})-parameter space.

These two points, which depend only on the performance characteristics of system A, completely determine the *geometry of sufficiency* in this framework. Thus, knowledge of the performance characteristics of the reference system can be used to define immediately the region S (as well as the other regions I and S').

In the *CR framework,* with the performance characteristics ρ_{11} and ρ_{10}, the condition that system A is sufficient for system B leads to the following inequalities on the performance characteristics of system B:

$$\rho_{10}^A \leq \rho_{10}^B < p_1 < \rho_{11}^B \leq \rho_{11}^A \tag{35}$$

or

$$\rho_{10}^A \leq \rho_{11}^B < p_1 < \rho_{10}^B \leq \rho_{11}^A \tag{36}$$

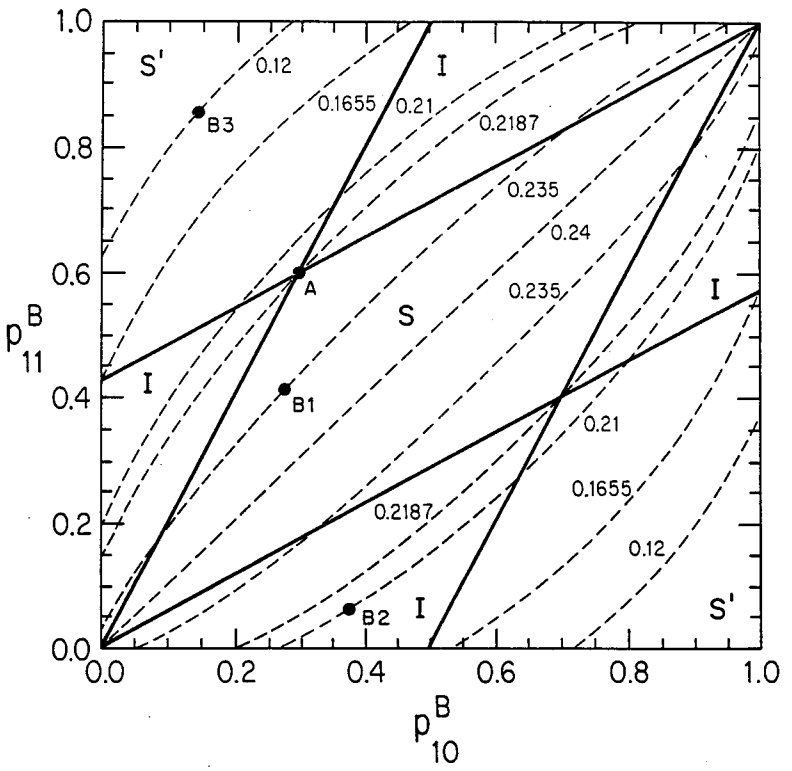

FIG. 1. The geometry of sufficiency and isopleths of expected half-Brier score (dashed curves) for binary-event primitive probabilistic forecasts, described in terms of performance characteristics associated with the likelihood-base rate factorization. A denotes the reference forecasting system and B1, B2, and B3 denote three alternative forecasting systems. The regions in which A is sufficient for B, A and B are insufficient for each other, and B is sufficient for A are denoted by *S*, *I*, and *S'*, respectively.

[see (21) and (22)]. It has been assumed here that $\rho^A_{11} > p_1$, which implies that $\rho^A_{11} > \rho^A_{10}$ (since $p_1 = \pi_1\rho_{11} + \pi_0\rho_{10}$; see below); otherwise, the inequalities in (35) and (36) must be reversed. Equation (35) corresponds to Lemma 2.1 in Krzysztofowicz and Long (1987), which states that if A is sufficient for B then (35) holds. The fact that (36) also holds means that, when system A is sufficient for system B the former is also sufficient for any system B with $\rho^B_{10} > p_1 > \rho^B_{11}$ (i.e., systems for which the event of concern is more likely to occur when $F = 0$ than when $F = 1$).

In this framework, the performance characteristics of reference system A define the geometry of sufficiency in (ρ_{11}, ρ_{10})-space, by means of the inequalities in (35) and (36). Specifically, they define the *rectangular* region *S* containing the points corresponding to the systems for which A is sufficient.

For the hypothetical forecasting systems introduced in section 2, the region *S* (and also *I* and *S'*; see section 5) and the points corresponding to systems A, B1, B2, and B3 are depicted in Fig. 2. Since A is sufficient for B1, the point B1 falls in region *S* (see also Tables 1 and 2). Combinations of values of ρ_{11} and ρ_{10} associated with the lower-left and upper-right corners of this diagram do not represent feasible forecasting systems because these combinations of values violate the equation $p_1 = \pi_1\rho_{11} + \pi_0\rho_{10}$, which is a consequence of Bayes' theorem [see (13) and (15)].

When two forecasting systems are compared within the *JD framework,* as described by (26) and (27), the joint probabilities a and b represent convenient performance characteristics. In this framework, the condition that system A is sufficient for system B implies that the following four inequalities must be satisfied simultaneously by B's performance characteristics a^B and b^B:

$$\mathrm{PR}_1(\mathrm{A}) \leq b^B/a^B \leq \mathrm{PR}_0(\mathrm{A}), \tag{37}$$

$$\mathrm{PR}_1(\mathrm{A}) \leq (p_0 - b^B)/(p_1 - a^B) \leq \mathrm{PR}_0(\mathrm{A}), \tag{38}$$

where $\mathrm{PR}_1(\mathrm{A}) = b^A/a^A$ and $\mathrm{PR}_0(\mathrm{A}) = (p_0 - b^A)/(p_1 - a^A)$ are the *joint probability ratios* characterizing system A. In developing these inequalities it was assumed that $\mathrm{PR}_1(\mathrm{A}) < p_0/p_1$ (which is equivalent to assuming that $\rho^A_{11} > p_1$); otherwise the inequalities in (37) and (38) must be reversed.

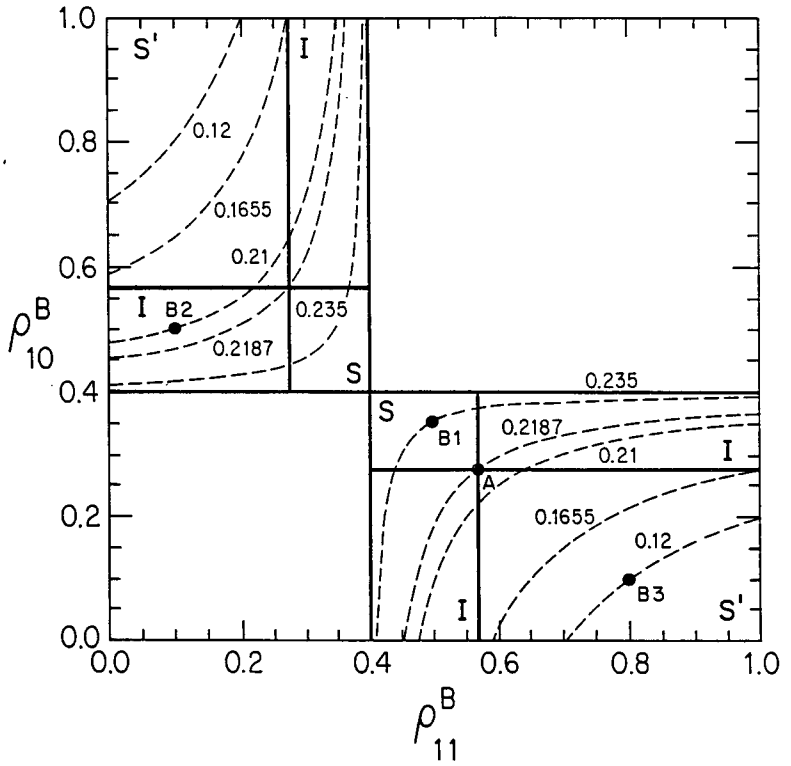

FIG. 2. As in Fig. 1 but for performance characteristics associated with the calibration-refinement factorization.

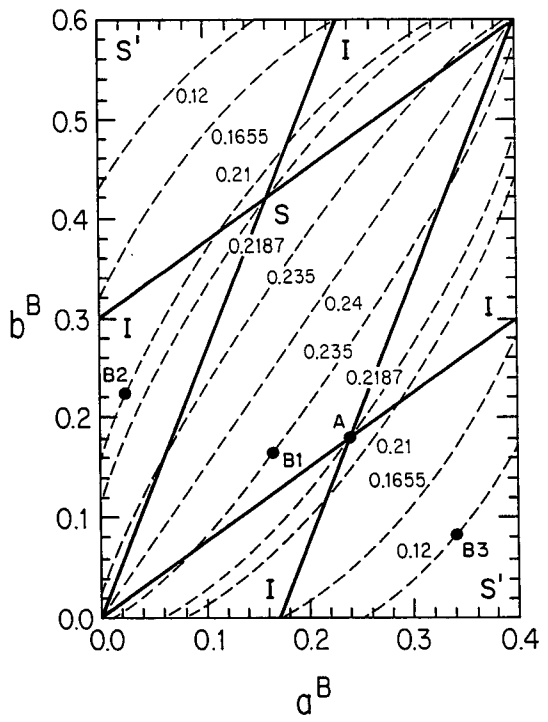

FIG. 3. As in Fig. 1 but for performance characteristics associated with the joint distribution of forecasts and observations.

Thus, in the JD framework, (37) and (38) define those systems B for which a given reference system A is sufficient. Specifically, all systems for which A is sufficient are found in a convex region S bounded by two pairs of parallel lines; these lines are obtained from (37) and (38) when b^B is expressed as a function of a^B with the inequalities being replaced by equalities. The points of intersection of the nonparallel lines have the coordinates $(a^B, b^B) = (p_1 - a^A, p_0 - b^A)$ and $(a^B, b^B) = (a^A, b^A)$ and are found in a manner analogous to that discussed in connection with the LBR framework. Therefore, the performance characteristics of the reference system A, in association with the climatological probability p_1, once again completely determine the geometry of sufficiency in (a, b)-space, as in the cases of the LBR and CR frameworks.

The shape of the region S in the JD framework is depicted in Fig. 3 (b^B plotted against a^B), once again using system A, with $\mathrm{PR}_1(\mathrm{A}) = 3/4$ and $\mathrm{PR}_0(\mathrm{A}) = 21/8$ (see Table 1) as the reference system. The points A, B1, B2, and B3 represent the corresponding forecasting systems in (a, b)-space; the points of intersection have the coordinates $(a^B, b^B) = (0.16, 0.42)$ and $(a^B, b^B) = (0.24, 0.18)$. Consistent with the results for this example in the LBR and CR frameworks, system B1 falls in the parallelogram S, indicating that A is sufficient for B1.

b. Implications for forecast quality (performance measures)

What are the implications of the condition "system A is sufficient for system B" for performance measures such as the Brier score? DeGroot and Fienberg (1986) showed that the expected Brier score can be decomposed into two terms, one of which measures the calibration of the forecasts and the other of which measures the refinement of the forecasts. For well-calibrated probabilistic forecasts, the first term in this decomposition vanishes. In the context of this paper, the assumption of well-calibrated forecasts implies that $f_1 = \Pr(X = 1 \mid F = f_1)$ and $f_0 = \Pr(X = 1 \mid F = f_0)$. That is, the relative frequencies with which $X = 1$ when the respective forecasts are used are equal to these forecast probabilities. In this regard, it can be seen from (17) and (18) that, under the assumption of well-calibrated forecasts, $f_1 = \rho_{11}$ and $f_0 = \rho_{10}$; that is, the forecast probabilities are equal to the conditional probabilities of event-occurrence/nonoccurrence given the respective forecasts. We adopt the assumption of well-calibrated forecasts here, but it should be noted that this assumption is used *only* when exploring the relationship between sufficiency and Brier scores.

In the case of well-calibrated primitive probabilistic forecasts, the expected half Brier score can be written as follows:

$$\mathrm{BS} = \pi_1[\rho_{11}(1 - \rho_{11})^2 + (1 - \rho_{11})(\rho_{11} - 0)^2] + (1 - \pi_1)[\rho_{10}(1 - \rho_{10})^2 + (1 - \rho_{10})(\rho_{10} - 0)^2] \quad (39)$$

(see also Murphy 1986). This measure of performance has a negative orientation, ranges from zero for perfect forecasts ($\rho_{11} = 1$, $\rho_{10} = 0$) to $p_1(1 - p_1)$ for climatological forecasts ($\rho_{11} = \rho_{10} = p_1$), and provides a complete ordering on the class of well-calibrated forecasting systems. Since the Brier score is a strictly proper scoring rule (Winkler and Murphy 1968), it can be shown (Theorems 9 and 4 in DeGroot and Fienberg 1986) that if forecasting system A is sufficient for forecasting system B then $\mathrm{BS}^A \leqslant \mathrm{BS}^B$, with strict inequality holding unless the two systems are identical. Thus, since system A is sufficient for system B1 in the example considered here, it follows that $\mathrm{BS}^A = 0.2187 < \mathrm{BS}^{B1} = 0.2350$. Since the converse of this theorem does not hold, conditions regarding sufficiency between two forecasting systems generally cannot be determined solely from their respective Brier scores. The implications for sufficiency of relationships between performance measures—and performance characteristics—of two forecasting systems will be investigated in section 5.

c. Implications for forecast value

As noted initially in section 1, an important feature of the sufficiency concept for comparative forecast verification is that if system A is sufficient for system B then the forecasts produced by system A are of greater economic value than the forecasts produced by system B for all users (i.e., regardless of the user's payoff structure). This property will be illustrated here in the context of the basic cost–loss ratio situation. For a recent description and discussion of this decision-making "model," see Murphy and Ehrendorfer (1987).

The expected value (per unit loss) of an imperfect weather forecast, VF, in this context can be defined as the difference in expected expenses between the situation involving only climatological information and the situation involving imperfect forecasts. Then VF can be calculated from

$$\mathrm{VF} = \min(C/L, p_1) - \pi_1 \min(C/L, \rho_{11}) - \pi_0 \min(C/L, \rho_{10}), \quad (40)$$

where $C/L (0 < C/L < 1)$ is the so-called cost–loss ratio (Murphy and Ehrendorfer 1987, p. 250). [In this model, C is the cost of taking protective action against adverse weather (i.e., the event occurrence) and L is the loss that is incurred if adverse weather occurs and protective action is not taken.] Note that VF in (40) depends only on the performance characteristics of the forecasting system and the cost–loss ratio C/L which characterizes the payoff structure of the user of the forecasts (assuming that p_1 is given).

The fact that a system A is sufficient for another system B implies that $\mathrm{VF}^A \geqslant \mathrm{VF}^B$ for all C/L. That is, sufficiency implies that this ordinal value-relationship holds for all users, regardless of their respective payoff structures. Moreover, this same relationship holds when more complex models are required to describe the users' decision-making problems.

To illustrate this result, VF was calculated for different values of C/L using (40), for the four forecasting systems introduced in section 2b. For example, $\mathrm{VF}^A(C/L = 0.48) = \min(0.48, 0.4) - 0.42 \min(0.48, 0.571) - 0.58 \min(0.48, 0.276) = 0.4 - 0.42*0.48 - 0.58*0.276 = 0.03832$. Figure 4 summarizes the results of these computations. The following relationships are immediately clear: (i) $\mathrm{VF}^{B3} > \mathrm{VF}^A$, since B3 is sufficient for A (see section 3b); (ii) VF^{B3} also exceeds VF^{B1} and VF^{B2}, since the former is also sufficient for the latter two systems; (iii) $\mathrm{VF}^{B2} \geqslant \mathrm{VF}^{B1}$, since B2 is sufficient for B1; (iv) $\mathrm{VF}^A > \mathrm{VF}^{B1}$, since A is sufficient for B1; and (v) no ordinal relationship exists between VF^A and VF^{B2}, since these two systems are insufficient for each other.

These illustrations, using the familiar cost–loss ratio situation as a basis for assessing the value of forecasting systems, provide tangible evidence of the claims made for the sufficiency concept. Namely, a forecasting system that is sufficient for another forecasting system will be preferred to this latter system by all users because of its greater expected economic value.

5. Implications of quality and accuracy for sufficiency

a. Implications of forecast quality (performance characteristics)

The problem addressed in this section is to decide, on the basis of a complete set of performance charac-

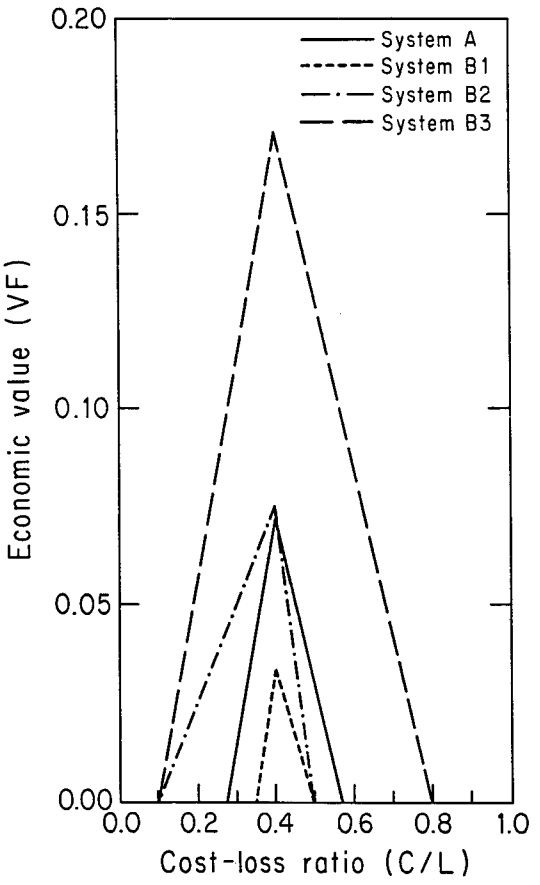

FIG. 4. The expected economic value VF [see Eq. (40)] of the hypothetical forecasting systems introduced in section 2b (see Tables 1–4) as a function of the cost–loss ratio C/L. For further details, see section 4c.

teristics for two forecasting systems, whether one system is sufficient for the other system or whether they are insufficient for each other. In other words, we want to determine the implications for the sufficiency relationship between two forecasting systems that can be inferred from knowledge of their relative quality (as measured by performance characteristics).

This question of sufficiency can be investigated within any of the three basic perspectives (i.e., LBR, CR, or JD) provided by the general framework for verification. For the binary-event primitive probabilistic forecasts considered in this paper, the sufficiency "problem" can be conveniently described in a two-dimensional diagram with coordinate axes corresponding to the performance characteristics of the chosen perspective. Since the specific values of the performance characteristics of a reference system determine completely the "geometry" within this two-dimensional space, it is conceptually reasonable to arbitrarily define one system as the reference system. The geometry implied by this reference system provides a means of determining the presence or absence of sufficiency with regard to the second system.

To examine the sufficiency relationship with respect to the systems A, B1, B2, and B3 (see section 2b) from the perspective of the LBR framework, we will again consider system A as the reference system. Its performance characteristics—$p_{11}^{A} = 0.6$ and $p_{10}^{A} = 0.3$—completely determine the regions identified in the two-dimensional space (see Fig. 1). The sufficiency relationship between system A and each of the B-systems can now be addressed directly by identifying the point in this space associated with the forecasting system B of interest. The situation can be summarized as follows: (i) if system B falls in region S, then system A is sufficient for system B (see section 4a); (ii) if system B falls in one of the four regions denoted by I, then systems A and B are insufficient for each other (A is not sufficient for B and B is not sufficient for A); and (iii) if system B falls in one of two regions denoted by S', then system B is sufficient for system A. The locations of the points in Fig. 1 corresponding to the alternative systems B1, B2, and B3 reveal that the presence or absence of sufficiency with regard to the reference system A, as described by this geometry, is in agreement with the analytical results presented in section 3. In summary, then, knowledge of the complete set of performance characteristics of two forecasting systems (e.g., p_{11} and p_{10} in the LBR framework) makes it possible to determine whether one forecasting system is sufficient or insufficient with respect to the other forecasting system, by reference to a simple geometrical framework.

Analogous results are obtained in the CR and JD frameworks (see Figs. 2 and 3, respectively). The designations associated with the regions (S, I, and S') in these frameworks have the same interpretations as in the case of the LBR framework. Regions in Fig. 2 without designations have been discussed in section 4a.

The geometry in the CR framework reveals an interesting result, summarized here in terms of the following lemma: *Lemma 1:* Let A and B represent two matched binary-event primitive probabilistic forecasting systems. Then system A is sufficient for system B if and only if either (35) or (36) holds (assuming without loss of generality that $\rho_{11}^{A} > \rho_{10}^{A}$). A proof of this lemma, which strengthens Lemma 2.1 reported by Krzysztofowicz and Long (1987) into an "if and only if" statement, is given in appendix A. Lemma 1 is quite useful because, given information on the relative magnitudes of a complete set of performance characteristics in the CR framework, it is possible to state immediately whether system A is sufficient for system B, system B is sufficient for system A, or the two systems are insufficient for each other.

For example, application of Lemma 1 shows immediately that system B2 is sufficient for system B1, since $\rho_{11}^{B2} = 0.1 \leqslant \rho_{10}^{B1} = 0.35 < p_1 = 0.4 < \rho_{11}^{B1} = 0.5 \leqslant \rho_{10}^{B2} = 0.5$. This result can also be verified using (21) and (22), with solution $u = 0$ and $v = 4/9$. [In this case, the assumption under which (35) and (36) are stated must be reversed; that is, $\rho_{11}^{B2} < \rho_{10}^{B2}$.]

b. Implications of forecast accuracy (performance measures)

What can be inferred about the sufficiency relationship between two systems when only the numerical values of a performance measure such as the Brier score are known? In general, the following statement can be made: If $BS^B \leq BS^A$, then system A cannot be sufficient for system B. That is, system A cannot be sufficient for system B if B has a "better" score than A. This result follows from Theorems 9 and 4 of DeGroot and Fienberg (1986). However, it is not possible to determine, solely from an ordinal relationship between the respective Brier scores, whether system B is sufficient for system A or whether the two systems are insufficient for each other. In this regard, note that $BS^A = 0.2187 < BS^{B1} = 0.235$ and, in this case, system A is sufficient for system B. However, $BS^{B2} = 0.21 < BS^A = 0.2187$ and, as shown in section 3, systems A and B2 are insufficient for each other.

Notwithstanding the statements made in the previous paragraph, it is possible to infer sufficiency from performance measures such as the Brier score when these scores satisfy certain additional conditions. In this regard, isopleths of expected half Brier score for well-calibrated forecasting systems have been included in Figs. 1–3. [Note: Knowledge of a complete set of performance characteristics in each framework makes it possible to specify BS; see (39).] Most isopleths of BS pass through both regions in which system B is sufficient for system A (i.e., S') *and* regions in which the two systems are insufficient for each other (I). For forecasting systems that yield such scores, it is not possible to determine, simply from the scores themselves, whether system B is sufficient for system A or whether A and B are insufficient for each other.

However, it can also be seen from these figures that some isopleths of BS are contained entirely within the region in which system B is sufficient for system A (regions denoted by S'). Let BS_c^A denote a critical Brier score, such that it is the largest value of BS for which the corresponding isopleth lies entirely in region S'; note that BS_c^A is a function of system A. Then the following lemma, for which a proof is provided in appendix B, can be stated: *Lemma 2:* Let A and B represent two matched binary-event primitive probabilistic forecasting systems. Without loss of generality, assume that $\rho_{11}^A > \rho_{10}^A$. Then if $BS^B < BS_c^A$ [see (41)], system B is sufficient for system A.

The critical value BS_c^A can be determined by introducing the two critical forecasting systems C1 and C2 defined by the following characteristics in the CR framework: $\rho_{11}^{C1} = 1$, $\rho_{10}^{C1} = \rho_{10}^A$ and $\rho_{11}^{C2} = \rho_{11}^A$, $\rho_{10}^{C2} = 0$. Then BS_c^A is determined by the following expression:

$$BS_c^A = \min(BS^{C1}, BS^{C2}) = \min[p_0\rho_{10}^A, p_1(1 - \rho_{11}^A)]. \quad (41)$$

Note that this critical Brier score necessarily depends on both performance characteristics of system A. For the reference forecasting system A associated with the example in this paper, $BS_c^A = \min(0.1655, 0.1716) = 0.1655$. Since $BS^{B3} = 0.1200 < 0.1655$, it is immediately evident that system B3 is sufficient for system A. If, however, an alternative system A′ with $\rho_{11}^{A'} = 0.9$ and $\rho_{10}^{A'} = \rho_{10}^A = 0.276$ were considered (note that system A′ is constructed by "shifting A to the right"), then only systems with Brier scores less than $BS_c^{A'} = \min(0.1655, 0.04)$ would be sufficient for A′. Obviously, B3 does not satisfy this condition.

The criterion set forth in Lemma 2 partitions the region S' (containing all of the forecasting systems B that are sufficient for A) into a region containing those systems that satisfy (41) and another region containing the remaining systems (which are still sufficient for system A). For the latter systems it is not possible to determine from the scores alone whether or not system B is sufficient for system A. This situation illustrates a serious deficiency in the practice of comparing forecasting systems on the basis of one-dimensional performance measures.

6. Discussion and conclusion

The present paper has investigated the concept of sufficiency and its implications for comparative evaluation of binary-event primitive probabilistic weather forecasting systems from three different but equivalent perspectives. Sufficiency is an important concept in this context because it provides an unambiguous—although partial—preference order on alternative forecasting systems. Specifically, if a reference forecasting system A is sufficient for an alternative forecasting system B, then all users of such forecasts will prefer the former to the latter.

As is evident from sections 4 and 5, comparison of such forecasting systems with reference to sufficiency is greatly facilitated by two-dimensional geometrical displays in which the systems are described in terms of a complete set of performance characteristics. Three types of convex regions constitute this two-dimensional space: (i) regions in which system A is sufficient for system B; (ii) regions in which system B is sufficient for system A; and (iii) regions in which the two systems are insufficient for each other. This geometry of sufficiency is completely determined by the performance characteristics of the forecasting system arbitrarily chosen as the reference system. Thus, sufficiency implies certain ordinal relationships on performance characteristics (quality) and knowledge of complete sets of performance characteristics of forecasting systems provides a basis for determining the presence or absence of sufficiency. The perspective provided by the calibration-refinement framework seems particularly useful here, since sufficiency can be determined immediately from an ordering of the numerical values of the respective (calibration-refinement) performance characteristics (see Lemma 1).

With regard to performance measures such as the Brier score (a strictly proper measure of accuracy), the

fact that system A is sufficient for system B implies that A's forecasts will achieve a better expected score than B's forecasts, under the assumption of well-calibrated forecasting systems. Conversely, however, if only the Brier scores of two forecasting systems are known, then only weak statements can generally be made vis-à-vis sufficiency. However, a result was reported according to which it is possible to infer the sufficiency of system B for system A when the accuracy of the former exceeds some critical value that depends on the performance characteristics of the latter (see Lemma 2).

With regard to extensions of this work, it should be noted that the forecasting systems considered here were restricted to binary-event primitive probabilistic forecasts. It would obviously be desirable to extend these results to situations involving (unrestricted) probabilistic forecasting systems—both well calibrated and not well calibrated. In this regard, if certain modeling assumptions are made regarding the likelihood functions, then it is possible to specify simple criteria under which one probabilistic forecasting system is sufficient for another system of this type (Krzysztofowicz and Long 1987). Nevertheless, this topic warrants further study under more general conditions, and the implications of sufficiency for performance characteristics of alternative systems also should be investigated. It seems desirable as well to explore the extension of these results to situations involving multiple events (as opposed to binary events), a topic that has also been addressed by DeGroot and Fienberg (1986).

The principal practical implications of this work relate to the fact that overall performance measures such as the Brier score do not completely describe forecast quality even for the very simple and restricted types of forecasts considered here. As a result, these measures cannot, in general, be used to determine sufficiency. Forecast quality must be described in terms of a complete set of performance characteristics to provide a basis for establishing a preference order (if it exists) between two or more forecasting systems using the concept of sufficiency. This fact implies that the practice of forecast verification should focus on the use of complete sets of performance characteristics in the context of comparative evaluation, a point recently emphasized as well by Chen et al. (1987) and Murphy and Ehrendorfer (1987).

Finally, according to Lemma 2 set forth in section 5b, it is possible to infer sufficiency from Brier scores alone when the accuracy of the alternative forecasts exceeds some relatively high threshold value. However, if users (or others) base their decisions regarding the choice of forecasting systems on such a criterion, then they might well reject alternative forecasting systems that are in fact sufficient for the reference system. To avoid such undesirable situations, it is necessary to compare forecasting systems in terms of their basic performance characteristics.

Acknowledgments. We acknowledge the valuable comments of Robert T. Clemen, Edward S. Epstein, and an anonymous reviewer on earlier versions of this paper. This research was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grants ATM-8507495 and ATM-8714108.

APPENDIX A

Proof of Lemma 1

Lemma 1. Let A and B be two matched binary-event primitive probabilistic forecasting systems. Without loss of generality, assume that $\rho_{11}^{A} > \rho_{10}^{A}$. Then system A is sufficient for system B if and only if one of the following sets of inequalities holds:

$$\rho_{10}^{A} \leq \rho_{10}^{B} < p_1 < \rho_{11}^{B} \leq \rho_{11}^{A} \tag{A1}$$

or

$$\rho_{10}^{A} \leq \rho_{11}^{B} < p_1 < \rho_{10}^{B} \leq \rho_{11}^{A}. \tag{A2}$$

1) *Proof that, if system A is sufficient for system B, then it follows that (A1) or (A2) holds.* Using the solution for u and v from (21) and (22) and applying the condition that $0 \leq u \leq 1$ (to ensure sufficiency), the following set of inequalities holds (with no loss of generality, it is assumed hereafter that $\rho_{11}^{B} > \rho_{10}^{B}$):

$$p_1(\rho_{10}^{A} - \rho_{11}^{B}) \leq \rho_{10}^{B}(\rho_{10}^{A} - \rho_{11}^{B}) \leq \rho_{10}^{B}(\rho_{10}^{A} - p_1) + \rho_{10}^{A}(p_1 - \rho_{11}^{B}). \tag{A3}$$

Since $\rho_{11}^{B} > \rho_{10}^{B}$, it follows that $\rho_{11}^{B} > p_1 > \rho_{10}^{B}$ and the inequality on the left in (A3) can hold only if $\rho_{10}^{A} \leq \rho_{11}^{B}$. Cancelling common terms in the inequality on the right in (A3) reduces this inequality to

$$\rho_{11}^{B}(\rho_{10}^{A} - \rho_{10}^{B}) \leq p_1(\rho_{10}^{A} - \rho_{10}^{B}). \tag{A4}$$

However, for (A4) to hold, it follows that $\rho_{10}^{A} \leq \rho_{10}^{B}$. These results can be combined into the following set of inequalities:

$$\rho_{10}^{A} \leq \rho_{10}^{B} < p_1 < \rho_{11}^{B}. \tag{A5}$$

Applying similar arguments to the set of inequalities arising from the condition that $0 \leq v \leq 1$ (to ensure sufficiency), it follows that

$$\rho_{10}^{B} < p_1 < \rho_{11}^{B} \leq \rho_{11}^{A}. \tag{A6}$$

Combining (A5) and (A6) establishes (A1). A similar argument can be used to establish (A2), under the assumption that $\rho_{11}^{B} < \rho_{10}^{B}$.

2) *Proof that, if (A1) holds, then system A is sufficient for system B.* [A completely analogous proof, assuming (A2) holds, will be omitted to conserve space.] To prove that system A is sufficient for system B, it suffices to show that the values of u and v [as solutions to (21) and (22)] both lie in the closed unit interval. (i) From (A1), it can be seen that

$$c(p_1 - \rho_{10}^{B}) \leq c(\rho_{11}^{B} - \rho_{10}^{B}) \tag{A7}$$

for any non-negative number c. Adding the positive term $(p_1 - \rho_{10}^B)(\rho_{11}^B - \rho_{10}^B)$ to both sides of (A7) and rearranging terms yields

$$(p_1 - \rho_{10}^B)(\rho_{11}^B - \rho_{10}^B + c) \leqslant (\rho_{11}^B - \rho_{10}^B)(c + p_1 - \rho_{10}^B). \quad \text{(A8)}$$

Setting $c = \rho_{10}^B - \rho_{10}^A \geqslant 0$, (A8) becomes

$$(p_1 - \rho_{10}^B)(\rho_{11}^B - \rho_{10}^A) \leqslant (\rho_{11}^B - \rho_{10}^B)(p_1 - \rho_{10}^A). \quad \text{(A9)}$$

Dividing both sides of (A9) by the positive quantity $(\rho_{11}^B - \rho_{10}^B)(p_1 - \rho_{10}^A)$ and making use of (A1) establishes that $0 \leqslant u \leqslant 1$, since $u = [(p_1 - \rho_{10}^B)(\rho_{11}^B - \rho_{10}^A)]/[(\rho_{11}^B - \rho_{10}^B)(p_1 - \rho_{10}^A)]$ [from (21) and (22)]. (ii) From (A1), it can be seen that

$$\rho_{11}^B > p_1 - d[1 - (p_1 - \rho_{10}^B)/(\rho_{11}^B - \rho_{10}^B)] \quad \text{(A10)}$$

for any non-negative number d (the expression in square brackets is greater than zero). From (A10), it follows that

$$(\rho_{11}^B - p_1 + d) > d[(p_1 - \rho_{10}^B)/(\rho_{11}^B - \rho_{10}^B)]. \quad \text{(A11)}$$

Setting $d = \rho_{11}^A - \rho_{11}^B \geqslant 0$, (A11) becomes

$$(\rho_{11}^A - p_1) > [(\rho_{11}^A - \rho_{11}^B)/(\rho_{11}^B - \rho_{10}^B)](p_1 - \rho_{10}^B). \quad \text{(A12)}$$

It can be immediately seen from (A12) that $0 \leqslant v \leqslant 1$, since $v = [(p_1 - \rho_{10}^B)(\rho_{11}^B - \rho_{11}^A)]/[(\rho_{11}^B - \rho_{10}^B)(p_1 - \rho_{11}^A)]$ [see (21) and (22)].

APPENDIX B

Proof of Lemma 2

Lemma 2. Let A and B represent two matched binary-event primitive probabilistic forecasting systems. Without loss of generality, assume that $\rho_{11}^A > \rho_{10}^A$. Now if $BS^B < BS_c{}^A$, then system B is sufficient for system A, where

$$BS_c{}^A = \min(BS^{C1}, BS^{C2}) = \min[p_0\rho_{10}^A, p_1(1 - \rho_{11}^A)]. \quad \text{(B1)}$$

Proof: $BS_c{}^A$ can be expressed by the Brier scores of two critical systems C1 and C2 with $\rho_{11}^{C1} = 1$, $\rho_{10}^{C1} = \rho_{10}^A$ and $\rho_{11}^{C2} = \rho_{11}^A$, $\rho_{10}^{C2} = 0$. From Lemma 1 it is clear that both C1 and C2 are sufficient for A. Consider now a system B satisfying $BS^B < BS_c{}^A$; that is, $BS^B < BS^{C1}$ *and* $BS^B < BS^{C2}$. From these ordinal relationships, as stated in section 5b, neither C1 nor C2 can be sufficient for B as a consequence of Theorems 9 and 4 in DeGroot and Fienberg (1986). That is, B is either sufficient for C1 or B and C1 are insufficient for each other *and* B is either sufficient for C2 or B and C2 are insufficient for each other. This statement reveals the existence of four distinct cases, each of which will be shown to imply that B is sufficient for A. It is now assumed, with no loss of generality, that $\rho_{11}^B > \rho_{10}^B$; otherwise, we would have to consider four other, completely symmetric statements.

1) B is sufficient for C1 *and* B is sufficient for C2. These conditions imply, from Lemma 1, that $\rho_{11}^B = 1$ and $\rho_{10}^B = 0$. The latter represent perfect forecasts, obviously indicating that system B is sufficient for system A.

2) B is sufficient for C1 *and* B and C2 are insufficient for each other. These conditions are only possible (from Lemma 1) if $\rho_{11}^B = 1$ and $0 < \rho_{10}^B \leqslant \rho_{10}^A$, from which it is evident that system B is sufficient for system A.

3) B and C1 are insufficient for each other *and* B is sufficient for C2. These conditions are only possible (from Lemma 1) if $\rho_{10}^B = 0$ and $\rho_{11}^A \leqslant \rho_{11}^B < 1$, from which it is evident that system B is sufficient for system A.

4) B and C1 are insufficient for each other *and* B and C2 are insufficient for each other. These conditions are possible (according to Lemma 1) only with the following ordering on the relevant conditional probabilities: $0 = \rho_{10}^{C2} \leqslant \rho_{10}^B \leqslant \rho_{10}^{C1} = \rho_{10}^A < \rho_{11}^{C2} = \rho_{11}^A \leqslant \rho_{11}^B \leqslant \rho_{11}^{C1} = 1$. However, Lemma 1 implies that, with this ordering, system B is sufficient for system A, which completes the proof and establishes Lemma 2.

REFERENCES

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

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

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

Chen, Y.-S., M. Ehrendorfer and A. H. Murphy, 1987: On the relationship between the quality and value of forecasts in the generalized cost-loss ratio situation. *Mon. Wea. Rev.,* **115,** 1534–1541.

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,** 12–22.

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

Krzysztofowicz, R., and D. Long, 1987: To protect or not to protect: Bayes decisions with forecasts. Department of Systems Engineering, University of Virginia, 39 pp.

Murphy, A. H., 1986: Comparative evaluation of categorical and probabilistic forecasts: Two alternatives to the traditional approach. *Mon. Wea. Rev.,* **114,** 245–249.

——, and M. Ehrendorfer, 1987: On the relationship between the accuracy and value of forecasts in the cost–loss ratio situation. *Wea. Forecasting,* **2,** 243–251.

——, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Winkler, R. L., and A. H. Murphy, 1968: "Good" probability assessors. *J. Appl. Meteor.,* **7,** 751–758.