Evaluation of Prototypical Climate Forecasts: The Sufficiency Relation

MARTIN EHRENDOERFER
Institute for Meteorology and Geophysics, University of Vienna, Austria

ALLAN H. MURPHY*
UCAR Visiting Scientist Program, National Meteorological Center, National Weather Service, NOAA, Washington, D.C.

(Manuscript received 27 September 1990, in final form 4 November 1991)

ABSTRACT

The sufficiency relation, originally developed in the context of the comparison of statistical experiments, provides a sound basis for the comparative evaluation of forecasting systems. The importance of this relation resides in the fact that if forecasting system A can be shown to be sufficient for forecasting system B, then all users will find A's forecasts of greater value than B's forecasts regardless of their individual payoff structures.

In this paper the sufficiency relation is applied to the problem of comparative evaluation of prototypical climate forecasting systems. The primary objectives here are to assess the basic applicability of the sufficiency relation in this context and to investigate the implications of this approach for the relationships among the performance characteristics of such forecasting systems.

The results confirm that forecasting system A is sufficient for forecasting system B when the former uses more extreme probabilities more frequently than the latter. Further, in terms of the relatively simple forecasting systems considered here, it is found that system A may be sufficient for system B even if the former uses extreme forecasts less frequently, provided that A's forecasts are—to a certain degree—more extreme than B's forecasts. Conversely, system A cannot be shown to be sufficient for system B if the former uses less extreme forecasts more frequently than the latter. The advantages of the sufficiency relation over traditional performance measures in this context are also demonstrated.

Several issues related to the general applicability of the sufficiency relation to the comparative evaluation of climate forecasts are discussed. Possible extensions of this work, as well as some implications of the results for verification procedures and practices in this context, are briefly described.

1. Introduction

The sufficiency relation originated in the fundamental work of Blackwell (1953) on the comparison of statistical experiments. It was introduced into the forecasting literature by DeGroot and Fienberg (1982), who explored the relationship between the sufficiency concept and the concepts of calibration and refinement in the context of probabilistic forecasting. The sufficiency relation is of particular interest in comparative evaluation of forecasting systems (or forecasters), since it embodies conditions under which one system can be unambiguously judged to be of higher quality and greater value than another system.

Ehrendorfer and Murphy (1988) introduced the sufficiency relation into the meteorological literature.

* Permanent affiliation: Departments of Atmospheric Sciences and Statistics, Oregon State University, Corvallis, Oregon.

Corresponding author address: Dr. Martin Ehrendorfer, Institute for Meteorology and Geophysics, University of Vienna, Hohe Warte 38, A-1190 Vienna, Austria.

© 1992 American Meteorological Society
between these prototypical forecasting systems is discussed in section 4, and the results of a series of numerical experiments involving the comparative evaluation of such systems are presented in section 5. The discussion in section 6 emphasizes the need for future studies that consider more general climate forecasting systems and that investigate other issues related to the applicability of the sufficiency relation in this context. Section 7 contains a brief summary of the results and their implications as well as some concluding remarks.

2. The sufficiency relation

Let \( T = \{ f_1, \ldots, f_r \} \) denote a finite set of \( r \) vectors that represents the union of all distinct forecasts used by two forecasting systems, say \( A \) and \( B \). Each \( f_i \) (\( i = 1, \ldots, r \)) represents a unique set of coherent forecast probabilities defined on a set of \( s \) mutually exclusive and collectively exhaustive events \( \theta_j \) (\( j = 1, \ldots, s \)). Let \( p(f_i | \theta_j) \) denote the conditional probability of forecast \( f_i \) given event \( \theta_j \). Then system \( A \) is sufficient for system \( B \) if a stochastic transformation \( h(f_k | f_i) \) exists such that

\[
\sum_{i=1}^{r} h(f_k | f_i) p^B(f_i | \theta_j) = p^A(f_k | \theta_j),
\]

\((k = 1, \ldots, r; j = 1, \ldots, s) \tag{1}\)

(see also DeGroot and Fienberg 1986). The function \( h_{ki} = h(f_k | f_i) \) qualifies as a stochastic transformation if \( 0 \leq h_{ki} \leq 1 \) for all \( i \) and \( k \) and \( \sum_{k=1}^{r} h_{ki} = 1 \) for each \( i \). Note that, under the assumption that the marginal probabilities \( p_i \) of the events \( \theta_j \) are known, the likelihoods \( p_i = p(f_i | \theta_j) \) represent a complete characterization of the quality of the forecasts (Murphy and Winkler 1987). In a meteorological context, these marginal probabilities are the sample climatological probabilities.

Since the sufficiency relation defined by (1) indicates that \( B \)'s likelihoods can be obtained by a stochastic transformation—or randomization—of \( A \)'s likelihoods, the former obviously contains greater uncertainty than the latter. As a result, the quality of system \( A \)'s forecasts is clearly superior to the quality of system \( B \)'s forecasts. The importance of this relation resides in the fact that if it can be shown that system \( A \) is sufficient for system \( B \), then \( A \)'s forecasts are of greater value than \( B \)'s forecasts to all users regardless of their individual payoff structures. However, due to this strong implication on the relationship between two forecasting systems, it may not be possible in all cases to show that \( A \) is sufficient for \( B \), or vice versa. In such cases, \( A \) and \( B \) are said to be insufficient for each other (Ehrendorfer and Murphy 1988); thus, the sufficiency relation is a quasi order (e.g., Krzysztofowicz and Long 1991a). The range of conditions under which forecasting systems are insufficient for each other is an important issue in applications of the sufficiency relation.

3. Climate forecasting systems: Description and characterization

We are interested here in prototypical forecasting systems that produce forecasts similar in format to operational climate forecasts. Specifically, it is assumed that the situations of concern involve three possible events (i.e., \( s = 3 \)) representing below-normal (\( \theta_1 \)), near-normal (\( \theta_2 \)), and above-normal (\( \theta_3 \)) conditions (e.g., temperatures or precipitation amounts). Each forecast vector in the set \( T \) consists of a coherent set of probabilities assigned to these three events. Moreover, we restrict our attention in this paper to forecasting systems that use the same three distinct forecast vectors; that is, it is assumed that \( r = 3 \) and \( T = \{ f_1, f_2, f_3 \} \). Finally, to be consistent with the monthly and seasonal forecasts issued by the NWS, it is assumed that the events are defined in such a way that their climatological probabilities are \( p = (p_1, p_2, p_3) = (0.3, 0.4, 0.3) \).

The dimensionality of forecasting systems can be defined as the number of independent parameters required to reconstruct the joint distribution of forecasts and observations (Murphy 1991). Since the climatological probabilities are known, specification of certain combinations of six of the nine conditional probabilities \( p_{ij} \) —for example, \( p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, \) and \( p_{23} \)—uniquely characterizes a forecasting system in this context (i.e., the system possesses six degrees of freedom). For the purposes of this paper, several additional assumptions are made to reduce the dimensionality of the systems of interest and to ensure that the forecasts possess characteristics similar to short-range climate forecasts.

Although the sufficiency relation can be described in a straightforward manner in terms of the \( p_i \) [see Eq. (1)], the characterization of the forecasting systems under consideration here is more convenient in terms of the \( \rho_{ji} = p(\theta_j | f_i) \) and the \( \pi_i = p(f_i) \), where \( \rho_{ji} \) is the conditional probability of event \( \theta_j \) given forecast \( f_i \), and \( \pi_i \) is the predictive probability of forecast \( f_i \). The relationship between these probabilities (and the climatological probabilities \( p_i \)) is embodied in Bayes' theorem:

\[
\rho_{ji} = \frac{p_{ij}\pi_j}{\pi_i}. \tag{2}\]

Under the assumptions made at the beginning of this section, specification of the six conditional probabilities \( p_{ij} \) \((i = 1, 2; j = 1, 2, 3)\) can be shown to be exactly equivalent to specification of the six independent quantities \( p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, \) and \( \pi_1 \). Given the latter, consistency requirements can be used to determine the remaining parameters—namely, \( p_{23}, p_{31}, p_{32}, p_{33}, \pi_2, \) and \( \pi_3 \). The six independent probabilities—as well as the dependent probabilities—are defined here in terms of three independent parameters, \( \delta, \pi, \) and \( \alpha \), as described in Table 1. Thus, for the prototypical fore-
casting systems considered in this paper, the conditional probabilities $p_{ij}$ are determined from (2), with $p = (p_1, p_2, p_3) = (0.3, 0.4, 0.3)$ and the $p_{ij}$ and $\pi_i$ given in terms of $\delta$, $\pi$, and $\alpha$. As a result, the dimensionality of the forecasting systems of interest is reduced from six to three.

Examination of the expressions in Table 1 reveals that the parameter $\delta$ represents the departure of the probability $p_{11}$ (and $p_{13}$) from its climatological value of 0.3. The parameter $\alpha$ characterizes the asymmetry of the forecasts. When $\alpha = 0$, the forecasts are symmetric in the sense that $p_{11} = p_{33}$ and $p_{13} = p_{31}$. Otherwise (i.e., $\alpha \neq 0$) the forecasts are asymmetric. The parameters $\delta$ and $\alpha$ may be chosen arbitrarily subject only to the condition $-0.3 \leq \delta (1 + \alpha) \leq 0.3$. Finally, in light of the parameterization of the $p_{ij}$ in Table 1, consistency allows only one of the $\pi_i$'s to be chosen arbitrarily. Here we specify that $\pi = (\pi_1, \pi_2, \pi_3) = (\pi, 1 - 2\pi, \pi)$, subject to the condition that $0 \leq \pi \leq 0.5$. The parameter $\pi$ represents the predictive probability—or relative frequency of use—of the forecast $f_1$ (and $f_3$).

Specification of the parameters $\delta$, $\alpha$, and $\pi$ completely determines the characteristics of the forecasting systems of interest in this study. This representation is intuitively appealing in the case of well-calibrated forecasts, since the $p_{ij}$ then describe the actual appearance of the forecast vectors. In this case, for example, $f_1 = (0.3 - \delta, 0.4 - \alpha \delta, 0.3 + \delta + \alpha \delta)$ (see Table 1). When $\delta$ and $\alpha$ are both positive (or $\delta > 0$ and $\alpha = 0$), $f_1(f_3)$ can be interpreted as a forecast of above-normal (below-normal) conditions. Obviously, the forecast $f_2$ is a climatological forecast, since the probabilities associated with this forecast are identical to the climatological probabilities.

Under the assumption that the prototypical forecasting systems considered here are well calibrated, the accuracy of these forecasts can be assessed appropriately by the expected ranked probability score (ERPS) (Epstein 1969; Murphy 1971). The definition of the ERPS for the forecasting systems considered in this study and its calculation from $\delta$, $\pi$, and $\alpha$ is described in appendix A.

In the remainder of this paper, attention is focused on the question of primary concern here, namely, the ranges of values of the parameters $\delta$, $\pi$, and $\alpha$ for which alternative forecasting systems $B$ are sufficient for a reference forecasting system $A$ with fixed values of these parameters. In the next section, we describe the way in which the sufficiency relation, as defined in (1), is used to answer this question. Before closing the description of the forecasting systems considered, it must be emphasized that the aforementioned assumption of perfect calibration is necessary only in the context of the interpretation of the forecast vectors and the ERPS, and not for the investigation of the sufficiency relation (see also section 5).

4. Use of sufficiency relation

In order to use the sufficiency relation defined in (1) to compare two forecasting systems $A$ and $B$, it is necessary to determine whether or not a function $h_{ij}$ exists that qualifies as a stochastic transformation. When $r = s = 3$ (three events and three distinct forecasts), the sufficiency relation can be rewritten in terms of the following matrix equation:

$$Sh = c,$$

where the $9 \times 9$ matrix $S$ is

$$S = \begin{pmatrix} P^d & O & O \\ O & P^d & O \\ O & O & P^d \end{pmatrix},$$

in which $O$ is the $3 \times 3$ zero matrix and the $3 \times 3$ matrix $P^d$ is defined as

$$P^d = \begin{pmatrix} p_{11}^d & p_{21}^d & 1 - p_{11}^d - p_{21}^d \\ p_{12}^d & 1 - p_{12}^d - p_{22}^d \\ p_{13}^d & 1 - p_{13}^d - p_{23}^d \end{pmatrix}. $$

The vectors $h$ and $c$ in (3) are defined as follows:

$$h = (h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33})',$$

and

$$c = (p_{11}^p, p_{12}^p, p_{13}^p, p_{21}^p, p_{22}^p, p_{23}^p, 1 - p_{11}^p - p_{21}^p, 1 - p_{12}^p - p_{22}^p, 1 - p_{13}^p - p_{23}^p)' ,$$

where a prime (') denotes transposition. In addition, the requirement that $\sum_{k=1}^{2} h_{ki} = 1$ for each $i$ (see section 2) is expressed as

$$(1, 1, 1)^t h = (1, 1, 1)^t,$$
in which I represents the $3 \times 3$ identity matrix. The equations in (3) and (8) constitute a linear system of 12 equations to be solved for the nine components of the vector $h$. Since three of these equations can be shown to be linearly related to the other nine equations, they can be eliminated from consideration. The equations eliminated here are the last three equations of (3). [For example, the last of the equations in (3) can be constructed by subtracting the third and sixth equation of (3) from the sum of $p^{1}$ times the first equation of (8) plus $p^{2}$ times the second equation of (8) plus $(1 - p^{1} - p^{2})$ times the third equation of (8).] The resulting system of equations to be solved for the vector $h$ can then be written as follows:

$$Qh = b,$$

where the $9 \times 9$ matrix $Q$ is now defined as

$$Q = \begin{pmatrix} P' & O & O \\ O & P' & O \\ I & I & I \end{pmatrix},$$

and the vector $b$ is given by

$$b = (p^{1}_{1} \quad p^{1}_{2} \quad p^{1}_{3} \quad p^{2}_{1} \quad p^{2}_{2} \quad p^{2}_{3} \quad I \quad I \quad 1)' .$$

By means of (9), the sufficiency relation defined in (1) has been rewritten as a linear algebraic system of equations in nine unknowns (i.e., the $h_{k}$) for the specific case in which $r = s = 3$. Given the conditional probabilities $p_{ij}$ that define the performance of two competing forecasting systems and that determine the components of the matrix $Q$ and the vector $b$, this system can be solved for the components of the vector $h$. The only remaining condition that must be satisfied by the components of $h$ is that they should be nonnegative. If these components are nonnegative, then system $A$ is sufficient for system $B$.

The parameterization of the forecasting systems of interest here (namely, expressing the $p_{ij}$ in terms of $\delta$, $\pi$, and $\alpha$; see section 3) leads to a singular matrix $Q$, which in turn means that Eq. (9) cannot be solved unambiguously for $h$. To use (9) to determine whether or not system $A$ is sufficient for system $B$, we employ an algorithm that has been found to be rather insensitive to this ambiguity. This algorithm is described in appendix B.

5. Numerical experiments

The sufficiency relation in the form developed in section 4 is used in a series of numerical experiments to answer the question of primary interest in this study; namely, to determine the range of values of the parameters $\delta$, $\pi$, and $\alpha$ for which alternative systems $B$ are sufficient for a reference system $A$ with fixed values of these parameters.

a. Experimental design

A list of the experiments performed and the associated parameter values is presented in Table 2. Once the reference system $A$ is specified in terms of the parameters $\delta^{A}$, $\pi^{A}$, and $\alpha^{A}$, as indicated in Table 2 (columns 2–4), the sufficiency relation between system $A$ and all possible alternative systems $B$ in ($\delta$, $\pi$, $\alpha$) space is investigated.

Two dimensions of this parameter space are completely covered by scanning over a two-dimensional grid of feasible values of $\delta^{B}$ and $\pi^{B}$ of the alternative systems using a given step size (always taken here as 0.01). The third dimension of the parameter space is eliminated by assigning to the value of the parameter $\alpha^{B}$ the same value as that specified for the reference system $A$. This assumption reduces the cost of the search and yet allows consideration of asymmetric forecasts ($\alpha \neq 0$). However, due to this simplification the direct influence of the parameter $\alpha$ on the sufficiency relation between two systems cannot be assessed. In addition, this specification implies that the range of possible values of $\delta^{B}$ is limited by the choice of $\alpha^{A}$ (see section 3 and column 5 of Table 2).

For each combination of the reference system $A$ and a particular alternative system $B$, the algorithm described in appendix B is used to determine whether $A$ is sufficient for $B$ or not. Interchanging the roles of systems $A$ and $B$ and once again applying this algorithm permits identification of the alternative systems $B$ that are sufficient for the reference system $A$. The results obtained by this process for a particular experiment can be displayed in a two-dimensional sufficiency diagram (note that $\alpha$ is kept fixed).

In all experiments except E5, the reference system $A$ is defined by $\delta^{A} = 0.1$ and $\pi^{A} = 0.15$. The parameter $\alpha^{A}$, on the other hand, takes on different values in different experiments. Thus, assuming well-calibrated forecasts (see section 3), the reference system $A$ of experiment E1 ($\alpha^{A} = 0$) is characterized by the forecast vectors $f_{1}^{A} = (0.4, 0.4, 0.2)$, $f_{2}^{A} = (0.3, 0.4, 0.3)$, and $f_{3}^{A} = (0.2, 0.4, 0.4)$, with the predictive probabilities $\pi^{A} = (0.15, 0.70, 0.15)$. The ERPS (see appendix A) of this reference system is 0.414 (see Table 2). Assum-

---

**Table 2.** List of experiments. This table displays for each experiment the values $\delta^{A}$, $\pi^{A}$, and $\alpha^{A}$ characterizing the performance of the reference system $A$. The permissible range of values of $\delta^{A}$ for the alternative systems $B$ is restricted by the bound indicated in column 5 of the table. (Note that we take $\alpha^{B} = \alpha^{A}$, and that $\pi^{A}$ is unrestricted; see sections 3 and 5.) The last column of the table gives the ERPS of the reference system $A$.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>$\delta^{A}$</th>
<th>$\pi^{A}$</th>
<th>$\alpha^{A}$</th>
<th>$\delta^{B}$-bound</th>
<th>ERPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>-0.1</td>
<td>0.15</td>
<td>0</td>
<td>0.3</td>
<td>0.41400</td>
</tr>
<tr>
<td>E2</td>
<td>-0.1</td>
<td>0.15</td>
<td>1</td>
<td>0.15</td>
<td>0.40500</td>
</tr>
<tr>
<td>E3</td>
<td>-0.1</td>
<td>0.15</td>
<td>0.5</td>
<td>0.20</td>
<td>0.41025</td>
</tr>
<tr>
<td>E4</td>
<td>-0.1</td>
<td>0.15</td>
<td>2</td>
<td>0.10</td>
<td>0.39000</td>
</tr>
<tr>
<td>E5</td>
<td>-0.2</td>
<td>0.30</td>
<td>0</td>
<td>0.30</td>
<td>0.37200</td>
</tr>
</tbody>
</table>
ing that $\alpha^A \neq 0$ in specifying the reference systems of experiments E2–E4, while maintaining the same values for $\delta^A$ and $\pi^A$, leads to slightly more accurate reference forecasting systems as revealed by the numerical values of the ERPS for these systems. Clearly, the reference system chosen for experiment E5 is more accurate than the one used in experiment E1 (see Table 2).

b. Experimental results

The results of experiment E1 (base case) are displayed in Fig. 1. In this experiment $\alpha^A (=\alpha^B) = 0$; from the considerations set forth in section 3 a range of possible values of $\delta^B$ from $-0.3$ to $0.3$ is implied (see Table 2). In Fig. 1 the diamonds identify the systems $B$ for which system $A$ has been found to be sufficient (this region is denoted in the figure and in the following discussion by $S$) and the crosses identify the systems $B$ that are sufficient for system $A$ (denoted by $S'$). Regions in this diagram without diamonds or crosses represent combinations of values of $\delta$ and $\pi$ for which systems $A$ and $B$ are insufficient for each other (denoted by $I$). The reference system $A$ is denoted by a large dot, and its parameter values ($\delta^A = -0.1$, $\pi^A = 0.15$, $\alpha^A = 0$) define the geometry of the sufficiency diagram. Recall that since $\alpha^A = \alpha^B = 0$, the forecasts are symmetric in this case.

Also included in Fig. 1 are isopleths of the ERPS; these isopleths can be determined from $\delta$, $\pi$, and $\alpha$, assuming that the forecasting systems under consideration are well calibrated (see appendix A). It is important to note that the assumption of perfect calibration is required only in the context of proper interpretation of the ERPS and that it is not necessary to assume perfect calibration in the context of the sufficiency relation (see also section 3). The ERPS is chosen here as an appropriate one-dimensional measure of forecast accuracy in order to be able to illustrate the relationship between such scores and the sufficiency relation. The isopleths shown in Fig. 1 are drawn at unequal intervals (refer to the figure legend for the respective numerical values) with increasing values of the ERPS toward the middle of the figure indicating less accurate forecasts.

Examination of the sufficiency regions in Fig. 1 re-

![Fig. 1. Sufficiency diagram for experiment E1 (for parameter values see Table 2). The reference system is denoted by a large dot. Diamonds identify alternative systems for which the reference system is sufficient (region denoted by $S$), whereas crosses denote alternative systems that are sufficient for the reference system (regions denoted by $S'$). Regions without diamonds and crosses denote parameter combinations for which the reference system and alternative system are insufficient for each other (regions denoted by $I$). Lines included in the diagram are isopleths of the ERPS. These isopleths are drawn at unequal intervals for the numerical values 0.25, 0.30, 0.35, 0.372, 0.39, 0.405, 0.41025, 0.412, 0.414, 0.415, 0.416, 0.417, 0.418, 0.419, 0.4196, 0.4199, with increasing values from the lower- and upper-right corners toward the middle of the diagram (note the symmetry around the horizontal line $\delta = 0$). The bold isopleth denotes the ERPS of the reference system, namely, 0.414. For additional details, see section 5.](image-url)
reveals that system \( A \) is sufficient for all systems \( B \) with (absolutely) smaller values of \( \delta \) and \( \pi \) and for some systems \( B \) with larger values of \( \pi \) given that they possess smaller values of \( \delta \). However, system \( A \) is not sufficient for any systems \( B \) with larger values of \( \delta \). Conversely, systems \( B \) with larger values of \( \delta \) and \( \pi \) are sufficient for system \( A \), as are some systems \( B \) with smaller values of \( \pi \) given that they possess larger values of \( \delta \). Analogously, no systems \( B \) with smaller values of \( \delta \) are sufficient for \( A \). These results can be summarized by indicating that the boundaries in the sufficiency diagram between the various regions (\( S, S', \) and \( I \)) consist of one horizontal line and one nonlinear curve. (In general, two nonlinear curves might be expected.) Both boundaries may be interpreted as parametric curves with \( \delta = \delta(\pi) \). Whereas the first boundary assumes a rather special form, namely, \( \delta = \pm 0.1 \), the second convex boundary is characterized by a complicated nonlinear relationship between \( \delta \) and \( \pi \).

The shape of the boundaries of the regions in the sufficiency diagram indicates that a smaller value of \( \pi \) for the alternative system can be compensated by a larger (absolute) value of \( \delta \) such that the alternative system is sufficient for the reference system. However, the results in Fig. 1 show that the converse is not possible; a smaller (absolute) value of \( \delta \) for the alternative system cannot be compensated by a larger value of \( \pi \).

In other words, with respect to the sufficiency of \( A \) for \( B \) (or \( B \) for \( A \)), a trade-off exists in terms of the values of \( \pi \) but not in terms of the values of \( \delta \).

The ERPS isopleths included in Fig. 1 provide an opportunity to discuss in some detail the relationship between the sufficiency relation and a one-dimensional performance measure when both are applied in the context of comparative evaluation. It is first noted that the ERPS isopleths are convex curves; that is, accuracy is increasing, as expected, for larger (absolute) values of \( \delta \) and \( \pi \). Second, as is evident from the fact that the ERPS isopleths are drawn at unequal intervals, the accuracy as measured by the ERPS can be increased much more by improving the performance characteristics of a rather accurate system than by improving by the same amount the performance characteristics of a rather inaccurate system. In other words, the ERPS "terrain" is rather flat in the vicinity of low-accuracy systems and rather steep in the vicinity of high-accuracy systems. Third, it is remarkable that in the case of the ERPS a trade-off exists in terms of both \( \delta \) and \( \pi \) in the sense described above; namely, a smaller (absolute) value of \( \delta \) for an alternative system can be compensated by a larger value of \( \pi \) to obtain an alternative system that is at least as accurate as the reference system (see the preceding paragraph).

The most important point that becomes evident when the sufficiency regions in Fig. 1 are contrasted with the ERPS isopleths is that these isopleths may cross both regions \( S' \) and \( I \) (or regions \( S \) and \( I \), respectively). This fact illustrates that it is generally impossible to ascertain from the value of the ERPS alone whether an alternative system \( B \) is sufficient for a reference system \( A \) or whether these systems are insufficient for each other (whether reference system \( A \) is sufficient for an alternative system \( B \) or whether these systems are insufficient for each other, respectively), unless the numerical values of the ERPS for the two systems differ by a certain (large) amount. In other words, an alternative system with a better score can be sufficient for the reference system, or both systems can be insufficient for each other.

This major deficiency common to all one-dimensional performance measures relates directly to the fact that by necessity these scores are unable to capture fully the multidimensional character of forecast quality, whose consideration is generally required for a truly comprehensive approach to comparative evaluation (see Murphy 1991). However, it is also clear from Fig. 1 that due to the convexity of both the boundaries of the sufficiency regions and the ERPS isopleths an alternative system with a score worse than that of the reference system can never be sufficient for the reference system. It is surmised that this property holds only for so-called proper scoring rules, such as the ERPS (e.g., Murphy and Daan 1985).

In Figs. 2–4 we present the results of three experiments (E2–E4) involving a reference system \( A \), with \( \alpha = -0.1, \pi = 0.15, \) and \( \alpha' \) assuming different values. When comparing these results it is important to keep in mind that different values of \( \alpha' \) lead to different reference systems. Moreover, \( \alpha' \) affects the range of \( \delta' \) given in column 5 of Table 2 (see section 3 and recall that we investigate only alternative systems \( B \) with \( \alpha'' = \alpha' \)). Comparison of the results of E2 (Fig. 2, \( \alpha = 1 \)) and E1 (Fig. 1, \( \alpha = 0 \)) reveals that inclusion of asymmetric forecasts has reduced the size of region \( S' \) and shifted the ERPS isopleths to the left since \( \alpha \neq 0 \) represents more accurate systems; otherwise, the results of the two experiments are identical.

Figures 3 and 4 present the results of E3 and E4 with \( \alpha = 0.5 \) and \( \alpha = 2 \), respectively. Apart from a change in the size of region \( S' \) and slight shifts of the ERPS isopleths, the results are identical with those of E1. However, it can be seen that smaller (larger) values of the parameter \( \alpha' \) are associated with smaller (larger) reductions in the size of the region \( S' \). In the case of E4, the restrictions on the values of \( \delta' (0.1 < \delta' \leq 0.1) \) are such that no systems \( B \) exist that are sufficient for system \( A \). Overall, the results of experiments E2–E4 indicate that the introduction of the asymmetry parameter \( \alpha \) does not change the general appearance of the sufficiency diagram (cf. Figs. 2–4 and Fig. 1). In particular, even when \( \alpha \neq 0 \) the boundaries between the sufficiency regions still consist of one horizontal line and one nonlinear curve.

Finally, Fig. 5 displays the sufficiency diagram for a situation (experiment E5) in which the reference system is defined by the parameters \( \delta = -0.2, \pi = 0.3, \)
Fig. 2. Same as Fig. 1 except for experiment E2. The same numerical values of the isopleths as in Fig. 1 are used, except that values start with 0.35. The bold isopleth denotes the ERPS of the reference system, namely, 0.405.

Fig. 3. Same as Fig. 1 except for experiment E3. The same numerical values of the isopleths as in Fig. 1 are used, except that values start with 0.30. The bold isopleth denotes the ERPS of the reference system, namely, 0.4125.
FIG. 4. Same as Fig. 1 except for experiment E4. The same numerical values of the isopleths as in Fig. 1 are used, except that values start with 0.35. The bold isopleth denotes the ERPS of the reference system, namely, 0.39.

FIG. 5. Same as Fig. 1 except for experiment E5. The bold isopleth denotes the ERPS of the reference system, namely, 0.372.
and $\alpha^d = 0$ with $\text{ERPS} = 0.372$ (symmetric forecasts). Since the reference system in E5 has larger values of both $\delta^d$ and $\pi^d$ than the reference system in E1 ($\delta^d = -0.1$, $\pi^d = 0.15$), the former is sufficient for (and unambiguously superior to) the latter (see Fig. 5). Comparison of Figs. 1 and 5 reveals that the region $S$ in E5 is substantially larger than the region $S$ in E1, with the opposite relationship holding for the region $S'$. It is expected and confirmed by this experiment that the general appearance of the sufficiency diagram is the same in both E1 and E5, with its geometry being determined by the performance characteristics of the reference system.

The remarks made with regard to the ERPS in the discussion of Fig. 1 remain valid for the experimental results shown in Figs. 2–5. Most importantly, ERPS isoline levels crossing both the regions $I$ and $S$ (or $S'$) always exist. This result clearly illustrates the inadequacy of one-dimensional performance measures in the context of comparative evaluation.

6. Discussion: General applicability of sufficiency relation

The applicability of the sufficiency relation in the context of comparative evaluation has been demonstrated for the case of prototypical climate forecasting systems by the experimental results described in section 5. In particular, these results indicate that the $(\delta, \pi)$ space—the space defined by the basic parameters of these forecasting systems—can be subdivided into three regions, with the size and shape of these regions being determined primarily by the performance characteristics of the reference system. Since the sufficiency regions are defined in terms of (all of) the relevant dimensions of forecast quality, they represent an appropriate framework within which to compare alternative climate forecasting systems. As indicated in section 5, several features of the results—especially those features related to the shape of the sufficiency regions and the relationship between these regions and one-dimensional performance measures—are of considerable interest in the context of comparative evaluation. The important features of these results, as they relate to the prototypical climate forecasting systems considered in this paper, are summarized in section 7.

It seems desirable here to address several issues related to the applicability of the sufficiency relation to the more general climate forecasting systems encountered in the real world. Recall that the prototypical systems considered here involved restrictions of several types. The major restriction related to the number of distinct forecast vectors, each of which specified the likelihood of occurrence of three events. Only three distinct forecasts were allowed to be made by the prototypical forecasting systems considered in this paper, whereas the operational short-range climate forecasts issued by the U.S. National Weather Service may involve three or four times as many distinct forecasts. Consideration of a larger number of distinct forecasts does not present a difficulty in principle with regard to the applicability of the sufficiency relation, but it may lead to formidable practical problems in terms of determining the existence of a stochastic transformation or in terms of interpreting the results. (These practical problems in some sense already arise in the context of the rather simple forecasting systems considered here; see section 4 and appendix B.) Similar comments would apply if consideration were given to increasing the number of events from three to five (or more).

Situations involving a larger number of distinct forecasts or a larger number of events, or both, entail problems of dimensionality (e.g., Murphy 1991). Specifically, the increase in dimensionality resulting from the consideration of more complicated forecasting systems may make it difficult to construct, and interpret, sufficiency diagrams (or "nomograms"). However, it may be possible to reduce the dimensionality of such situations by using statistical models to parameterize the performance characteristics of the forecasting systems (e.g., Krzysztofowicz and Long 1991b). The nomograms could then be constructed in terms of parameters of the relevant distributions, thereby achieving a substantial reduction in dimensionality. Notwithstanding the practical (e.g., computational) difficulties that may be encountered in applying the sufficiency relation in real-world situations, it must be emphasized that this relation is applicable under very general conditions and only requires that the quality of the forecasting systems be described in terms of the conditional probabilities that enter into Eq. (1).

Several general issues arise in conjunction with comparative evaluation of forecasting systems and it seems appropriate to discuss these issues briefly here, with particular reference to the applicability of the sufficiency relation. One such issue is that of sampling variability. In the studies reported here, it has been assumed that the parameters of the forecasting systems (namely, $\delta$, $\pi$, and $\alpha$) are known without error. We believe that this assumption is reasonable in the context of exploratory studies of prototypical forecasting systems. In general, however, the parameters that characterize forecasting performance are not known without error but must be estimated from a sample of forecasts and observations. Thus, studies involving the application of the sufficiency relation to real-world climate forecasting systems should include investigations of the impact of sampling variability on (for example) the identification of the sufficiency and insufficiency regions and the determination of whether one forecasting system is sufficient for another forecasting system.

Another problem that can arise in the context of comparative evaluation is that the forecasting systems of interest may possess quite different performance
characteristics under different conditions. For example, it is not unreasonable to assume that short-range climate forecasts might possess different performance characteristics in situations involving different types of climate anomalies. In such situations, it might be appropriate to characterize forecasting performance "locally" (under different conditions) rather than "globally" (under all conditions). If sufficiency can be demonstrated under a particular set of conditions, then system A could be said to be sufficient for system B (or vice versa) under these conditions. Although local sufficiency is clearly not as strong a result as global sufficiency, it presumably would be of considerable interest and value. The conditions under which it might be possible to infer global sufficiency from local sufficiency require further study.

In summary, the general applicability of the sufficiency relation in the context of comparative evaluation is not limited by any restrictions inherent in the relation itself. This fact has been illustrated by applying the sufficiency relation to comparative evaluation of prototypical climate forecasting systems and demonstrating that no limiting assumptions were made in the practical implementation of the theory. The discussion of two general issues—not inherently linked to the sufficiency relation—that may arise in comparative evaluation simply reflects the need (common to all evaluation methodology) to investigate the practical utility of this relation in complex real-world situations.

7. Summary and Conclusions

The results presented in section 5 reveal that the sufficiency space—the two-dimensional region defined by the ranges of values of the parameters δ and π—is covered largely by regions representing alternative forecasting systems for which the reference system is sufficient, or vice versa. That is, relatively small portions of this space correspond to regions in which the two systems are insufficient for each other. This result has important implications for the practical applicability of the sufficiency relation.

It is obvious that forecasting systems with larger values of both δ and π (i.e., larger or smaller probabilities used more frequently) are preferred to forecasting systems with smaller values of both of these parameters (smaller absolute values in the case of δ). In addition, the results in section 5 demonstrate that a trade-off exists in terms of the parameter π (but not in terms of the parameter δ). Specifically, a system with smaller value of δ can be sufficient for some systems with larger values of π, provided that the value of δ associated with the former is larger (in absolute value) than the value of δ associated with the latter. This trade-off occurs because the boundary between the sufficiency regions is nonlinear in π.

No equivalent trade-off exists in terms of δ due to the linear and horizontal boundary between the sufficiency regions. Thus, if the value of δ for an alternative system B is smaller (in absolute value) than the value of δ for the reference system A, then this "deficiency" cannot be offset by a value of π that is larger for B than for A. In fact, the experimental results show that if δᵦ is smaller than δᵦ, then B is either insufficient for A or A is sufficient for B, depending on the magnitude of δᵦ (and πᵦ).

With regard to the climate forecasting systems of interest here, this result seems to imply that it would be more desirable to produce alternative systems with larger values of δ and (if necessary) smaller values of π rather than vice versa, since the possibility then exists that these systems would be sufficient for the reference system. However, this tentative conclusion obviously should be treated cautiously in light of the assumptions made in modeling the climate forecasts and the relatively modest number of experiments conducted to date.

When the sufficiency relation is used in the real world to compare highly competitive climate forecasting systems, the results in some cases may be inconclusive, in the sense that these systems may be found to be insufficient for each other. Such results imply that some users would prefer one system, whereas other users would prefer another system, depending on their individual payoff structures. The temptation to compare these systems in an evaluation framework of lower dimensionality (e.g., by using a one-dimensional measure of accuracy such as the ERPS) must be resisted, however, since such an analysis can yield misleading results (see section 5). In this regard, forecasts with a larger value of the ERPS can actually be of greater economic value to some users (see Murphy and Ehrendorfer 1987).

In spite of the simplicity of the forecasting systems considered in this study, the results obtained here clearly demonstrate the applicability and potential usefulness of the sufficiency relation in the context of comparative evaluation of climate forecasting systems. The usefulness of this relation for comparative evaluation stems in particular from the fact that this approach accounts fully for the multidimensional character of forecast quality. The possibility then exists of establishing an unambiguous preference order among forecasting systems on the basis of the sufficiency relation.

Future work with the sufficiency relation in the area of climate forecasts must concentrate on its applicability to the comparative evaluation of more general climate forecasting systems (e.g., removing the restrictions on the parameter α in the context of the prototypical forecasting systems considered here or, more generally, consideration of forecasting systems of higher dimensionality; see also section 6). The applicability of the sufficiency relation in these more complex situations could still be studied within a framework of reasonable dimensionality by using appropriate
parametric statistical models to characterize forecast quality.

Acknowledgments. This work was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grants ATM-8714108 and SES-9106440. The comments of two anonymous reviewers, as well as the availability of computing time on the IBM 3090-400E at the University of Vienna, are gratefully acknowledged.

APPENDIX A

Expected Ranked Probability Score

The accuracy of the short-range climate forecasting systems introduced in section 3 can be measured by the ranked probability score (Epstein 1969; Murphy 1971). Assuming that these forecasting systems are well calibrated in the sense that the jth component of the ith forecast vector f equals the conditional probability ρij, it is possible to compute the expected ranked probability score (ERPS).

In terms of the ρij and the πj, the ERPS for the prototypical forecasting systems considered in this study can be written as

\[ \text{ERPS} = \pi_1 [\rho_{11}((\rho_{11} - 1)^2 + (\rho_{11} + \rho_{21} - 1)^2 + (\rho_{11} + \rho_{21} + \rho_{31} - 1)^2)] + \pi_2 [\rho_{12}((\rho_{12} - 1)^2 + (\rho_{12} + \rho_{22} - 1)^2 + (\rho_{12} + \rho_{22} + \rho_{32} - 1)^2)] + \pi_3 [\rho_{13}((\rho_{13} - 1)^2 + (\rho_{13} + \rho_{23} - 1)^2 + (\rho_{13} + \rho_{23} + \rho_{33} - 1)^2)] \]

where W is a diagonal matrix containing the singular values of Q and the matrices U and V satisfy the following relationship:

\[ U'U = V'V = I. \]

The vector h₀ is computed from

\[ h_0 = WV^{-1}U'b, \]

where the two zero singular values enter as zeroes in W⁻¹. If h₀ satisfies Eq. (9), then (9) has a solution. In this case, we first check to see whether the components of h₀ qualify as a stochastic transformation. If so, then system A is sufficient for system B. If not, then the general solution of Eq. (9) is constructed by

\[ h = h_0 + \lambda_1v_1 + \lambda_2v_2, \]

where λ₁ and λ₂ are any real numbers between ±∞ and v₁ and v₂ are the columns of V corresponding to the zero singular values of Q.

The solution space described by (B4) is scanned by allowing different combinations of λ₁ and λ₂ over a finite two-dimensional grid. Each new solution h is checked as to whether or not it qualifies as a stochastic transformation. If, in a finite time, this search does not yield a vector h with components qualifying as a stochastic transformation, then system A is judged as not being sufficient for system B.

Sensitivity experiments performed with this algorithm show in particular that it suffices to let both λ₁ and λ₂ vary between ±1 with a step width of 0.01 to ensure that a valid stochastic transformation is found if it exists. Therefore, with these specifications, this algorithm can be used to determine on the basis of Eq. (9) the sufficiency relation between two systems A and B that possess the prototypical characteristics described in section 3.
REFERENCES


