

NOTES AND CORRESPONDENCE

Climatology, Persistence, and Their Linear Combination
as Standards of Reference in Skill Scores

ALLAN H. MURPHY

Departments of Atmospheric Sciences and Statistics, Oregon State University, Corvallis, Oregon

8 February 1992 and 13 July 1992

ABSTRACT

Skill scores measure the accuracy of the forecasts of interest relative to the accuracy of forecasts based on naive forecasting methods, with either climatology or persistence usually playing the role of the naive method. In formulating skill scores, it is generally agreed that the naive method that produces the most accurate forecasts should be chosen as the standard of reference. The conditions under which climatological forecasts are more accurate than persistence forecasts—and vice versa—were first described in the meteorological literature more than 30 years ago. At about the same time, it was also shown that a linear combination of climatology and persistence produces more accurate forecasts than either of these standards of reference alone. Surprisingly, these results have had relatively little if any impact on the practice of forecast verification in general and the choice of a standard of reference in formulating skill scores in particular.

The purposes of this paper are to describe these results and discuss their implications for the practice of forecast verification. Expressions for the mean-square errors of forecasts based on climatology, persistence, and an optimal linear combination of climatology and persistence—as well as expressions for the respective skill scores—are presented and compared. These pairwise comparisons identify the conditions under which each naive method is superior as a standard of reference. Since the optimal linear combination produces more accurate forecasts than either climatology or persistence alone, it leads to lower skill scores than the other two naive forecasting methods. Decreases in the values of the skill scores associated with many types of operational weather forecasts can be anticipated if the optimal linear combination of climatology and persistence is used as a standard of reference. The conditions under which this practice might lead to substantial decreases in such skill scores are identified.

1. Introduction

Skill in weather forecasting is generally defined as the accuracy of the forecasts of interest relative to the accuracy of forecasts based solely on some naive forecasting method (Brier and Allen 1951; Murphy and Daan 1985). In this context, a naive forecasting method represents a *standard of reference* (i.e., it establishes a zero point on the scale on which skill is measured). Undoubtedly, the two most widely used standards of reference in the field of forecast verification are climatology and persistence (e.g., see Murphy and Daan 1985; Stanski et al. 1989).

In choosing a naive forecasting method as a standard of reference in a particular context, it seems appropriate to adopt the general rule that the method that produces the most accurate forecasts in that context should be selected. Although this rule is both sensible and reasonable, it is seldom explicitly applied in practice. At

best, general guidelines such as “persistence is a more appropriate standard of reference in measuring skill for short-range forecasts and climatology is a more appropriate standard of reference in measuring skill for medium-range and long-range forecasts” are followed. However, an approach based on these guidelines is necessarily qualitative rather than quantitative in nature, and it may be quite misleading in some situations (see section 5).

It is of particular interest here to note that the relative accuracy of forecasts based on climatology and persistence—where accuracy is measured by the mean-square error—can be shown to depend solely on the magnitude of the correlation between the initial (i.e., persistence) and final (i.e., observed) values of the variable of concern and that this result was first reported in the meteorological literature more than 30 years ago (Gringorten and Sissenwine 1960). Moreover, it has also been known for more than 30 years that a linear combination of climatology and persistence generally produces more accurate forecasts than either of these two standards of reference alone (Buell 1958). These results provide the basis for a quantitative approach in which the rule of choosing the most accurate standard

Corresponding author address: Allan H. Murphy, Department of Atmospheric Sciences, Oregon State University, Strand Agricultural Hall 326, Corvallis, OR 97331-2209.

of reference in defining skill scores can be applied in a rational manner.

Despite their potential significance, the author of this paper cannot recall any substantive discussion of these results in the context of choosing a standard of reference to measure the skill of weather forecasts. [Daan (1980) represents a notable exception.] Moreover, the results appear to have had little if any effect on the practice of forecast verification. For example, rarely has the choice between climatology and persistence in this context been based on a quantitative assessment of their relative accuracy, and the linear combination of climatology and persistence has seldom if ever been used as a standard of reference in formulating skill scores for operational weather forecasts.

The purposes of this paper are to present the relevant results and to discuss their implications for the practice of forecast verification. Section 2 contains definitions of climatology and persistence as forecasting methods, and introduces the linear combination of these two naive forecasting methods. The measures of accuracy and skill employed in this paper are the mean-square error and a skill score based on the mean-square error, and section 3 includes expressions for these measures in the cases of reference forecasts based on climatology, persistence, and the optimal linear combination of climatology and persistence. These expressions are compared in section 4, with particular emphasis on the conditions under which one of the standards of reference produces more accurate forecasts than the others and on the relative magnitudes of various differences in accuracy and skill. Section 5 consists of a discussion of the practical implications of these results and section 6 contains a brief summary and some concluding remarks.

2. Standards of reference: Forecasts

It is assumed here that the verification process, including the computation of skill scores involving various standards of reference, is based on a sample of data consisting of n pairs of forecasts and observations. This verification data sample is denoted by the set $\{(f_i, x_i); i = 1, \dots, n\}$, where f_i and x_i represent the forecast and observation, respectively, on the i th forecasting occasion. In this section we describe forecasts based on climatology, persistence, and their linear combination in terms of the elements of the verification data sample and its statistics (e.g., the mean of the x_i).

a. Climatological forecasts

A climatological forecast is a forecast based solely on the mean or average value of the variable of interest, where the average is computed over an appropriate data sample. Such a forecast is the same on all forecasting occasions in the verification data sample (at least for those data samples, or portions of data samples, for which statistical stationarity is a reasonable as-

sumption). Thus, a forecast based on climatology indicates that the value of the variable at the valid time of the forecast will be equal to this average value.

A climatological forecast is usually based on historical data; however, under the assumption that the verification data sample is representative and relatively large, the mean of the sample generally represents a good estimate of the mean of the historical data. Here, we take the climatological forecast to be a forecast based on the mean of the observations in the verification data sample. In this case, $f_i = \bar{x}$ for all i , where the overbar denotes an average. That is, $\bar{x} = (1/n) \sum_i x_i$ ($i = 1, \dots, n$). The pros and cons of using sample climatology as a standard of reference in defining skill scores are discussed by Murphy (1973).

b. Persistence forecasts

A persistence forecast is a forecast based solely on the value of the variable of interest at an appropriate "initial" time. Such a forecast indicates that the value of the variable at the valid time of the forecast will be equal to this initial value. If we denote the initial value on the i th forecasting occasion by x_i^o , then a persistence forecast on this occasion can be expressed as $f_i = x_i^o$ ($i = 1, \dots, n$).

c. Forecasts based on a linear combination of climatology and persistence

In considering naive forecasting methods that might produce more accurate forecasts than either climatology or persistence alone, it is quite natural to consider a combination of these two standards of reference. Linear combinations of climatology and persistence have been investigated by several meteorologists (e.g., Buell 1958; Gringorten and Sissenwine 1960; Daan 1980; Fraedrich and Leslie 1988). Here we let f_i , where

$$f_i = kx_i^o + (1 - k)\bar{x}, \quad (1)$$

denote a convex linear combination of the persistence forecast (x_i^o) and climatological forecast (\bar{x}) on the i th occasion ($i = 1, \dots, n$), where k represents a constant ($0 \leq k \leq 1$). When $k = 0$ the combined forecast is identical to a climatological forecast, when $k = 1$ the combined forecast is identical to a persistence forecast, and when $0 < k < 1$ the combined forecast is a linear combination of climatology and persistence.

3. Standards of reference: Mean-square errors and skill scores

A skill score measures the accuracy of the forecasts of interest relative to the accuracy of forecasts produced by a standard of reference. Thus, to measure skill, it is necessary to choose a measure of accuracy (as well as a standard of reference). In this paper the mean-square error (MSE) serves as the basic measure of forecast

accuracy. The MSE can be written in terms of the elements of the verification data sample as follows:

$$MSE = (1/n) \sum_i (f_i - x_i)^2. \quad (i = 1, \dots, n). \quad (2)$$

A skill score based on the MSE is usually defined as the improvement in the MSE of the forecasts of interest over the MSE of the reference forecasts (e.g., see Murphy and Daan 1985). Thus, if we denote a generic skill score by SS, then

$$SS = 1 - (MSE_f/MSE_r), \quad (3)$$

where MSE_f denotes the MSE of the forecasts of interest and MSE_r denotes the MSE of the reference forecasts. As defined in (3), $SS > 0$ when $MSE_f < MSE_r$, $SS = 0$ when $MSE_f = MSE_r$, and $SS < 0$ when $MSE_f > MSE_r$. Further, $SS = 1$ when $MSE_f = 0$ (perfect forecasts).

The primary focus of this paper is the impact that the choice of a particular standard of reference has on the value of the skill score SS. Specifically, SS in (3) is considered to be a function of MSE_r , with MSE_f fixed (moreover, attention is restricted to situations in which $SS \geq 0$, or $0 \leq MSE_f \leq MSE_r$). In this regard, viewed as a function of MSE_r , SS decreases (increases) as MSE_r decreases (increases). That is, the skill of the forecasts of interest decreases (increases) as the accuracy of the forecasts produced by the standard of reference increases (decreases).

In this section we present expressions for the MSEs of forecasts based on climatology, persistence, and the optimal linear combination of climatology and persistence. These expressions were first reported (to the author's knowledge) by Gringorten and Sissenwine (1960) [see also Daan (1980) and Fraedrich and Smith (1989)]. The SS's corresponding to these MSEs are also defined.

a. Climatology

In the case of climatological forecasts based on the mean of the verification data sample, $f_i = \bar{x}$ for all i (see section 2a). If we denote the MSE of such forecasts by MSE_c , substitution of $f_i = \bar{x}$ into (2) immediately yields

$$MSE_c = s_x^2, \quad (4)$$

where s_x^2 represents the variance of the observations x_i ($i = 1, \dots, n$). That is, the MSE of climatological forecasts is simply the variance of the observed values of the variable of interest. The corresponding skill score SS_c is, from (3) and (4),

$$SS_c = 1 - (MSE_f/s_x^2) \quad (5)$$

($MSE_r = MSE_c$). Note that $SS_c > 0$ when $MSE_f < s_x^2$, $SS_c = 0$ when $MSE_f = s_x^2$, and $SS_c < 0$ when $MSE_f > s_x^2$.

b. Persistence

In the case of persistence forecasts, $f_i = x_i^o$ for all i (see section 2b). Let the MSE of persistence forecasts be denoted by MSE_p . Then, under the assumption that any end effects associated with the computation of the statistics of the two data series (i.e., the series consisting of the x_i^o and the series consisting of the x_i) are negligible, it is relatively easy to show that substitution of $f_i = x_i^o$ into (2) yields

$$MSE_p = 2(1 - r)s_x^2, \quad (6)$$

where r is the correlation coefficient describing the strength of the linear relationship between x_i^o and x_i ($i = 1, \dots, n$). (The assumption of negligible end effects implies that $\bar{x}_i^o = \bar{x}$ and $s_{x^o}^2 = s_x^2$.) Thus,

$$r = s_{x^o x} / s_{x^o} s_x = s_{x^o x} / s_x^2, \quad (7)$$

where

$$s_{x^o x} = (1/n) \sum_i (x_i^o - \bar{x})(x_i - \bar{x}) \quad (i = 1, \dots, n) \quad (8)$$

is the covariance between the initial values x_i^o and the observations x_i . For convenience, we generally refer to r as the persistence correlation coefficient (it can also be viewed as a first-order autocorrelation coefficient).

Values of MSE_p are shown in Table 1 for selected values of r , under the assumption that $s_x^2 = 1$. The expressions for MSE_c and MSE_p in (4) and (6), respectively, are compared in section 4a. The skill score corresponding to MSE_p is SS_p , where

$$SS_p = 1 - [MSE_f/2(1 - r)s_x^2] \quad (9)$$

($MSE_r = MSE_p$). Note that $SS_p > 0$ when $MSE_f < 2(1 - r)s_x^2$, $SS_p = 0$ when $MSE_f = 2(1 - r)s_x^2$, and $SS_p < 0$ when $MSE_f > 2(1 - r)s_x^2$.

TABLE 1. Values of MSE_c , MSE_p , MSE_{cp} , and $DMSE_{cp}$ for selected values of the persistence correlation r , under the assumption that $s_x^2 = 1$. See text for additional details.

Persistence correlation r	Climatology MSE_c	Persistence MSE_p	Linear combination MSE_{cp}	Decrease in MSE $DMSE_{cp}$
0.00	1.00	2.00	1.00	0.00
0.10	1.00	1.80	0.99	0.01
0.20	1.00	1.60	0.96	0.04
0.30	1.00	1.40	0.91	0.09
0.40	1.00	1.20	0.84	0.16
0.50	1.00	1.00	0.75	0.25
0.60	1.00	0.80	0.64	0.20
0.70	1.00	0.60	0.51	0.15
0.80	1.00	0.40	0.36	0.10
0.90	1.00	0.20	0.19	0.05
1.00	1.00	0.00	0.00	0.00

c. Optimal linear combination of climatology and persistence

To determine the value of the constant k (in the expression for the linear combination of climatology and persistence) that minimizes the MSE, it is first necessary to substitute the expression for f_i in (1) into the expression for the MSE in (2). Then differentiating the resulting expression with respect to k yields the solution $k = r$. That is, the optimal linear combination of climatology and persistence is

$$f_i = rx_i^o + (1 - r)\bar{x} \tag{10}$$

(to obtain this expression it is necessary to make use of the assumption of negligible end effects). The expression for f_i in (10) makes intuitive sense, since $f_i = \bar{x}$ when $r = 0$ and $f_i = x_i^o$ when $r = 1$ ($i = 1, \dots, n$).

Let MSE_{cp} denote the MSE of forecasts produced by the optimal linear combination of climatology and persistence in (10). Then, under the assumption of negligible end effects, substitution of (10) into (2) yields (after some algebraic manipulation)

$$MSE_{cp} = (1 - r^2)s_x^2. \tag{11}$$

Values of MSE_{cp} are shown in Table 1 for selected values of r (once again assuming that $s_x^2 = 1$). The expression for MSE_{cp} in (11) is compared with the expressions for MSE_c and MSE_p in section 4b. Finally, the skill score corresponding to MSE_{cp} is SS_{cp} , where

$$SS_{cp} = 1 - [MSE_f / (1 - r^2)s_x^2] \tag{12}$$

($MSE_f = MSE_{cp}$). Note that $SS_{cp} > 0$ when $MSE_f < (1 - r^2)s_x^2$, $SS_{cp} = 0$ when $MSE_f = (1 - r^2)s_x^2$, and $SS_{cp} < 0$ when $MSE_f > (1 - r^2)s_x^2$.

4. Standards of reference: Comparison of MSEs and SS's

a. Comparison of MSE_c (SS_c) and MSE_p (SS_p)

Under what conditions is a climatological forecast more accurate than a persistence forecast (and vice versa)? Comparison of (4) and (6) reveals that

$$MSE_p = 2(1 - r)MSE_c. \tag{13}$$

Thus, $MSE_c < MSE_p$ if $r < 1/2$, $MSE_c = MSE_p$ if $r = 1/2$, and $MSE_c > MSE_p$ if $r > 1/2$. (Since the persistence correlation—or first-order autocorrelation—of most weather variables is positive, we restrict our attention here to values of r between 0 and 1 inclusive.) If the persistence correlation is relatively low ($r < 1/2$), climatology outperforms persistence. On the other hand, if r is relatively high ($r > 1/2$), persistence outperforms climatology. Thus, if the rule of choosing the most accurate standard of reference is followed, then this result suggests that climatology should be used as the standard of reference when $r < 1/2$ and persistence should be used as the standard of reference when $r > 1/2$. For

further discussion of the implications of this result, see section 5.

Combining the expressions for SS_c and SS_p in (5) and (9), respectively, reveals that

$$SS_p = 1 - [(1 - SS_c) / 2(1 - r)]. \tag{14}$$

Examination of this expression reveals that $SS_c < SS_p$ when $r < 1/2$, $SS_c = SS_p$ when $r = 1/2$, and $SS_c > SS_p$ when $r > 1/2$. These relationships are, of course, implicit in the relationship between MSE_c and MSE_p in (13) and the basic definition of the skill score SS in terms of MSEs [see (3)].

b. Comparison of MSE_{cp} (SS_{cp}) with MSE_c (SS_c) and MSE_p (SS_p)

Comparison of MSE_{cp} in (11) with MSE_c and MSE_p in (4) and (6) reveals that

$$MSE_{cp} = (1 - r^2)MSE_c \tag{15}$$

and

$$MSE_{cp} = [(1 + r) / 2]MSE_p, \tag{16}$$

respectively. Thus, $MSE_{cp} \leq MSE_c$ and $MSE_{cp} \leq MSE_p$, with equality only when $r = 0$ and $r = 1$, respectively. Since $0 < r < 1$ in general, the optimal linear combination of climatology and persistence always outperforms climatology or persistence alone.

Since $MSE_c < MSE_p$ for $r < 1/2$ and $MSE_c > MSE_p$ for $r > 1/2$, and a choice is traditionally made between climatology and persistence, it is of interest to compare MSE_{cp} with $\min(MSE_c, MSE_p)$. This comparison is accomplished here by computing $DMSE_{cp}$, where

$$DMSE_{cp} = 1 - [MSE_{cp} / \min(MSE_c, MSE_p)]. \tag{17}$$

Substituting the expressions for MSE_c , MSE_p , and MSE_{cp} from (4), (6), and (11), respectively, into (17), it follows that

$$DMSE_{cp} = \begin{cases} r^2 & \text{if } 0 \leq r \leq 1/2 \\ (1/2)(1 - r) & \text{if } 1/2 \leq r \leq 1. \end{cases} \tag{18}$$

As defined, $DMSE_{cp}$ represents the fractional decrease in the MSE achieved by using the optimal linear combination instead of climatology or persistence alone (whichever produces the more accurate forecasts).

The values of $DMSE_{cp}$ in (18) are shown in Table 1 for selected values of the persistence correlation r . Note that $0 \leq DMSE_{cp} \leq 0.25$, with $DMSE_{cp} = 0$ for $r = 0$ or 1 and $DMSE_{cp} = 0.25$ for $r = 1/2$. The optimal linear combination achieves the largest reductions in MSE when the value of r is near $1/2$ (i.e., when reference forecasts based on climatology or persistence alone are relatively inaccurate). These reductions exceed 0.15 (or 15%) when $0.4 \leq r \leq 0.6$.

It is also of interest to note that $DMSE_{cp}$ is a non-linear function of r when $r < 1/2$ (i.e., when $MSE_c < MSE_p$) and a linear function of r when $r > 1/2$ (i.e., when $MSE_c > MSE_p$). That is, $DMSE_{cp}$ increases linearly as r decreases from 1 toward $1/2$, whereas $DMSE_{cp}$ increases nonlinearly as r increases from 0 toward $1/2$. In particular, the initial rate of decrease in the values of $DMSE_{cp}$ is greater as r decreases from $1/2$ toward 0 than it is as r increases from $1/2$ toward 1.

Expressions describing the relationships between the skill scores—in particular, between SS_c and SS_{cp} and between SS_p and SS_{cp} —can be obtained by combining (5) and (12) and (9) and (12), respectively. These expressions are

$$SS_{cp} = 1 - [(1 - SS_c)/(1 - r^2)] \quad (19)$$

and

$$SS_{cp} = 1 - [2(1 - SS_p)/(1 + r)]. \quad (20)$$

Since $MSE_{cp} \leq \min(MSE_c, MSE_p)$, it follows that $SS_{cp} \leq \min(SS_c, SS_p)$. That is, the skill score involving a standard of reference based on the optimal linear combination is always less than or equal to the skill score involving a standard of reference based on climatology or persistence alone.

The values of SS_{cp} for selected values of $\min(SS_c, SS_p)$ and the persistence correlation r are shown in Table 2. (Note that $SS_c < SS_p$ when $r < 1/2$ and $SS_c > SS_p$ when $r > 1/2$.) As indicated previously, $SS_{cp} \leq \min(SS_c, SS_p)$ for all r ($0 \leq r \leq 1$). The difference between $\min(SS_c, SS_p)$ and SS_{cp} is generally less than 0.05 (5%) when $\min(SS_c, SS_p) \geq 0.8$ and/or $r \leq 0.2$ or $r \geq 0.9$. However, the difference between MSE_{cp} and $\min(MSE_c, MSE_p)$ is larger—in some cases, considerably larger—when $\min(SS_c, SS_p) < 0.8$ and $0.2 < r < 0.9$. Combinations of values of $\min(SS_c, SS_p)$ and r for which this difference exceeds 0.10 (10%) are underlined. Moreover, SS_{cp} is negative for some combi-

nations of values of $\min(SS_c, SS_p) (\geq 0)$ and r . Further discussion of the contents of Table 2 and their implications is postponed until section 5.

5. Discussion: Implications of results

What are the implications of the results presented in sections 3 and 4 for the practice of forecast verification in general and the choice of a standard of reference in defining skill scores in particular? In discussing these implications, we adopt the general—and previously mentioned—rule that the naive standard of reference that performs best in an MSE sense should be chosen from the set of available standards. The implications of using measures of accuracy other than the MSE are also examined briefly.

First we consider situations in which the choice of a standard of reference is restricted to either climatology or persistence alone. In such situations, it follows from the results presented in sections 3 and 4 that climatology should be selected when the persistence correlation is less than one-half and persistence should be selected when the persistence correlation is greater than one-half. To what extent is this rule actually applied in practice and what would be the result of a strict application of such a rule?

Since the persistence correlation is seldom if ever reported in studies involving the use of skill scores, we can only assume that the choice between these two standards of reference is at best based on informal estimates of the strength of the persistence relationship. In this regard, general guidelines and/or common perceptions regarding the relative performance of climatology and persistence may be quite misleading in some situations. For example, even in the case of short-range weather forecasts, the persistence correlation may be relatively low (i.e., $r < 1/2$) in situations in which 6–12-h forecasts are made in the presence of strong diurnal variability and/or situations in which 12–36-h

TABLE 2. The skill score for the optimal linear combination of climatology and persistence, SS_{cp} , for selected values of $\min(SS_c, SS_p)$ and the persistence correlation r . Values of SS_{cp} (rounded to two decimal places) for which $\min(SS_c, SS_p) - SS_{cp} \geq 0.10$ are underlined. See text for additional details.

	$\min(SS_c, SS_p)$										
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.0	0.000	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.800	0.900	1.000
0.1	-0.010	0.091	0.192	0.293	0.394	0.495	0.596	0.697	0.798	0.899	1.000
0.2	-0.042	0.062	0.167	0.271	0.375	0.479	0.583	0.688	0.792	0.896	1.000
0.3	-0.099	0.011	0.121	0.231	0.341	0.451	0.560	0.670	0.780	0.890	1.000
0.4	-0.190	-0.071	0.048	0.167	0.286	0.405	0.524	0.643	0.762	0.881	1.000
0.5	-0.333	-0.200	-0.067	0.067	0.200	0.333	0.467	0.600	0.733	0.867	1.000
0.6	-0.250	-0.125	0.000	0.125	0.250	0.375	0.500	0.625	0.750	0.875	1.000
0.7	-0.176	-0.059	0.059	0.176	0.294	0.412	0.529	0.647	0.765	0.882	1.000
0.8	-0.111	0.000	0.111	0.222	0.333	0.444	0.556	0.667	0.778	0.889	1.000
0.9	-0.053	0.053	0.158	0.263	0.368	0.474	0.579	0.684	0.789	0.895	1.000
1.0	0.000	0.100	0.200	0.300	0.400	0.500	0.600	0.700	0.800	0.900	1.000

forecasts are made in the presence of substantial day-to-day variability. In any case, it is quite likely that this informal approach will sometimes lead to the selection of the inferior standard of reference. Thus, strict application of the rule of choosing the superior standard of reference can be expected to result in reductions in the values of MSE-based skill scores in at least some situations.

When the set of available standards of reference is extended to include the optimal linear combination of climatology and persistence, it is clear from the results presented in sections 3 and 4 that the latter should always be chosen in preference to climatology or persistence alone. However, with the exception of the study by Daan (1980), the author cannot recall a single instance in which this practice has been followed explicitly. Certainly, the use of a linear combination of climatology and persistence as a standard of reference in measuring skill is the exception rather than the rule.

At this point, it seems appropriate to ask about the impact that choosing the optimal linear combination rather than climatology or persistence alone will have on MSE-based skill scores. The answer to this question is provided by the contents of Table 2. These results indicate that the impact will be relatively small when the skill scores based on climatology and persistence (SS_c and SS_p) are both very high or the persistence correlation (r) is very high or very low. However, substantially larger impacts occur in situations involving lower values of SS_c or SS_p and intermediate values of r . Reductions in skill—from the situation in which skill is measured by SS_c or SS_p to the situation in which skill is measured by SS_{cp} (the skill score based on the optimal linear combination)—exceed 10% for a considerable range of values of the “parameters” $\min(SS_c, SS_p)$ and r (see underlined entries in Table 2). It should also be noted that when either SS_c or SS_p is relatively small and r takes on intermediate values, the use of SS_{cp} instead of SS_c or SS_p may “transform” positive skill scores (SS_c or SS_p) into negative skill scores (SS_{cp}).

Two further comments regarding these results are warranted. First, many situations encountered in the real world involve relatively small values of $\min(SS_c, SS_p)$ and intermediate values of r [on the other hand, situations involving large values of $\min(SS_c, SS_p)$ are relatively rare]. Thus, it is quite likely that the replacement of $\min(SS_c, SS_p)$ with SS_{cp} will lead to substantial reductions in the skill (as measured by skill scores) of some forecasts. [Trial calculations made at the Royal Netherlands Meteorological Institute more than 10 years ago support this tentative conclusion (H. Daan, personal communication, 1992).] Second, this discussion of the differences between the situation in which the standard of reference is climatology or persistence alone and the situation in which the standard of reference is the optimal linear combination has been based

on the assumption that the choice between climatology and persistence is made on the basis of the standard of reference that produces the most accurate forecasts (according to the MSE). However, if the choice is not optimal in this sense, then the aforementioned differences between $\min(SS_c, SS_p)$ and SS_{cp} are in reality lower bounds on the real differences. That is, a sub-optimal choice between climatology and persistence generally will lead to even larger reductions in skill (than those indicated in Table 2) when climatology or persistence is replaced by the optimal linear combination.

Thus, if the optimal linear combination of climatology and persistence were to be used as the standard of reference in measuring forecast skill, then the values of MSE-based skill scores of many operational weather forecasts would be reduced. In effect, these skill scores would be measuring relative forecast accuracy with respect to a superior standard of reference. What are the pros and cons of such a change in current verification practices? Understandably, this prospect might be viewed with considerable consternation by many operational meteorologists. On the other hand, making the process of selecting a standard of reference more rational and choosing a standard of reference that is truly representative of the best naive forecasting methods are both important goals within the overall framework of forecast verification. Moreover, the use of such a standard of reference should provide a more realistic assessment of the incremental contributions (in terms of a reduction of the MSE) of numerical and/or statistical models and human judgment.

The results presented and discussed in this paper relate to the choice of a standard of reference for skill scores in situations in which the underlying measure of accuracy is the mean-square error. Of course, skill scores based on other measures of accuracy are used in some situations. What can be said about the choice among climatology, persistence, and their optimal linear combination in these situations? Daan (1980) investigated this problem for the case of a linear measure of accuracy (i.e., the mean absolute error) and obtained qualitatively similar results, although skill scores based on this measure of accuracy were found to be somewhat less sensitive to the choice of a standard of reference than those based on the MSE. Clearly, the extent to which the results presented here are—or are not—generalizable to situations involving other measures of forecast accuracy is a topic worthy of further study.

6. Conclusions

This paper has described various results related to the use of climatology, persistence, and an optimal linear combination of climatology and persistence as standards of reference in determining the skill (i.e., relative accuracy) of weather forecasts. Expressions for

the mean-square errors (MSEs) of forecasts produced by these three naive forecasting methods—and the corresponding MSE-based skill scores (SS's)—have been presented and compared. The differences among the respective MSEs (and SS's) depend on the correlation between the persistence forecasts and the corresponding observations, with climatology (persistence) producing more accurate forecasts than persistence (climatology) when the persistence correlation is less (greater) than one-half. The optimal linear combination of climatology and persistence provides more accurate forecasts than either of these two naive forecasting methods alone, and achieves fractional reductions in the MSE (over its closest competitor) exceeding 15% for persistence correlations between 0.4 and 0.6. Moreover, sets of forecasts produced in situations in which skill (according to skill scores based on climatology or persistence alone) is low or moderate and the persistence correlation takes on intermediate values could “experience” substantial decreases in their MSE-based skill scores if the optimal linear combination is accepted as the appropriate standard of reference.

These results appear to have important implications for the choice of a standard of reference when MSE-based skill scores are used to evaluate weather forecasts. In particular, they provide a rational basis for choosing between climatology and persistence in this context. More importantly, they demonstrate that the optimal linear combination of climatology and persistence always outperforms either climatology or persistence alone. It follows that the use of the optimal linear combination as a standard of reference in MSE-based skill scores will lead to reductions in the values of these scores (in effect, the zero point on the scale on which skill is measured will be raised). Of course, this change in traditional practices might be viewed as inappropriate and/or undesirable in some quarters. However, it would lead to a more realistic assessment of forecast skill, where skill refers to the difference between the accuracy of state-of-the-art forecasts and the accuracy of forecasts based solely on the best available naive forecasting method.

This paper has considered the linear combination of climatology and persistence only in the context of identifying the most appropriate naive standard of reference for use in defining MSE-based skill scores. In particular, it has not been concerned with the general problem of combining forecasts produced by different forecasting methods. The differences between the MSE of the optimal linear combination of climatology and persistence and the respective MSEs of climatology or persistence alone, however, provides a graphic example of the potential benefits of combining forecasts. The fact that combining forecasts from different methods can lead to improvements in forecasting performance has been demonstrated recently in a variety of situations (e.g., see Clemen and Murphy 1986; Fraedrich

and Leslie 1988; Fraedrich and Smith 1989; Thompson 1977). A comprehensive review of the so-called combining literature in many different fields has been provided by Clemen (1989).

In order to obtain realistic and credible estimates of the skill (i.e., relative accuracy) of weather forecasts, appropriate standards of reference must be chosen to define the scale (in particular, the zero point) on which skill is measured. The basic results described in this paper and first reported in the meteorological literature more than 30 years ago provide the quantitative information needed to make a rational choice among climatology, persistence, and the optimal linear combination of these two common standards of reference, at least in those situations in which the mean-square error is the underlying measure of accuracy. Thus, these results—and their application in operational and experimental contexts—should help to place the measurement of skill, an important overall characteristic of forecasting performance, on a sound scientific basis.

Acknowledgments. The comments of Harald Daan and three anonymous reviewers on earlier versions of this paper are gratefully acknowledged. This research was supported in part by the National Science Foundation (Division of Social and Economic Sciences) under Grant SES-9106440.

REFERENCES

- Brier, G. W., and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology*, T. F. Malone, Ed., Amer. Meteor. Soc., 841–848.
- Buell, C. E., 1958: Meaning of combined climate and persistence forecasts. *J. Meteor.*, **15**, 564–565.
- Clemen, R. T., 1989: Combining forecasts: A review and annotated bibliography. *Int. J. Forecasting*, **5**, 559–583.
- , and A. H. Murphy, 1986: Objective and subjective precipitation probability forecasts: Some methods for improving forecast quality. *Wea. Forecasting*, **1**, 213–218.
- Daan, H., 1980: Climatology and persistence as reference forecasts in verification studies. *WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*. Geneva, Switzerland, World Meteorological Organization, 195–201.
- Fraedrich, K., and L. M. Leslie, 1988: Real time short-term forecasting of precipitation at an Australian tropical station. *Wea. Forecasting*, **3**, 104–114.
- , and N. R. Smith, 1989: Combining predictive schemes in long-range forecasting. *J. Climate*, **2**, 291–294.
- Gringorten, I. I., and N. Sissenwine, 1960: Correlation coefficients and prediction accuracy. *J. Meteor.*, **17**, 462–463.
- Murphy, A. H., 1973: Hedging and skill scores for probability forecasts. *J. Appl. Meteor.*, **12**, 215–223.
- , and H. Daan, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.
- Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification measures in meteorology. Atmospheric Environment Service, Downsview, Ontario, Canada, Research Report No. 89-5, 113 pp.
- Thompson, P. D., 1977: How to improve accuracy by combining independent forecasts. *Mon. Wea. Rev.*, **105**, 228–229.