Skill Scores and Correlation Coefficients in Model Verification

ALLAN H. MURPHY

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

EDWARD S. EPSTEIN

Climate Analysis Center, National Weather Service, Washington, D.C.

(Manuscript received 31 May 1988, in final form 19 September 1988)

ABSTRACT

Attributes of the anomaly correlation coefficient, as a model verification measure, are investigated by exploiting a recently developed method of decomposing skill scores into other measures of performance. A mean square error skill score based on historical climatology is decomposed into terms involving the anomaly correlation coefficient, the conditional bias in the forecast, the unconditional bias in the forecast, and the difference between the mean historical and sample climatologies. This decomposition reveals that the square of the anomaly correlation coefficient should be interpreted as a measure of *potential* rather than actual skill.

The decomposition is applied to a small sample of geopotential height field forecasts, for lead times from one to ten days, produced by the medium range forecast (MRF) model. After about four days, the actual skill of the MRF forecasts (as measured by the "climatological skill score") is considerably less than their potential skill (as measured by the anomaly correlation coefficient), due principally to the appearance of substantial conditional biases in the forecasts. These biases, and the corresponding loss of skill, represent the penalty associated with retaining "meteorological" features in the geopotential height field when such features are not predictable. Some implications of these results for the practice of model verification are discussed.

1. Introduction

Numerical weather prediction (NWP) models in several meteorological centers routinely produce forecasts of geopotential height fields—and other upper air and near-surface parameters—for lead times up to ten days. In order to evaluate model performance, insofar as the circulation of the atmosphere is concerned, it is common practice to compare these two-dimensional forecasts of geopotential height, at various levels, with the corresponding observed (or analyzed; or initialized) height fields. This comparison generally is accomplished by computing an overall measure of the degree of association or correspondence between the two fields, such as an anomaly (with respect to climatology) correlation coefficient or a root mean square error (Hollingsworth et al. 1980; Arpe et al. 1985).

Both correlation coefficients and mean square error measures have their supporters and detractors in the context of forecast verification. For example, Brier and Allen (1951, p. 845) note that the correlation coefficient "is insensitive to any bias or error in scale." It is because of such deficiencies that correlation coefficients are seldom used to verify *weather* forecasts. Arpe et al. (1985) obviously represent a quite different perspective when they state "Unlike the anomaly correlation coefficient, the rms score has the disadvantage that it favors forecasts which underestimate atmospheric variability." The fact that correlation coefficients ignore biases and errors in scale suggests that verification measures such as the anomaly correlation coefficient, interpreted at their face value, may overestimate the performance of NWP models. This tendency toward overestimation is explicitly recognized in the practice, based on synoptic experience, of using an arbitrary anomaly correlation coefficient value of 0.6 as a lower limit for "useful" medium range forecasting (e.g., Hollingsworth et al. 1980). On the other hand, as Arpe et al. (1985) indicate, the "asymptotic value (of the rms score) after the loss of predictive skill varies considerably with synoptic situation, so it is not terribly useful for defining a limit of useful predictive skill."

The purpose of this paper is to investigate the deficiencies in the anomaly correlation coefficient as a model verification measure. To realize this objective, we exploit a method of decomposing skill scores into other measures of performance—a method used recently in the context of *weather* forecasting to explore relationships between the basic (product moment) correlation coefficient and various mean square error skill scores (Murphy 1988). Specifically, a mean square error skill score based on historical climatology is decomposed into terms involving (*inter alia*) the anomaly

Corresponding author address: Prof. Allan H. Murphy, Dept. of Atmospheric Sciences, Oregon State University, Corvallis, OR 97331.

© 1989 American Meteorological Society

correlation coefficient, thereby providing a means of comparing the former and the latter as overall measures of *model* performance. The decomposition of this skill score is described and discussed in section 2. In section 3 the decomposition is applied to forecasts of the 1000 and 500 mb geopotential height fields produced by the medium range forecast (MRF) model, currently run operationally for lead times from one to ten days at the U.S. National Meteorological Center (NMC). Section 4 summarizes the results of this study and briefly discusses some implications of these results for the practice of model verification.

2. Skill score and anomaly correlation coefficient

a. Mean square error skill score

Skill generally is defined as the accuracy of the forecasts of interest relative to the accuracy of forecasts produced by some reference procedure—or *standard of reference*—such as climatology or persistence. The basic measure of accuracy employed in this paper is the mean square error (MSE). Let f_i and x_i denote the forecast and analyzed (or initialized) geopotential heights, respectively, at the ith of n grid points defining the respective two-dimensional fields at a particular "point" in time. Then MSE (f, x) at this time can be expressed as follows:

$$\mathrm{MSE}(f, x) = \langle (f_i - x_i)^2 \rangle, \tag{1}$$

where the angled brackets denote a mean over the n grid points. Thus, forecast accuracy here is concerned with the correspondence between forecasts and observations over a geographical domain at a specific time. Note that $\mathrm{MSE}(f, x) \geq 0$, with equality only for perfect correspondence between the two fields (i.e., $f_i = x_i$ for all i).

The standard of reference of interest here is historical or long-term climatology. Let c_i denote the long-term climatological value of the geopotential height field at the ith grid point. Then MSE for the reference forecast can be expressed as follows:

$$\mathrm{MSE}(c, x) = \langle (c_i - x_i)^2 \rangle. \tag{2}$$

Skill scores usually are defined as the improvement in the accuracy of the forecasts of interest over the reference forecasts, relative to the total possible improvement in accuracy (e.g., see Murphy and Daan 1985). Thus, a climatological skill score (SS) based on the mean square error can be defined as follows:

$$\mathrm{SS}(f, c, x) = [\mathrm{MSE}(c, x) - \mathrm{MSE}(f, x)]/[\mathrm{MSE}(c, x) - 0] \tag{3}$$

[recall that $\mathrm{MSE}(f, x) = 0$ for a perfect forecast] or

$$\mathrm{SS}(f, c, x) = 1 - [\mathrm{MSE}(f, x)/\mathrm{MSE}(c, x)]. \tag{4}$$

Note that $\mathrm{SS}(f, c, x)$ is positive (negative) when the accuracy of the forecasts is greater (less) than the accuracy of the reference forecasts. Moreover, $\mathrm{SS}(f, c, x) = 1$ when $\mathrm{MSE}(f, x) = 0$ (perfect forecasts) and $\mathrm{SS}(f, c, x) = 0$ when $\mathrm{MSE}(f, x) = \mathrm{MSE}(c, x)$. $\mathrm{SS}(f, c, x)$ can be translated into a measure of percentage improvement in accuracy simply by multiplying the right-hand side (rhs) of (4) by 100. For convenience, we will sometimes refer to $\mathrm{SS}(f, c, x)$ as the *climatological skill score.*

Since historical climatology (i.e., c) is known only imperfectly, difficulties may arise in connection with the application of the climatological skill score. This fact can affect the values of $\mathrm{SS}(f, c, x)$, especially when $\mathrm{MSE}(c, x)$ is small, and it makes comparison of results involving different climatologies tenuous. On the other hand, in comparing the forecasts produced by different forecasting procedures, the ordinal relationship among the skill scores generally will not be affected by the choice of a particular climatology.

b. Decomposition of mean square error skill score

To facilitate the decomposition of the climatological skill score, we first decompose $\mathrm{MSE}(f, x)$ and $\mathrm{MSE}(c, x)$. The decomposition of $\mathrm{MSE}(f, x)$ of interest here can be derived by adding and subtracting c_i within the parentheses on the rhs of (1). Initially, we obtain

$$\mathrm{MSE}(f, x) = \langle [(f_i - c_i) - (x_i - c_i)]^2 \rangle \tag{5}$$

or

$$\mathrm{MSE}(f, x) = \langle (f'_i - x'_i)^2 \rangle, \tag{6}$$

where $f'_i = f_i - c_i$ and $x'_i = x_i - c_i$ are the anomalies in the forecast and analyzed heights, respectively, at the ith grid point. Adding and subtracting the mean anomalies within the parentheses on the rhs of (6) yields

$$\mathrm{MSE}(f, x) = \langle [(f'_i - \langle f' \rangle) - (x'_i - \langle x' \rangle) + (\langle f' \rangle - \langle x' \rangle)]^2 \rangle, \tag{7}$$

where $\langle f' \rangle$ is the sample mean anomaly in the forecast field and $\langle x' \rangle$ is the sample mean anomaly in the analyzed field. Completing the squaring process within the square brackets on the rhs of (7) and averaging over the n grid points yields

$$\mathrm{MSE}(f, x) = (\langle f' \rangle - \langle x' \rangle)^2 + s^2_{f'} + s^2_{x'} - 2s_{f'x'}, \tag{8}$$

where $s^2_{f'}$ $[= \langle f'^2 \rangle - \langle f' \rangle^2]$ is the sample variance of the anomalies in the forecast field, $s^2_{x'}$ $[= \langle x'^2 \rangle - \langle x' \rangle^2]$ is the sample variance of the anomalies in the analyzed field, and $s_{f'x'}$ $[= \langle f'x' \rangle - \langle f' \rangle \langle x' \rangle]$ is the sample covariance between the anomalies in the forecast and analyzed fields. Moreover, since $s_{f'x'} = s_{f'} s_{x'} r_{f'x'}$, $\mathrm{MSE}(f, x)$ in (8) can also be expressed

as

$$\mathrm{MSE}(f, x) = (\langle f' \rangle - \langle x' \rangle)^2 + s_{f'}^2 + s_{x'}^2 - 2 s_{f'} s_{x'} r_{f'x'}, \quad (9)$$

where $r_{f'x'}$ is the sample (product moment) coefficient of correlation between the anomalies in the forecast and analyzed fields. That is, the latter is the *anomaly correlation coefficient*.

An analogous decomposition of $\mathrm{MSE}(c, x)$ can be derived in a similar manner. In this case, $\langle x' \rangle$ is added and subtracted within the parentheses on the rhs of (2), yielding

$$\mathrm{MSE}(c, x) = \langle [(x_i' - \langle x' \rangle) + \langle x' \rangle]^2 \rangle. \quad (10)$$

Completing the squaring process within the square brackets on the rhs of (10) and averaging over the n grid points yields

$$\mathrm{MSE}(c, x) = s_{x'}^2 + \langle x' \rangle^2. \quad (11)$$

Thus, the decomposition of $\mathrm{MSE}(c, x)$ involves only the sample variance of the anomalies in the analyzed field and the square of the mean anomaly in this field.

The decomposition of the climatological skill score $\mathrm{SS}(f, c, x)$ of interest here can now be obtained by making use of the decompositions of $\mathrm{MSE}(f, x)$ and $\mathrm{MSE}(c, x)$. In particular, substituting (9) and (11) into (4) yields

$$\mathrm{SS}(f, c, x) = \{2(s_{f'}/s_{x'}) r_{f'x'} - (s_{f'}/s_{x'})^2 - [(\langle f' \rangle - \langle x' \rangle)/s_{x'}]^2 + (\langle x' \rangle/s_{x'})^2\}/[1 + (\langle x' \rangle/s_{x'})^2] \quad (12)$$

or

$$\mathrm{SS}(f, c, x) = \{r_{f'x'}^2 - [r_{f'x'} - (s_{f'}/s_{x'})]^2 - [(\langle f' \rangle - \langle x' \rangle)/s_{x'}]^2 + (\langle x' \rangle/s_{x'})^2\}/[1 + (\langle x' \rangle/s_{x'})^2]. \quad (13)$$

It is immediately evident that the decomposition of $\mathrm{SS}(f, c, x)$ in (13) provides an analytical relationship between the climatological skill score and the anomaly correlation coefficient. The terms in this decomposition are interpreted and discussed in section 2c.

c. *Interpretation and discussion*

For convenience we will refer to the terms in the numerator on the rhs of (13) as A, B, C, and D. Specifically,

$$A = r_{f'x'}^2, \quad (14)$$

$$B = [r_{f'x'} - (s_{f'}/s_{x'})]^2, \quad (15)$$

$$C = [(\langle f' \rangle - \langle x' \rangle)/s_{x'}]^2, \quad (16)$$

$$D = (\langle x' \rangle/s_{x'})^2. \quad (17)$$

The term A is the square of the anomaly correlation coefficient, and it necessarily lies in the closed unit interval [0, 1]. Terms B and C are both nonnegative quantities that enter negatively into (13). On the other hand, the term D enters positively and appears in both numerator and denominator. As a result, D has a positive influence on the skill score. Thus, since $\mathrm{SS}(f, c, x) \leq 1$, it follows that $A - B - C \leq 1$.

In interpreting the individual terms in (13), it is useful to recall that the joint distribution of forecast and analyzed (or observed) anomalies in the geopotential heights contains all of the non-time-dependent information relevant to verification (Murphy and Winkler, 1987). With this perspective in mind, it is of interest to note that all of the terms in (13) are defined in terms of summary measures of the empirical joint and marginal distributions of the respective anomalies (i.e., in terms of means, variances or standard deviations, and a covariance or correlation of the anomalies). A linear regression model, in which the anomalies in the analyzed height field are regressed on the anomalies in the forecast height field, provides a potentially convenient and useful means of describing this joint distribution. The regression model can be represented by the following equation:

$$E(x' | f') = a + bf', \quad (18)$$

where $E(x' | f')$ is the expected (or mean) value of the anomalies in the analyzed heights given a particular anomaly in the forecast heights and a and b are estimates of the (unknown) regression coefficients. Specifically, a and b represent estimates of the intercept and slope, respectively, of this linear equation. Note that if $b \neq 1$, then the error in the forecast—namely, $f' - E(x' | f')$—depends systematically on f', and the forecast is *conditionally* biased. If the mean error, $\langle f' \rangle - \langle x' \rangle = \langle f' \rangle - \langle E(x' | f') \rangle \neq 0$, then the forecast is *unconditionally* biased. This latter condition implies that $a = (1 - b)\langle f' \rangle$ for an unconditionally unbiased forecast.

Now D is the only term in (13) that is independent of the forecast heights—it depends solely on the analyzed and climatological height values. In particular, D is the square of the ratio of the mean (over the entire field or that portion of the field being verified) anomaly in the analyzed heights to the standard deviation of these anomalies. As such, it represents the square of the coefficient of variation of the anomalies in the analyzed height field. This term should be small (compared to unity) unless the verification is performed over a very limited domain.

Note that the term C is proportional to the difference between the mean anomaly in the forecast heights and the mean anomaly in the analyzed heights, suitably normalized by dividing by the standard deviation of the anomalies in the analyzed heights. Thus, this term

TABLE 1. Anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 1000 mb forecasts verifying on 1 January 1987. *Key to symbols*: LT—lead time (days); $r_{f'x'}$—anomaly correlation coefficient; A, B, C, D—terms in decomposition of climatological skill score [see section 2 and equations (13)–(17)]; SS—climatological skill score; MSE(f, x)—mean square error of forecasts; MSE(c, x)—mean square error of climatological forecasts; $s_{f'}$—standard deviation of forecast anomalies; $s_{x'}$—standard deviation of initialized anomalies.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.961	0.923	0.001	0.000	0.923	430.	0.988
2	0.922	0.850	0.001	0.000	0.851	833.	0.956
3	0.877	0.768	0.001	0.003	0.769	1295.	0.902
4	0.735	0.540	0.025	0.000	0.523	2672.	0.892
5	0.563	0.317	0.142	0.000	0.189	4544.	0.939
6	0.588	0.345	0.094	0.001	0.263	4128.	0.895
7	0.419	0.176	0.306	0.019	−0.129	6329.	0.972
8	0.375	0.141	0.467	0.028	−0.332	7463.	1.059
9	0.539	0.290	0.216	0.010	0.080	5158.	1.004
10	0.204	0.042	0.511	0.071	−0.514	8486.	0.919
				D = 0.017	MSE(c, x) = 5605.		$s_{x'}$ = 74.2

is a nondimensional measure of the unconditional (i.e., overall) bias in the forecast anomalies. It vanishes only for unconditionally unbiased forecast anomalies (i.e., $\langle f' \rangle = \langle x' \rangle$).

The term B represents the square of the difference between the anomaly correlation coefficient and the ratio of the standard deviation of the anomalies in the forecast heights to the standard deviation of the anomalies in the analyzed heights. In the context of the regression model, the estimated slope of the line, b, is equal to $(s_{x'}/s_{f'})r_{f'x'}$. It is evident, then, that the term B vanishes only when $b = 1$, an obviously desirable characteristic of such a line in the context of forecast verification. When $b \neq 1$, the conditional expected values of the anomalies in the analyzed heights are not equal to the corresponding anomalies in the forecast heights, implying that the latter are *conditionally* biased. Thus, the term B is a nondimensional measure of the conditional bias in the forecast anomalies.

Alternatively, given a linear relationship between f'_i and x'_i ($i = 1, \cdots, n$) of specified strength (i.e., given $r_{f'x'}$) and an analyzed field of anomalies with a known variance (i.e., given $s^2_{x'}$), B indicates the amount of variability in the forecast field of anomalies—namely, $s^2_{f'} = s^2_{x'}r^2_{f'x'}$—that is required to eliminate the conditional bias. Most NWP models—and certainly all climate models used for extended range prediction—are designed with the intent of maintaining natural atmospheric variability (i.e., with the objective of continuing to look "meteorological"). As a result, $s_{f'}$ is approximately equal to $s_{x'}$, which necessarily leads to the introduction of conditional bias as $r_{f'x'}$ decreases from one toward zero.

As noted earlier, the term A is the square of the anomaly correlation coefficient. Except for the slight correction associated with the term D, A is the skill score that would be achieved if both the conditional and unconditional biases could be eliminated. It is in this sense that $r^2_{f'x'}$ is referred to as a measure of *potential* rather than actual skill.

The numerical values of the terms in (13) will be influenced by the particular choice of a climatology. This choice can also affect the "contributions" of the various terms to the overall skill score. However, since the relative magnitudes of skill scores in comparative verification are determined solely by the respective

TABLE 2. Anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts verifying on 1 January 1987. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.975	0.951	0.000	0.001	0.951	433.	0.990
2	0.950	0.902	0.000	0.003	0.900	879.	0.936
3	0.878	0.770	0.000	0.020	0.755	2157.	0.892
4	0.752	0.566	0.028	0.019	0.530	4144.	0.918
5	0.638	0.407	0.093	0.028	0.303	6145.	0.943
6	0.624	0.389	0.066	0.023	0.316	6030.	0.882
7	0.444	0.197	0.235	0.069	−0.082	9543.	0.929
8	0.437	0.191	0.350	0.080	−0.210	10673.	1.029
9	0.518	0.268	0.319	0.061	−0.086	9573.	1.083
10	0.324	0.105	0.719	0.102	−0.676	14778.	1.172
				D = 0.024	MSE(c, x) = 8817.		$s_{x'}$ = 92.8

TABLE 3. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 1000 mb forecasts verifying in January 1987. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.967	0.936	0.004	0.001	0.932	434.	1.022
2	0.918	0.843	0.015	0.001	0.830	1097.	1.024
3	0.838	0.705	0.033	0.003	0.676	2107.	0.999
4	0.745	0.559	0.071	0.005	0.492	3360.	0.986
5	0.628	0.401	0.134	0.007	0.272	4771.	0.965
6	0.506	0.265	0.220	0.010	0.053	6226.	0.943
7	0.373	0.162	0.367	0.014	−0.197	7919.	0.941
8	0.244	0.088	0.578	0.018	−0.477	9817.	0.965
9	0.148	0.064	0.738	0.027	−0.665	11064.	0.963
10	0.091	0.039	0.814	0.038	−0.774	11801.	0.963
				D = 0.023	MSE(c, x) = 6665.		$s_{x'}$ = 80.2

mean square errors, which do not depend on climatology, the skill score SS(f, c, x) rather than its components on the rhs of (13) should be preferred in this context.

Before discussing the calculation of skill scores and anomaly correlation coefficients based on operational NWP forecasts, it is of interest to briefly examine equation (13) under the following conditions: (a) $\langle x' \rangle \ll s_{x'}$ (i.e., the mean anomaly in the analyzed field is close to zero); (b) $\langle f' \rangle - \langle x' \rangle \ll s_{x'}$ (i.e., the unconditional bias is negligible); and (c) $s_{f'} \approx s_{x'}$ (i.e., a realistic atmospheric variability is maintained in the forecast height field). Under these conditions, (13) reduces to

$$\mathrm{SS}(f, c, x) \approx r_{f'x'}^2 - (r_{f'x'} - 1)^2 \qquad (19)$$

or

$$\mathrm{SS}(f, c, x) \approx 2r_{f'x'} - 1. \qquad (20)$$

Thus, under these conditions, skill remains positive whenever the anomaly correlation coefficient exceeds 0.5, and the traditional criterion that $r_{f'x'}$ must exceed 0.6 for useful forecasts implies a skill score greater than 0.2 (cf. Hollingsworth et al. 1980).

3. Skill scores and correlation coefficients for MRF model

We now present and discuss the results of some calculations of the mean square error skill score and anomaly correlation coefficient based on operational MRF model output at NMC. The purpose of these calculations is not to evaluate the performance of the MRF model (for which our data and analyses are inadequate), but to examine the relationship between these measures of skill in a realistic context.

The operational MRF model during the period from which our data are taken (i.e., 1987) was an 18-layer spectral model with enhanced vertical resolution near the surface and an extensive physics package. A reasonably complete description of the model, as well as of the changes that the model has undergone in recent years, is given by White (1988).

Calculations reported here are limited to an R12 truncation of both the forecast fields and the initialized fields with which the former are compared. Forecast calculations were carried out at an R40 truncation. The climatology is the 5-year (1982–86) spectral climatology described by Epstein (1988). For all calculations the spatial domain has been limited to the latitude belt between 20° and 80°N.

TABLE 4. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts verifying in January 1987. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.982	0.965	0.001	0.002	0.962	497.	0.994
2	0.945	0.984	0.004	0.007	0.885	1535.	0.988
3	0.881	0.779	0.011	0.015	0.755	3322.	0.974
4	0.796	0.638	0.033	0.026	0.582	5694.	0.957
5	0.680	0.471	0.085	0.038	0.354	8786.	0.942
6	0.554	0.321	0.161	0.052	0.115	12115.	0.920
7	0.428	0.202	0.266	0.068	−0.122	15483.	0.914
8	0.304	0.113	0.412	0.083	−0.370	18978.	0.915
9	0.198	0.069	0.560	0.094	−0.573	21858.	0.909
10	0.132	0.051	0.670	0.108	−0.711	23750.	0.912
				D = 0.009	MSE(c, x) = 13945.		$s_{x'}$ = 117.0

TABLE 5. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts verifying in January 1987 over quadrant 1. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.989	0.978	0.001	0.007	0.972	424.	1.004
2	0.959	0.920	0.005	0.014	0.909	1432.	1.012
3	0.892	0.798	0.021	0.024	0.773	3585.	1.005
4	0.795	0.640	0.052	0.032	0.591	6504.	0.988
5	0.662	0.462	0.122	0.065	0.334	10427.	0.957
6	0.525	0.314	0.202	0.105	0.085	14382.	0.907
7	0.387	0.195	0.338	0.147	−0.186	18598.	0.887
8	0.249	0.115	0.465	0.190	−0.423	22463.	0.854
9	0.099	0.078	0.580	0.201	−0.572	25714.	0.798
10	0.021	0.080	0.650	0.248	−0.673	27525.	0.748
				D = 0.084	MSE(c, x) = 16723.		$s_{x'}$ = 122.0

Initially we examine the scores for the ten forecasts with lead times ranging from one to ten days, all of which verified on 1 January 1987. The anomaly correlation coefficients, mean square error skill scores, terms in the decomposition of the skill score, and other relevant statistics for the 1000 and 500 mb geopotential height field forecasts are presented in Tables 1 and 2. For the forecasts verifying on this particular day, the anomaly correlation coefficient falls below 0.5 on day 7 (i.e., seven days in advance) at both the 1000 and 500 mb levels. The skill score first becomes negative at exactly this same lead time. As measured by SS(f, c, x), a very steep drop in skill occurs between days 4 and 5 at the 1000 mb level; at the 500 mb level, the skill remains above 0.3 through day 6. The anomaly correlation coefficient falls below 0.6 on day 5 at 1000 mb and on day 7 at 500 mb. Standard deviations of the forecast fields ($s_{f'}$) vary somewhat from day to day but they do not depart systematically from the standard deviation of the initialized fields ($s_{x'}$). Since the unconditional bias terms (C) are not large (especially through day 4), the expression for SS(f, c, x) in (20) represents a very close approximation to the actual relationship between the skill score and the anomaly correlation coefficient for these data.

The results in Tables 1 and 2 refer to a single day. We now consider the average values of these same quantities over the entire month of January 1987. These results are shown in Tables 3 and 4 for the 1000 and 500 mb levels. Apparently the results for 1 January were not entirely atypical. The month's average anomaly correlation coefficient, at both levels, is above 0.6 through day 5 and above 0.5 through day 6; the average skill score is above 0.2 through day 5 and remains positive through day 6. Conditional biases contribute most strongly to the difference between potential skill, represented by A, and actual skill, represented by SS, especially after day 4 and more noticeably at the lower level. The unconditional bias is larger at 500 mb; it is a relatively minor factor at 1000 mb. In both cases this latter bias remains sufficiently small, and $s_{f'}$ remains sufficiently close to $s_{x'}$, that the simple expression in (20) is a good approximation to the relationship between SS(f, c, x) and $r^2_{f'x'}$, at least as long as some indication exists of substantial actual or potential skill.

The calculations on which Tables 1–4 have been based involve data covering most of the Northern Hemisphere. They could equally well have dealt with the entire globe. It is perfectly feasible to make such calculations over all or part of the domain for which

TABLE 6. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts verifying in January 1987 over quadrant 2. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.984	0.967	0.002	0.007	0.963	308.	0.982
2	0.950	0.902	0.009	0.021	0.890	925.	0.985
3	0.880	0.779	0.020	0.042	0.753	2060.	0.981
4	0.804	0.652	0.048	0.101	0.555	3732.	0.985
5	0.687	0.481	0.113	0.156	0.292	6059.	0.984
6	0.565	0.342	0.199	0.233	0.016	8474.	0.960
7	0.433	0.238	0.375	0.317	−0.274	11008.	0.972
8	0.315	0.173	0.664	0.414	−0.654	13634.	0.994
9	0.158	0.104	0.953	0.503	−1.031	16776.	0.999
10	0.123	0.102	1.069	0.489	−1.085	17596.	1.018
				D = 0.132	MSE(c, x) = 8735.		$s_{x'}$ = 88.0

TABLE 7. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts verifying in January 1987 over quadrant 3. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.978	0.956	0.002	0.004	0.954	608.	0.992
2	0.941	0.887	0.007	0.009	0.878	1592.	0.984
3	0.892	0.799	0.018	0.018	0.775	2877.	0.978
4	0.825	0.691	0.060	0.030	0.622	4703.	0.973
5	0.736	0.554	0.113	0.033	0.440	6950.	0.958
6	0.648	0.444	0.211	0.044	0.230	9491.	0.945
7	0.551	0.341	0.340	0.067	−0.010	12405.	0.955
8	0.449	0.254	0.533	0.078	−0.285	15744.	0.984
9	0.369	0.206	0.699	0.074	−0.490	18420.	0.995
10	0.278	0.187	0.894	0.088	−0.706	21597.	1.004
				D = 0.069	MSE(c, x) = 14275.		$s_{x'}$ = 114.0

forecasts and analyses are available. Subregions of the complete domain may be defined in physical space (e.g., a hemisphere or a continent) or in phase space (e.g., groups of spectral coefficients or zonal wavenumbers). The process of subdividing the domain has some predictable effects on the nature of the decomposition of the skill score and on the comparison of this measure of performance with the anomaly correlation coefficient.

First we consider a geographical subdivision of the domain. For this purpose, the Northern Hemisphere between 20° and 80°N was subdivided into four equal quadrants, each of 90° longitudinal extent. Tables 5–8 contain the average statistics for the 500 mb level for January 1987 for the four separate regions. These results are based on the same forecasts and analyses that were employed in producing Table 4.

The terms involving mean values of the anomalies in the forecasts or the analyses—specifically, B, C, and D—are generally larger for the regional subdivisions than for the entire region. Note that individual values of the skill score for the separate regions are influenced by the ability of the MRF model to respond to interregional differences in mean values, but the anomaly correlation coefficients are concerned only with intraregional behavior. For higher quality forecasts—that is, for forecast lead times of three days or less—this difference is of relatively little concern because the conditional and unconditional biases (i.e., B and C) remain sufficiently small and their effects are somewhat offset by larger values of D. In this set of data, although the meteorological character of the four quadrants differs substantially (e.g., monthly averages of $s_{x'}$ vary between 88 and 122), the nature of the decrease in skill with increasing lead time is substantially the same in all four domains.

This latter result no longer holds when the subdivision is made in phase space according to zonal wavenumber. Tables 9–12 give the relevant statistics for zonal wavenumbers 0, 1–3, 4–7, and 8–12, respectively. (The mean square error over all wavenumbers 0–12 is the sum of the mean square errors of the individual waves. In the case of the subdivision in physical space, the mean square error for the entire domain is the average of the mean square errors for the four quadrants of equal area.) Zonal wavenumber 0 represents the zonal mean. All other waves have mean values that are identically equal to zero. Thus, for wavenumbers greater than zero, C and D are identically equal to zero; for wavenumber 0, these mean values become quite large. It can be seen in Table 9 that, for the day 4 forecast, the skill score is negative in spite of the fact

TABLE 8. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts verifying in January 1987 over quadrant 4. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.980	0.961	0.002	0.011	0.949	626.	0.987
2	0.926	0.867	0.014	0.033	0.828	2103.	0.963
3	0.834	0.715	0.025	0.053	0.650	4514.	0.924
4	0.728	0.548	0.052	0.076	0.442	7445.	0.888
5	0.598	0.391	0.154	0.118	0.156	11145.	0.882
6	0.461	0.258	0.313	0.168	−0.173	15222.	0.888
7	0.348	0.175	0.448	0.201	−0.422	18722.	0.886
8	0.209	0.125	0.682	0.252	−0.741	22533.	0.885
9	0.133	0.098	0.813	0.293	−0.928	24796.	0.889
10	0.106	0.088	0.887	0.361	−1.072	26369.	0.916
				D = 0.033	MSE(c, x) = 15085.		$s_{x'}$ = 117.3

TABLE 9. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts of zonal wavenumber 0 verifying in January 1987. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.984	0.969	0.016	0.066	0.921	54.	0.976
2	0.947	0.898	0.089	0.212	0.728	171.	0.984
3	0.862	0.770	0.133	0.534	0.393	368.	0.980
4	0.762	0.658	0.253	0.927	−0.025	608.	0.988
5	0.562	0.554	0.585	1.261	−0.585	946.	1.040
6	0.485	0.545	0.875	1.696	−1.182	1343.	1.107
7	0.414	0.497	1.275	2.084	−1.900	1810.	1.218
8	0.412	0.481	1.806	2.533	−2.747	2281.	1.329
9	0.389	0.440	2.131	2.738	−3.131	2640.	1.460
10	0.380	0.378	2.897	3.070	−3.913	3019.	1.607
				D = 0.300	MSE(c, x) = 816.		$s_{x'}$ = 25.2

that, on average, $r_{f'x'} = 0.76$ for January 1987. This result is due in large measure to the quite substantial and persistent value of the unconditional bias. A tendency also exists for the model to generate, with time, much more variance in the zonal mean than is found in the atmosphere, and this tendency contributes to the excessive growth of the conditional bias (i.e., the term B).

For the other wavenumbers, for which $C = D = 0$, we are left with the very simple relationship that SS $= A - B$. That is, the mean square error skill score is the square of the anomaly correlation coefficient minus the measure of conditional bias. For wavenumbers 1–3 some skill in the forecasts is retained out to day 7, aided by a decrease in $s_{f'}$ so that the term B remains relatively small. In the case of wavenumbers 4–7, positive skill is retained, on the average, out to day 6, in spite of the fact that the monthly mean value of $r_{f'x'}$ is only about 0.43 at that lead time. Such a result is possible because the model lost variance in those waves during the month of January 1987.

Note that for the shortest waves (wavenumbers 8–12—see Table 12) the model's loss of variance for the first two days of the forecast is so severe that $s_{f'}/s_{x'} < r_{f'x'}$. Statistically, the model underpredicts the wave amplitudes. This result implies that the forecast of these wavenumbers should be *inflated* to achieve a higher mean square error skill score. If such underprediction is a persistent feature of the forecasts, then modelers are indeed justified in their concern for the loss of energy at higher wavenumbers. Elsewhere, however, the model overpredicts the wave amplitudes, at least from the perspective of what is actually predictable, and it is possible to increase the skill score by *deflating* the wave amplitudes.

4. Conclusion

In this paper we have compared the mean square error skill score and the anomaly correlation coefficient, as model verification measures, by exploiting a recently developed method of decomposing such skill scores. Specifically, a climatological skill score was decomposed into terms involving the anomaly correlation coefficient, the conditional bias in the forecast, the unconditional bias in the forecast, and the difference between the long-term and sample climatologies. This decomposition suggests that it is reasonable to interpret the square of the anomaly correlation coefficient as a measure of *potential* rather than actual skill. Applica-

TABLE 10. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts of zonal wavenumbers 1–3 verifying in January 1987. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.986	0.972	0.002	0.000	0.970	217.	1.007
2	0.958	0.919	0.006	0.000	0.912	670.	1.004
3	0.914	0.837	0.012	0.000	0.824	1390.	0.991
4	0.855	0.732	0.026	0.000	0.706	2337.	0.980
5	0.763	0.587	0.067	0.000	0.520	3848.	0.976
6	0.637	0.418	0.142	0.000	0.275	5809.	0.959
7	0.505	0.276	0.247	0.000	0.030	7788.	0.950
8	0.364	0.163	0.408	0.000	−0.245	10048.	0.951
9	0.230	0.110	0.575	0.000	−0.465	11979.	0.926
10	0.162	0.088	0.660	0.000	−0.571	12813.	0.908
				D = 0.000	MSE(c, x) = 8097.		$s_{x'}$ = 89.4

TABLE 11. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts of zonal wavenumbers 4–7 verifying in January 1987. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.982	0.964	0.014	0.000	0.963	144.	0.983
2	0.936	0.878	0.006	0.000	0.871	490.	0.972
3	0.845	0.724	0.019	0.000	0.702	1165.	0.947
4	0.711	0.524	0.067	0.000	0.457	2166.	0.916
5	0.555	0.342	0.137	0.000	0.205	3155.	0.865
6	0.427	0.231	0.206	0.000	0.025	3924.	0.825
7	0.308	0.155	0.316	0.000	−0.161	4701.	0.795
8	0.201	0.112	0.439	0.000	−0.327	5401.	0.772
9	0.139	0.090	0.508	0.000	−0.418	5830.	0.761
10	0.014	0.066	0.658	0.000	−0.591	6535.	0.757
				D = 0.000	MSE(c, x) = 4170.		$s_{x'}$ = 63.9

tion of the decomposition to MRF model output for the 1000 and 500 mb geopotential height fields indicates that after about four days the actual skill of the forecasts, as measured by the climatological skill score, is appreciably less than their potential skill, as measured by the anomaly correlation coefficient. This result is due principally to the appearance of substantial conditional biases in the forecasts. These biases, and the corresponding loss of skill, can be interpreted as the penalty associated with retaining "meteorological" features in the geopotential height field when such features are not predictable.

In terms of comparative verification, the climatological skill score and anomaly correlation coefficient are largely but not entirely equivalent, except in special circumstances (see section 3). That is, certain ranges of values of the anomaly correlation coefficient are generally associated with certain ranges of values of the climatological skill score. For example, anomaly correlation coefficients in the range from 0.5 to 0.6 usually correspond to climatological skill scores in the range from zero to 0.2. In addition, the relative values of the anomaly correlation coefficients can depend on the particular choice of climatology.

However, in the following absolute sense, the skill score provides a more "honest" assessment of performance than the anomaly correlation coefficient. First of all, as indicated by the decomposition of SS(f, c, x) in (13), it is the square of the anomaly correlation coefficient rather than the anomaly correlation coefficient itself that describes the (potential) level of performance. That is, an anomaly correlation coefficient of 0.6 corresponds to a potential skill level of 0.36, where unity represents perfection. Moreover, potential skill (as measured by the square of the anomaly correlation coefficient) generally represents an upper limit on actual skill (as measured by the climatological skill score), thereby implying that the latter may be considerably less than the former in some cases. The existence of substantial differences between actual and potential skill is supported by the results presented in this paper. Thus, an anomaly correlation coefficient of 0.6 may be "translated" into a skill score of approximately 0.2 in some cases, revealing that, in reality, the forecast is only 20% (rather than 60%) of the way toward a perfect forecast.

In a related vein, the decomposition of the climatological skill score in (13) indicates that model verification performed solely in terms of the anomaly correlation coefficient is "incomplete." That is, such a verification necessarily ignores the conditional and unconditional biases in the forecasts, as well as any difference between the mean anomaly in the analysis and its climatological norm. At a minimum, it would seem

TABLE 12. Averages of anomaly correlation coefficients, climatological skill scores, terms in decomposition of skill scores, and other statistics for 500 mb forecasts of zonal wavenumbers 8–12 verifying in January 1987. See legend of Table 1 for key to symbols.

LT	$r_{f'x'}$	A	B	C	SS	MSE(f, x)	$s_{f'}/s_{x'}$
1	0.946	0.895	0.006	0.000	0.888	80.	0.907
2	0.862	0.745	0.012	0.000	0.733	193.	0.852
3	0.718	0.530	0.045	0.000	0.485	368.	0.844
4	0.576	0.364	0.127	0.000	0.237	533.	0.848
5	0.401	0.220	0.282	0.000	−0.061	767.	0.837
6	0.224	0.148	0.432	0.000	−0.284	927.	0.796
7	0.128	0.107	0.520	0.000	−0.413	1034.	0.778
8	0.102	0.103	0.575	0.000	−0.473	1055.	0.748
9	−0.001	0.073	0.732	0.000	−0.658	1194.	0.770
10	0.042	0.103	0.691	0.000	−0.588	1143.	0.767
				D = 0.000	MSE(c, x) = 742.		$s_{x'}$ = 27.0

appropriate to compute these additional terms, as well as the anomaly correlation coefficient, to obtain a more complete and realistic assessment of model performance.

In conclusion, this paper has focused solely on the anomaly correlation coefficient and its relationship to a mean square error skill score based on a climatological standard of reference. However, it also is possible to decompose such a skill score into terms involving (*inter alia*) the tendency correlation coefficient. (The tendency correlation is defined as the correlation between the forecast and analyzed deviations from the initial values of the variable of interest.) This decomposition would contain terms that are similar in appearance and interpretation to the terms that were involved in the decomposition described here.

Acknowledgments. We acknowledge the helpful comments of H. van den Dool, H. J. Thiebaux, C.-H. Yang, and an anonymous reviewer on an earlier version of the manuscript. This research was supported in part by the National Science Foundation (Division of Atmospheric Sciences) under Grant ATM-8714108.

REFERENCES

Arpe, K., A. Hollingsworth, M. S. Tracton, A. C. Lorenc, S. Uppala and P. Kallberg, 1985: The response of numerical weather prediction systems to FGGE level IIb data. Part II: Forecast verifications and implications for predictability. *Quart. J. Roy. Meteor. Soc.*, **111**, 67–101.

Brier, G. W., and R. A. Allen, 1951: Forecast verification. *Compendium of Meteorology*. T. F. Malone, Ed., Amer. Meteor. Soc., 843–851.

Epstein, E. S., 1988: A spectral climatology. *J. Climate*, **1**, 88–107.

Hollingsworth, A., K. Arpe, M. Tiedtke, M. Capaldo and H. Savijaervi, 1980: The performance of a medium range forecast model in winter—impact of physical parameterizations. *Mon. Wea. Rev.*, **108**, 1736–1773.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.

——, and H. Daan, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*. A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

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

White, G. H., 1988: Systematic performance of NMC medium-range forecasts 1985–1988. Preprints, *Eighth Conf. on Numerical Weather Prediction*. Baltimore, Maryland, Amer. Meteor. Soc. 466–471.