Why Do Forecasts for "Near Normal" Often Fail?

HUUG M. VAN DEN DOOL AND ZOLTAN TOTH*

Cooperative Institute for Climate Studies, Department of Meteorology, University of Maryland, College Park, Maryland

(Manuscript received 28 May 1990, in final form 6 November 1990)

ABSTRACT

It has been observed by many that skill of categorical forecasts, when decomposed into the contributions from each category separately, tends to be low, if not absent or negative, in the "near normal" (N) category. We have witnessed many discussions as to why it is so difficult to forecast near normal weather, without a satisfactory explanation ever having reached the literature. After presenting some fresh examples, we try to explain this remarkable fact from a number of statistical considerations and from the various definitions of skill. This involves definitions of rms error and skill that are specific for a given anomaly amplitude. There is low skill in the N-class of a 3-category forecast system because a) our forecast methods tend to have an rms error that depends little on forecast amplitude, while the width of the categories for predictands with a near Gaussian distribution is very narrow near the center, and b) it is easier, for the verifying observation, to 'escape' from the closed N-class (2-sided escape chance) than from the open ended outer classes. At a different level of explanation, there is lack of skill near the mean because in the definition of skill we compare the method in need of verification to random forecasts as the reference. The latter happens to perform, in the rms sense, best near the mean. Lack of skill near the mean is not restricted to categorical forecasts or to any specific lead time.

Rather than recommending a solution, we caution against the over-interpretation of the notion of skill-by-class. It appears that low skill near the mean is largely a matter of definition and may therefore not require a physical-dynamical explanation. We note that the whole problem is gone when one replaces the random reference forecast by persistence.

We finally note that low skill near the mean has had an element of applying the notion forecasting forecast skill in practice long before it was deduced that we were making a forecast of that skill. We show analytically that as long as the forecast anomaly amplitude is small relative to the forecast rms error, one has to expect the anomaly correlation to increase linearly with forecast magnitude. This has been found empirically by Tracton et al. (1989).

1. Introduction

It has been noted for decades by many workers in long range weather forecasting that skill of categorical forecasts, when decomposed into contributions from each category separately, has a tendency to be low, if not absent, in the "near normal" (N) category. For instance Gilman (1986), in discussing categorical forecasts of winter mean temperature in three classes over the United States, noted how completely the skill of the forecasts is concentrated in the categories Below (B) and Above (A), see his Fig. 4. Similar kinds of observations were made by Namias (1964), Folland et al. (1986), Shabbar (1989), Epstein (1988), Toth (1989), Livezey et al. (1990), and privately by many others in several countries. This simple fact has had quite an impact on the way the official monthly and seasonal forecasts for the United States have been presented to the user and the public since the middle of 1982 (Gilman 1982; Lehman 1987). Livezey (1990) capitalized on the issue by presenting a map of skill over the United States calculated for only those cases when either A or B were forecast.

The lack of skill of forecasts for "near normal" weather (temperature mostly) may have been noted by many, but a credible discussion as to why this happens is missing in the literature. The problem has, to our knowledge, mostly been noted by long-range forecasters who have traditionally favored categorical statements, but as we shall see, it occurs at all lead times and is not restricted to just categorical forecasts.

Absence of skill in a given class (or portion of the frequency distribution of the predictand) would be a most interesting diagnostic: if we could understand why it happens, and find a cure to the problem, and if an increase of skill in the ill-performing class would lead to an overall improvement of the forecasts. The lack of skill in the N class is also interesting, because to

* *Present affiliation:* National Research Council, NOAA/NMC, Development Division, Washington DC.

On leave from: Central Institute for Weather Forecasting, Budapest, Hungary.

Corresponding author address: Dr. Huug van den Dool, Climate Analysis Center, National Meteorological Center, 5200 Auth Road, Camp Springs, MD 20746.

many laymen it must seem most unlikely that we do better at forecasting the outer classes than at forecasting near normal. It seems to defy logic, also because conservative regression forecasts dictate to us to forecast close to the mean.

In his essay Gilman (1985) explicitly states that we would very much like to forecast the extreme cases correctly, those "real terrible months," but that at present all we can do is to alter slightly the frequency distribution of expected weather categorized in fairly coarse classes. "The tails of the conditional frequency distribution are the most uncertain thing about it" (Gilman 1985). At first sight, this emphatic statement seems inconsistent with the notion that all skill resides in the A and B classes. For later reference we call this the Gilman paradox. From examples given in section 3, we get the impression that "skill" (as we measure it) increases monotonically away from the mean, so the paradox cannot be solved by concluding that all skill is in the shoulders of the distribution.

Our discussion obviously requires definitions of 'skill' and what we mean by "better." Right from the outset we must be prepared to find that the whole problem of low skill in a given class could be a matter of a) definition, b) framing the forecasts, c) the particulars of the verification method and d) over-interpretation of the notion "skill by class." It is possible, however, that there are intrinsic properties in the atmospheric dynamics/physics that would make forecasts for certain types of anomalous weather easy, and forecasts of non-descript, near-normal weather more difficult. In synoptic terms, it may well be that once the circulation is strongly anomalous, it is relatively easy to forecast a continuation of the outer weather class. While close to the mean, the success of the forecast depends too much on details whose predictability (under all circumstances) is limited.

Note also the following: the skill of forecasts is quoted to be high when the *forecast* calls for anomalous weather, i.e., there has been a clear element of forecasting forecast skill to it long before it was deduced that we were making a forecast of that skill. Forecasting forecast skill only recently has become a popular topic because of efforts made in numerical weather prediction to understand why certain forecasts are very good and others bad. We argue that Branstator (1986), who found that 72-h Northern Hemisphere height forecasts score higher anomaly correlation when the forecast height anomaly itself is large, was dealing with a similar issue. Nap et al. (1981) found the following to be true as well: whenever the *observed* weather itself turns out to be anomalous, the skill of categorical forecasts is found to be high.

In this article we first define the phenomenon (section 2), then give some unpublished well-documented examples (section 3) and discuss likely explanations for the lack of skill in the N class (section 4). The primary purpose is to increase our understanding as to why it happens. Conclusions and discussion are presented in section 5.

2. The problem

Suppose we make a total of M categorical forecasts of a weather element, say temperature. The frequency distribution of observed temperature can be used to define three climatologically equiprobable classes named A(bove), N(ormal), and B(elow). The (overall) skill score S may be defined as

$$S = \frac{H - E}{M} * 100, \qquad (1)$$

where H is the number of correct forecasts and E is the number of correct forecasts expected a priori by chance, i.e., $M/3$ for three classes. S measures precisely how many hits per 100 forecasts we score over and above the hits expected by chance. A forecast is a hit when the forecast class is correct. Although measures of skill are somewhat arbitrary, they always have in common a comparison of the forecast method in need of verification to a standard of reference (Murphy and Daan 1985). For (1) the chosen reference to determine E could be the random forecast, and a forecast method that scores more/less hits than random forecasts is said to have positive/negative skill score S.

S is an overall skill score measure and can be decomposed as follows:

$$S = S_{\mathrm{A}} + S_{\mathrm{N}} + S_{\mathrm{B}}, \qquad (2)$$

where $S_{\mathrm{A}} = \dfrac{H_{\mathrm{A}} - E_{\mathrm{A}}}{M} * 100$ etc.

Here H_{A} is the number of hits in the A class, E_{A} is the number of hits expected by chance for the A class, and M remains as before the total number of forecasts. If M_{A} is the number of forecasts landing in the A class, we often tacitly assume $E_{\mathrm{A}} = M_{\mathrm{A}}/3$ (which would'be $M/9$ if the frequency distribution of the forecasts equals that of the observations).

The problem noted by Gilman (1986) and investigated here can be stated as follows:

$$S \approx S_{\mathrm{A}} + S_{\mathrm{B}}, \quad \text{and} \quad S_{\mathrm{N}} \approx 0 \quad \text{or small.}$$

Obviously, if we could somehow elevate S_{N}, we are, perhaps naively, entitled to expect a higher overall S.

In the United States, the Heidke skill score is often used:

$$SS = \frac{H - E}{M - E} * 100, \qquad (3)$$

where all symbols have the same meaning as in (1). Using SS instead of S we can state the problem in nearly the same way. We will mostly use S, however, because skill decomposition is slightly more complicated when using SS, particularly for non-equiprobable classes.

Eq. (2) can be applied straightforwardly to any number of classes of any imaginable width. For convenience we will use one more skill measure Q, defined as

$$Q_I = \frac{H_I - E_I}{M_I} * 100, \qquad (4)$$

which reads like a legitimate skill score for class I (I = A, B or N). Obviously we can write

$$S = \frac{M_A}{M} Q_A + \frac{M_N}{M} Q_N + \frac{M_B}{M} Q_B.$$

If the frequency distribution of the forecasts equals that of the observations, the information gathered from inspecting the Q_Is is identical to that obtained from the S_Is in (2), i.e., $S_I/Q_I = M_I/M$. For instance $S_I = Q_I/3$ for three equiprobable classes (I = A, B, N).

With this straightforward framework in mind, we now proceed to present three examples of forecasts that lack skill in the N class. It is only for later reference that we point out here two important facts about skill as expressed by (1) and (4). The first is that two-class errors are no more damaging to the skill scores than one-class errors. The second is a difficulty in determining E_I in (4)—unless stated otherwise, E_I will be $M_I/3$ for three equiprobable classes.

3. Examples

Three examples will be presented in some detail.

The first is a specification experiment. Suppose we knew the atmospheric circulation in advance, how much skill would we have regarding forecasts of the daily maximum and minimum surface air temperature? In order to investigate this, we used 30 years of daily minimum and maximum temperatures at Budapest (1 January 1951–31 December 1980) and the daily circulation catalog developed for Hungary by Peczely (1983). For each month (i.e., January) a conditional anomaly temperature frequency distribution in five equiprobable classes was made for each circulation type using daily data of all 30 Januaries. Then, for each day in all 30 Januaries a temperature forecast is made as follows. First we identify the Peczely circulation type valid on the day in question. The categorical temperature forecast is then the most likely of the five classes in the conditional frequency distribution associated with that particular Peczely type. These forecasts of instantaneous weather do not refer to any lead time in particular—using them gives an upper limit to forecast skill provided we know the circulation type in advance. Note that we did not apply our procedure to independent data, but to the developmental 1951–80 period. Table 1a gives the verification of these forecasts, namely the contingency table, the skill score S, and the contributions per class calculated as gross-averages over all years and months. It is immediately clear that, by our skill measure S, almost all skill resides in the outer classes. In fact, skill has a tendency to increase with distance from the center, i.e., the Much Above and Much Below classes contribute heaviest to S. We note only in passing by how low the overall skill is. The results for maximum and minimum temperature are consistent. We repeated all calculations using an objective definition of circulation types over the Atlantic-European region (Bartholy et al. 1984). This gave us essentially the same results. One specific reason here as to the small contribution of the N class to S (less than 1.0) is that the N class is rarely the most likely category associated with a given circulation type, i.e., we rarely forecast the N class. But also when we measure skill per class, using Q, we find the outer classes to perform better by a factor of more than 2.

For comparison, we show in Table 1b the results of *operational* forecasts made at the National Meteorological Center (NMC) during fall 1957 to summer 1963 for the mean temperature over the United States averaged over day 2 through day 6. These forecasts have a skill rather comparable to S reported in Table 1a and are likewise verified in a 5 class system (although not quintiles). Although the details are quite different, we see a major similarity in that most skill is contributed by the outer classes, particularly when viewing it through the Q measure.

The second example is based on the 3000 12-h 500-mb height forecasts made by a limited area analogue method described in Van den Dool (1989). The observed initial heights at 38°N, 80°W from 27 January

TABLE 1a. The contingency table, skill and skill decomposition of specification of maximum and minimum temperature at Budapest provided that we have perfect knowledge of the Peczely circulation type ahead of time. Five equiprobable classes are used. For minimum temperature the contingency table has been omitted.

	Maximum temperature					
F\O	MB	B	N	A	MA	Σ_F
MB	11.4	7.3	5.4	3.8	1.9	29.9
B	3.6	5.5	3.4	2.9	1.7	17.2
N	1.4	1.9	2.4	1.7	1.0	8.4
A	1.5	2.3	3.3	4.6	3.6	15.4
MA	1.9	3.0	5.5	6.8	11.7	29.1
Σ_O	19.9	20.1	20.0	19.9	20.0	100.0%
	MB	B	N	A	MA	Overall
Q_I	18	12	8	10	20	
S_I	5.5	2.0	0.7	1.5	5.9	S = 15.6
	Minimum temperature					
Q_I	15	10	6	9	15	
S_I	4.9	1.2	0.8	1.3	4.4	S = 12.6
	MB	B	N	A	MA	Overall

The five classes are MB = much below normal, B = below normal, N = normal, A = above normal and MA = much above normal.

TABLE 1b. As Table 1a but now for the NMC's operational 2–6 day mean temperature forecast during Fall 1957–Summer 1963. The extreme categories have 12.5% climatological probability, the three center categories, 25% each (Courtesy of Dr. D. L. Gilman).

F\O	MB	B	N	A	MA	Σ_F
MB	3.9	3.6	1.2	0.6	0.1	9.4
B	5.3	11.5	7.5	5.4	1.2	31.0
N	1.7	6.2	6.7	6.4	2.1	23.1
A	0.9	4.7	7.1	10.5	3.0	28.3
MA	0.1	0.6	1.3	3.0	3.1	8.2
Σ_O	11.9	26.7	23.9	25.9	11.6	100.0%

	MB	B	N	A	MA	Overall
Q_I	29	10	5	11	26	
S_I	2.8	3.2	1.2	3.2	2.1	S = 12.5

0000 UTC through 5 February 1200 UTC during 1963–77 are used to determine three equiprobable classes empirically. The original point height forecasts are, for the current purposes, transformed to categorical forecasts simply by checking in which class the point forecasts fell. The results are summarized in Table 2a. The overall skill ($S = 42.6$) is quite high but the decomposition shows that 75% of the skill is due to A and B. Note that 2-class errors are virtually absent, a feature not honored in the way we calculate skill via (1). Each of the possible 1-class errors occurs about 6% of the time. While forecasts for N may fail on both sides (i.e., the observations "escape" to A or B), forecasts for A or B have only one-sided fail chances.

Because the frequency distribution of observed and forecast heights are almost the same, there is no need to discuss the Q-measure.

For comparison, Table 2b shows a skill analysis of 3000 persistence forecasts verifying at the same place and times as the analogue forecasts. Although S(37.0) is slightly lower than that for the analogue forecasts, the results in Tables 2a and 2b are surprisingly similar. Apparently even a forecast as simple as persistence suffers from a lack of skill for near normal weather. The decrease in overall skill (from analogue to persistence methods) is most clearly seen in a decrease of skill in the N-class. By increasing the 1-class errors by about 1%, the skill in the N class obviously suffers more than the skill in the outer classes.

TABLE 2a. The contingency Table, the skill score and its decomposition for 3000 12-h forecasts of 500-mb height at 38°N, 80°W using a limited area analogue method. Three nearly equiprobable classes are used. In order to avoid artificial skill we carry the exact frequency of occurrence for each class as indicated in the lines where the calculation of S_A etc. is explicitly given.

F\O	A	N	B	Σ_F
A	27.6	5.7	0.1	33.4
N	5.9	22.9	6.7	35.5
B	0.1	5.5	25.5	31.1
Σ_O	33.7	34.0	32.3	100.0%

$S_A = 27.6 - 33.4*33.7/100.0 = 16.4$

$S_N = 22.9 - 35.5*34.07/100.0 = 10.8$

$S_B = 25.5 - 31.1*32.3/100.0 = 15.4$

	A	N	B	Overall
Q_I	49	30	50	
S_I	16.4	10.8	15.4	$S = 42.6$

TABLE 2b. As Table 2a but no persistence forecasts.

F\O	A	N	B	Σ_F
A	25.7	7.3	0.3	33.3
N	8.0	19.0	6.3	33.3
B	0.0	7.7	25.7	33.3
Σ_O	33.7	34.0	32.3	100.0%

	A	N	B	Overall
Q_I	43	23	45	
S_I	14.4	7.7	14.9	$S = 37.0$

The third, last, and lengthiest example is a constructed case, amenable to easy further analysis, that we will rely heavily upon in section 4 for the explanation. We will also use this example to move from skill for discrete classes to continuous measures of skill. Assume that successive atmospheric observations x_i are generated by a linear first order Markov process

$$x_{i+1} = \rho x_i + \epsilon_i \qquad (5)$$

where the random number ϵ_i is drawn from a normal distribution with zero mean and standard deviation 1. The autocorrelation is denoted by ρ, and $0 < \rho < 1$. The expected value of x_i is zero and its standard deviation (sd) is $(1 - \rho^2)^{-1/2}$. An "observer," who knows all past x_i without error, is asked to make a forecast for x_{i+1}. Since ϵ_i is not known in advance, the best forecast (certainly in root-mean-square (RMS) error sense) is given by

$$f_{i+1} = \rho x_i. \qquad (6)$$

For the purpose of our paper, (6) represents the most instructive forecast scheme which we name damped persistence (DP).

Assuming that ρ is perfectly known, the forecast error (fe) of DP will be

$$\mathrm{fe}_{i+1} = f_{i+1} - x_{i+1} = -\epsilon_i. \qquad (7)$$

Hence the expected RMS value of the forecast error is 1.

The most important feature to note in (7) is that the error does not depend on x_i itself. Independent of forecast magnitude (i.e., the absolute value of f_{i+1}), the expected rms error will be 1. By lower case rms we denote from now on the expected root-mean-square error as a continuous function of f_{i+1}, while the upper case "RMS" error is the root-mean-square error integrated over all possible forecast magnitudes (i.e., the regular definition). Because the expected rms error is always 1 we can say that DP is a forecast scheme with *uniform rms error*. (This remains true for $\rho \rightarrow 0$, but not for $\rho = 0$, a singular limit). Other schemes (a random forecast, and to a lesser extent, persistence) have rms errors that increase monotonically with $|f_{i+1}|$. (We shall argue in the discussion that many forecast schemes, including operational ones, have more or less uniform rms error.)

Taking the rms error as our sacred verification tool we could tell the user of the forecast that the DP method has uniform accuracy and is equally reliable for extreme (large $|f_{i+1}|$) and close to normal ($|f_{i+1}| \approx 0$) weather events.

But now consider a verification of a three-class categorical forecast in the same DP setting. To this end we have generated 50 000 x_is through (5), choosing $\rho = 0.40$. The tercile borders are chosen as ±0.48 (i.e., 0.43∗sd) to create three nearly equiprobable classes for x_i. Forecasts f_{i+1}, made by (6) are said to be for the A, N, or B category if the f_{i+1} (a point forecast) fell in that class. The verification results are given in Table 3a. While the observed frequencies are, by construction, close to ⅓ for each class, the forecast frequencies are obviously very much biased towards the N class. From Table 3a we calculate $S = 8.0$ and, because it is obvious that most skill resides in the outer classes, we would be tempted to agree with Gilman (1986). We now have an apparent contradiction in that forecasts made by DP have uniform accuracy and yet very little skill in the N class.

In Table 3a most forecasts are for the N class. A common procedure to avoid forecasting N too often is called forecast inflation (Klein et al. 1959). Table 3b is given to show that if we "jack up" the forecasts by reducing the tercile borders for f_{i+1} to ±0.22 we increase the total skill (consistent with Glahn and Allen 1966) but have not changed the picture regarding low skill in the N class (in fact it decreased). We have repeated the above experiments for a) additional samples of 50 000 forecasts, b) for a variety of ρ values, c) using the Heidke skill score and found the result (low skill in the N class) to be absolutely robust.

TABLE 3a. As Table 2a but now for data generated by a first order Markov process and damped persistence as the forecast method. The autocorrelation is 0.4 and the tercile class limits (both for forecasts and observations) are + and −0.48 (i.e., 0.43∗sd).

F\O	A	N	B	Σ_F
A	7.8	4.1	1.8	13.7
N	23.5	25.1	23.4	72.0
B	1.8	4.0	8.4	14.2
Σ_O	33.1	33.3	33.5	100.0%
	B	N	A	Overall
Q_l	24	1.6	25	
S_l	3.3	1.1	3.6	$S = 8.0$

TABLE 3b. As Table 3a but now the tercile class limits are + and −0.22 for the forecasts and + and −0.48 for the "observations."

F\O	A	N	B	Σ_F
A	16.2	10.6	5.7	32.5
N	11.0	12.3	11.3	34.6
B	5.9	10.4	16.5	32.9
Σ_O	33.1	33.3	33.5	100.0%
	B	N	A	Overall
Q_l	17	2.2	17	
S_l	5.4	0.8	5.5	$S = 11.7$

In another set of experiments we widened the N class (for observations and forecasts alike) at the equal expense of A and B. The results are given in Table 3c. Although an N class covering slightly over 50% of the x_is achieves the goal of making S_N more or less equal to S_A and S_B, we also found the overall S to have gone down. We assume that high S is the first priority. From Daan (1985) we know that it may be very difficult, if not impossible, to compare S for different class configurations. Another indication for that uncomfortable circumstance is that Table 3b has higher S than Table 3a, even though an identical set of point forecasts are used. The inflated forecasts (Table 3b) have higher rms error and yet a higher skill score S. Using a wider N class leaves the rms error unchanged but reduces S, as indicated in Table 3c. (If maximizing S is a worthwhile goal, either inflated forecasts or shrinking the N class [down to a two class system] is recommendable.)

In spite of the message in Tables 3a and 3b, it would be counterintuitive to conclude that DP is particularly good for forecasting extreme weather. All we do is damp the initial anomaly back towards zero. We get high skill in the outer classes because extreme weather has a large predictable component (ρx_i), while success of forecasts following zero initial weather depends entirely on ϵ_i. Gilman's paradox may well concern our inability to forecast extreme weather when the current conditions are not extreme. This appears certainly the case for the analogue forecasts in Tables 2a and 2b. If we calculate the expected percent of hits from persistence as the reference forecast, we find $S = 5.6$ to decompose

TABLE 3c: As Table 3a (without contingency tables) but now for 5 choices of the tercile class limits. Going down in the table the N-class is widened.

Tercile limits	% of observations in N-class	S_B	S_N	S_A	S
±0.48	33.3	3.3	1.1	3.6	8.0
±0.60	40.9	2.3	1.2	2.6	6.1
±0.70	46.9	1.6	1.1	1.7	4.5
±0.80	52.8	1.1	0.9	1.1	3.1
±0.90	58.3	0.7	0.7	0.8	2.2

as 2.0, 3.1, and 0.5 for the A, N, and B classes, respectively, thus presenting evidence that, in some sense, we have more or less uniform skill. (This decomposition cannot exactly be recovered from the information supplied in Tables 2a and 2b because we need to know the scores of persistence of the initial state for all cases where the analogue forecast landed in a specific class. But to good approximation, the skill decomposition can be estimated from the difference of the diagonal elements in the matrices in Table 2a and 2b.)

4. Explanation

A generic definition of skill always involves a comparison of the forecast method to a reference which by common sense is considered to represent the zero skill level. When we use the most common of all attributes, the mean square error (MSE), the definition of skill would be (Murphy and Epstein 1989)

$$S = \frac{\text{MSE}_r - \text{MSE}_m}{\text{MSE}_r}, \tag{8}$$

i.e., the percent improvement in the MSE by our forecast method (index m) over the reference (index r). In order to write skill as a continuous function of distance (f) to the norm ($f = 0$), we introduce a local skill $s(f)$

$$s(f) = \frac{\text{mse}_r - \text{mse}_m}{\text{mse}_r}, \tag{9}$$

where the lower case quantities are thought to result from the infinite ensemble of forecasts having a specific forecast amplitude f. Assuming mse_m to be uniform and noting that mse_r for random forecasts ($f_{i+1} = R$) can be written (in the notation of the last example in section 3)

$$\text{mse}_{R(andom)} = \langle (R - x_{i+1})^2 \rangle = (\langle R^2 \rangle + \text{sd}^2)$$

we can write (9), for $R = f$ as

$$s(f) = 1 - \frac{\text{mse}_m}{f^2 + \text{sd}^2} \tag{9a}$$

where the angled brackets stand for the expectation and sd is the standard deviation of the x_is. (Expectation is an operator very much like averaging.) For the infinite collection of random forecasts that accidently land at the norm ($f = 0$), mse_R is expected to be sd^2; i.e., equal to the error of always forecasting climatology. For $|f| > 0$ mse_R increases monotonically with $|f|$, and so does skill defined by (9a), all the way up to $s = 1$ for very large forecast amplitude. (Eq. [9a] should not be construed to imply that we can reach $s = 1$ by simply amplifying the forecast.) In Fig. 1 we have sketched the situation for two levels of uniform rmse_m. The most important thing to note is that skill always increases away from the center monotonically. This is because the rms error of the reference method increases away from the center. The lack of skill near the center is caused by the tough competition of random forecasts landing near the norm. If $\text{rmse}_m >$ standard deviation, we may see negative skill near the origin even if the overall skill S is positive. If $\text{rmse}_m < \text{sd}$, we have a minimum in skill, although positive, right at the norm.

The explanation is, so far, based on mse as the attribute and may not be very relevant to the examples given in section 3. How does this carry over to skill based on actual and expected hits in a categorical verification system such as defined by (1)? Using the lower case convention, local (i.e., amplitude specific) skill, as in (4), is defined now by

$$q = \frac{h - e}{m} * 100, \tag{10}$$

where all variables refer to class I. One can imagine K equiprobable classes where K is very large allowing us to write q, h, and e as functions of forecast amplitude. Under those circumstances, e is uniformly m/K for random reference forecasts. We further suspect that $h(f)$ is proportional to the width of the class ($w(f)$), to m, and inversely proportional to rmse_m, in the way explained in Fig. 2. Hence, (10) can be written approximately as

$$q = \frac{cmw/\text{rsme}_m - m/K}{m} * 100, \tag{11}$$

where c is a proportionality constant. (The expression for h is strictly valid only as long as $\text{w}/\text{rmse}_\text{m}$ is small!) The constant c can be determined by first realizing that e ($= m/K$) can also be written, like h, as cmw/rmse_R. Since right at the center $\text{rmse}_R = \text{sd}$, we can evaluate $cmw(f = 0)/\text{sd}$ as m/K, and hence c, for all closed classes, is determined by $c = \text{sd}/(Kw(f = 0))$. So (11) can be written

$$q(f) = \frac{\dfrac{\text{sd}}{\text{rmse}_m} \times \dfrac{w(f)}{w(f = 0)} - 1}{K} * 100, \tag{11a}$$

q is a function of forecast amplitude f only through the argument $w(f)$. In this derivation, non-uniform q is seen to be associated by non-uniform w. Since, for a normal distribution, w increases away from the mean, q will likewise increase. This is, with reference to Fig. 2, because the observations (given the forecast) escape more easily from a narrow than from a wide class. Right at the center, q is negative if and only if $\text{rmse}_m > \text{sd}$, even if the overall skill is positive. This is the likely explanation of Livezey et al.'s (1990) negative skill in the N class (their Table 1). The fact that rmse_m is even larger than sd in operational methods, is a consequence of the habit of jacking up the forecast (Klein et al. 1959).

The derivation of (10)–(11a) was based on choosing K equiprobable classes, which, in (10), leaves only h

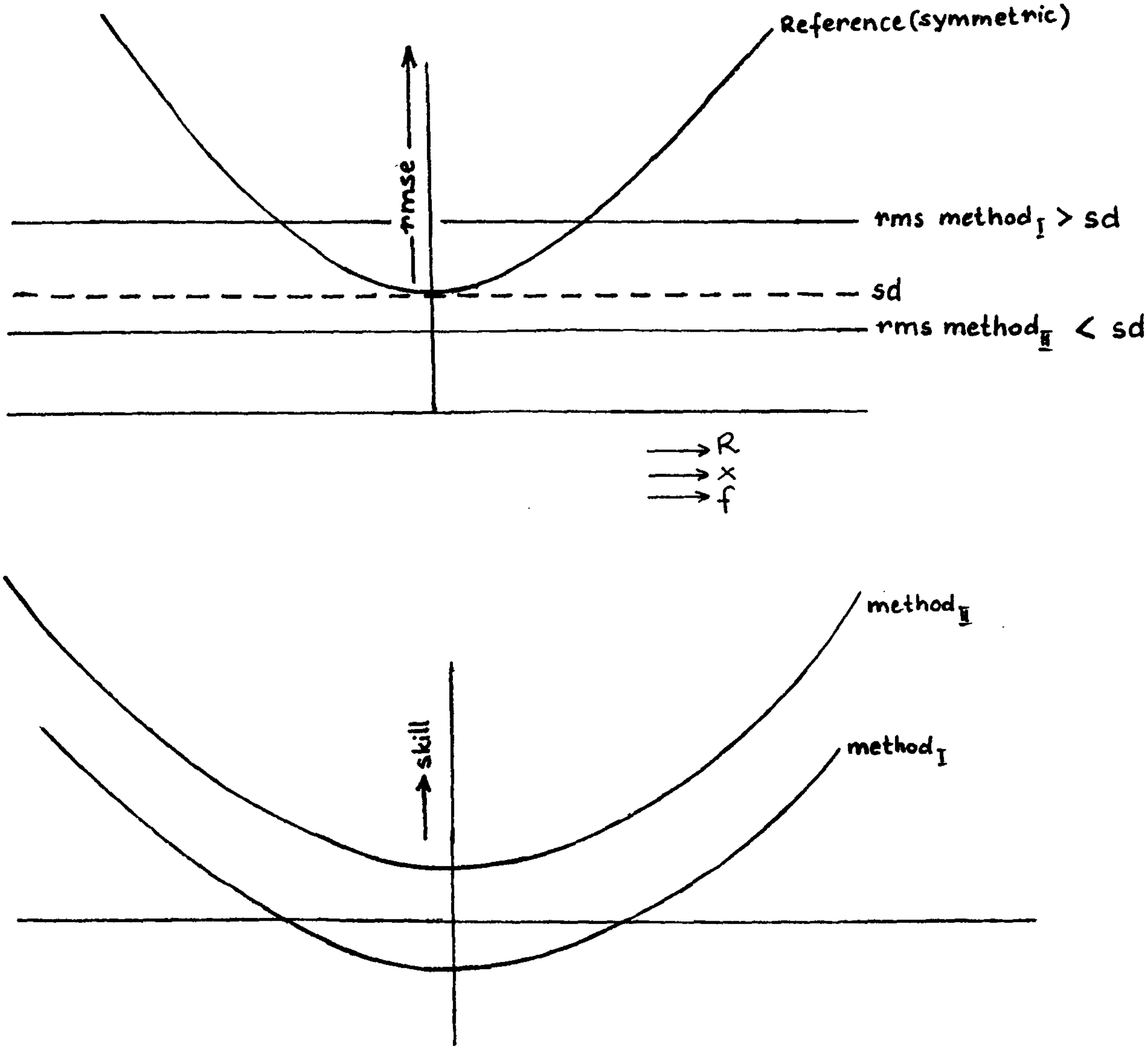

FIG. 1. A sketch of local skill using a definition based on the rms error as attribute, see Eq. (9a). Along the x-axis we have anomaly amplitude for observations (x), forecasts (f), and random forecasts (R). Presented are methods with an rms error smaller and bigger, respectively, than the standard deviation of x.

as a function of f. (This is a choice to make it easy to derive (11a), not an assumption.) If one chooses, alternatively, the classwidth $w(f)$ such that h, in (10), becomes independent of f, then the dependence of q on f is transferred to e, without any change in results.

The explanations based on hits and mse as attribute are somewhat similar but not identical. The only common factor is the importance of the ratio rmse$_m$/sd in determining whether skill near the center goes negative. In (11) we have low skill near the norm because w is small where the probability density is high, which is near the norm for a Gaussian distribution. It follows, therefore, that if the forecast variable had a uniform frequency distribution, local skill, as defined in, (11) would be uniform as well, and if the forecast variable had a U-shaped frequency distribution (example: tomorrow's cloud amount) we would find a puzzling abundance of skill in the N class. On the other hand the explanation based on (9) holds for any frequency distribution. This is because the rmse keeps increasing for bigger and bigger errors while a miss is only a miss no matter how large the actual error. Some definitions of skill punishing 2-class errors more severely than 1-class errors would require a mix of (9) and (11) to explain the lack of skill near the center.

It seems to us that (11a) explains, to a large extent, the lack of skill for near normal temperature or height in all examples given in section 3, and for almost all examples quoted in the Introduction. There is, however, one more very important detail. While c in (11)

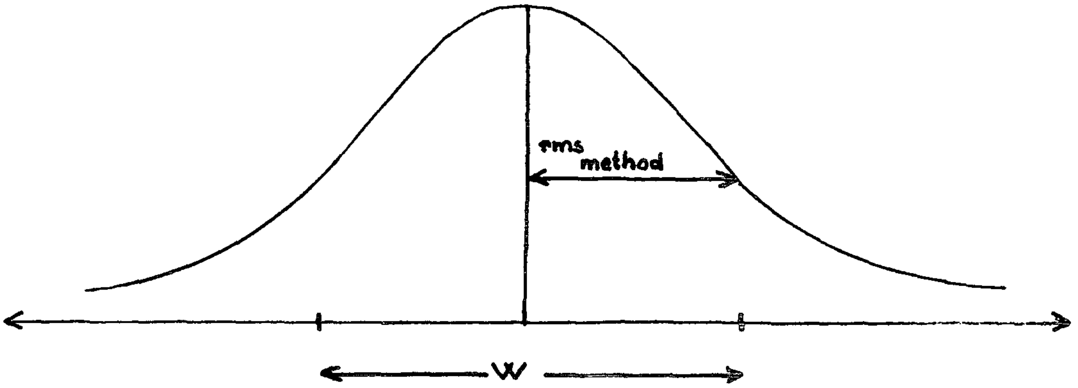

FIG. 2. A sketch of the notion escape chance. Along the x-axis we have class I of width w and portions of class I − 1 and I + 1. The curve represents the frequency distribution of observations given a forecast for class I. The verifying observations x_{i+1} will be drawn from this distribution which has $\langle x_{i+1} - f_{i+1} \rangle = 0$ and standard deviation $\mathrm{rmse}_{\mathrm{method}}$. The chance of a hit is clearly proportional to the ratio of w to $\mathrm{rmse}_{\mathrm{method}}$.

is the same for all closed classes, it is larger for the extreme open ended classes (to be precise: by a factor of two in the last example of section 3). As long as K goes to infinity this is a minor detail, but in practice $K = 5$ or $K = 3$ only. Therefore, in a 3-class system, even if the forecast variable has a uniform frequency distribution, the skill will be lower in the N class than in the outer classes.

5. Conclusions and discussion

It is not possible to say briefly why skill has been observed to be low in the near normal class and high in the outer classes. Superficially, it is because a) class width w increases relative to $\mathrm{rmse}_{m(ethod)}$ away from the mean, i.e., the likelihood of observations escaping the forecast class becomes smaller for wider classes and b) the likelihood of observations escaping the forecast class is considerably higher for a closed (N) than for an open-ended class (A and B). With respect to a), one might add at the next level of explanation that the low skill near the mean occurs because c) we use forecast methods that generally turn out to have more or less uniform rmse_m while d) the observations (at least temperature, height) tend to have frequency distributions peaked near the normal value. At yet another level of explanation we have low skill near the center because e) we compare an attribute of the method in need of verification to a similar attribute obtained by random forecasts as the reference. On top of all this, one can define skill in more than one way so as to make reasons a) to e) relevant in slightly different ways. An interesting conclusion is also that skill near the center goes negative if $\mathrm{rmse}_m > \mathrm{sd}$.

The above explanation has validity as long as rmse_m is more or less uniform and the random forecast is the reference. Uniform rmse is likely in all examples quoted in the above because, in our efforts to maximize overall skill S, we have to minimize RMSE_m, which is achieved by uniform rmse. Statistical forecasts would generally work that way. For example, regressing seasonal mean temperature in the United States against antecedent Southern Oscillation Index results in uniform rms error and the smallest possible RMSE in the temperature forecast.

The usage of the random forecast as the reference is our choice and could be replaced by any other. While for overall S in (1) it rarely matters (exception noted by Radok 1988) whether the reference forecast is a) random forecast, b) always climatology, i.e., always N-class, or c) always above (A) or always below (B). These choices become critically important when decomposing skill. At first sight we have no choice, in (9) and (11), but to use the random forecast as the reference. After all, if we always forecast climatology as reference then there is no mse_r to compare the mse_m to away from the mean. But then why do we want to compare mses (or hits) for the same distance ($|R| = |f_{i+1}|$) to the norm anyway? The reference forecast is not supposed to know what we are doing under the "method." An extreme pathological example: If "'always climatology" is the reference, and the method is a random forecast, then the total skill $S = 0$ will decompose into $S_{\mathrm{N}} = -\frac{2}{9}$ and $S_{\mathrm{A}} = S_{\mathrm{B}} = +\frac{1}{9}$. In order to prevent such pathological behavior, a linkage is often made between the method and the reference. For example, in (4) E_{I} would be assumed to be proportional to M_{I} (Livezey et al. 1990). If we let E_{I} be determined by persistence there is no linkage assumption, and yet the pathological behavior of skill by class is gone.

Of course there are "cures" for the problem of low skill in the N class (such as narrowing the N class to zero: a two class system, or transforming the predictand to have a uniform frequency distribution). None of these cures necessarily make for better forecasts, since the problem is largely definitional. We only mention that the use of persistence as the reference method (instead of the random forecast) makes skill rather uniform.

It is easy to criticize the notion of skill-by-class (or skill as a function of forecast amplitude) on several accounts. More than likely we are over-extending and over-interpreting a definition of skill never designed to be decomposed. Comparing skill in neighboring classes may be as meaningless and confusing as comparing over-all skill (even of a set of identical forecasts) as determined by, say, a 3- and a 5-class system (Daan 1985; Wagner 1989). Although it is factually true that skill in the N-class is low (as measured by (2) and (4)), it may not imply anything very interesting about our capability to forecast in general or about the physics of the atmosphere. The pathological example mentioned above (decomposing zero overall skill into $-2/9$ for N and $+1/9$ for A and B, respectively) is a case in point here. We add another example to show how questionable the decomposition is. Using Eq. (5), we generated data that ought to characterize one single unique red noise process in which the autocorrelation does not depend on forecast amplitude. Nevertheless, we can empirically calculate the autocorrelation (ρ_c; index c for calculated) from a long set of x_is, separately for those initial x_is that fall in the N-class and for those that fall in either A or B. The empirical ρ_c turns out large for A and B ($\rho_c > \rho$) and low ($\rho_c < \rho$) for the N-class. This is a bizarre result because from (5) we know that a single ? has been used.

An underlying problem of definitions like (1) is that only the diagonal elements of the contigency matrix (refer Tables 1, 2, and 3) are used to measure skill. A truly complete verification would take into account also the off-diagonal elements, and would for instance note that while the outer classes register more hits than the N-class they also suffer from more 2-class errors.

Some potentially good aspects of skill by class are a) it is one step in the direction of a more complete verification, and b) it is a practical application of forecasting forecast skill. A great difficulty is that the verification tools have an impact on the forecast, or at least on the way we present the forecast. Taking the RMS scores to be sacred, we are discouraged to make large amplitude forecasts (in fact at present skill levels, monthly and seasonal forecasts would nearly always fall in the N class), while the hit/miss based scores (using Eq. (1)) encourage or allow a more daring forecast. Discussions along these lines were published by Klein et al. (1959) and Glahn and Allen (1966), but it seems impossible to settle these issues.

A similar problem of low skill for flows with small anomalies has recently been encountered in the developing field of forecasting forecast skill. Branstator (1986) reports a modest correlation between the forecast anomaly amplitude and the anomaly correlation of dynamically produced 72-h Northern Hemisphere 500-mb height forecasts, thus implying a modest capability to know in advance when the prediction skill of atmospheric flow will be below/above average. As was shown by Murphy and Epstein (1989) the anomaly correlation behaves rather similar to the skill score used in (8). This led us to derive a local anomaly correlation as a function of forecast amplitude. The detailed derivation in the Appendix leads to

$$ac(f) = \frac{f/\mathrm{rmse}_m}{(1 + f^2/\mathrm{mse}_m)^{1/2}}. \qquad (12)$$

As a result we see that when $f^2/\mathrm{mse}_m \ll 1$, the anomaly correlation depends linearly on forecast magnitude as long as forecast accuracy is uniform. Therefore, as with the definitions of skill, we see the lowest *ac* near the center no matter how accurate the forecast. For larger f, *ac* depends less and less on f as far as definitional constraints go.

The analytically obtained relation (12) between *ac* and f is consistent with Tracton et al. (1989) who, in trying to elaborate on Branstator's result, were disappointed to find empirically that the relation between forecast amplitude and anomaly correlation score only holds up for small anomalies and low anomaly correlation. They blamed this on the "signal to noise ratio," implying that for small anomalies the verifying observations are uncertain due to observational error. From the above we can see that their explanation is only partly correct. It is a signal to noise ratio problem but the noise is not observational error (that is negligible) but forecast error itself. What we conclude here is that a forecast for a certain departure from normal becomes credible only if that departure has a favorable ratio to the forecast's rms error.

Acknowledgments. The authors thank the following colleagues for their helpful comments: D. L. Gilman, A. G. Barnston, A. J. Wagner, A. Johansson, W. E. Klein, E. S. Epstein, M. Cai, R. E. Livezey, P. Csapo, and L. Gandin. The research was supported by NOAA under grant NA89-AA-H-MC066.

APPENDIX

Derivation of Local Anomaly Correlation

As in Murphy and Epstein (1989) we start from

$$\mathrm{mse}_{m(ethod)} = \langle (f_{i+1} - x_{i+1})^2 \rangle,$$

where, in lower case convention, we consider the infinite ensemble of forecasts that has a specific forecast amplitude f. Obviously

$$\mathrm{mse}_m = f^2 + \langle x_{i+1}^2 \rangle - 2\langle f_{i+1}x_{i+1} \rangle.$$

A local *ac* is generically defined by

$$ac(f) = \frac{\langle f_{i+1}x_{i+1} \rangle}{f\langle x_{i+1}^2 \rangle^{1/2}}.$$

Even though the $\langle\ \rangle$ operator is taken over all cases of identical f, the covariance is not zero since we consider departures from the climate rather than departures from the mean f. In view of the above, $ac(f)$ can be written

$$ac(f)=\frac{(f^2+\langle x_{i+1}^2\rangle-\mathrm{mse}_m)/2}{f\langle x_{i+1}^2\rangle^{1/2}}.$$

Assuming locally unbiased forecasts, we obtain $\langle x_{i+1}^2\rangle = f^2+\mathrm{mse}_m$, which is valid exactly in the example in section 3, but holds only for reasonably small f in real forecast situations. It then follows

$$ac(f)=\frac{f}{(f^2+\mathrm{mse}_m)^{1/2}}=\frac{f/\mathrm{rmse}_m}{(1+f^2/\mathrm{mse}_m)^{1/2}}.$$

REFERENCES

Bartholy, J., P. Ambrozy and O. Gulyas, 1984: A system of seasonal macrocirculation patterns for the Atlantic-European region. *Idojaras,* **88,** 121–133.

Branstator, G., 1986: The variability in skill of 72-h global-scale NMC forecasts. *Mon. Wea. Rev.,* **114,** 2628–2639.

Daan, H., 1985: Sensitivity of verification scores to the classification of the predictand. *Mon. Wea. Rev.,* **113,** 1384–1392.

Epstein, E. S., 1988: Long-range weather prediction: Limits of predictability and beyond. *Wea. Forecasting,* **3,** 69–75.

Folland, C. K., A. Woodcock and L. D. Varah, 1986: Skill of the monthly forecasts. *Meteor. Mag.,* **115,** 377–395.

Gilman, D. L., 1982: The new look of the "Monthly and Seasonal Weather Outlook." *Proceedings of the Seventh Annual Climate Diagnostics Workshop*. U.S. Department of Commerce, p. 482.

——, 1985: Long-range forecasting: The present and the future. *Bull. Amer. Meteor. Soc.,* **66,** 159–164.

——, 1986: Expressing uncertainty in long-range forecasts. *Namias Symposium*. John O. Roads, Ed., Scripps Institution of Oceanography reference series 86-17, 174–187.

Glahn, H. R., and R. A. Allen, 1966: A note concerning "inflation" of regression forecasts. *J. Appl. Meteor.,* **5,** 124–126.

Klein, W. H., B. M. Lewis and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.,* **16,** 672–682.

Lehman, R. L., 1987: A model for decision-making based on NWS monthly temperature outlooks. *J. Climate. Appl. Meteor.,* **26,** 263–274.

Livezey, R. E., 1990: Variability of skill of long range forecasts and implications for their use and value. *Bull. Amer. Meteor. Soc.,* **71,** 300–309.

——, A. G. Barnston and B. K. Neumeister, 1990: Mixed analog/persistence prediction of United States seasonal mean temperatures. *Int. J. Climatology,* **10,** 329–340.

Murphy, A. H., 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 E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.,* **117,** 572–581.

Namias, J., 1964: A 5-year experiment in the preparation of seasonal outlooks. *Mon. Wea. Rev.,* **92,** 449–464.

Nap, J. L., H. M. van den Dool and J. Oerlemans, 1981: A verification of long range weather forecasts in the seventies. *Mon. Wea. Rev.,* **109,** 306–312.

Peczely, G., 1983: Catalogue of the macrosynoptic types for Hungary (1881–1983). In Hungarian. Meteorological Reports No. 53. Hungarian Meteorological Service, Budapest, 116 pp.

Radok, U., 1988: Chance behavior of skill scores. *Mon. Wea. Rev.,* **116,** 489–494.

Shabbar, A., 1989: Verification of Canadian monthly temperature forecasts. *Proceedings of the 13th Climate Diagnostic Workshop,* 31 Oct–4 Nov 1988, Cambridge, 468–474.

Toth, Z., 1989: Long-range weather forecasting using an analog approach. *J. Climate,* **2,** 594–607.

Tracton, M. S., K. Mo, W. Chen, E. Kalnay, R. Kistler and G. White, 1989: Dynamical extended range forecasting (DERF) at the National Meteorological Center. *Mon. Wea. Rev.,* **117,** 1604–1635.

Van den Dool, H. M. 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.,* **117,** 2230–2247.

Wagner, A. J. 1989: Medium- and long-range forecasting. *Wea. Forecasting,* **4,** 413–426.