Use of the ''Odds Ratio'' for Diagnosing Forecast Skill

DAVID B. STEPHENSON

Laboratoire de Statistiques et Probabilités, Université Paul Sabatier, Toulouse, France

(Manuscript received 22 March 1999, in final form 9 November 1999)

ABSTRACT

This study investigates ways of quantifying the skill in forecasts of dichotomous weather events. The odds ratio, widely used in medical studies, can provide a powerful way of testing the association between categorical forecasts and observations. A skill score can be constructed from the odds ratio that is less sensitive to hedging than previously used scores. Furthermore, significance tests can easily be performed on the logarithm of the odds ratio to test whether the skill is purely due to chance sampling. Functions of the odds ratio and the Peirce skill score define a general class of skill scores that are symmetric with respect to taking the complement of the event. The study illustrates the ideas using Finley's classic set of tornado forecasts.

1. Motivation

Forecasts of the future state of our environment are possible due to the fundamental conservation of energy and momentum. For example, operational weather forecasts have been made routinely since 1950 based on numerical approximations of the dynamical equations governing the atmosphere. More recently, there has been a growing interest in forecasting climatic variations seasons in advance using numerical coupled ocean–atmosphere models.

In order to assess the ability of such forecasts, it is necessary to have accurate ways of quantifying ''forecast skill.'' Forecast skill, also sometimes referred to as forecast ''accuracy'' or ''quality,'' is an overall measure of how well previous forecasts were associated with previous observations (Murphy and Daan 1985; Murphy 1993). Forecast evaluation and verification is confounded by the many possible skill measures that can be used to summarize the complex behavior that occurs in even quite simple forecasts. Forecasts of M distinct categories require $M^2 - 1$ numbers to fully describe the joint probability distribution between the forecasts and observations. For several categories, this leads to many possible measures of forecast skill (''curse of dimensionality''), and it is not obvious which measures are most suitable for comparing forecasts with observations (Murphy 1991). One guiding principle is that skill measures should be as invariant/constant as possible, so as to provide robust measures that are less prone to being manipulated.

It is also important to know the sampling distribution of the skill score under no-skill conditions so that the skill score can be tested for statistical significance. The old saying that ''a measurement without an error estimate is meaningless'' is applicable to skill scores. This aspect has not received much attention from meteorologists and climate researchers yet is a necessary and vital part of forecast verification. Furthermore, it is important to distinguish between ''skill'' and ''value/utility'' of a forecast. Skill measures the general association between the forecasts and observations, whereas value focuses on user-specific costs (or utilities) that are expected to arise from using the forecasts. Significant skill does not necessarily imply useful value for any particular user, neither does useful value in certain situations imply any significant overall skill.

This study focuses on measuring the skill of forecasts of a discrete number of events, referred to as ''categorical forecasts.'' Furthermore, only the case of yes/no type dichotomous forecasts will be considered, for example, forecasts of whether or not a tornado will occur later in the same day. No account will be made of possible ordering of the categories, and it will be assumed that the number of forecast trials is fixed in advance. Other experimental designs can also be imagined, for example forecasts continued until a certain score is achieved, yet this is not usually the case in practice.

After more than 100 yr of research, fresh approaches to categorical skill scores are still possible as will be shown in this study. The following section will present a description of the example forecasts used in this study. Section 3 will then briefly describe some useful concepts from signal detection theory. Section 4 will introduce

Corresponding author address: Dr. D. B. Stephenson, Department of Meteorology, University of Reading, Earley Gate, P.O. Box 243, Reading RG6 6BB, United Kingdom.
E-mail: d.b.stephenson@reading.ac.uk

TABLE 1. Contingency table for Finley's original tornado forecasts.

Forecast	Tornado observed		
	Yes	No	Total
Yes	28	72	100
No	23	2680	2703
Total	51	2752	2803

the central idea of odds/risk. A brief comparison of skill scores will be presented in section 5, and the following section will examine the statistical significance of various scores. Section 7 will consider some more theoretical issues concerned with the sensitivity and invariance of the various skill scores. Section 8 concludes the article with a brief summary and some possible future applications.

2. Finley's tornado forecasts

a. Finley's original tornado forecasts

Sergeant John Finley's twice daily forecasts of tornados provide a useful historical dataset for illustrating the advantages and disadvantages of different forecast evaluation methods (Finley 1884; Murphy 1996). Using telegraphed synoptic information, Sgt. Finley issued forecasts at 0700 EST and 1500 EST each day stating whether tornadoes would form in 18 regions east of the Rocky Mountains. In common with many other environmental phenomena of human interest, tornadoes occur infrequently, yet can incur major loss and damage. By counting the number of successful forecasts of both ''tornado'' and ''no-tornado'' events, Sgt. Finley claimed that his forecasts were 96.6% correct. Gilbert (1884) pointed out a ''serious fallacy'' in Finley's measure of accuracy in that it took no account of the rare occurrence of tornado events, and that an even higher skill of 98.2% could have been obtained by forecasting no tornado every time! By considering different skill measures and their statistical significance, it will be shown that Finley's forecasts did have some real skill at reproducing the observations.[1] The total number of events in Finley's original forecasts are given in Table 1.

The numbers in each category will be represented by the symbols given in Table 2. In this study, the columns are used to denote the observed variable, while the rows are reserved for the predicted variable. Note that other conventions have sometimes been used, for example, Stanski et al. (1989) in which columns and not rows represented the forecast events.

[1] But this does not imply that there was any useful value for any forecast users!

TABLE 2. Schematic contingency table for categorical forecasts of a binary event. The symbols a–d represent the different number of events observed to occur in each category.

Event forecast	Event observed		
	Yes	No	Total
Yes	a (hit)	b (false alarm)	$a + b$
No	c (miss)	d (correct rejection)	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = n$

b. Hedging to obtain unbiased forecasts

A simple question that should be asked is whether the same fraction of events were forecast as were observed. The ''bias'' of the forecasts is given by the fraction of events forecast to those observed,

$$B = \frac{a + b}{a + c}, \tag{1}$$

and should be unity (unbiased) for a perfect forecasting system that aims to offer valuable forecasts to many diverse users. However, in practice, it generally differs from unity due to the presence of systematic biases (errors) in the forecasting model or observing system. Sometimes such biases are introduced intentionally in order to minimize the risk of missing potentially costly events (J. E. Thornes 1999, personal communication). Finley forecast nearly twice as many tornados as were actually reported leading to a bias B of 1.96 (=100/51). This bias is most likely due to the sparseness of the observing network rather than to a severe bias in the forecasting method (Finley 1884). Finley could have obtained unbiased forecasts by randomly rejecting 49% of the cases when he forecast a tornado to occur. He could have done this approximately by flipping a coin each time he had forecast a tornado, and then keeping the forecast only if the coin showed heads. Such a procedure entails moving a fraction $\alpha = (b - c)/(a + b)$ of events from *each* box in the upper row of the contingency table to the *corresponding* box in the lower row. In other words, it transforms the number of events (a, b, c, d) into the number of events $(a - \alpha a, b - \alpha b, c + \alpha a, d + \alpha b)$, and reduces the bias from B to 1.

Table 3 gives the contingency table for the unbiased forecasts obtained with $\alpha = 0.49$. Such an adjustment procedure can be thought of as a way of correcting systematic bias by ''hedging'' the forecasts toward the most frequently observed category (climatology). Hedg-

TABLE 3. Contingency table for the unbiased tornado forecasts obtained by hedging Finley's original forecasts.

Forecast	Tornado observed		
	Yes	No	Total
Yes	14	37	51
No	37	2715	2752
Total	51	2752	2803

TABLE 4. Contingency table constructed for random tornado forecasts having the same marginal totals as Finley's original forecasts.

Forecast	Tornado observed		
	Yes	No	Total
Yes	2	98	100
No	49	2654	2703
Total	51	2752	2803

ing can be considered to be either an optimal adjustment procedure, or a mild form of cheating that can in principle be used to obtain higher forecast skills (Gandin and Murphy 1992).

c. Random no-skill forecasts

A simple no-skill benchmark is provided by ''random forecasts,'' in which each event is forecast randomly but with the constraint that the marginal totals of both the forecasts and the observations in the contingency table remain the same as the marginal totals in the original verification table. Note that the phrase climatological forecast (without the word random) is commonly used to describe constant forecasts of the climatologically most likely category. For a random forecast, the expected number of events is given by $a' = np(o)p(f) = (a + c)(a + b)/n$, $b' = np(\overline{o})p(f) = (b + d)(a + b)/n$, $c' = np(o)p(\overline{f}) = (a + c)(c + d)/n$, and $d' = np(\overline{o})p(\overline{f}) = (b + d)(c + d)/n$. The ''base rate'' $p(o) = (a + c)/n$ is an estimate of the probability that the event will occur, whereas $p(\overline{o}) = 1 - p(o)$ is an estimate of the probability that the event will not occur. Despite the base rate for rare catastrophic events such as tornadoes being vanishingly small, it nevertheless plays an important role in determining the useful ''value'' of such forecasts (Matthews 1996). Table 4 shows a contingency table constructed in this manner based on Finley's original forecasts. A small error is introduced into the random forecasts by rounding the cell counts to integers. For example, a' should be 1.82 but is instead rounded to 2 in Table 4. By construction, the rows and columns of contingency tables for random forecasts are completely independent of one another and there is no association between the forecasts and the observations.

3. Detection of signals

How can the overall skill of Finley's tornado forecasts be diagnosed? With a fixed total number of events, as is normally the case for forecast trials, three degrees of freedom are needed to fully describe the four values in a 2×2 contingency table. One quantity has already been introduced, namely, the bias B of the forecast and two others remain to be chosen. Ideas from signal detection theory suggest two other useful quantities: the ''hit rate'' and the ''false alarm rate.'' Many diverse disciplines such as radio communications, medical imaging, medical diagnosis, and psychology use signal detection theory to optimally detect and diagnose signals (Swets 1973, 1988; Macmillan and Creelman 1991; Green and Swets 1996; Swets and Pickett 1982). Signal detection theory was first applied to the verification of meteorological forecasts in the pioneering studies of Mason (1980, 1982) and provides a universal framework for evaluating the joint probability distribution of forecasts and observations (Stanski et al. 1989; Harvey et al. 1992; Mason 1997; and references therein). As explained in Murphy and Winkler (1987), the joint distribution can be factorized by either stratifying on the observations (likelihood-base rate factorization) or on the forecasts (calibration-refinement factorization). Both these stratifications will now be considered.

a. Likelihood-base rate factorization

The hit rate (H) gives the relative number of times an event was forecast when it occurred, whereas the false alarm rate (F) gives the relative number of times the event was forecast when it did not occur. False alarms are nicely illustrated by the Grimm brothers' story about a boy who cries (shouts) wolf when there is none. A tragedy finally happens when the boy encounters a wolf (the observed event) because people in the village no longer believe the boy's shouts (the forecast) due to his numerous previous false alarms. The rates are usually estimated by the ratios of the number of events,

$$H = \frac{a}{a + c} = \hat{p}(f \mid o) \tag{2}$$

$$F = \frac{b}{b + d} = \hat{p}(f \mid \overline{o}), \tag{3}$$

and provide ''frequentist'' estimates of the conditional likelihoods $p(f \mid o)$ and $p(f \mid \overline{o})$. The hit rate is sometimes referred to as the ''probability of detection'' in the earlier literature. Borrowing terminology that is used in statistical hypothesis testing, the false alarm rate is interpreted as the rate of making a ''type I error'' whereas the ''miss rate,'' equal to one minus the hit rate, measures the chance of making a ''type II error.'' The two types of error often have very different consequences; for example, failing to forecast a tornado that then occurs (type II error—a miss) is generally more damaging than forecasting a tornado that does not then appear (type I error—a false alarm). For catastrophic events, the expected loss incurred by the forecast user generally depends more on the hit rate than on the false alarm rate. Note that a complete analogy with hypothesis testing is not possible since hypothesis testing is concerned with making decisions between unknown ''parameters'' whereas categorical forecasting is concerned with making decisions between possible processes.

Improved estimates may be obtained by using Bayesian methods that incorporate prior information about

TABLE 5. Signal detection statistics for the different tornado forecasts.

Statistic	Symbol	Range	Finley	Hedged	Random
Hit rate	H	[0, 1]	0.549	0.275	0.039
Odds of a hit	$H/(1 - H)$	[0, ∞]	1.217	0.378	0.041
False alarm rate	F	[0, 1]	0.026	0.014	0.036
Odds of a false alarm	$F/(1 - F)$	[0, ∞]	0.027	0.014	0.037
Odds ratio	θ	[0, ∞]	45.31	27.76	1.11
Log odds ratio	$\log\theta$	[−∞, ∞]	3.81 ± 0.31	3.32 ± 0.36	0.10 ± 0.73
Degrees of freedom	n_h	[0, ∞]	10.70	7.95	1.88

possible uncertainty in model bias, etc. For example, an improved Bayesian estimate of the hit rate can be obtained using the simple ''rule of succession'' based on a uniform prior (Fisher 1990). In other words, $\hat{p}(f|o)$ can be estimated using the Bayesian expression $(a + 1)/(a + c + 2)$ (''add one hit and one miss'') rather than the more common frequentist expression $a/(a + c)$. Such an estimate is slightly closer to 0.5 and avoids either under- or overestimating the rate especially when the sample size is small.

b. Calibration-refinement factorization

One can also estimate alternative rates by stratifying on the forecasts:

$$H' = \frac{a}{a + b} = \hat{p}(o|f) \tag{4}$$

$$F' = \frac{c}{c + d} = \hat{p}(o|\overline{f}). \tag{5}$$

The ratio $1 - H'$ is sometimes referred to as the ''false alarm ratio,'' which should not be mistaken with the previously discussed false alarm rate (e.g., Wilks 1995, chap. 7). The ratio F' is a ''conditional miss rate.'' The likelihood-base rate and calibration–refinement are related to one another by expressions derived from Bayes' theorem such as

$$p(o|f) = \frac{p(f|o)p(o)}{p(f|o)p(o) + p(f|\overline{o})p(\overline{o})} \tag{6}$$

and can be expressed in terms of one another as

$$H' = \frac{H}{B} \tag{7}$$

$$F' = \frac{F(1 - H)}{F - H + B(1 - F)}. \tag{8}$$

They therefore contain equivalent information.

Table 5 gives the hit and false alarm rates calculated for the different forecasts. It can be seen the hit rate is greater in the Finley forecasts (0.549) than in the hedged forecasts (0.275). The Finley and hedged forecasts correctly predicted the event (a tornado) on more than 25% of the occasions when a tornado actually occurred. The false alarm rate is less than 4% for all three forecasts, suggesting that very few tornadoes were forecast when none occurred. When forecasts of a particular event (tornado, wolf, etc.) are rare, the conditional probability $p(f|\overline{o}) = p(f \text{ and } \overline{o})/p(\overline{o})$ is generally small. For example, for no-skill random forecasts the false alarm rate is given by $p(f)$, which is small for events that are forecast rarely. Because of the rarity of wolves, most ''normal'' children do not cry wolf very often and so generally have a low false alarm rate. The rarity of the forecast event often implies a low false alarm rate, yet conversely, a low false alarm rate cannot be entirely attributed to the rarity of the event since a low false alarm rates can also result from high forecast skill.

TABLE 6. Categorical forecast totals expressed in terms of the bias $B = (a + b)/(a + c)$, the hit rate $H = a/(a + c)$, and the false alarm rate $F = b/(b + d)$. The multiplier $m = n/(B - H + F)$ multiplies all the totals and does not therefore contribute to ratios of any of the totals.

Event forecast	Event observed		
	Yes	No	Total
Yes	$F\,Hm$	$F(B - H)m$	$F\,Bm$
No	$F(1 - H)m$	$(1 - F)(B - H)m$	$(F - H + (1 - F)B)m$
Total	Fm	$(B - H)m$	$(B - H + F)m = n$

c. The BHF representation

For a fixed total number of events, the three quantities, B (bias), H (hit rate), and F (False alarm rate), completely describe the numbers of events in the contingency table. The numbers a, b, c, and d can be expressed in terms of B, H, and F as shown in Table 6.

This provides a useful representation for describing dichotomous forecasts. The bias B compares the marginal probabilities of the forecasts and observations, whereas H and F are conditional probabilities that completely describe the joint conditional distribution. As explained in Murphy and Winkler (1987), it is useful to factor the joint distribution in such a way especially when the base rates (i.e., climatological probabilities) are quite dissimilar. Note that the quantity m becomes singular when B is exactly equal to $H - F$, yet this is likely to never occur in practice.

A useful visual representation is obtained by marking

(F, H) values in the unit square. When a control parameter such as the threshold for the event is varied, a locus of points is traced in the (F, H) plane. In the medical and psychological literature, this curve is referred to as a ''receiver operating characteristic'' (ROC), or less commonly as a ''relative operating characteristic.'' The ROC provides a useful diagnostic summary of the discriminatory capability of the forecast system, and should not be confused with the operating characteristic (OC) curve that is widely used to test between different statistical hypotheses/parameters.

4. Odds assessment of forecasts

Forecasting is an inherently risky business that involves making predictions about which events are most likely to occur in the future.[2] Fortunately, in weather forecasting it is possible to produce skillful forecasts by making use of the physical laws that determine the evolution of the universe. This section will discuss the central concept of odds for assessing the overall risk involved in making forecasts.

a. Odds and risk

The ''odds'' or ''risk'' of an event is the ratio of the probability that the event occurs to the probability that the event does not occur. In other words, the odds of an event that has a probability p of occurring is given by $p/(1 - p)$, and ranges from zero to infinity. For example, an event with probability of 0.8 of occurring has an odds of $0.8/(1 - 0.8) = 4$ (or 4 to 1 ''on/for'' in bookmaker's jargon). Odds and probability/chance differ because of the denominator, which becomes important for more frequent events. An interesting property of odds is that the odds for the complement of an event (i.e., not the event) is the reciprocal of the odds for the event. For example, an event with probability of $0.2 = 1 - 0.8$ of occurring has an odds of $0.2/(1 - 0.2) = ¼$ (or 4 to 1 ''against'' in bookmaker's jargon). Hit and false alarm rates can be interpreted in terms of odds. For example, the odds of Finley's forecasts correctly predicting a tornado (a hit) given that one occurred is given by $H/(1 - H) = 0.549/(1 - 0.549) = 1.22$ and so the odds of a correct tornado forecast is 1.22 (or about 6 to 5 for), which is close to ''evens'' (odds of 1.0).

b. The ''odds ratio''

Forecast skill can be judged by comparing the odds of making a good forecast (a hit) to the odds of making a bad forecast (a false alarm). In other words, by using the ''odds ratio'':

$$\theta = \frac{H}{1 - H}\left(\frac{F}{1 - F}\right)^{-1}. \tag{9}$$

This ratio is greater than one when the hit rate exceeds the false alarm rate. For Finley's tornado forecasts, the ratio of odds for the hit rate (1.22) to the odds for the false alarm rate (0.027) is greater than 1 and equals 45.31. The odds ratio is equal to the ''cross-product ratio'':

$$\theta = \frac{ad}{bc} = \frac{p(f\,|\,o)p(\overline{f}\,|\,\overline{o})}{p(\overline{f}\,|\,o)p(f\,|\,\overline{o})}, \tag{10}$$

which can easily be calculated from a contingency table. The odds ratio is unity when the forecasts and observations are independent and provides a good way of summarizing the ''association'' in the joint probability distribution. Note that it depends solely on the conditional joint probabilities and not on the marginal probabilities; it is therefore independent of any bias between the observations and the forecasts. It is widely used in medical trials for testing the associations between clinical drug treatment and side effects (Agresti 1996).

The difference of the odds ratio from unity is equal to the weighted difference between the hit and false alarm rate:

$$\theta - 1 = \frac{H - F}{F(1 - H)}. \tag{11}$$

From this it can be seen that the odds ratio is unity when the hit and false alarm rates are identical. Associated variables gives odds ratios larger than unity and can be easily tested for significance by considering the natural logarithm of the odds ratio referred to as ''log odds'':

$$\log\theta = \log a + \log d - \log b - \log c, \tag{12}$$

which is approximately Gaussian distributed for large enough a, b, c, and d (each at least greater than 5). Rather than being a weighted sum of the the raw counts as is generally assumed for scores (Gandin and Murphy 1992; Potts et al. 1996), log odds is a weighted sum of the *logarithms* of the counts and thereby accounts to some extent for the larger sampling uncertainties in the larger numbers of events.

c. Odds ratio parameterization of ROC curves

The definition of odds ratio [Eq. (9)] can be used to obtain the hit rate as a function of the false alarm rate:

$$H = \frac{\theta F}{1 + (\theta - 1)F}. \tag{13}$$

Examples of isopleth curves obtained using this expression for different values of odds ratio are shown in Fig. 1. They closely resemble the ROC curves that have been found in previous weather forecasting studies (e.g., Mason 1982; Harvey et al. 1992). The interesting similarity between isopleths of θ and empirical ROC curves

[2] Prophesy is a good line of business, but it is full of risks—Mark Twain, *Following the Equator: A Journey Around the World.*

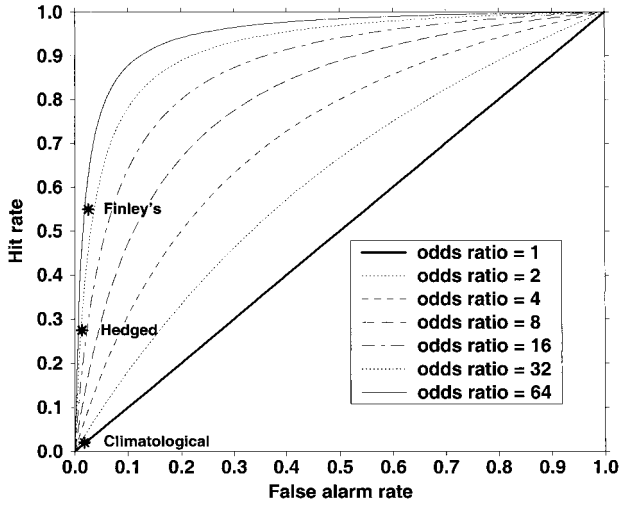

FIG. 1. ROC curves for different values of odds ratio. Tornado forecasts are also marked on the diagram as asterisks.

has also been noted in the nonmeteorological context (Swets 1986). The odds ratio is almost invariant with decision threshold and may provide a threshold-independent skill score and an effective way to parameterize empirical ROC curves.

5. Comparison of various scores

This section will briefly review some commonly used scores and compare their performance on the tornado forecasts (Table 7). Differences between the skill become most apparent in biased real-world cases such as Finley's original tornado forecasts.

a. Proportion correct (PC)

Finley (1884) judged forecast accuracy by considering the simple matching coefficient based on the ''proportion'' of total ''correct'' hits and rejections (PC):

$$\mathrm{PC} = \frac{a + d}{n} = \frac{B - H - FB + 2HF}{B + F - H}. \quad (14)$$

The scores are given in Table 7 where it can be seen that all the forecasts including even the random forecasts give high scores. Furthermore, the hedged forecasts have a higher score than the original forecasts. In this particular case, hedging the forecasts toward the most observed category has increased the proportion correct. As explained in Gandin and Murphy (1992), PC is not an ''equitable'' score since it can be improved by forecasting more frequently the most observed category. This can be undesirable because it may encourage some forecasters to hedge their forecasts away from forecasting less likely yet important events.

TABLE 7. Scores for the different tornado forecasts. Except for PC and GSS, all the scores would have been exactly zero for the random no-skill forecasts if the cell counts had not been rounded to whole numbers in Table 4.

Skill score		Range	Finley	Hedged	Random
Proportion correct	PC	[0, 1]	0.966	0.974	0.948
Heidke	HSS	[−1, 1]	0.365	0.261	0.002
Gilbert	GSS	[0, 1]	0.228	0.159	0.013
Peirce	PSS	[−1, 1]	0.523	0.261	0.004
Yule's Q	ORSS	[−1, 1]	0.957	0.931	0.050
Pearson	X^2/n	[0, 1]	0.142	0.068	0.000
Likelihood	G^2/n	[0, 1]	0.045	0.020	0.000

b. Heidke skill score (HSS)

By comparing the proportion correct PC to that obtained for no-skill random forecasts $\mathrm{PC}_0 = (a' + d')/n$, it is possible to construct the widely used Heidke skill score (HSS):

$$\mathrm{HSS} = \frac{\mathrm{PC} - \mathrm{PC}_0}{1 - \mathrm{PC}_0} = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}, \quad (15)$$

which is the number of correct hits and rejections standardized so that random forecasts have zero skill (Doolittle 1888; Heidke 1926). The Heidke skill score ranges from −1 to 1, and in *BHF* representation has the rather cumbersome form

$$\mathrm{HSS} = \frac{2(B - H)(H - F)}{F - H + B(1 + B - H - F)}, \quad (16)$$

which depends on the bias. Table 7 shows that a larger skill score of 0.365 is obtained for the Finley forecasts than the skill score of 0.261 obtained for the hedged forecasts. This is because the proportion correct (0.964) for random forecasts based on the hedged forecasts is slightly larger than the proportion correct (0.948) for random forecasts based on the original forecasts. By standardizing relative to random forecasts, the Heidke skill score becomes a more ''equitable'' measure than the proportion correct score.

c. Gilbert skill score (GSS)

In his critique of Finley's scoring procedure, Gilbert (1884) proposed an alternative verification statistic for forecasts of rare events, which is referred to as either the ''critical success index,'' the ''threat score'' (Schaefer 1990), or the Jaccard coefficient. The Gilbert skill score (GSS) is defined as

$$\mathrm{GSS} = \frac{a}{a + b + c} = \frac{H}{1 + B - H} \quad (17)$$

and so takes no account of the false alarm rate. It completely ignores the large number of frequent events (d) when a tornado was not observed and was not forecast

to occur. This can be an advantage in the large number of forecast situations where it is difficult or impossible to define d with any certainty. If a tornado were forecast correctly on every occasion ($c = 0$), then GSS would give H'. It can be seen from Table 7 that the GSS avoids the problem noted for PC in which the hedged forecasts have a higher score than the unhedged forecasts. Unfortunately, however, the Gilbert skill score has the disadvantage of not being zero even for no-skill climatological or random forecasts (Wilks 1995). The GSS can be a useful index but only if supplemented with additional information such as the frequency of occurrence of the event (Mason 1989).

d. Peirce skill score (PSS)

A simple reliable measure of skill is obtained by taking the difference between the hit rate and the false alarm rate:

$$\text{PSS} = H - F = \frac{ad - bc}{(a + c)(b + d)}. \tag{18}$$

Stimulated by Gilbert's remarks on the accuracy of Finley's tornado forecasts, Peirce (1884) offered this alternative as a ''measure of the science of the method.'' It has since been rediscovered and renamed several times: ''Hanssen–Kuipers discriminant'' (Hanssen and Kuipers 1965), ''Kuipers' performance index'' (Murphy and Daan 1985), and the ''true skill statistic'' (Flueck 1987). To respect its original discovery, it will be referred to as the Peirce skill score (PSS) in this article. When the score is greater than zero, the hit rate exceeds the false alarm rate and this can then be used to infer that there is some forecast skill. For example, a boy who cries wolf frequently when there is none, should not be ignored if it can be shown that he actually cries wolf at a more frequent rate when there is one! Most people behave less rationally than this and would tend to ignore the boy because of his previous high false alarm rate as is illustrated in the Grimm story-tale when a wolf eventually does visit the village.

The Peirce skill score is larger for the Finley forecasts than for the hedged forecasts (Table 7). For the Finley forecasts, it is also larger than the other scores. The majority of this skill comes from the high hit rate based on the number of tornadoes forecast when tornadoes were actually observed. In other words, the skill is coming from the two small numbers in the first column of the contingency table ($a = 28$ and $c = 23$), and the other numbers of events (b and d) make a negligible contribution. It is a weakness of the Peirce skill score that when one cell count in the contingency table is large (e.g., d), then the other cell count in the same column is almost completely disregarded (e.g., b).

e. Odds ratio skill score (ORSS)

A simple skill score ranging from -1 to $+1$ can be obtained from the odds ratio by the transformation

$$\text{ORSS} = \frac{\theta - 1}{\theta + 1} = \frac{H - F}{H + F - 2HF}. \tag{19}$$

This score was proposed long ago as a ''measure of association'' by the statistician G. U. Yule (Yule 1900) and is referred to as Yule's Q. Despite its wide use for measuring association in contingency tables (Agresti 1996), until now, it has never been applied for verifying meteorological forecasts. It is based entirely on the joint conditional probabilities, and so is not influenced in any way by the marginal totals.

The odds ratio skill scores for the tornado forecasts are presented in Table 7. Finley's original and the hedged forecasts have high skill scores close to 1, whereas the random forecasts have an odds ratio skill score (ORSS) close to zero. Because ORSS is independent of the marginal distribution, it strongly discriminates between the cases with and without association even when the different contingency tables appear to have similar cell counts. This is in contrast to other scores such as the proportion correct, which gave similar scores for all three sets of forecasts. However, one should not be misled into thinking that high values of ORSS imply significant amounts of skill. To test for real skill or real differences in skill, it is essential that careful significance testing is performed on the skill scores as will be discussed in more detail in section 6. Smaller skill scores based on the odds ratio can be obtained if so desired by using simple functions of ORSS such as ORSS to some power.

f. Chi-squared measures of association

Although rarely used in evaluating meteorological forecasts, association can also be tested by using the ''Pearson'' and the ''likelihood ratio'' chi-squared distributed measures of fit:

$$\chi^2 = \sum_{i,j=1}^{2} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \tag{20}$$

$$G^2 = 2 \sum_{i,j=1}^{2} n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right), \tag{21}$$

where n_{ij} are the observed cell counts and $\mu_{ij} = n_{i\cdot}n_{\cdot j}/n$ are the counts expected from climatology ($n_{i\cdot}$ is the total count for the ith row, etc.). For independent events, both measures are asymptotically chi-squared distributed with one degree of freedom. These statistics normalized by the total number of events are given in Table 7 and are significantly different from zero at 99.9% confidence ($\chi^2 > 10.83$) for Finley's original and the hedged tornado forecasts. For 2×2 contingency tables, the measure χ^2/n is equal to

$$\frac{(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} = (1 - H')F'(1 - H)F(\theta - 1)^2 \tag{22}$$

and provides a squared correlation measure r^2 of two-sided association in the contingency table that was first proposed for use in forecast verification by Doolittle (1885). It has been used for thunderstorm forecast verification by Pickup (1982). From Eq. (22), it can be seen that χ^2/n contains a factor of $(\theta - 1)^2$ that depends directly on the odds ratio. Since this is generally the most variable factor in the chi-squared skill score, the odds ratio can be used instead of chi squared to provide a simpler and more direct measure of association (Yule 1900; Agresti 1996).

6. Do forecasts have any real skill?

Skill scores compiled from contingency tables are ''sample estimates'' of past performance and, therefore, contain sampling uncertainties. Impressively good scores can sometimes be obtained purely by chance, especially if the score has been compiled over an insufficient number of independent events. For example, it would be grossly misleading to claim that a coupled model forecasting system had skill based on the successful forecast of only one El Niño event. Statistical significance testing can be used to reject the null hypothesis that good scores occurred simply by chance sampling fluctuations. With the exception of only a few studies, the rather dull yet important business of testing the significance of skill scores has received relatively little attention by meteorologists (Woodcock 1976; Seaman et al. 1996). The sampling distributions are not even known for most of the frequently used skill scores. Furthermore, for skill scores such as Heidke's, that have quite complicated dependence on the number of events, the sampling distribution is likely to be difficult if not intractable to calculate analytically. This section will briefly discuss how statistical error estimates (confidence intervals) can be used to judge both the hit and false alarm rates, and the Peirce and odds ratio skill scores.

a. Confidence intervals for hit and false alarm rates

The ''score confidence interval'' (Agresti and Coull 1998) for proportions such as hit rates and miss rates is given by

$$\frac{\hat{p} + \dfrac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{[\hat{p}(1 - \hat{p}) + z_{\alpha/2}^2/4n]/n}}{1 + z_{\alpha/2}^2/n}. \quad (23)$$

The statistic $\hat{p}$ is the estimated hit or false alarm rate and n is the total number of events used to estimate the rate (Agresti and Coull 1998). The parameter $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution used to determine the $100(1 - \alpha)\%$ confidence interval. For example, the hit rate of Finley's tornado forecasts is 0.549 (=28/51) calculated with $n = 51$ events and so has an estimated 95% confidence interval of ± 0.13 ($z_{\alpha/2} = z_{0.0275} = 1.96$). The hit rate is therefore significantly different from zero at 95% confidence. The score confidence interval enables error bars (confidence intervals) to be added to ROC plots to give a better idea of possible uncertainties.

b. Standard error of the Peirce skill score

Assuming independence of the hit and false alarm rates, the standard error in the Peirce skill score is simply the square root of the sum of the squared standard errors in the hit and false alarm rates. For large enough samples, the standard error is approximated by

$$\sqrt{\frac{H(1 - H)}{n_H} + \frac{F(1 - F)}{n_F}}, \quad (24)$$

where $n_H = a + c$ and $n_F = b + d$. For Finley's forecasts, this expression gives an estimated standard error of 0.069 in the Peirce score of 0.523, and hence the Peirce score is significantly different from a zero score no-skill forecast.

c. Significance testing of the odds ratio

The odds ratio can be easily tested for significance by considering the natural logarithm of the odds ratio, which is asymptotically Gaussian distributed with a standard error given by $1/(n_h)^{1/2}$, where n_h is the effective number of degrees of freedom (d.o.f.'s) $1/n_h = 1/a + 1/b + 1/c + 1/d$ (Agresti 1996). The d.o.f. takes into account the number of events in each category and can never exceed the smallest cell count. To test whether there is any forecast skill, one can test against the null hypothesis that the forecast and observations are independent with a log odds of zero. For Finley's tornado forecasts, log odds is 3.81 with an asymptotic standard error of 0.31 (Table 6) and therefore the log odds is more than 1.96 standard errors away from zero implying that there is less than 5% chance that the skill could be due to pure chance. At more than 95% confidence, Finley's tornado forecasts were not independent of the observations and therefore had some skill (but not necessarily any useful value!). Log odds is simply twice the Fisher z transform of the ORSS measure of association; that is, $\log \theta = \log(1 + \text{ORSS})/(1 - \text{ORSS})$. An alternative score $\Phi(\sqrt{n_h}) \log\theta)$ can be constructed that incorporates the sampling error. It is approximately equal to $\theta^{(n_h)^{1/2}}/(\theta^{(n_h)^{1/2}} + 1)$ and can be conveniently interpreted as the probability that the forecasts and the observations are positively associated. An easy-to-use lookup table for assessing the significance of the odds ratio skill score can be constructed.

Singular behavior occurs when any one of the numbers a, b, c, or d is zero. If either b or c becomes zero, then ORSS = 1 indicating perfect association. If either a or d becomes zero, then ORSS = −1 indicating perfect negative association. Because the odds ratio can be unity

for forecasts that are not completely ''perfect'' (i.e., both b and c are zero), Woodcock (1976) argued that the odds ratio was unsuitable for use in forecast evaluation. However, when any one of the cell counts is zero, the asymptotic standard error in log odds becomes infinite and the odds ratio can no longer be meaningfully tested for significance (Agresti 1996). By taking into account the significance of the score, it is possible to avoid Woodcock's criticism and thereby use the odds ratio for forecast evaluation. If all the boxes in any of the rows or columns have small counts equal or close to zero, the 2×2 verification problem becomes rank deficient and lower-dimension verification should be considered. In summary, the odds ratio is no longer a meaningful measure of association when any of the cell counts are zero. Furthermore, care should be exercised in testing the significance of the odds ratio when any of the cell counts become particularly small (i.e., less than 5). In such cases, exact significance tests should be performed numerically using software such as StatXact. A comprehensive account of significance testing for various measures of association is given in Bishop et al. (1975).

7. Invariance of skills

Skill scores are measures of similarity between the forecasts and observations, and can be chosen in many different ways. To be useful overall measures, skill scores should not depend strongly on the way the forecaster decides to define the categories etc. Skill scores that do not depend on such choices are in principle less easily manipulated and are, therefore, more powerful than other less invariant measures. For example, spatial correlations made using the Mahalanobis metric are invariant under linear transformations and, therefore, remain the same regardless of linear mapping of the variables onto different spatial grids (Stephenson 1997). This section will discuss various transformation properties of categorical skill scores.

a. Improvement by hedging towards climatology?

It is desirable that overall measures of skill should not be improvable by hedging the forecasts toward the most frequent category, otherwise forecasters may be discouraged from making forecasts of the possibly more useful yet less frequent category (e.g., tornados). Hedging toward the most frequent category, which is represented in our example by the lower row in the contingency table, transforms the number of events as follows:

$$(a, b, c, d) \rightarrow (a, b, c, d)' = (a - \alpha a, b - \alpha b, c + \alpha a, d + \alpha b). \quad (25)$$

Hedging toward the most frequent category reduces $B \rightarrow (1 - \alpha)B$, $H \rightarrow (1 - \alpha)H$, and $F \rightarrow (1 - \alpha)F$. The reduction in the hit and false alarm rates is due to there being less chance of making a hit or a false alarm if less events are being forecast.

Since the Peirce skill score is the difference $H - F$, it is also reduced by the same factor $\text{PSS} \rightarrow (1 - \alpha)\text{PSS}$. An unscrupulous forecaster would therefore be unable to improve their Peirce skill score by hedging their forecasts toward climatology! The Peirce skill score cannot be improved by forecasting a particular class of events—it is an ''equitable'' score (Gandin and Murphy 1992). Furthermore, Gandin and Murphy (1992) demonstrated that the Peirce skill score is the *only* equitable *linear* score for binary forecasts.

Since the odds ratio is a linear combination of the logarithm of the number of events rather than being a linear combination of the number of events, it is outside the general class of linear scores considered by Gandin and Murphy (1992). It is therefore necessary to treat the odds ratio skill score as a special case. Under hedging toward climatology, the odds ratio skill score transforms to

$$\text{ORSS} = \frac{H - F}{H + F - 2HF} \rightarrow \text{ORSS}' = \frac{H - F}{H + F - 2HF + 2\alpha HF}. \quad (26)$$

The positive quantity $2\alpha HF$ is added to the denominator, which causes the transformed ORSS to always be smaller than the unhedged ORSS. For small amounts of hedging, a Taylor expansion in α gives

$$\text{ORSS}' \approx \text{ORSS}\left(1 - \frac{2\alpha HF}{H + F - 2HF}\right). \quad (27)$$

The relative reduction $2\alpha HF/(H + F - 2HF)$ is less than that of α obtained for the Peirce skill score, when either the hit or false alarm rate are small. In such cases, the odds ratio skill score is less sensitive to hedging as can be noted, for example, in the similar ORSS values obtained for Finley's original and the hedged forecasts in Table 7. This is an advantage since it means that the ORSS is less sensitive to small changes in bias that can easily occur due to changes in either the forecast model or the base period climatologies. The ORSS provides a robust measure of association that is independent of such changes in the marginal distributions.

b. Complement symmetry

Instead of choosing the event to be ''tornado occurs,'' it would have been equally possible to choose the event to be ''tornado does not occur.'' A complementary contingency table would then have been obtained having swapped rows, and swapped columns, in other words, $(a, b, c, d) \rightarrow (a, b, c, d)' = (d, c, b, a)$. Which skill scores give the same value for the complementary table as for the original table? It is easily verified that the proportion correct, Heidke, and Peirce skill scores are

all invariant under this operation, and so do not depend on the subjective choice of the event or its complement. The Gilbert skill score, however, depends on the subjective choice of what is the event and the nonevent. The event is invariably chosen to be the rarer outcome (e.g., tornado rather than no tornado), yet additional information about base rates should also be supplied (Mason 1989).

Hit rates and false alarm rates transform to $H \rightarrow 1 - F$ and $F \rightarrow 1 - H$ and, therefore, also depend on the choice of event and nonevent. This transformation corresponds to a reflection of the points about the line $H = 1 - F$ in the (F, H) plane. Because the Peirce score is the special combination $H - F$ it remains invariant under such reflections. One might suspect that the ratio H/F could also be a suitable measure of forecast skill. However, this quantity transforms to the different value of $(1 - F)/(1 - H)$, and so depends on whether one chooses the event or its complement. For example, for Finley's tornado forecasts H/F is 20.99 if one chooses the event to be tornado occurs but is 2.16 if the event is chosen to be no tornado occurs. However, unlike the ratio of rates, the ratio of odds $\theta = ad/bc$ is invariant under taking the complement and so, therefore, are all functions of the odds ratio such as log odds and ORSS.

All other complement symmetric combinations of hit and false alarm rate can be expressed as functions of the Pierce and odds ratio skill scores. Because the Jacobian

$$\frac{2H(1 - H) - 2F(1 - F)}{(H + F - 2HF)^2} \tag{28}$$

is never singular except on the line $H = F$, it is possible to express any complement symmetric function of (H, F) in terms of the the transformed variables (PSS, ORSS). Explicit expressions for H and F can be found by using the definitions of θ and PSS given in Eqs. (9) and (19). By simultaneously solving these equations, one obtains $H = F + \text{PSS}$, and $x = 1/F$ as a solution to the quadratic equation:

$$\text{PSS}x^2 + (1 - \text{PSS})(1 - \theta)x - (1 - \theta) = 0. \tag{29}$$

This then allows θ and PSS to be substituted for H and F. In other words, the isopleths of PSS and ORSS in the (H, F) plane can be used as curvilinear coordinates for describing any complement symmetric quantity. As an example, consider the skill score:

$$\text{SS} = \frac{(ad - bc)(a + c)(b + d)}{(ad + bc)(ad + bc)} = \frac{H - F}{(H + F - 2HF)^2}, \tag{30}$$

which is a complement symmetric function of the hit and false alarm rate. A little algebra reveals that this complement symmetric score can be rewritten in terms of PSS and ORSS as $\text{PSS}^{1-k}\text{ORSS}^k$ with $k = 2$. Functions of the Pierce and odds ratio skill score define the *entire* class of complement symmetric skill scores based on just the hit and false alarm rate.

c. Transpose symmetry

Suppose that the computer files containing the observed results and the forecast results were inadvertently mixed up. Which of the skill scores would still give the same values? Swapping the forecast and observations corresponds to transposing the contingency table $(a, b, c, d) \rightarrow (a, c, b, d)$. For unbiased forecasts, $b = c$ and so this transformation would not change the contingency table.[3] However, for biased forecasts the contingency table and scores based on it may change. Certain scores such as the proportion correct, Heidke, and odds ratio skill scores remain invariant under this transformation. However, the Peirce skill score transforms to

$$\text{PSS} = \frac{ad - bc}{(a + c)(b + d)} \rightarrow \frac{ad - bc}{(a + b)(c + d)} = \frac{ad - bc}{B(a + c)[B(b + d) + (1 - B)n]} \tag{31}$$

and is therefore not invariant under transposing the forecasts and observations when the table is biased. For example, when Finley's forecasts and observations are transposed, the PSS of 0.523 predominantly based on a and c transforms to the much smaller value of 0.272 based predominantly on a and b. In other words, the Peirce skill score is defined in terms of the likelihood-base rate factorization and not the calibration–refinement factorization and so requires prior information about which are the forecasts and which are the observations. It should be noted that this distinction may be important for value decisions based on forecasts and that there is no inherent merit in a score being transpose symmetric. Interestingly, the invariant score obtained by taking the product of both Peirce skill scores is identical to the squared correlation derived from the Pearson chi-squared statistic ($r^2 = \chi^2/n$).

8. Concluding remarks

The 2 × 2 problem was stated by M. H. Doolittle (1888) as follows:

> "Having given the number of instances respectively in which things are both thus and so, in which they are thus but not so, in which they are so but not thus, and in which they are neither thus nor so, it is required to eliminate the general quantitative relativity inhering in the mere thingness of the things, and to determine the special

[3] This invariance of unbiased forecasts can be tested using McNemar's simple test (Agresti 1996, p. 227).

> quantitative relativity subsisting between the thusness and the soness of the things.''

These remarks give an indication of the possible complexity involved when quantifying even small 2×2 contingency tables. It is amazing how such an apparently simple problem can prove to be so complicated and controversial as evidenced by the proliferation of association measures and skill scores (Goodman and Kruskal 1979). After more than a century of heated debates, there are still simmering arguments about whether it is important to condition on the margins of the table (Yates 1984).

In this study, it has been argued that the ''odds ratio'' can provide a useful new measure of association for verifying binary forecasts. A simple and powerful skill score, ORSS, can easily be constructed from the odds ratio that has the following useful properties:

1) It is simple to calculate and is easily interpreted in terms of signal detection theory quantities; it is the ratio of the odds of making a hit given that the event occurred to the odds of making a false alarm given that the event failed to occur.
2) It is a single measure that summarizes the $(M - 1)^2$ degrees of freedom in the *conditional* joint probability distribution. It does not depend on the marginal totals and so is an ''equitable'' score that cannot be easily hedged.
3) It can easily be used to test whether the forecast skill is significant (i.e., not due to chance sampling). This is achieved by testing if the Gaussian distributed log odds is zero.
4) It is complement symmetric and so measures the skill of forecasting both the event and its complement.
5) The score does not distinguish between which are the forecasts and which are the observations, and so is a transpose symmetric measure for comparing the forecasts with observations.
6) It becomes indeterminate if any of the rows or columns in the contingency table are completely zero. This is reasonable since 2×2 contingency tables are no longer appropriate when all the forecasts or all the observations fall into only one particular category.

The independence of the odds ratio with respect to the marginal totals makes it a valuable quantity for summarizing the joint conditional probability distribution of diagnostic systems such as weather forecasts. For example, the log odds can provide a reliable measure of how well the system discriminates between hits and false alarms. This study has shown that the odds ratio and the Peirce skill score can be used to completely summarize the joint *conditional* distribution of 2×2 categorical forecasts [i.e., (F, H) behavior]. However, these two scores provide no information about the marginal distributions of the forecasts and observations, which instead can be compared by considering the bias (B). In other words, the triplet ORSS, PSS, and B form a useful complete set for describing the three degrees of freedom in 2×2 categorical forecasts. In addition to using PSS and bias to summarize forecasts, more use should be made of the odds ratio in forecast verification. However, care should be exercised when any of the cell counts are very small (i.e., less than about 5), in which case the odds ratio may become unreliable. The odds ratio is no longer a suitable measure of skill for testing hypotheses if any of the cell counts become zero (Woodcock 1976).

To understand how certain factors control the skill of forecasts, a regression can be performed of the skill on the various possible factors. Unlike bounded skill scores or probabilities, log odds is a an asymptotically Gaussian distributed quantity that is suitable for regressions such as $\log\theta = ax + b$, where x is a possible factor and a and b are regression parameters to be determined. This type of regression using log odds as the dependent variable is known as ''logistic regression'' and is widely used in medical trials for assessing the factors that control risk (Agresti 1996). The approach is justified by elegant theoretical arguments concerning generalized linear models (GLMs). Regression of log odds provides a natural way of quantifying the influence of various factors on forecast skill, and it would be interesting to use it to investigate the effect of model resolution, model parameters, forecaster stress, etc. on the forecast performance of an operational weather forecasting system.

It is important to realize that forecast skill does not necessarily imply anything about the possible utility or value of the forecasts. For rare catastrophic events such as tornadoes, the value comes from correctly forecasting the rare events (tornadoes) and not the nonevents (no tornadoes). Skill scores are measures of overall association between the forecasts and observations, and do not give the same information as forecast value, which depends on the particular needs of the forecast user. For example, Finley's tornado forecasts have a significant association with the observations, yet are of generally little useful value except perhaps to the rare individual who might incur a substantial loss if a tornado did not happen! The purpose of skill scores is to quantify the overall agreement between the forecasts and the observations, and so by definition should not depend on what the user considers to be important (e.g., tornado rather than not tornado as the event). Certain skill scores can, however, be useful in specific value calculations; for example, the Peirce skill score is of direct use in simple cost–loss decision models (Mason 1980).

This study has shown that the odds ratio is a useful measure for evaluating the skill of binary yes/no forecasts. The odds ratio can also, however, be used for verifying ''probabilistic'' forecasts in which forecasts are used to estimate the probabilities of a future event. By making ensembles of forecasts, it is possible to estimate the probability that a tornado might occur for each event. An $m \times 2$ contingency table can be compiled

over many such forecasts that consists of two columns for whether or not a tornado was observed and m rows for the number of times forecast probabilities fell into m distinct probability ranges, for example, $p = 0.0$–0.1, 0.1–0.2, . . . , 0.9–1.0 [as explained in more detail in Harvey et al. (1992)]. The odds ratio can then be calculated for different probability thresholds by accumulating the number of events in the probability classes into two classes: one above and one below the threshold. This could be a promising new direction for evaluating the overall forecast skill of probabilistic forecasts.

Acknowledgments. I am grateful to Mike Harrison for arousing my interest in categorical forecast verification and to Ian Mason for interesting discussions and articles concerning the introduction of signal detection ideas into forecast verification. I also thank Francisco Doblas-Reyes, Daniel Rousseau, and John Thornes for stimulating discussions concerning the subtleties and the often nontrivial interpretation of forecast trials. For helpful and expert remarks on statistical matters, I am indebted to Alain Agresti, Philippe Besse, Antoine Falguerolles, and Ian Jolliffe. I also wish to thank Philippe Besse for etymological discussions concerning the rather odd word ''odds.''[4]

REFERENCES

Agresti, A., 1996: *An Introduction to Categorical Data Analysis.* John Wiley and Sons, 290 pp.

——, and B. A. Coull, 1998: Approximation is better than ''exact'' for interval estimation of binomial proportions. *Amer. Stat.,* **52,** 1–7.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland, 1975: *Discrete Multivariate Analysis: Theory and Practice.* The MIT Press, 557 pp.

Doolittle, M. H., 1885: The verification of predictions. *Amer. Meteor. J.,* **2,** 327–329.

——, 1888: Association ratios. *Bull. Philos. Soc. Wash.,* **10,** 83–96.

Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.,* **1,** 85–88.

Fisher, R. A., 1990: *Statistical Methods, Experimental Design, and Scientific Inference.* Oxford University Press, 155 pp.

Flueck, J. A., 1987: A study of some measures of forecast verification. Preprints, *10th Conf. on Probability and Statistics in Atmospheric Sciences,* Edmonton, AB, Canada, Amer. Meteor. Soc., 69–73.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.,* **120,** 2709–2712.

Gilbert, G. F., 1884: Finley's tornado predictions. *Amer. Meteor. J.,* **1,** 166–172.

Goodman, L. A., and W. H. Kruskal, 1979: *Measures of Association for Cross Classifications.* Springer-Verlag, 146 pp.

Green, D. M., and J. A. Swets, 1966: *Signal Detection Theory and Psychophysics.* Wiley, 155 pp. (Reprinted by Robert E. Krieger Publishing Co., 1974.)

Hanssen, A. W., and W. J. A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Meded. Verh.,* **81,** 2–15.

Harvey, L. O., Jr., K. R. Hammond, C. M. Lusk, and E. F. Moss, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.,* **120,** 863–883.

Heidke, P., 1926: Berechnung des Erfolges und der Güte der Windstärkevorhersagen im Sturmwarnungdienst (Calculation of the success and goodness of strong wind forecasts in the storm warning service). *Geogr. Ann. Stockholm,* **8,** 301–349.

Macmillan, N. A., and C. D. Creelman, 1991: *Detection Theory: A User's Guide.* Cambridge University Press, 155 pp.

Mason, I. B., 1980: Decision-theoretic evaluation of probabilistic predictions. *Proc. WMO Symp. on Probabilistic and Statistical Methods in Weather Forecasting,* Nice, France, WMO, 219–228.

——, 1982: A model for the assessment of weather forecasts. *Aust. Meteor. Mag.,* **30,** 291–303.

——, 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

——, 1997: The weather forecast as a statistical decision: An outline of signal detection theory and ROC analysis in assessment of forecast quality. *Proc. Forecast and Warning Verification Workshop,* Melbourne, Australia, Australian Bureau of Meteorology, 1–10.

Matthews, R. A. J., 1996: Base-rate errors and rain forecasts. *Nature,* **382,** 766.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.,* **119,** 1590–1601.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting,* **8,** 281–293.

——, 1996: The Finley affair: A signal event in the history of forecast verification. *Wea. Forecasting,* **11,** 3–20.

——, 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.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science,* **4,** 453–454.

Pickup, M. N., 1982: A consideration of the effect of 500mb cyclonicity on the success of some thunderstorm forecasting techniques. *Meteor. Mag.,* **111,** 87–97.

Potts, J. M., C. K. Folland, I. T. Jolliffe, and D. Sexton, 1996: Revised ''LEPS'' scores for assessing climate model simulations and long-range forecasts. *J. Climate,* **9,** 34–53.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Seaman, R., I. Mason, and F. Woodcock, 1996: Confidence intervals for some performance measures of yes/no forecasts. *Aust. Meteor. Mag.,* **45,** 49–53.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. WMO/TD-No. 358, World Meteorological Organization, Geneva, Switzerland, 114 pp.

Stephenson, D. B., 1997: Correlation of spatial climate/weather maps and the advantages of using the Mahalanobis metric in predictions. *Tellus,* **49A,** 513–527.

Swets, J. A., 1973: The relative operating characteristic in psychology. *Science,* **182,** 990–1000.

——, 1986: Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychol. Bull.,* **99,** 100–117.

——, 1988: Measuring the accuracy of diagnostic systems. *Science,* **240,** 1285–1293.

——, and R. M. Pickett, 1982: *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* Academic Press, 253 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 465 pp.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.,* **104,** 1209–1214.

Yates, F., 1984: Tests of significance for 2×2 contingency tables. *J. Roy. Stat. Soc.,* **147A,** 426–463.

Yule, G. U., 1900: On the association of attributes in statistics. *Philos. Trans. Roy. Soc. London,* **194A,** 257–319.

[4] The word ''odds'' is derived from the Old Norse word odda meaning point or angle and is difficult to translate literally into French. The French word ''la cote'' is generally substituted.