Q. J. R. Meteorol. Soc. (2006), **132**, pp. 2905–2923 doi: 10.1256/qj.06.25

Measuring forecast skill: is it real skill or is it the varying climatology?

By THOMAS M. HAMILL[1]* and JOSIP JURAS[2]
[1]*NOAA Earth System Research Laboratory, Boulder, Colorado, USA*
[2]*Geophysical Institute, University of Zagreb, Croatia*

(Received 14 February 2005; revised 16 May 2006)

SUMMARY

It is common practice to summarize the skill of weather forecasts from an accumulation of samples spanning many locations and dates. In calculating many of these scores, there is an implicit assumption that the climatological frequency of event occurrence is approximately invariant over all samples. If the event frequency actually varies among the samples, the metrics may report a skill that is different from that expected. Many common deterministic verification metrics, such as threat scores, are prone to mis-reporting skill, and probabilistic forecast metrics such as the Brier skill score and relative operating characteristic skill score can also be affected.

Three examples are provided that demonstrate unexpected skill, two from synthetic data and one with actual forecast data. In the first example, positive skill was reported in a situation where metrics were calculated from a composite of forecasts that were comprised of random draws from the climatology of two distinct locations. As the difference in climatological event frequency between the two locations was increased, the reported skill also increased. A second example demonstrates that when the climatological event frequency varies among samples, the metrics may excessively weight samples with the greatest observational uncertainty. A final example demonstrates unexpectedly large skill in the equitable threat score of deterministic precipitation forecasts.

Guidelines are suggested for how to adjust skill computations to minimize these effects.

KEYWORDS: Brier skill score Contingency tables Ensemble forecasting Equitable threat score Forecast verification Probabilistic weather forecasts Relative operating characteristic

1. INTRODUCTION

This article will demonstrate that many commonly used systems of measurement ('metrics') in weather forecast verification are capable of reporting positive forecast skill in situations where the meteorologist would assume none truly exists, or the metrics may report different skill from that expected. Depending on the metric and the situation, this effect can be large or small. The unexpected skill is a consequence of inappropriately pooling data over subsets with different climatological event frequencies.

Our interest in this topic resulted from using conventional verification metrics and diagnosing unexpectedly large skill. For example, the first author used a common probabilistic metric, the relative operating characteristic, in a comparison of ensemble forecast methods (Hamill *et al.* 2000, Fig. 13). The author reported a relative operating characteristic curve for wind speed forecasts at five days lead that indicated a highly skilful forecast, different from what experience would suggest for this lead time. Juras (2000) discussed an unexpectedly large forecast skill, in a comment on an article by Buizza *et al.* (1999). It was indicated that chosen metrics might report unexpectedly large skill if climatological event frequencies varied within the verification area. This issue was also raised by Mason (1989) and less directly by other authors, including Buizza (2001, p. 2335), Stefanova and Krishnamurti (2002, p. 543), Atger (2003), Glahn (2004, p. 770), and Göber *et al.* (2004). Still, there are many authors who have applied common verification metrics, incorrectly assuming that the conventional method of calculation would result in zero skill for the reference, which is commonly assumed to be a random draw from the observed climatological distribution.

Here, section 2 will provide a brief review of the three chosen metrics that may be subject to 'mis-estimating' skill, the Brier skill score (Wilks 2006), the relative operating

* Corresponding author: NOAA Earth System Research Laboratory, Physical Sciences Division, R/PSD 1, 325 Broadway, Boulder, CO 80305-3328, USA. e-mail: tom.hamill@noaa.gov

characteristic (Swets 1973; Harvey *et al.* 1992) skill score, and the equitable threat score (Schaefer 1990). Many other metrics, such as the ranked probability skill score (Epstein 1969; Murphy 1971; Wilks 2006, p. 302), economic value diagrams (Richardson 2000; Palmer *et al.* 2000; Richardson 2001b; Zhu *et al.* 2002 and Buizza *et al.* 2003) and other contingency-table based threat scores, will not be discussed but can be assumed to be subject to the same effect. In addition to describing the conventional method of calculation of these metrics, section 2 will describe possible improved methods of calculation. Section 3 follows with two simple examples of how unexpected skill can be diagnosed from synthetic weather data when using the conventional methods of calculation. Section 4 demonstrates how large the mis-estimation effect can be for a common real-weather verification problem, the threat scores of short-range precipitation forecasts. Section 5 concludes with a discussion of the implications and how to adapt verification strategies to minimize or avoid this effect.

2. Review of three common verification metrics

Below, three general verification metrics are reviewed, the Brier skill score, relative operating characteristic (ROC) skill score, and the equitable threat score.

The long-used Brier score (Brier 1950) is a measure of the mean-square error of probability forecasts for a dichotomous (two-category) event, such as the occurrence/non-occurrence of precipitation. Wilks (2006, p. 284) has provided a review, and references to provide further background. The Brier score is often converted to a skill score, its value normalized by the Brier score of a reference forecast such as climatology (ibid). A Brier skill score (BSS) of 1.0 indicates a perfect probability forecast, while a BSS of 0.0 should indicate the skill of the reference forecast (see Mason (2004) for further discussion of whether a BSS of 0.0 indicates no skill).

The relative operating characteristic (ROC) has gained widespread acceptance in the past few years as a tool for probabilistic weather forecast verification. The ROC has been used for decades in engineering, biomedical, and psychological applications. The ROC measures the hit rate of a forecast against its false-alarm rate as the decision threshold (perhaps a quantile of a probabilistic forecast) is varied. It also can be understood as a graph of the tradeoff of Type I vs. Type II statistical errors in a hypothesis test (Swets 1973). The ROC's application in meteorology was proposed by Mason (1982), Stanski *et al.* (1989), and Harvey *et al.* (1992). The ROC was recently made part of the World Meteorological Organization's verification standard (WMO 1992). Characteristics of the ROC have been discussed by Buizza *et al.* (1998), Mason and Graham (1999, 2002), Juras (2000), Wilson (2000), Buizza *et al.* (2000a,b), Wilks (2001), Kheshgi and White (2001), Kharin and Zwiers (2003), Mason (2003), and Marzban (2004). The technique has been used to diagnose ensemble forecast accuracy, for example by Buizza and Palmer (1998), Buizza *et al.* (1999), Hamill *et al.* (2000), Palmer *et al.* (2000), Richardson (2000, 2001a,b), Wandishin *et al.* (2001), Ebert (2001), Mullen and Buizza (2001, 2002), Bright and Mullen (2002), Yang and Arritt (2002), Legg and Mylne (2004), Zhu *et al.* (2002), Toth *et al.* (2003), and Gallus and Segal (2004). Harvey *et al.* (1992) provided a thorough review of the concepts underlying the ROC. In subsequent discussion, we will discuss the skill score 'ROC*SS*' derived from the ROC.

The equitable threat score (ETS) provides one of many ways of summarizing the ability of a deterministic prediction to forecast a dichotomous (two-category) event correctly. The ETS will produce a score of 1.0 for a perfect forecast, and random forecasts should be assigned a value of 0.0. The ETS is commonly used to evaluate

the skill of forecasts, especially precipitation. See, for example, Rogers *et al.* (1995, 1996), Hamill (1999), Bayler *et al.* (2000), Stensrud *et al.* (2000), Xu *et al.* (2001), Ebert (2001), Gallus and Segal (2001), Chien *et al.* (2002), and Accadia *et al.* (2003).

The methods for computing these metrics are now discussed, starting with the probabilistic metrics. The BSS and ROC will be generated from ensemble forecasts, though they can be generated from any probabilistic forecast.

Start by defining a dichotomous event of interest, such as occurrence/non-occurrence of precipitation, or temperature above or below a threshold. Let $\mathbf{X}_e(j) = [X_1(j), \ldots, X_n(j)]$ be an n-member ensemble forecast of the relevant scalar variable (again, precipitation or temperature) for the jth of m samples (taken over many case days and/or locations). The ensemble at that day and location is first sorted from lowest to highest. This sorted ensemble is then converted into an n-member binary forecast $\mathbf{I}_e(j) = [I_1(j), \ldots, I_n(j)]$ indicating whether the event was forecast (= 1) or not forecast (= 0) by each member. The observed weather is also converted to binary, denoted by $I_o(j)$.

(*a*) *Brier skill scores*

Assuming that each member forecast is equally likely, a forecast probability $p_f(j)$ for the jth sample is calculated from the binary ensemble forecasts:

$$p_f(j) = \frac{1}{n}\sum_{i=1}^{n} I_i(j). \tag{1}$$

The Brier score of the forecast BS_f is calculated as

$$BS_f = \frac{1}{m}\sum_{j=1}^{m}\{p_f(j) - I_o(j)\}^2. \tag{2}$$

A Brier skill score (BSS) is commonly calculated as

$$\mathrm{BSS} = 1 - \frac{BS_f}{BS_c}, \tag{3}$$

where BS_c is the Brier score of the reference probability forecast, commonly the probability of event occurrence from climatology.

Ideally, the climatological probabilities would be determined from independent data, but commonly they are calculated from the sample observed data. In the conventional method of calculation, an average climatology p_c is used:

$$p_c = \frac{1}{m}\sum_{j=1}^{m} I_o(j), \tag{4}$$

in which case the reference Brier score of climatology used in Eq. (3) is

$$BS_c = \frac{1}{m}\sum_{j=1}^{m}\{p_c - I_o(j)\}^2. \tag{5}$$

The conventional method of calculation of the BSS in Eqs. (1)–(5) may report a score that differs from what the meteorologist may expect if the climatological event frequency is known to vary among the m samples (section 3). Consequently, we propose

some alternative methods of formulation of the scores and shall discuss the change in skill that was reported under the new calculations.

Suppose the samples could be split up into n_c subsets, each with a distinct climatological event frequency. Let $p_c(k)$ be the climatological event frequency in the kth of the n_c subsets. Also, let there be $n_s(k)$ samples in this subset, and let $\mathbf{r}_k = [r(1), \ldots, r(n_s(k))]$ be the associated set of sample indices from the m samples. Then suppose the Brier score of climatology is calculated separately for each subset with a different climatology:

$$\widehat{BS}_c(k) = \frac{1}{n_s(k)} \sum_{j=1}^{n(k)} [p_c(k) - I_o\{r(j)\}]^2. \tag{6}$$

A possible alternative calculation of the Brier score of climatology would then be to calculate a sample weighted average:

$$\widehat{BS}_c = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} \widehat{BS}_c(k). \tag{7}$$

The BSS would then calculated following Eq. (3), replacing BS_c with $\widehat{BS}_c$.

A third possible alternative for calculating the BSS would be to calculate the Brier Score of the forecasts separately for the different subsets with climatological event frequencies, just as was done with the climatological forecast in Eq. (6):

$$\widehat{BS}_f(k) = \frac{1}{n_s(k)} \sum_{j=1}^{n_s(k)} [p_f\{r(j)\} - I_o\{r(j)\}]^2. \tag{8}$$

Then the BSS would be computed as a sample-weighted average of the skill scores for each distinct climatological regime:

$$\overline{BSS} = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} \left\{ 1 - \frac{\widehat{BS}_f(k)}{\widehat{BS}_c(k)} \right\}. \tag{9}$$

This may better conform with forecaster intuition, e.g., if two locations with equal numbers of samples have BSSs of 0.0 and 1.0, a skill of 0.5 will be reported.

(*b*) *ROC diagrams and the ROC skill score*

For ensembles, the ROC is a curve that indicates the relationship between hit rate and false alarm rate as different sorted ensemble members are used as decision thresholds. The area under the ROC curve can be used in the calculation of a probabilistic skill score. The conventional method of calculation of the ROC from ensembles typically starts with the population of 2×2 contingency tables, with separate contingency tables tallied for each sorted ensemble member. The contingency table (Table 1) has four elements: $\Gamma_i = [a_i, b_i, c_i, d_i]$. These elements indicate the proportion of hits, false alarms, misses and correct rejections, respectively, when the value of the ith sorted member is used as the forecast. The contingency table is populated using data over all m samples.

The hit rate (*HR*) for the ith sorted member forecast is defined as

$$HR_i = \frac{a_i}{a_i + c_i}. \tag{10}$$

TABLE 1. CONTINGENCY TABLE FOR THE iTH OF n SORTED MEMBERS AT THE jTH LOCATION

		Event observed?	
		YES	NO
Event forecasted by the ith member?	YES	a_i	b_i
	NO	c_i	d_i

Entries in the four cells of the table denote the proportions of trials which result in hits (a_i), false alarms (b_i), misses (c_i), and correct rejections (d_i).

Similarly, the false alarm rate is defined as

$$FAR_i = \frac{b_i}{b_i + d_i}. \tag{11}$$

This prototypical ROC is a plot of HR_i (ordinate) vs. FAR_i (abscissa), $i = 1, \ldots, n$. A ROC curve that lies along the diagonal $HR = FAR$ line is commonly believed to indicate no skill; a curve that sweeps out maximal area, as far toward the upper left corner as possible, is believed to indicate maximal skill. The ROC is commonly summarized through the integrated area under the ROC curve, or AUC. A perfect forecast $AUC_{\text{perf}} = 1.0$, and forecasts that are random draws from climatology are presumed to provide an $AUC_{\text{clim}} = 0.5$. In order to calculate the forecast area AUC_{f}, for the n-member ensemble let us assume the existence of fictitious zeroth and $(n + 1)$th ensemble members to provide boundary conditions $HR_0 = 0.0$, $FAR_0 = 0.0$, $HR_{n+1} = 1.0$, and $FAR_{n+1} = 1.0$. Then an approximate integral AUC_{f} can be calculated as

$$AUC_{\text{f}} = \sum_{i=1}^{n+1} \frac{(FAR_i - FAR_{i-1})(HR_i + HR_{i-1})}{2} \tag{12}$$

(there are other valid methods of calculation). Commonly, a skill score ROCSS is calculated from AUC_{f} (Wilks 2006, p. 295):

$$\text{ROC}SS = \frac{AUC_{\text{f}} - AUC_{\text{clim}}}{AUC_{\text{perf}} - AUC_{\text{clim}}} = \frac{AUC_{\text{f}} - 0.5}{1.0 - 0.5} = 2\,AUC_{\text{f}} - 1. \tag{13}$$

As will be demonstrated in section 3, the conventional method of calculation of the ROC and ROCSS can result in an estimation of skill where none was expected if the climatological event frequency varies among samples. Hence, we outline a possible alternative method of calculation of ROC area and skill. Assume, as with the BSS, that we can divide up the samples into n_{c} subsets with distinct climatological event frequencies. Then an alternative method for calculation of the ROC area would be to calculate it separately for each subgroup and produce a weighted-average ROC area, which we shall call $\overline{AUC}_{\text{f}}$. Using the $n_{\text{s}}(k)$ samples in the kth subset, the hit rates and false alarm rates for the kth climatology are

$$\widehat{HR}_i(k) = \frac{a_i(k)}{a_i(k) + c_i(k)} \tag{14}$$

and

$$\widehat{FAR}_i(k) = \frac{b_i(k)}{b_i(k) + d_i(k)}. \tag{15}$$

From this, the area under the ROC curve for the kth subset can be calculated in a manner analogous to Eq. (12), providing $\widehat{AUC}_{\rm f}(k)$. Then, as was done with the BSS, a sample-weighted $\overline{AUC}_{\rm f}$ is calculated according to

$$\overline{AUC_{\rm f}} = \sum_{k=1}^{n_{\rm c}} \frac{n_{\rm s}(k)}{m} \widehat{AUC_{\rm f}}(k), \tag{16}$$

and a skill score is calculated using Eq. (13), substituting $\overline{AUC}_{\rm f}$ for $AUC_{\rm f}$.

(*c*) *Equitable threat score*

Assume now that we are evaluating deterministic forecasts rather than ensembles. The conventional method of calculating the ETS assumes Table 1 is populated with all the samples available (here we drop the $_i$ subscript in Table1 denoting the ensemble member number). The equation for the ETS is

$$\text{ETS} = \frac{a - a_{\rm r}}{a + b + c - a_{\rm r}}, \tag{17}$$

where a_r is the expected fraction of hits for a random forecast

$$a_{\rm r} = \frac{(a + c)(a + b)}{a + b + c + d}. \tag{18}$$

As with the other scores, we shall show in sections 3 and 4 that this conventional method of calculation will produce an unexpectedly high estimate in situations where the climatology varies. An alternative method of calculation of the ETS respects the possibility of different regions with different climates. Again, assume we have $n_{\rm c}$ contingency tables, each associated with samples with a distinct climatological event frequency. For the kth distinct climatology we thus construct a separate contingency table and calculate the threat score $\widehat{ETS}(k)$. An alternative, sample-weighted ETS is then calculated as

$$\overline{ETS} = \sum_{k=1}^{n_{\rm c}} \frac{n_{\rm s}(k)}{m} \widehat{ETS}(k). \tag{19}$$

3. Example of skill overestimation: synthetic data at two independent locations

Using synthetic data, we now illustrate two general problems with verification metrics calculated in the conventional manner. Firstly, they may report skill even when the forecasts are samples from climatology. This may occur when the overall sample is comprised of subsamples that are drawn from different climatological distributions. Secondly, when the climatological uncertainty of event occurrence varies among samples, the skill scores may reflect an uneven weighting of the sample data.

(*a*) *Positive skill diagnosed from reference climatological forecasts*

Let us suppose that a hypothetical planet is covered by a global ocean interrupted only by two small, isolated islands, and that island weather forecasting is utterly impossible on this planet; the best one can do is to forecast the (stationary) climatological probability distribution appropriate to each island. Given that the weather appears random to residents on each island, one would expect a skill score to report zero skill, a desired attribute that is part of the property known as 'equitability' (Gandin and Murphy 1992; Wilks 2006, p. 274).

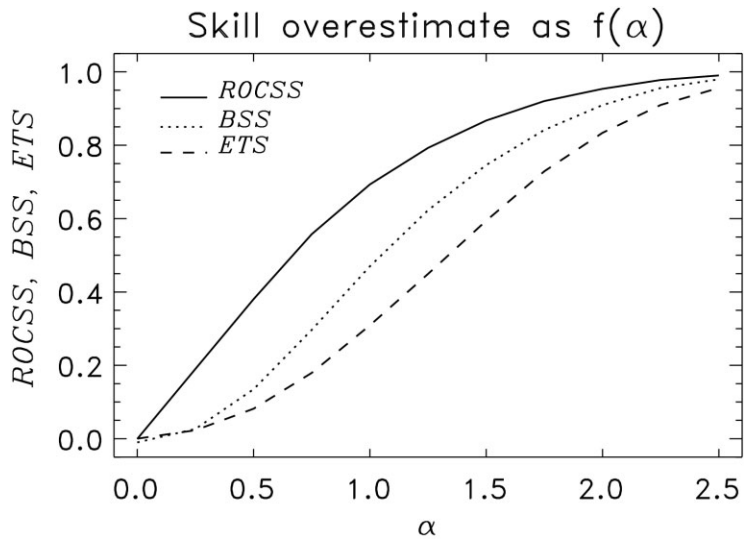

Figure 1. Three conventional verification metrics as functions of the parameter α: relative operating characteristic skill score (ROC*SS*), Brier skill score (BSS) and equitable threat score (ETS). The parameter α describes the mean observed climatological value (say of temperature) at island 1, while island 2 has a mean value of $-\alpha$.

To simulate this scenario, assume that at island 1, on each day the observed daily maximum temperature was a draw from a normal distribution with a mean of α and a standard deviation of 1.0 (the specific units of temperature are unimportant in this thought experiment). We denote this normal distribution by $\sim N(+\alpha, 1)$. We also generated a 100-member ensemble each day to calculate the BSS and ROC*SS* and a single-member deterministic forecast to calculate the ETS. In each instance the forecasts were also $\sim N(+\alpha, 1)$ and were uncorrelated with each other and with the observation. On island 2, each day's observed daily maximum value $\sim N(-\alpha, 1)$, and again a 100-member ensemble and deterministic forecast were drawn from $\sim N(-\alpha, 1)$, with uncorrelated forecasts and observations. We will consider the event that the temperature was greater than zero. Forty-thousand days of forecasts and observations were generated for island and each value of α, and we examine the skill scores as α increases from zero, that is, as the two islands' climatologies grow increasingly different.

Figure 1 synthesizes the overestimate of the scores by each metric as a function of α when the BSS was calculated by computing the pooled samples by Eqs. (4) and (5), the ROC*SS* was calculated using Eqs. (10)–(13), and the ETS was calculated using Eqs. (17) and (18). Hereafter, these will be called 'the conventional methods' of calculation. The expected skill should be zero regardless of the value of α, for forecasts were always drawn from the reference climatological distribution appropriate to that island. However, as α was increased, the diagnosed forecast skill also increased.

What was the source of the skill estimates being larger than expected? On island 1, the climatological probability of the observed being greater than zero increased with α, while on island 2 it decreased. The probabilities estimated from the ensemble behaved similarly, increasing with α on island 1 and decreasing on island 2. However, each of the conventional methods of calculating the scores implicitly assumed that the reference climatological event probability for all forecasts at all α was a fixed 0.5, a consequence of pooling the data from both islands together. Hence as α increased, the randomly

drawn forecasts became increasingly sharp and accurate relative to this nonspecific composite climatology. The random forecasts from each island were awarded higher and higher scores based merely on the increasing differences in the two islands' mean values, not through any intrinsic improvement in forecast skill. This illustrates that these scores may report unexpectedly large skill in situations where the climatologies differ among the samples used to populate the contingency tables; they credit a forecast with having skill when the climatologies of the individual samples are different from the climatology of the combined samples. In this example, the more the climatologies differed, the larger the diagnosed skill.

Though not shown here, an overestimation of skill would still have occurred even if forecasts on each island were positively correlated with the observed value and thus skilful. In such a situation, the actual skill would have been inflated by an additional amount due to the compositing of the two islands' climatology. This inflation of skill also makes it more difficult to evaluate potential forecast improvements. When α was very large, a forecast was scored as nearly perfect regardless of whether or not the forecast actually was nearly perfect. The difference between good and mediocre forecasts is thus shrunk, complicating the task of evaluating whether one model was better than another.

Consequently, the preferred course of action when the underlying climatology varies among samples is to analyse the data separately for each distinct climatological regime. A similar and more general conclusion was arrived at in the classic paper on 'Simpson's Paradox' (Simpson 1951; see comment 7 on second-order interactions*). Cochran (1954) also is unambiguous with regards to inferences from contingency tables:

> 'One method that is sometimes used is to combine all the data into a single 2×2 table . . . this procedure is legitimate only if the probability p of an occurrence (on the null hypothesis) can be assumed to be the same in all the individual 2×2 tables. Consequently, if p obviously varies from table to table, or we suspect that it may vary, this procedure should not be used.'

Cochran also proposed a statistical test to determine if contingency table data can be added; Mantel and Haenszel (1959) proposed a related test, and Agresti (2002, p. 231) provided a summary. Unfortunately, the Cochran and Mantel–Haenszel tests may be difficult to apply in meteorological verification, for one of the underlying assumptions is that the samples used to populate the contingency tables are independent. In meteorological verification, two samples may come from adjacent grid points that will in fact have correlated errors.

The meteorological statistician may sometimes still desire a single-number summary of the skill of the forecast, especially if the sample size of the forecasts is limited in each region with a different climatological event frequency. To preserve the desirable property of ensuring that random draws from the no-skill reference are evaluated as having null skill, the method of calculating the skill scores could be reformulated or the problem could be transformed to eliminate the effect of the varying climatology. For example, had the BSS been calculated with Eqs. (3), (6), and (7) or Eqs. (6), (8), and (9), had the ROC*SS* been calculated with Eqs. (13)–(16), and had the ETS been calculated separately at each island and then averaged using Eq. (19), the reported scores would have been zero within sampling error. Another way to report the expected

* Simpson actually asserts something even more rigorous: contingency-table data can be added only when there are no 'second-order interactions' in the contingency tables. These interactions may occur due to differences in climatological event frequency, but they also may occur in situations where the forecast skills were different between the subsets.

zero skill would be to change the test threshold to one where the climatological event frequencies were identical among sub-samples. For example, change the test threshold for temperature from 'greater than zero' to 'exceeding the 50th percentile of each individual island's climatological distribution'. Of course, reformulating the verification problem in this manner may not address the underlying question asked by the researcher.

(*b*) *Skill contributions weighted toward samples with larger observed uncertainty*

This experimental setup will illuminate how samples with different underlying climatological uncertainty* can be unequally weighted, affecting the computation of skill. Our two-island scenario is now altered; consider the event that the daily maximum temperature was greater than 2.0. Island 1's observed maximum temperature was randomly drawn from a $\sim N(0, 1)$ distribution, the forecasts were also $\sim N(0, 1)$, and forecast and observations were uncorrelated. On island 2, the observed and forecasted temperatures were drawn from $N(0, \beta)$ distributions, and these two values of temperature were correlated at 0.9. The value of β varied between 1 and 3. Other aspects such as the ensemble size and number of days were the same as in the previous experiment.

As β increased, at island 2 the forecasted and observed event frequency increased (Fig. 2). Ideally, the reported composite skills using the conventional methods would not change much as β changed, for the forecasted–observed correlation never changed even though island 2's spread changed.

Figures 3 (a, b and c) show that on island 1, skills remained near zero in each of the three metrics when using the conventional methods. On island 2, skill was near 1.0, and increased (BSS and ETS) or decreased (ROC*SS*) slightly with increasing β. When combined over both islands, the overall skill increased as β increased, as did the climatological event frequency and the uncertainty. Hence, the conventional methods apparently more heavily weighted the contribution from island 2 as β increased.

An examination of contingency tables for deterministic forecasts illuminates why the overall ETS was more heavily weighted toward island 2's contribution (Tables 2–4). Table 2 reports island 1's contingency table, Table 3 reports island 2's when $\beta = 1$, and Table 4 reports island 2's when $\beta = 3$. The ETS for island 1 alone was -0.0022, the ETS for island 2 and $\beta = 1$ was 0.4195; the combined ETS when $\beta = 1$ was 0.193 which gives nearly equal weight to the contribution of each island. Note that when $\beta = 1$ the climatological event frequencies were very similar: 0.0232 at island 1 and 0.0288 at island 2. However, for $\beta = 3$, the climatological event frequency at island 2 was 0.26, and its ETS was 0.532. The combined ETS for $\beta = 3$ was 0.499 and so much closer to that of island 2 than 1. The unequal weighting is illuminated by considering the sums of the contingency tables. Note for example that the 'hits' in the combined table for $\beta = 3$ (combining Tables 2 and 4) were determined almost exclusively by the hits from island 2, which contributed more than 98%.

This second example showed another undesirable property of the conventional method of calculating verification scores, namely that the weighting of samples is related to the observed event uncertainty. This may distort the calculation skill of important variables like heavy precipitation (see the example in the appendix of the paper by Hamill and Whitaker (2006)). In locations where heavy precipitation is quite rare (observational uncertainty small), the climatological reference produces a low-error

* Uncertainty here refers to a measure of the intrinsic variability of the observations, as in the Brier score decomposition (Wilks 2006, p. 286). Given a climatological event probability p_c, the uncertainty is $p_c(1 - p_c)$, which is maximized when $p_c = 0.5$ and minimized when $p_c = 1.0$ or 0.0.

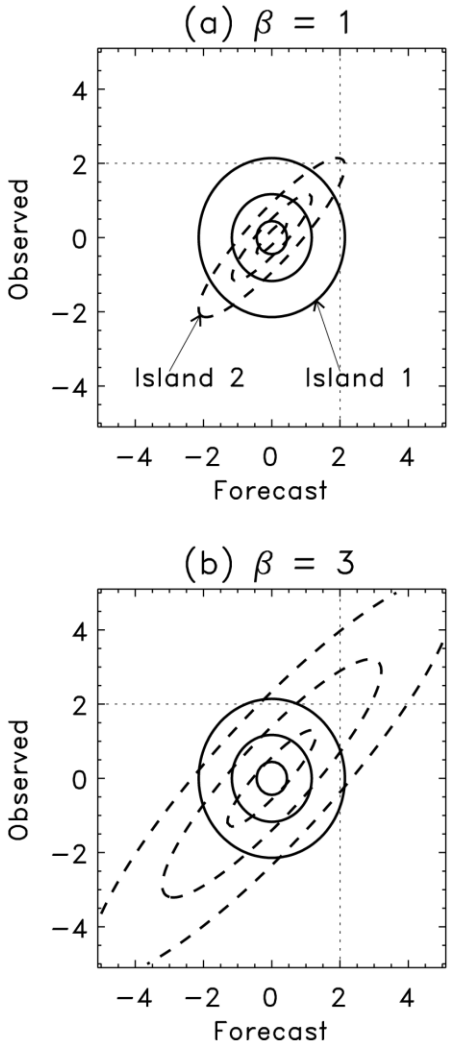

Figure 2. Illustration of experimental design (see subsection 3(b) of the text). Forecasts of temperature are simulated at two hypothetical islands. The first island has forecasts and observations which are uncorrelated; the second island has forecasts and observations correlated at 0.90. On both islands, the observed and forecasted values are both normally distributed about the mean, which is taken as zero. The standard deviation β for forecasts and observations is fixed at 1.0 in panel (a) for both islands. In panel (b), β is fixed at 1.0 for island 1 and 3.0 for island 2. Dotted lines indicate the event threshold of the forecasted or observed temperature being greater than 2.0.

forecast in most circumstances, and so a modest absolute forecast error can be evaluated as having negative skill relative to the climatology. Conversely, if heavy precipitation is more common (observational uncertainty larger), that same modest absolute forecast error may translate to a forecast with skill relative to the climatology, which is longer producing a low-error forecast in most circumstances. Hence the locations diagnosed as having more skill are commonly the ones with greater observational uncertainty; consequently, they may end up being more highly weighted in the calculation of the skill score, resulting in a skill larger than the average of skills at the constituent grid-points.

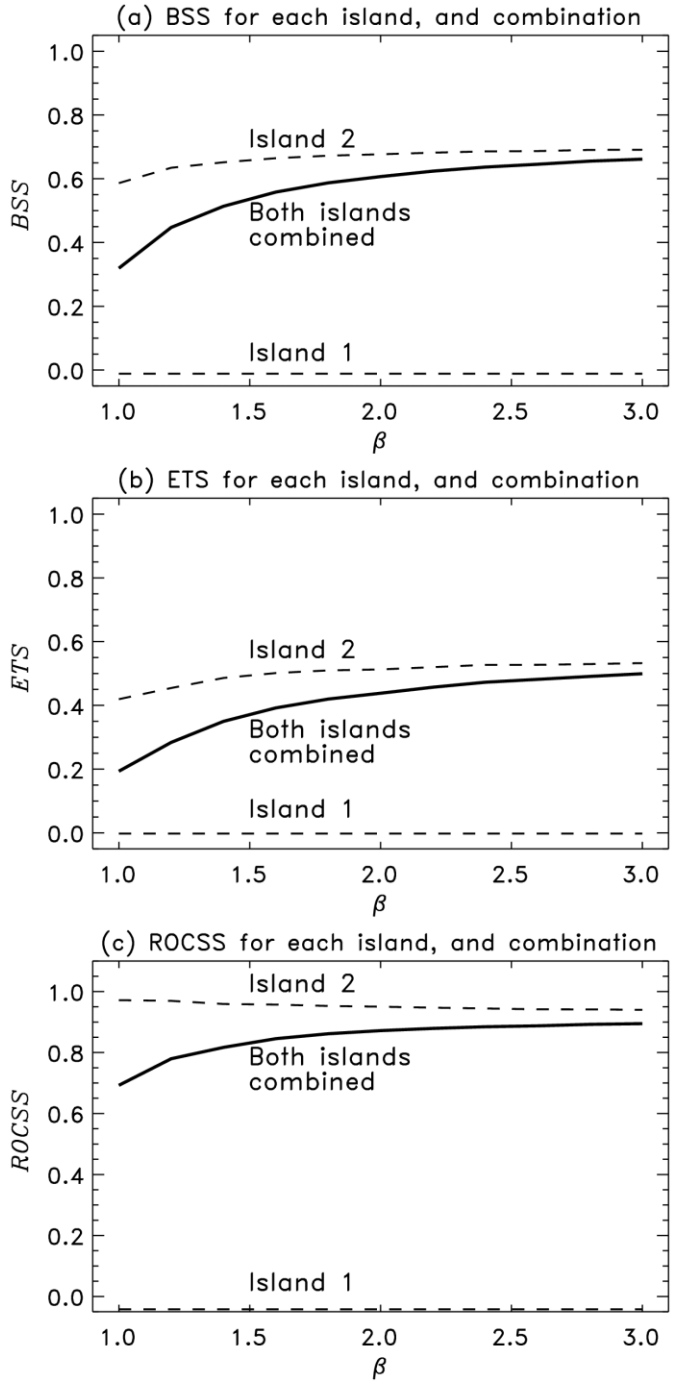

Figure 3. Scores for three verification metrics when applied separately to forecasts at islands 1 and 2, and when combined using the conventional methods of calculation: (a) Brier skill score (BSS); (b) equitable threat score (ETS) and (c) the ROC Skill Score (ROC*SS*). Scores are shown for a range of standard deviations β at island 2. See Figure 2 and subsection 3(b) for more on the experimental design.

TABLE 2. CONTINGENCY TABLE FOR ISLAND 1

		Event observed?	
		YES	NO
Event forecasted?	YES	0.004	0.0223
	NO	0.0228	0.954
	Total	0.0232	0.9763

See subsection 3(b). The observed event frequency = 0.0232 and the equitable threat score (ETS) = −0.0022.

TABLE 3. CONTINGENCY TABLE FOR HYPOTHETICAL ISLAND 2 WHEN $\beta = 1.0$

		Event observed?	
		YES	NO
Event forecasted?	YES	0.0171	0.0108
	NO	0.0117	0.9603
	Total	0.0288	0.9711

See subsection 3(b). The observed values of temperature have a standard deviation $\beta = 1.0$; the observed event frequency = 0.0288 and the equitable threat score (ETS) = +0.4195.

TABLE 4. CONTINGENCY TABLE FOR ISLAND 2 WHEN $\beta = 3.0$

		Event observed?	
		YES	NO
Event forecasted?	YES	0.2022	0.0597
	NO	0.0578	0.6802
	Total	0.2600	0.7399

See subsection 3(b). The observed values of temperature have a standard deviation $\beta = 3.0$; the observed event frequency = 0.2600 and the equitable threat score (ETS) = +0.5327.

It is conceptually possible that conventional methods could also underestimate skill. This would have happened, for example, had we repeated this experiment, but with forecasts and observations highly correlated at island 1 rather than at island 2. Practically, though, our experience suggests that skill tends to be more commonly overestimated (Hamill and Whitaker 2006).

The solutions proposed in the previous example may be useful here as well, with one exception. In this example, the calculation of the BSS cannot be fixed by defining BS_c using Eq. (7); it will yield a similar result to when Eq. (5) is used. Equation (7) will still effectively weight the samples with greater climatological uncertainty higher than samples with less climatological uncertainty. If Eq. (9) were used, the reported BSS would be a simple arithmetic average of the skill at the two islands. Similarly, the reported ROC*SS* will be an arithmetic average if calculated with Eqs. (13)–(16), as will be the ETS if calculated separately at each island and then averaged using Eq. (19).

4. Example of skill overestimation: equitable threat scores for numerical precipitation forecasts

Here we demonstrate that the ETS for real precipitation forecasts is subject to the same overestimation problem as with the synthetic data. The ETS is commonly used by the US National Weather Service to evaluate the skill of their deterministic precipitation forecasts. Typically, the ETS is estimated at fixed precipitation thresholds from a single contingency table populated over many days or months and over a wide geographic region such as the conterminous USA (CONUS).

To demonstrate the tendency to report a larger-than-expected ETS, a very large set of numerical forecasts was used. These forecasts were generated using the analogue forecast technique discussed by Hamill *et al.* (2006). The details of the forecast methodology can be found there but are not particularly important here. What is relevant is that a 25-year time series of gridded deterministic precipitation forecasts was produced, all using the same model and forecast technique. These forecasts have characteristics roughly similar to those of current operational forecasts. For the present demonstration, we limit ourselves to considering the ETS of the mean of a 5-member ensemble of analogue forecasts over the CONUS for January and February from 1979 to 2003. Both the forecast and the verification data (from the North American Regional Reanalysis, Mesinger *et al.* (2006)) are on a $\sim$32 km grid. We consider the 5 mm precipitation threshold.

Figure 4(a) illustrates the geographic dependence of the ETS on forecast location. Contingency tables and ETS were calculated separately for each grid point. The ETSs were much larger in the south-east USA and along the west coast than in the north-western Great Plains. Figure 4(b) provides the (climatological event) frequency of more than 5 mm rain falling during routine 24 h measuring periods. Note the strong relationship between the ETS and the event frequency, a characteristic previously described for a similar skill score by Mason (1989) and for the ETS by Göber *et al.* (2004). Since observational uncertainty is thus typically larger at grid points with higher ETS, we might expect to see the effect demonstrated in subsection 3(b), whereby an ETS calculated from the sum of all contingency tables across the CONUS will unduly weight the influence of the forecasts with the higher skill. Indeed, the ETS calculated from the contingency table sum using Eq. (17) was approximately 0.415. However, from Fig. 4(a) it is apparent that the large majority of grid points have ETS much below 0.415. When calculated using Eq. (19) after binning the climate into six categories*, the weighted-average ETS was much smaller, viz. $\sim$0.28 (Fig. 5).

The ETS estimation technique of Eq. (19) has drawbacks. Notably, the climatological event probability was defined by the sample event probability $(a + c)/(a + b + c + d)$. This assumption is reasonable in the present example of more than twenty years of winter forecasts. If the verification period is very short, on the other hand, then this sample event probability may be a poor estimate of the true long-term event probability. Ideally, a climatology should be estimated from a long time series of independent data, if available. If this is not possible, cross-validation techniques could be used to isolate the data being verified from the data being used to define the climatological event frequency. Nonetheless, these details should not obscure the main point: a substantially larger-than-expected threat score is possible when contingency table values are summed across grid points with different climatologies.

* Further subdivision into a greater number of categories did not increase the ETS appreciably.

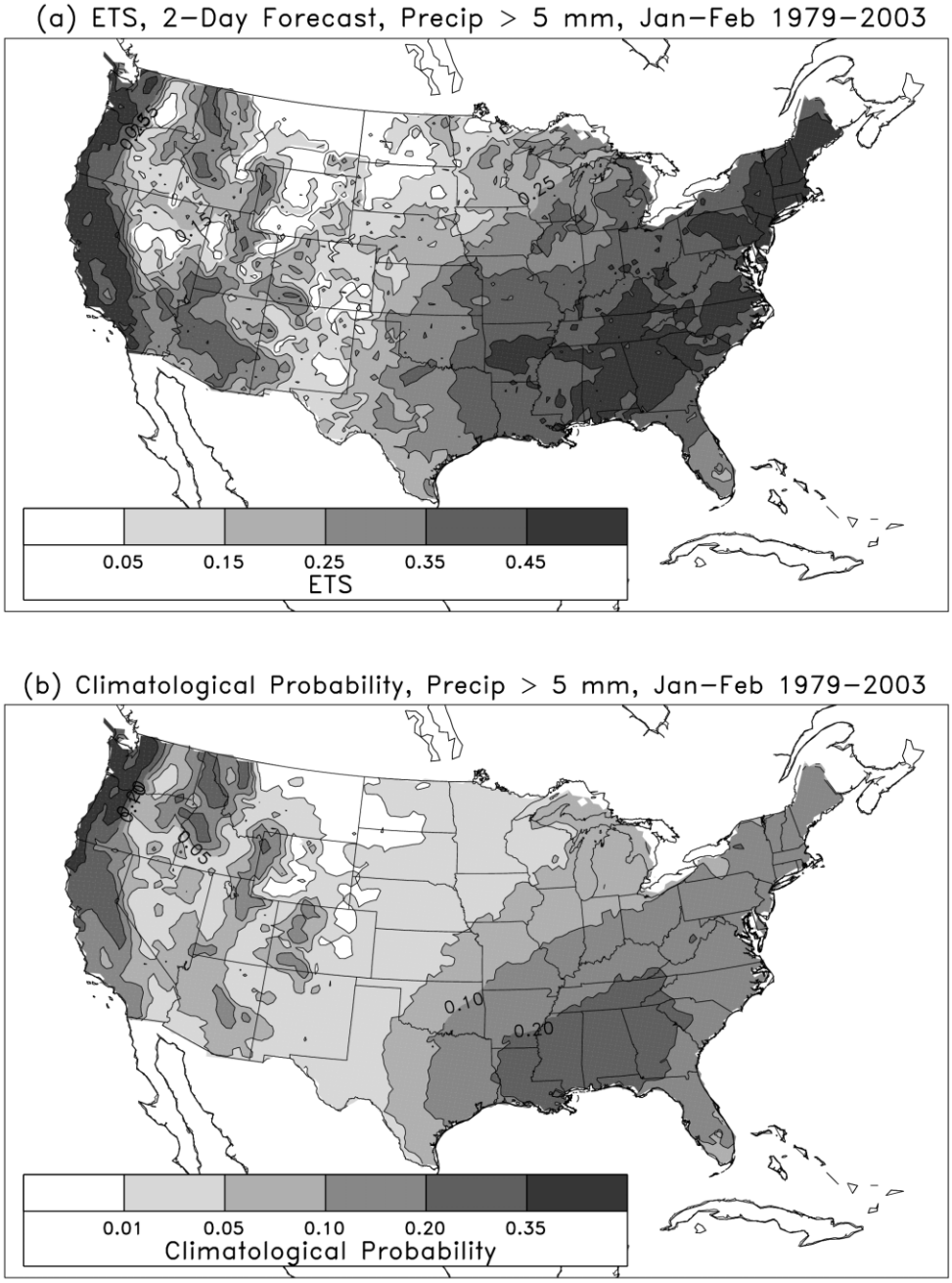

Figure 4. Equitable threat score (ETS) for 1–2 day (24–48 h) precipitation forecasts, using January and February 1979–2003 forecasted and analysed data, and (b) the climatological probability of precipitation greater than 5 mm for January and February.

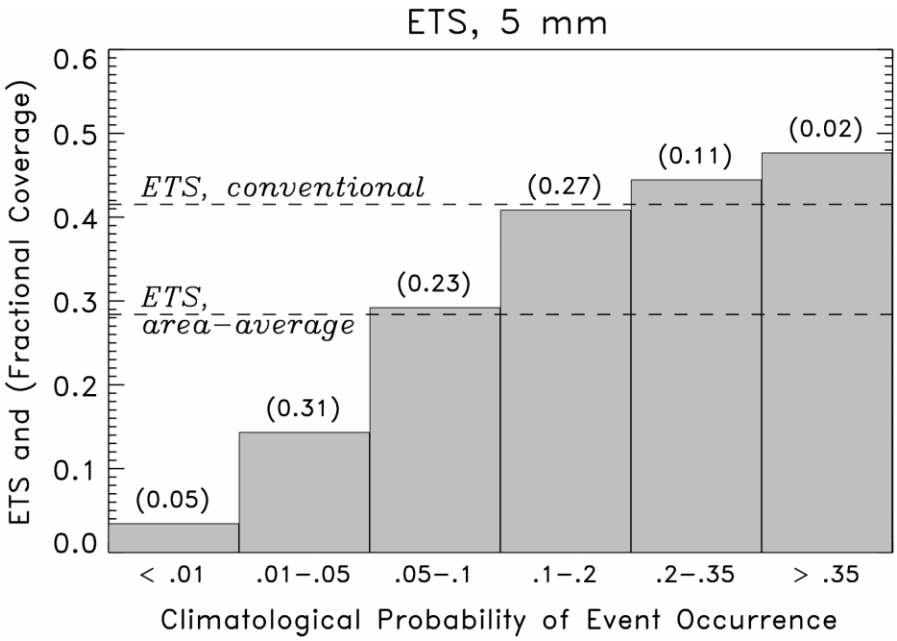

Figure 5. Equitable threat scores (ETSs) of forecasts of precipitation exceeding 5 mm when divided into six categories based on the climatological probability of the event occurring. Forecasted and analysed data are from January and February 1979–2003 over the conterminous USA. Brackets contain the proportion of grid points occurring in each category. The upper dashed line shows the ETS calculated using the conventional method (Eq. (17)) and the lower dashed line the area-averaged (population-weighted) ETS (Eq. (19)).

5. Conclusions

The preceding examples have demonstrated that the Brier skill score, relative operating characteristic, and the equitable threat score must be interpreted with care when verifying weather forecasts. In situations where the climatological event frequency differs between sample locations, these metrics may report skill which is different from that expected. The more the event frequencies differ, the more the skill may be wrongly estimated. By logical extension, skill may also be estimated wrongly if the verification samples are composited when they span different seasons or even different times of the day with different climatologies. Other scores, such as the ranked probability skill score and other contingency-table based scores, can be assumed to be subject to the same tendencies. These 'mis-estimates' can complicate the evaluation of model performance. Are two models nearly equal in their high degree of skill because they both provide high-quality forecasts? Or are they actually less skilful, and are differences in skill obscured by fictitious added skill from the varying climatology?

One primary reason why skill scores have been calculated as sums over sets with varying climatologies is that the sample sizes of the forecasts and observations are often small, and skills for the subsets may have a large sampling variability. A weighted-average skill over these subsets, as we have proposed, may not be resistant to outliers. Also, if independent observational data are not available to define the climatological event frequency for sub-samples, then this must be estimated from the same data used for model verification, potentially causing several additional problems. Firstly, the small sample size may result in large errors in estimating the climatological event frequency. Secondly, unless the observational data used to define the climatology are separated

from the observational data used for forecast verification to preserve independence (cross validation), the forecast skill may be underestimated; the error of the climatology will be diagnosed as smaller since it increasingly resembles the observed data as sample size decreases.

Clearly, talents of the statistical meteorologist will be put to the test when data are limited. While each situation may be different, one consideration should at least be to design the verification method to minimize the reported increase in skill introduced by varying climatologies, making at least relative inferences of skill (is model A more skilful than model B?) more trustworthy.

We propose two changes that both address the tendency to mis-estimate skill. First, if sample sizes are large enough, perform the calculations separately each for sub-sample with similar climatological event frequencies, as demonstrated for the equitable threat score in section 4. If the statistical meteorologist requires a single-number summary of the skill, consider weighted-average calculations similar to those proposed in section 2. Second, consider estimating skills for alternative events where the climatological event frequencies are the same for all samples, such as exceeding a quantile of the local climatological distribution (e.g., Zhu *et al.* (2002) or Fig. 5 of Buizza *et al.* (2003)). Then, regardless of whether the climatological means and variances are large or small, the fraction of events classified as 'yes' events is identical for different locations or times of the year.

We have two final recommendations: the *specific details* regarding how verification metrics are calculated should be fully described in journal articles and texts, since minor changes in the methodology can dramatically change the reported scores, and, whatever the chosen verification metric, it is prudent to verify that climatological forecasts report the expected no-skill result before proceeding.

ACKNOWLEDGEMENTS

Three anonymous reviewers are thanked for their constructive criticism. William 'Matt' Briggs (Cornell University) is thanked for his review and for pointing out the Mantel–Haenszel test. Dan Wilks (Cornell), Craig Bishop (US Navy/NRL), Beth Ebert (Australian BOM), Steve Mullen (U. Arizona), Simon Mason (IRI), Bob Glahn (NOAA/MDL), Neill Bowler and Ken Mylne (UK Met Office), Bill Gallus (Iowa State), Frederic Atger (Météo-France), Francois LaLaurette (EMCWF and Météo-France), Zoltan Toth (NCEP), and Jeff Whitaker (NOAA/ESRL/PSD) are thanked for their discussions and comments on an earlier version of this manuscript. This research was supported by National Science Foundation grants ATM-0130154 and ATM-0205612.

REFERENCES

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A. and Speranza, A. 2003 Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbour average method on high-resolution verification grids. *Weather and Forecasting*, **18,** 918–932

Agresti, A. 2002 *Categorical Data Analysis*, 2nd edn. Wiley Interscience, Hoboken, NJ, USA

Atger, F. 2003 Spatial and interannual variability of the reliability of the ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Weather Rev.*, **131,** 1509–1523

Bayler, G. M., Aune, R. M. and Raymond, W. H. 2000 NWP cloud initialization using GOES sounder data and improved modeling of non-precipitating clouds. *Mon. Weather Rev.*, **128,** 3911–3920

Brier, G. W. 1950 Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.*, **78,** 1–3

Bright, D. R. and Mullen, S. L.	2002	Short-range ensemble forecasts of precipitation during the southwest monsoon. *Weather and Forecasting*, **17,** 1080–1100
Buizza, R.	2001	Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Mon. Weather Rev.*, **129,** 2329–2345
Buizza, R. and Palmer, T. N.	1998	Impact of ensemble size on ensemble prediction. *Mon. Weather Rev.*, **126,** 2503–2518
Buizza, R., Petroliagis, T., Palmer, T. N., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A. and Wedi, N.	1998	Impact of model resolution and ensemble size on the performance of an Ensemble Prediction System. *Q. J. R. Meteorol. Soc.*, **124,** 1935–1960
Buizza, R., Hollingsworth, A., Lalaurette, F. and Ghelli, A.	1999	Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Weather and Forecasting*, **14,** 168–189
	2000a	Reply to comments by Wilson and by Juras. *Weather and Forecasting*, **15,** 367–369
Buizza, R., Barkmeijer, J., Palmer, T. N. and Richardson, D. S.	2000b	Current status and future development of the ECMWF ensemble prediction system. *Meteorol. Appl.*, **7,** 163–175
Buizza, R., Richardson, D. S. and Palmer, T. N.	2003	Benefits of increased resolution in the ECMWF ensemble prediction system and comparison with poor-man's ensembles. *Q. J. R. Meteorol. Soc.*, **129,** 1269–1288
Chien, F.-C., Kuo, Y.-H. and Yang, M.-J.	2002	Precipitation forecast of MM5 in the Taiwan area during the 1998 Mei-yu season. *Weather and Forecasting*, **17,** 739–754
Cochran, W. G.	1954	Some methods of strengthening the common χ^2 tests. *Biometrics*, **10,** 417–451
Ebert, E. E.	2001	Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Weather Rev.*, **129,** 2461–2480
Epstein, E. S.	1969	A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.*, **8,** 985–987
Gallus, W. A. Jr. and Segal, M.	2001	Impact of improved initialization of mesoscale features on convective system rainfall in 10-km Eta simulations. *Weather and Forecasting*, **16,** 680–696
	2004	Does increased predicted warm-season rainfall indicate enhanced likelihood of rain occurrence? *Weather and Forecasting*, **19,** 1127–1135
Gandin, L. S. and Murphy, A. H.	1992	Equitable skill scores for categorical forecasts. *Mon. Weather Rev.*, **120,** 361–370
Glahn, B.	2004	Discussion of verification concepts in 'Forecast verification: A practitioner's guide in atmospheric science'. *Weather and Forecasting*, **19,** 769–775
Göber, M., Wilson, C. A., Milton, S. F. and Stephenson, D. B.	2004	Fair play in the verification of operational quantitative precipitation forecasts. *J. Hydrol.*, **288,** 225–236
Hamill, T. M.	1999	Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, **14,** 155–167
Hamill, T. M. and Whitaker, J. S.	2006	Probabilistic quantitative precipitation forecasts based on reforecast analogues: theory and application. *Mon. Weather Rev.*, In press. Available at www.cdc.noaa.gov/people/tom.hamill/reforecast_analog_v2.pdf
Hamill, T. M., Snyder, C. and Morss, R. E.	2000	A comparison of probabilistic forecast from bred, singular-vector and perturbed observation ensembles. *Mon. Weather Rev.*, **128,** 1835–1851
Hamill, T. M., Whitaker, J. S. and Mullen, S. L.	2006	Reforecasts, an important dataset for improving weather predictions. *Bull. Am. Meteorol. Soc.*, **87,** 33–46
Harvey, L. O. Jr., Hammond, K. R., Lusk, C. M. and Mross, E. F.	1992	The application of signal detection theory to weather forecasting behavior. *Mon. Weather Rev.*, **120,** 863–883
Juras, J.	2000	Comments on 'Probabilistic predictions of precipitation using the ECMWF ensemble prediction system.' *Weather and Forecasting*, **15,** 365–366
Kharin, V. V. and Zwiers, F. W.	2003	On the ROC score of probability forecasts. *J. Climate*, **16,** 4145–4150
Kheshgi, H. S. and White, B. S.	2001	Testing distributed parameter hypotheses for the detection of climate change. *J. Climate*, **14,** 3464–3481

Legg, T. P. and Mylne, K. R.	2004	Early warnings of severe weather from ensemble forecast information. *Weather and Forecasting*, **19,** 891–906
Mantel, N. and Haenszel, W.	1959	Statistical aspects of the analysis of data from retrospective studies of disease. *J. National Cancer Institute*, **22,** 719–748
Marzban, C.	2004	The ROC curve and the area under it as performance measures. *Weather and Forecasting*, **19,** 1106–1114
Mason, I. B.	1982	A model for assessment of weather forecasts. *Aust. Meteorol. Mag.*, **30,** 291–303
	1989	Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteorol. Mag.*, **37,** 75–81
	2003	'Binary events'. Pp. 37–76 in *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Eds. I. T. Jolliffe and D. B. Stephenson. John Wiley, Hoboken, NJ, USA
Mason, S. J.	2004	On using 'climatology' as a reference strategy in the Brier and ranked probability skill scores. *Mon. Weather Rev.*, **132,** 1891–1895
Mason, S. J. and Graham, N. E.	1999	Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, **14,** 713–725
	2002	Areas beneath relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.*, **128,** 2145–2166
Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jovi, D., Woollen, J., Rogers, E., Berbery, E. H., Ek, M. B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D. and Shi, W.	2005	North American regional reanalysis. *Bull. Am. Meteorol. Soc*., **87,** 343–360
Mullen, S. L. and Buizza, R.	2001	Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Weather Rev.*, **129,** 638–663
	2002	The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Weather and Forecasting*, **17,** 173–191
Murphy, A. H.	1971	A note on the ranked probability score. *J. Appl. Meteorol.*, **10,** 155–156
Palmer, T. N., Branković, Č. and Richardson, D. S.	2000	A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Q. J. R. Meteorol. Soc.*, **126,** 2013–2033
Richardson, D. S.	2000	Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **126,** 649–667
	2001a	Ensembles using multiple models and analyses. *Q. J. R. Meteorol. Soc.*, **127,** 1847–1864
	2001b	Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol. Soc.*, **127,** 2473–2489
Rogers, E., Deaven, D. G. and DiMego, G. J.	1995	The regional analysis system for the operational 'early' Eta Model: original 80-km configuration and recent changes. *Weather Forecasting*, **10,** 810–825
Rogers, E., Black, T. L., Deaven, D. G., DiMego, G. J., Zhao, Q., Baldwin, M., Junker, N. W. and Lin, Y.	1996	Changes to the operational 'early' Eta analysis/forecast system at the National Centers for Environmental Prediction. *Weather and Forecasting*, **11,** 391–413
Schaefer, J. T.	1990	The critical success index as an indicator of warning skill. *Weather and Forecasting*, **5,** 570–575
Simpson, E. H.	1951	The interpretation of interaction in contingency tables. *J. R. Stat. Soc.*, **13,** 238–241
Stanski, H. R., Wilson, L. J. and Burrows, W. R.	1989	'Survey of common verification methods in meteorology'. Environment Canada Research Report 89-5. Available from Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin St., Downsview, Ontario, M3H 5T4, Canada

Stefanova, L. and Krishnamurti, T. N.	2002	Interpretation of seasonal climate forecast using Brier skill score, the Florida State University superensemble and the AMIP-I dataset. *J. Climate*, **15,** 537–544
Stensrud, D. J., Bao, J.-W. and Warner, T. T.	2000	Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Weather Rev.*, **128,** 2077–2107
Swets, J. A.	1973	The relative operating characteristic in psychology. *Science*, **182,** 990–1000
Toth, Z., Talagrand, O., Candille, G. and Zhu, Y.	2003	'Probability and ensemble forecasts'. Chapter 7 of *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, Hoboken, NJ, USA
Wandishin, M. S., Mullen, S. L., Stensrud, D. J. and Brooks, H. E.	2001	Evaluation of a short-range multimodel ensemble system. *Mon. Weather Rev.*, **129,** 729–747
Wilks, D. S.	2001	A skill score based on economic value for probability forecasts. *Meteorol. Appl.*, **8,** 209–219
	2006	*Statistical Methods in the Atmospheric Sciences*, 2nd edn. Academic Press, St Louis, MO, USA
Wilson, L. J.	2000	Comments on 'Probabilistic predictions of precipitation using the ECMWF ensemble prediction system'. *Weather and Forecasting*, **15,** 361–364
WMO	1992	*Manual on the Global Data Processing System*, section III, Attachments II.7 and II.8 (revised in 2002). World Meteorological Organization, Geneva, Switzerland. Available from http://www.wmo.int/web/www/DPS/Manual/WMO485.pdf
Xu, M., Stensrud, D. J., Bao, J.-W. and Warner, T. T.	2001	Applications of the adjoint technique to short-range ensemble forecasting of mesoscale convective systems. *Mon. Weather Rev.*, **129,** 1395–1418
Yang, Z. and Arritt, R. W.	2002	Tests of a perturbed physics ensemble approach for regional climate modeling. *J. Climate*, **15,** 2881–2896
Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne. K.	2002	The economic value of ensemble-based weather forecasts. *Bull. Am. Meteorol. Soc.*, **83,** 73–83