Q. J. R. Meteorol. Soc. (2002), **128**, pp. 2145–2166

Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation

By S. J. MASON* and N. E. GRAHAM
Scripps Institution of Oceanography, USA

(Received 5 November 2001; revised 23 April 2002)

SUMMARY

The areas beneath the relative (or receiver) operating characteristics (ROC) and relative operating levels (ROL) curves can be used as summary measures of forecast quality, but statistical significance tests for these areas are conducted infrequently in the atmospheric sciences. A development of signal-detection theory, the ROC curve has been widely applied in the medical and psychology fields where significance tests and relationships to other common statistical methods have been established and described. This valuable literature appears to be largely unknown to the atmospheric sciences where applications of ROC and related techniques are becoming more common. This paper presents a survey of that literature with a focus on the interpretation of the ROC area in the field of forecast verification. We extend these foundations to demonstrate that similar principles can be applied to the interpretation and significance testing of the ROL area. It is shown that the ROC area is equivalent to the Mann–Whitney U-statistic testing the significance of forecast event probabilities for cases where events actually occurred with those where events did not occur. A similar derivation shows that the ROL area is equivalent to the Mann–Whitney U-statistic testing the magnitude of events with respect to whether or not an event has been forecast. Because the Mann–Whitney U-statistic follows a known probability distribution, under certain assumptions it can be used to define the statistical significance of ROC and ROL areas and for comparing the areas of competing forecasts. For large samples the significance of either measure can be accurately assessed using a normal-distribution approximation.

KEYWORDS: Forecast verification Mann–Whitney U-test Probabilistic forecasts Signal-detection theory Student's t-test

1. INTRODUCTION

The relative operating characteristics (ROC) curve (Peterson and Birdsall 1953; Green and Swets 1966; Swets 1973, 1988; Mason 1982; Harvey *et al.* 1992; Mason and Graham 1999) is a useful method of representing the quality of deterministic and probabilistic detection and forecast systems. The ROC methodology, which was originally developed in the field of radar signal-detection theory as 'receiver operating characteristics' (Peterson and Birdsall 1953; Peterson *et al.* 1954), draws on pioneering work in statistical quality control theory from Bell Laboratories (e.g. the operational characteristics (OC) curve, Dodge and Romig 1929; Shewhart 1931; Dooley 2000) and shares basic attributes from the Neyman–Pearson lemma (Neyman and Pearson 1933) and the concept of the 'power' of a statistical test (e.g. Sokal and Rohlf 1973). The technique has been used extensively in the fields of psychological and medical test evaluation (e.g. Green and Swets 1966; Swets 1973, 1979, 1988, 1995; Egan 1975; Metz 1978; Falmagne 1985; Begg 1991; Swets *et al.* 2000)†, and is being applied increasingly in the atmospheric sciences (e.g. Mason 1982; Winston 1988; Buizza and Palmer 1998; Buizza *et al.* 1998, 1999; Mason *et al.* 1999; Graham *et al.* 2000; Palmer *et al.* 2000; Richardson 2000; Zhang and Casey 2000; Buizza 2001; Frogner and Iversen 2001; Mullen and Buizza 2001; Thorncroft and Pytharoulis 2001; Wandishin *et al.* 2001; Wilks 2001; Zhu *et al.* 2002). It has become part of the WMO Standardized Verification System for assessing the quality of forecasts (Stanski *et al.* 1989; WMO 2000).

* Corresponding author: Climate Research Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093-0230, USA. e-mail: smason@ucsd.edu

† An extensive list of references to the ROC and its application in the medical field is available from http://www.spl.harvard.edu:8000/ pages/ppl/zou/roc.html

The area under the ROC curve characterizes the quality of a forecast system by describing the system's ability to anticipate correctly the occurrence or non-occurrence of pre-defined 'events'. In constructing a ROC curve, forecasts are expressed in binary as 'warnings' or 'no warnings' indicating whether or not the defined event is expected to occur. In this context, the measure is based on the likelihood-base rate factorization according to the formalism of Murphy and Winkler (1987). For probabilistic forecast systems (in which forecasts are expressed in terms of the probability that an event will occur) the probability at which a warning is issued is varied across a range of thresholds (see Green and Swets 1966; Mason 1979). For each threshold, the correspondence between the forecasts (a sequence of warnings or non-warnings) and observations (a sequence of events or non-events) is examined. This correspondence is described by a two-component vector defined by the 'hit rate' (the proportion of events for which a warning was correctly issued) and the 'false-alarm rate' (the proportion of non-events for which a warning was incorrectly issued). The hit rate and false-alarm rate give a (two-dimensional) coordinate in ROC space and collectively define the ROC curve.

A related procedure, relative operating levels (ROL), which is a development of the response analysis characteristic of Swets and Birdsall (1967), and which is based on a calibration-refinement factorization (Murphy and Winkler 1987), has been proposed in which the 'intensity' of the events varies across an ideally continuous scale, and warnings are issued according to fixed criteria (Mason and Graham 1999). As with the ROC curve, forecasts are binary statements concerning whether or not a defined event is expected to occur. The observations are also expressed in binary terms by defining a given observation as an 'event' if the observation exceeds some pre-defined threshold, and as a 'non-event' otherwise. For the ROL curve, the intensity distinguishing events from non-events is varied across a range of thresholds. Exactly as with the ROC technique, the correspondence between the forecasts (a sequence of warnings and non-warnings) and the observations (a sequence of events and non-events) is examined for each threshold. This correspondence is again defined by a two-component vector, in this case composed of the 'correct-alarm ratio' (defined as the proportion of cases in which an event occurred following a warning) and 'miss ratio' (defined as the proportion of cases in which an event occurred following no warning). The correct-alarm and miss ratios then give coordinates in ROL space and collectively (for the set of event-definition thresholds) define the ROL curve. The ROL curve provides a measure of the skill with which a forecast system can predict that the predictand will exceed a set of defined thresholds. For example, the ROL curve has been used to detect and characterize the high predictability of extremely dry conditions during the March–May rainfall season over eastern Africa (Mason and Graham 1999).

It can be shown that when the forecast system has some skill the areas beneath the ROC and ROL curves will exceed 0.5 (Mason 1982; Mason and Graham 1999). Beyond this, it would be very useful to have a formal method for calculating the statistical significance of these areas under different numbers of cases, events and non-events, and warnings and non-warnings. Such significance tests for ROC areas have been calculated only rarely in the atmospheric sciences, and in these cases using Monte Carlo tests (Graham *et al.* 2000; Mullen and Buizza 2001; Thorncroft and Pytharoulis 2001) apparently because the existence of appropriate formal test procedures was not known. In any case, it is clear that much of the research on the ROC and the distribution properties of the area beneath the curve that has been published in the medical and psychology signal-detection fields is not well-known in the field of atmospheric science. This paper provides a review of some of the relevant findings as they apply to the ROC area, and develops an extension showing that the same tests can be applied to

the ROL area. In particular, it is shown that simple re-scalings of the ROC and ROL areas follow a Mann–Whitney U-distribution (Bamber 1975), and that both measures are equivalent to the probability of a correct decision in a two-alternative forced choice (2AFC) test (Green and Swets 1966). In the context of weather and climate forecasting, and from the 2AFC perspective, the ROC area can be interpreted as the following: given arbitrarily paired observations, one an event and one a non-event, the ROC area is the probability that the forecast probability assigned to the event is higher than to the non-event. Likewise, given arbitrarily paired forecasts, one for an event and one for a non-event (i.e. a warning and a non-warning), the ROL area defines the probability that the outcome is more intense when a warning has been issued than when not.

2. The ROC area and Mann–Whitney U-statistic

(*a*) *Continuous forecast probabilities (no ties)*

For this discussion we begin with the following nomenclature concerning a hypothetical forecast system:

n = the total number of cases (a forecast is issued for each case);
e = the number of pre-defined events;
$e' = n - e$ = the number of non-events.

If the ROC curve is constructed from probabilistic forecasts in which the forecast probabilities (FPs) for an event are different in each case (no ties), then the forecasts can be ranked uniquely in order of forecast probability (the principles are the same for the case of forecasts with tied probabilities, which is discussed below). Let the forecasts be ordered in descending order, i.e. with the forecasts giving the highest FP first*. If a forecast system is skilful there should be some correspondence between events and high FP. In contrast, if the system has no skill, the ordering of the events and non-events should be random with regard to FP (see further discussion in section 5). As an example, consider the set of probabilistic predictions for above-median March–May precipitation over north-east Brazil (see appendix for a description of the forecast model and methodology). The forecasts are shown, with the probabilities from an Atmospheric Model Intercomparison Project (AMIP)-style simulation, in Table 1. The ELVIS-ed probabilities (i.e. the Ensemble Likelihood Values from Inferred Statistics—see appendix) from the statistically inflated ensemble are used to avoid ties, and are sorted by FP in Table 2. Note that the events (actual cases of above-median March–May precipitation) are listed near the top of the table corresponding to the higher FPs, thus suggesting that the forecast model possesses some skill. In the following discussion the term 'hit' refers to those cases when an event is forecast and an event occurs, and 'false alarm' to those cases when an event is forecast and an event does not occur.

If the ROC curve is constructed at maximum resolution (i.e. each forecast is considered in turn and there are no ties in the FPs so that the number of points on the ROC curve is $n + 1$) the curve will be stepped. Note that area is gained whenever a hit has a higher associated FP than any of the false alarms, and no area is gained when a false alarm is issued. For each hit, the area gained is a simple function of the number of non-events having a higher FP than the current hit, f, and of the total number of events (e) and non-events (e', see Fig. 1):

$$\text{area gained} = \frac{(e' - f)}{e'e}. \tag{1}$$

* Note that the focus is on the forecast probabilities rather than on the absolute values of the individual ensemble-member predictions.

TABLE 1. PROBABILISTIC FORECASTS AND AMIP-STYLE SIMULATIONS OF ABOVE-MEDIAN MARCH–MAY PRECIPITATION OVER NORTH-EAST BRAZIL FOR 1981–95

Year	Precipitation index	Event (1)/ non-event (0)	Forecast probability	ELVIS-ed probability	AMIP probability
1981	−1.82	0	80.0	92.8	20.0
1982	−2.33	0	80.0	57.6	80.0
1983	−4.41	0	0.0	0.8	0.0
1984	1.96	1	100.0	94.4	100.0
1985	2.91	1	100.0	83.2	100.0
1986	3.22	1	60.0	81.6	100.0
1987	−0.97	0	40.0	13.6	60.0
1988	2.49	1	80.0	58.4	60.0
1989	3.58	1	0.0	3.2	100.0
1990	−2.28	0	0.0	1.6	40.0
1991	−0.48	0	20.0	28.0	100.0
1992	−3.07	0	0.0	2.4	0.0
1993	−3.46	0	0.0	0.0	20.0
1994	0.12	1	100.0	98.4	100.0
1995	1.50	1	100.0	95.2	80.0

A regional precipitation index is indicated, with positive values indicating above-median precipitation, with reference to a 1951–80 climatology. Forecast probabilities are derived from the five ensemble members, whilst the ELVIS-ed probabilities (see appendix) are derived using a statistically inflated ensemble size. All probabilities are given as percentages.

TABLE 2. FORECASTS FROM TABLE 1 SORTED BY DECREASING ELVIS-ED (SEE APPENDIX) FORECAST PROBABILITY

Year	Event (1)/ non-event (0)	Probability (%)
1994	1	98.4
1995	1	95.2
1984	1	94.4
1981	0	92.8
1985	1	83.2
1986	1	81.6
1988	1	58.4
1982	0	57.6
1991	0	28.0
1987	0	13.6
1989	1	3.2
1992	0	2.4
1990	0	1.6
1983	0	0.8
1993	0	0.0

From Eq. (1), the total ROC area, A, can be calculated as:

$$A = \frac{1}{e'e} \sum_{i=1}^{e} (e' - f_i), \tag{2}$$

which simplifies to:

$$A = 1 - \frac{1}{e'e} \sum_{i=1}^{e} f_i. \tag{3}$$

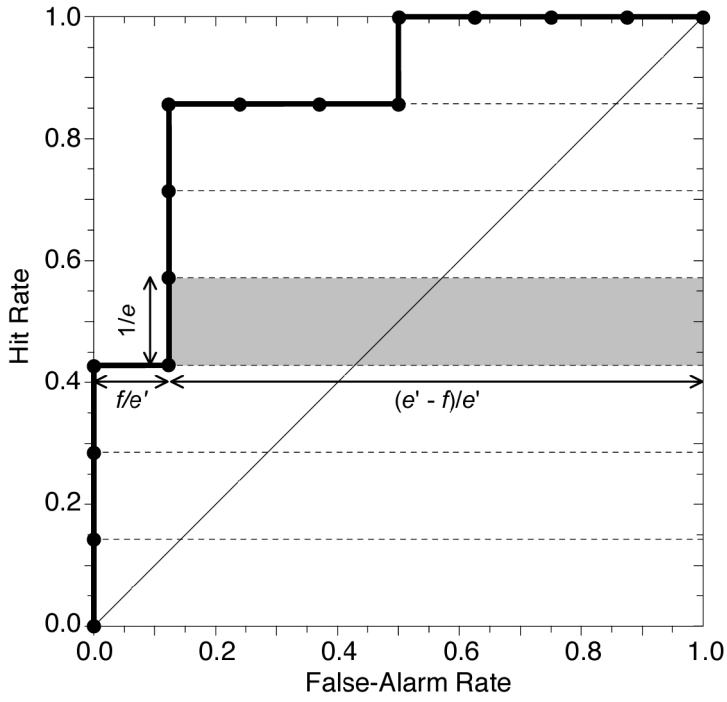

Figure 1. An example ROC diagram for a forecast system with continuous forecast probabilities. The example is based on the data in Table 2. Dotted horizontal lines mark the contribution of each hit to the ROC, and the contribution of the fourth hit is shaded. The bullets indicate individual points on the ROC curve. See text for further details.

Note that because the number of events and non-events is constant, the area is simply a function of the sum of false alarms with higher FPs than each hit* (DeLong *et al.* 1988). So, for example, in Table 2 the prediction issued in 1981 has a greater FP than predictions issued for four events (1985, 1986, 1988, 1989), while the predictions for 1982, 1991, and 1987 have higher FPs than 1989. For this example (Tables 1 and 2) and as illustrated in Fig. 1, $\sum f_i = 7$, and from Eq. (3) the ROC area is 0.875†. The sum of false alarms with higher FPs than each hit, F, can be obtained easily from the ranks of the forecasts corresponding to each hit, r_i:

$$F = \sum_{i=1}^{e} f_i = \sum_{i=1}^{e} (r_i - i) = \sum_{i=1}^{e} r_i - \frac{e(e+1)}{2}. \quad (4)$$

The minimum value of F is achieved when forecast skill is perfectly positive, so that all possible hits have higher FPs than all false alarms. In this case the value of F is 0 which, from Eq. (3), corresponds to a ROC area of 1.0. Conversely, the maximum value of F is achieved when forecast skill is perfectly negative, so that all false alarms have higher FPs than all hits. From Eq. (4), the maximum value of F is ee', corresponding to a ROC area of 0.0 (Eq. (3)).

* In the simple sequence of 10110, sorted by decreasing forecast probability, where $n = 5$ forecasts, and $e = 3$ events, a false alarm is ranked higher than two hits, while in 11001, two false alarms are ranked higher than one hit. In both cases, therefore, $\sum f_i = 2$, and the ROC areas are identical (0.667).

† Equation (3) is a computationally inefficient means of calculating the ROC area. See Hanley and McNeil (1982) for a more efficient procedure.

TABLE 3. FORECASTS FROM TABLE 1 SORTED BY DECREASING ELVIS-ED (SEE APPENDIX) FORECAST PROBABILITY, WITH FORECASTS FOR EVENTS AND NON-EVENTS SORTED SEPARATELY

Year	Event (1)/ non-event (0)	Probability (%)
1994	1	98.4
1995	1	95.2
1984	1	94.4
1985	1	83.2
1986	1	81.6
1988	1	58.4
1989	1	3.2
1981	0	92.8
1982	0	57.6
1991	0	28.0
1987	0	13.6
1992	0	2.4
1990	0	1.6
1983	0	0.8
1993	0	0.0

Although Eqs. (1)–(3) are illuminating, it is not immediately apparent how this approach relates to a significance test for the ROC area. An alternative and equivalent approach leads to more progress. First, the FPs for events and non-events are sorted separately by rank into two lists. These lists of ranked probabilities can then be concatenated (events first) as shown in Table 3. In a skilful forecast system the high probabilities should be concentrated near the top of the list with the events. Then F can be calculated as the summed number of cases where the FP for an event was exceeded by the FP for non-events. (Such reversals in order are termed 'inversions' in the literature on sorting algorithms (e.g. Knuth 1998) and developments from that field are used in the presentation that follows.) As an example, in Table 3 the FP for 1985 when an event occurred is less than the FP for 1981, when an event did not occur; this is one inversion. In the full list in Table 3 there is a total of seven inversions.

In theory, therefore, it is possible to obtain the statistical significance of the ROC area by cycling through all possible orderings of the forecasts and calculating how many times Eq. (4) is at most the observed number of inversions, i.e. by identifying what proportion of all possible hit and false-alarm combinations would generate a larger area. In practice, however, if the number of forecasts, n, is large, the total number of possible hit and false-alarm combinations, N, given by

$$N = \binom{n}{e} = \frac{n!}{e! \times e'!}, \tag{5}$$

can be large, and it becomes impractical to calculate all possible values of F. Fortunately, as noted above, the number of inversions is important in other applications such as sorting algorithms (Knuth 1998), and properties of its distribution are well-known. In particular, the number of inversions forms the basis of the Mann–Whitney U-test for differences in the central tendencies of two independent samples (Conover 1999; Sheskin 2000). Given two samples sized n_1 and n_2, the Mann–Whitney U-statistic is defined as:

$$U = \sum_{i=1}^{n_1} r_{1i} - \frac{n_1(n_1+1)}{2}, \tag{6}$$

where r_{1i} is the rank of the ith case of sample 1 (e.g. Wilks 1995; von Storch and Zwiers 1999). If sample 1 denotes the set of FPs for which an event occurred ($e = n_1$), and sample 2 denotes the set for which events did not occur (the set of non-events, $e' = n_2$), Eq. (6) becomes identical to Eq. (4). From Eqs. (3), (4), and (6) then

$$F = U = e'e(1 - A). \tag{7}$$

Thus the calculation of F is equivalent to that of the Mann–Whitney U-statistic (as first pointed out by Bamber 1975), and using Eq. (7) the ROC area can be re-scaled into a Mann–Whitney U-statistic, which has a known symmetric distribution (Kendall and Stuart 1977; Conover 1999). The U-distribution, $g(n_1, n_2, U)$, is defined by the recurrence relation:

$$g(n_1, n_2, U) = \frac{n_1}{n_1 + n_2} g(n_1 - 1, n_2, U) + \frac{n_2}{n_1 + n_2} g(n_2 - 1, n_1, U - n_1), \tag{8a}$$

$$\text{where } g(n_1, n_2, U) = 0 \text{ if } U < 0,$$

$$\text{and } g(n_1, 0, U) = g(0, n_2, U) = \begin{cases} 0 & \text{if } U \neq 0 \\ 1 & \text{if } U = 0 \end{cases} \tag{8b}$$

(Mann and Whitney 1947). The statistical significance of a given ROC area can, therefore, be calculated exactly by integrating the left-tail area of Eq. (8) (Odeh 1972; Dineen and Blakesley 1973; Harding 1984; Neumann 1988). Tables of Mann–Whitney U-statistics for small samples have been published in the atmospheric science literature (e.g. von Storch and Zwiers 1999). For larger samples, the Mann–Whitney U-distribution approximates the normal distribution (see section 3). The ROC area of 0.875, shown in Fig. 1, has a p-value of 0.007, indicating that the skill of the model is significantly high at a confidence level of greater than 99%. (See section 5 for a more detailed interpretation.)

An alternative means of calculating the significance of the ROC area by permutation tests has been shown to be equivalent to the Mann–Whitney U-test (Sheskin 2000). Further, because the assumptions for a standard permutation test (von Storch and Zwiers 1999) are the same as those required for the Mann–Whitney U-test (as detailed in section 5), the test offers no advantage if the interest is only in testing the statistical significance of the area under the entire ROC curve. Of course, permutation tests can provide additional information about individual segments of the curve (although there are alternative methods of analysing specific points on the curve (Metz 1978; Mann *et al.* 1992)), and more sophisticated randomization tests can be designed if the assumptions for a standard test are invalid.

(*b*) *Discrete forecast probabilities (ties allowed)*

In practice the ROC curve is often constructed from a set of probabilistic forecasts in which there are tied FPs (fourth column of Table 1). In this instance the curve is no longer stepped. However, as with the case of no ties, area is still gained whenever a hit has a higher FP than any of the false alarms, and the area gained by each hit is a function of the number of false alarms that have higher or equal FPs, and of the total number of events. The area gained by segment j can be obtained from the trapezium rule (Fig. 2) as:

$$\text{area gained} = \frac{h_j(2e' - f_j - \tilde{f}_j)}{2ee'}, \tag{9}$$

TABLE 4. FORECASTS FROM TABLE 1 SORTED BY DECREASING FORECAST PROBABILITY, WHERE THERE ARE TIED FORECAST PROBABILITIES FOR DIFFERENT CASES

Forecast	Event (1)/ non-event (0)	Probability (%)
1984	1	100.0
1985	1	100.0
1994	1	100.0
1995	1	100.0
1988	1	80.0
1981	0	80.0
1982	0	80.0
1986	1	60.0
1987	0	40.0
1991	0	20.0
1989	1	0.0
1983	0	0.0
1990	0	0.0
1992	0	0.0
1993	0	0.0

where h_j is the number of hits achieved in segment j, f_j is the total number of false alarms that have higher FPs than that associated with segment j, and $\tilde{f}_j$ is the total number of false alarms that have equal or higher FPs than that associated with segment j. From (9) the total ROC area for a p-pointed curve can be calculated as:

$$\tilde{A} = 1 - \frac{1}{2e'e} \sum_{j=1}^{p} h_j (f_j + \tilde{f}_j), \tag{10}$$

where $\tilde{A}$ represents the ROC in the case of ties.

Without any loss in accuracy, the ROC area can be calculated by integrating the area contributed by each hit (i.e. in a manner similar to that for continuous probabilities) dividing any tied false alarms between the corresponding tied hits. In this case, Eq. (10) can be restated as:

$$\begin{aligned} \tilde{A} &= 1 - \frac{1}{2e'e} \sum_{i=1}^{e} (f + \tilde{f}_i) \\ &= 1 - \frac{1}{e'e} \sum_{i=1}^{e} f_i - \frac{1}{2e'e} \sum_{i=1}^{e} (\tilde{f}_i - f_i) \end{aligned} \tag{11}$$

(DeLong *et al.* 1988). It is evident that Eq. (11) is the same as Eq. (3), but with an adjustment for the number of ties. As before, because e and e' are constant, the area is simply a function of the sum of false alarms with higher FP than each hit, and of the number of ties. And therefore, as in the case for continuous probabilities, the ROC area for discrete probabilities can be interpreted as a re-scaled Mann–Whitney U-statistic, but with an adjustment for the ties (Klotz 1966; Conover 1973, 1999). For an exact significance level of the ROC area in the case of discrete FPs, the ties need to be considered explicitly (see Neumann 1988), although if the number of ties is small Eq. (8) can be used with reasonable accuracy (Sheskin 2000).

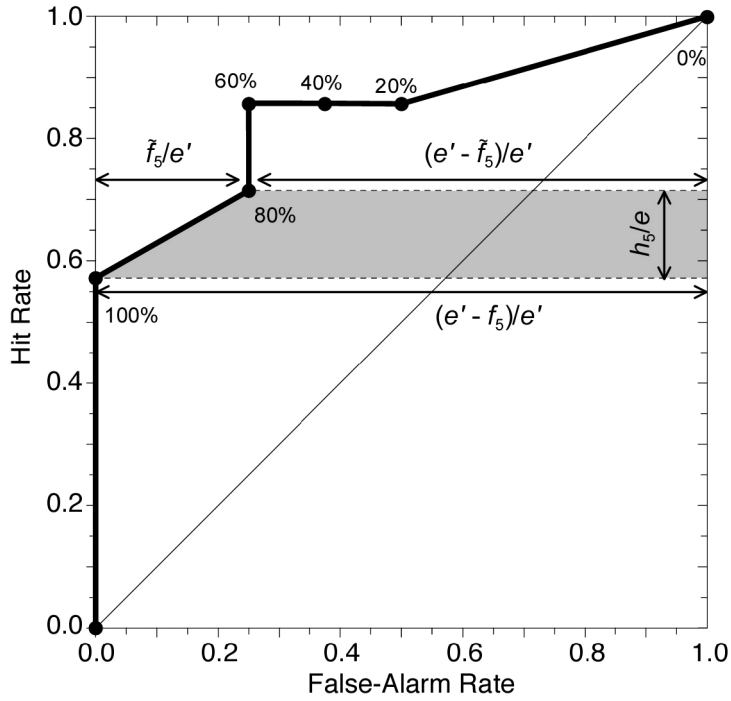

Figure 2. An example ROC diagram for a forecast system with discrete forecast probabilities, based on the data in Table 4. Dotted horizontal lines mark the contribution of the hits in each segment to the ROC area, and the contribution of the fourth segment is shaded. The bullets indicate individual points on the ROC curve, and the forecast probability threshold associated with each point is indicated. See text for further details.

If the FPs are taken from the fourth column of Table 1, a number of ties in probabilities result (Table 4). The ROC diagram for the predictions from Table 4 are illustrated in Fig. 2, and the ROC area is 0.839. This ROC area has a p-value of 0.011, indicating that the skill of the model is significantly high at a confidence level of 98%, but note that the area is less than for the case with no ties (Fig. 1). (See section 5 for a more detailed interpretation.)

3. Normal approximations of the probability distribution of the ROC area

For large sample sizes, U (and therefore F) can be approximated by a normal test statistic:

$$F \text{ and } U \sim N\left(\frac{ee'}{2}, \frac{ee'(n+1)}{12}\right) \tag{12}$$

(Mann and Whitney 1947; Conover 1999; Sheskin 2000). (Note that this normality approximation refers only to the distribution of the U-statistic, not to the underlying distributions of the FPs, as discussed below.) The definition of 'large' is open to some dispute (Sheskin 2000), but lower limits can be defined as cases in which the larger of e or $e' \geqslant 30$, and $(e + e') \geqslant 40$ (see also Metz *et al.* 1984). The resultant errors in significance levels from the normality assumption, after adjusting for continuity (Sheskin 2000), are indicated in Fig. 3 for a range of sample sizes and for different relative sample sizes. The errors are never larger than 10%, but are largest for the smallest sample sizes, and are not necessarily largest at the tails of the U-distribution.

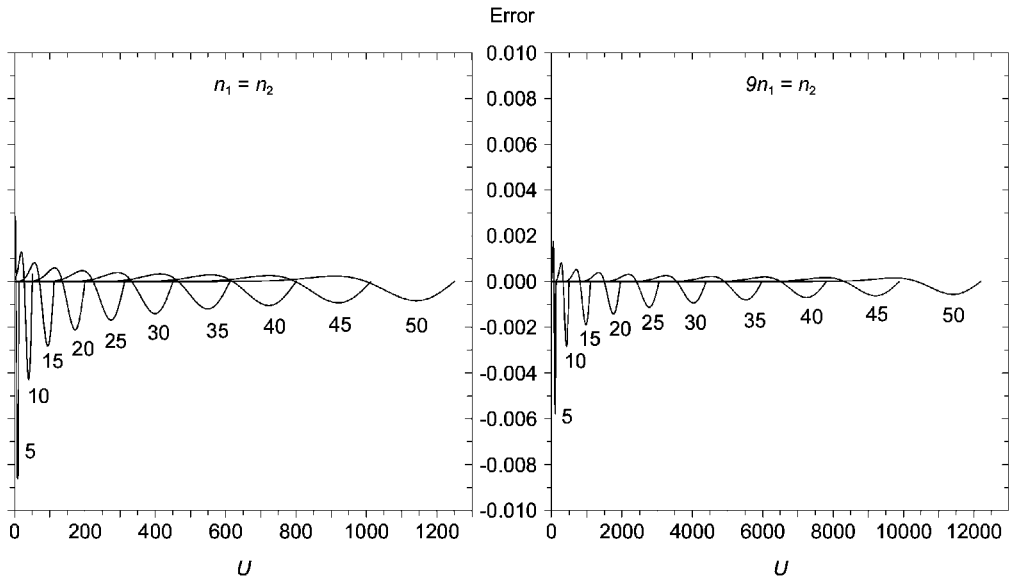

Figure 3. Errors in estimating the left-tail area of the Mann–Whitney U-distribution by using a normal-approximation assumption for cases when the two sample sizes, n_1 and n_2, are identical (left) and for when one sample is nine times larger than the other (right). Each line represents a different sample-size combination, and each point on the line represents a different value of the U-statistic. Only errors in the left-tail area are shown; since the normal and Mann–Whitney U-distributions are symmetric, the errors are identical in the right tail, but of the other sign.

For the example shown in Fig. 2 the approximated significance is 0.008, only slightly greater than the exact value of 0.007. If only one of the samples is large (when, for example, there is a long history of forecasts of rare events), a slightly more accurate approximation to the significance of the ROC area can be obtained by testing whether the ranks of the smaller sample come from a uniform distribution bounded by 1 and n (Buckle *et al.* 1969).

Although the errors in estimating the left-tail area of the U-distribution by using the normal approximation are small, as shown in Fig. 3, these errors can become large if there are many ties in the FPs. In the case of tied FPs, the variance of the U-statistic defined by Eq. (12) over-estimates the true variance. A more accurate approximation of the probability distribution of the U-statistic, $\tilde{U}$, is given by

$$\tilde{U} \sim N\left\{\frac{ee'}{2}, \frac{ee'(n+1)}{12} - \frac{ee'}{12n(n-1)}\sum_{j=1}^{\tau} t_j(t_j-1)(t_j+1)\right\}, \qquad (13)$$

where τ is the number of groups of ties, and t_j is the number of ties in group j (Conover 1999). If there are no ties in the FPs, Eq. (13) simplifies to Eq. (12). Even with the adjusted variance of Eq. (13), errors in the estimated significance of the U-statistic as obtained from the normal approximation can be noteworthy, although in the example given above the approximated significance of the ROC area with ties (0.014) is close to the exact value (0.011).

The normal approximation of the distribution of the U-statistic as defined in Eqs. (12) and (13) is distinct from the normal transformation of the axes of the ROC diagram, as proposed by Mason (1982) and Wilson (2000), and originally suggested by Dorfman and Alf (1969) and Grey and Morgan (1972). The normal approximation is

exploited in tests for comparing ROC areas, as discussed in section 6. A further distinction should be drawn between the normal transformation of the axes of the ROC diagram and the assumption that the underlying distributions of the FPs are normal, in which case standard tests could be used to test for the difference between mean FPs for events and non-events (Swets *et al.* 1961; Egan 1975; Winston 1988). Instead, the binormal transformation assumes only that the distributions of the FPs can be transformed monotonically to normal distributions (Nelson 1986; Hsieh and Turnbull 1996). This semi-parametric assumption will be valid in most cases* (Swets and Pickett 1982), and Mason's (1982) d-statistic is reasonably robust to violations of the assumption anyway (Swets and Pickett 1982; Hanley 1988). It is frequently preferred to the trapezoidal area beneath the ROC curve on linear axes (and to other summary measures of the ROC (Simpson and Fitter 1973; Centor and Schwartz 1985; Nelson 1986; Swets 1986; Hilden 1991)) because, compared to the d-statistic, the trapezoidal area is more sensitive to the number of points on the curve, and systematically underestimates the true area (assuming that the ROC curve is convex; Hanley and McNeil 1982; Swets and Pickett 1982; Centor and Schwartz 1985). The 'errors' in the estimated significance of the U-statistic as obtained from the normal approximation, described above, may then be at least partly offset by an improved estimate in the area beneath the curve. Whether or not the binormal transformation is adopted, the relationship between the ROC and the U-statistic provides a useful perspective for interpreting the ROC (section 5).

4. The ROL area and Mann–Whitney U-statistic

The principles outlined above for assessing the statistical significance of the ROC area can be applied in a similar manner to the ROL area. Given a total of n forecasts, let the total number of warnings be given by w, so that the number of non-warnings $w' = n - w$. If the ROL curve is constructed from observations that are measured on a continuous scale, and/or from a set of forecasts in which no two observations are the same, then the observations can be ranked in order of intensity. Let the observations be ordered in descending order, i.e. with the most intense events first. (Of course, this order will depend on the definition of 'intense', e.g. it will be reversed depending on whether one is concerned with very wet or very dry cases.) For example, by issuing a warning only when the FPs exceed 80%, the precipitation and warnings for March–May precipitation over north-east Brazil are as shown in Table 5. In this context it is important to note that the amount of expected precipitation does not necessarily increase monotonically with increasing probability for above normal precipitation. However, given a skilful forecast system, it is true that more precipitation should be expected in cases in which warnings (for above normal precipitation) are issued than in cases in which warnings are not issued. Thus for a skilful forecast system the marginal probability of event intensity is conditional on the presence of a warning. On the other hand, if the system has no skill the expected distribution of event intensity will be the same whether or not a warning has been issued.

If the ROL area is calculated at maximum resolution and there are no ties in the observations (i.e. each event is considered in turn so that the number of points on the curve is $n + 1$), the curve will be stepped, and area is gained whenever the event that occurs when a warning is issued is more intense than when there is no warning (Fig. 4). For each warning the area gained is a function of the number of non-warnings that have

* The assumption that the distributions of the FPs for events and non-events are Gaussian is unlikely to be valid in the contexts of weather and climate forecasts, and especially so for the shortest-range forecasts, whereas the assumption that the distributions can be transformed monotonically to normal is more reasonable.

TABLE 5. FORECASTS FROM TABLE 1 SORTED BY DECREASING INTENSITY OF THE OUTCOME

Year	Precipitation index	Warning (1)/ non-warning (0)
1989	3.58	0
1986	3.22	1
1985	2.91	1
1988	2.49	0
1984	1.96	1
1995	1.50	1
1994	0.12	1
1991	−0.48	0
1987	−0.97	0
1981	−1.82	1
1990	−2.28	0
1982	−2.33	0
1992	−3.07	0
1993	−3.46	0
1983	−4.41	0

Warnings are issued when the forecast probability exceeds 50%.

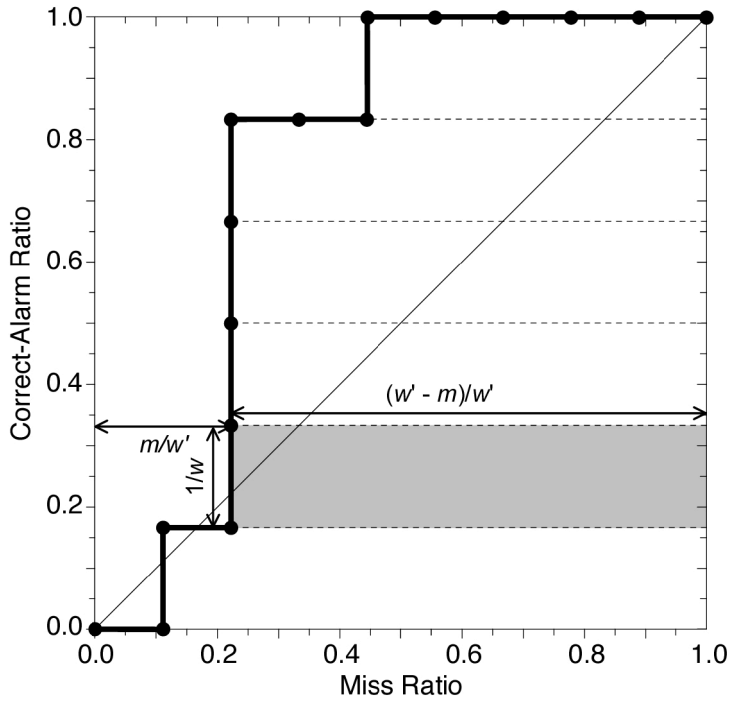

Figure 4. An example ROL diagram for a forecast system with continuous predictand. The example is based on the data in Table 5. Dotted horizontal lines mark the contribution of each hit to the ROL area, and the contribution of the second hit is shaded. The bullets indicate individual points on the ROL curve. See text for further details.

TABLE 6. FORECASTS FROM TABLE 1 SORTED BY DECREASING INTENSITY OF THE OUTCOME, WITH FORECASTS FOR WARNINGS AND NON-WARNINGS SORTED SEPARATELY

Year	Precipitation index	Warning (1)/ non-warning (0)
1986	3.22	1
1985	2.91	1
1984	1.96	1
1995	1.50	1
1994	0.12	1
1981	−1.82	1
1989	3.58	0
1988	2.49	0
1991	−0.48	0
1987	−0.97	0
1990	−2.28	0
1982	−2.33	0
1992	−3.07	0
1993	−3.46	0
1983	−4.41	0

more intense events, m, and of the total number of warnings:

$$\text{area gained} = \frac{(w' - m)}{w'w}. \tag{14}$$

From Eq. (14), the total ROL area, B, can be calculated as:

$$B = 1 - \frac{1}{w'w} \sum_{i=1}^{w} m_i. \tag{15}$$

Because the sum of w and w' is constant, the area is a function of the collective sum of the number of non-warnings that pair with more intense events than for the warnings. So, for example, in Table 5 the warning issued in 1986 was for a less intense event than 1989, when no warning was issued. For this example, shown in Fig. 3, $\sum m_i = 12$, and from Eq. (15) the ROL area is 0.778. The sum of non-warnings that had more intense events than for the warnings, M, can be obtained from the ranks of the observations corresponding to each warning, q_i:

$$M = \sum_{i=1}^{w} m_i = \sum_{i=1}^{w} (q_i - i) = \sum_{i=1}^{w} q_i - \frac{w(w+1)}{2}. \tag{16}$$

As with the ROC area, the problem can be cast in an alternative and equivalent form that leads to the relationship to the U-statistic. In this case the observations are ranked in descending order in two lists for the warning and non-warning cases separately. The lists are then concatenated, warnings first, as shown in Table 6. As discussed above, in a skilful forecast system the most intense events should be at the top of the list (with the warnings), and M can be calculated as the number of inversions in the observations. In Table 6, for example, the rainfall for 1989 (a non-warning) was more than for all the years with warnings, and the rainfall for 1988 (a non-warning) was more than for all years with warnings except 1985 and 1986. As in Table 5, there is a total of 12 inversions.

Just as the ROC area can be transformed to a U-statistic, so can the ROL area. By setting sample 1 as the set of observations for which a warning was issued ($w = n_1$),

and sample 2 as the set for which a warning was not issued ($w' = n_2$), Eq. (6) becomes identical to Eq. (15). So from Eqs. (15), (16), and (6)

$$M = U = w'w(1 - B), \tag{17}$$

i.e. M is equivalent to a Mann–Whitney U-statistic. The significance of the ROL area can then be obtained as the left-tail area of Eq. (8), and for the example shown in Fig. 4 the p-value is 0.044. The skill of the model (as measured by the ROL area) is therefore significantly high at a 95% level of confidence (see further discussion in section 5).

The extension to the case of discrete rather than continuous observations for the ROL area follows arguments parallel to those for the ROC area (section 2(b)). Details are not provided here but the principals are identical to those in section 2(b). Similarly, the normal-approximation assumption can be applied to the significance and comparison of the ROL area(s) in exactly the same way as to the ROC area(s).

5. Assumptions and interpretation

The assumptions implicit in, and interpretations of, the Mann–Whitney U-statistic can be applied to the ROC and ROL areas because of the relationships defined in Eqs. (7) and (17), respectively (Hanley and McNeil 1982). The most important assumption of the Mann–Whitney U-test of relevance when calculating the statistical significance of the ROC and ROL areas is the independence of the samples (Sheskin 2000). In a forecasting context, this translates to an assumption that the forecasts and/or that the observations are sequentially and/or spatially independent. Otherwise a chance forecast–observation agreement (or disagreement) at time t will affect the level of agreement at time $t \pm \delta t$, if δt is some time less than the effective sample time interval (e.g. Preisendorfer 1988), and the number of degrees of freedom (here the number of cases) will be overestimated. Sequential dependence could arise in many contexts ranging from daily weather forecasts to monthly sea-surface temperature (SST) forecasts from a coupled model. Similarly, spatial dependence is likely to be problematic if a set of forecasts sampled from different points in space is being verified. Standard randomization tests make the same assumption of independent forecasts, and techniques such as moving-block bootstrapping (von Storch and Zwiers 1999) would be required if the assumption is invalid.

An additional assumption of the Mann–Whitney U-test (and of permutation, but not bootstrap, tests) is that the variance of the two samples is the same (see Sheskin (2000) who notes that this assumption is frequently overlooked). In the case of the ROC area, the assumption is that the variance of the FPs (not that of the individual ensemble predictions per se) when non-events occurred is the same as when events occurred; for the ROL area the assumption is that the variance of observations when warnings were issued is the same as when warnings were not issued. This assumption is less serious than that of independent forecasts, partly because the Mann–Whitney U-test is reasonably robust to violations of homoscedasticity (Sheskin 2000). Violations of the assumption will not affect the interpretation of the areas; rather, as discussed below, they will affect the probability distribution of the U-statistic: the variance of the U-statistic will be decreased, and so significance tests will be conservative.

The relationship between the ROC/ROL areas and the Mann–Whitney U-statistic provides another perspective from which to interpret ROC and ROL areas. The Mann–Whitney U-test is most typically described as a non-parametric test for differences in the central tendency of two samples, i.e. a set of conditioned rankings is unlikely to occur by chance. It can also be shown that the Mann–Whitney U-test is appropriate to evaluate

the null hypothesis that there is a probability of 0.5 that a case drawn from sample 1 is greater than one from sample 2. In the medical and psychology literature, such a test involving the comparison of two cases drawn from each sample is known as the 'two alternative forced choice' test introduced in section 1 (Green and Swets 1966; Bamber 1975; Centor 1991; Lovell *et al.* 1996). Further, Green and Swets (1966) demonstrate that the ROC area is equivalent to the probability of correctly distinguishing an event from a non-event in a 2AFC test. In the context of forecast verification using a ROC curve, let two samples be represented by FPs for a set of events and non-events. Let X_1 represent the set of FPs issued when an event occurred, and X_2 the set of FPs when events did not occur. A 2AFC test can be defined as follows:

$$P_{2\mathrm{AFC}}(X_1, X_2) = P(x_1 \in X_1 > x_2 \in X_2) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(x_{i1}, x_{j2}), \tag{18a}$$

where

$$I(x_{i1}, x_{j2}) = \begin{cases} 0 & \text{if } x_{i1} < x_{j2} \\ 0.5 & \text{if } x_{i1} = x_{j2} \\ 1 & \text{if } x_{i1} > x_{j2} \end{cases}, \tag{18b}$$

and where n_1 is the number of events (e), and n_2 is the number of non-events ($n - e$). The relationship between $P_{2\mathrm{AFC}}(X_1, X_2)$ and the ROC area can be demonstrated by comparing Eqs. (18) and (3). From Eq. (3), the ROC area is calculated by sorting all the forecasts in descending order, then, for each i of the $n_1 = e$ events, descending from the top of the list until event i is reached and counting the number of false alarms, f_i. Note that in Eq. (3) one is simply counting the number of cases for which the FP for the event is greater than the FPs for the non-event group, i.e. $P(x_2 \in X_2 > x_1 \in X_1)$. From Eqs. (3), (4), (7), and (18), and from the fact that $P_{2\mathrm{AFC}}(X_1, X_2) = 1 - P_{2\mathrm{AFC}}(X_2, X_1)$ (assuming no ties),

$$\begin{aligned} P_{2\mathrm{AFC}}(X_1, X_2) &= 1 - P_{2\mathrm{AFC}}(X_2, X_1) \\ &= 1 - \frac{U}{e'e} = 1 - \frac{F}{e'e} = A, \end{aligned} \tag{19}$$

i.e. the ROC area is equivalent to a 2AFC test comparing the FPs of events and non-events; it gives the probability that the FP for an event is greater than for a non-event.

In the case of ties, $P_{2\mathrm{AFC}}(X_1, X_2) = 1 - P_{2\mathrm{AFC}}(X_2, X_1) - P(x_1 \in X_1 = x_2 \in X_2)$, and so from Eq. (11):

$$\begin{aligned} &P_{2\mathrm{AFC}}(X_1, X_2) + \frac{P(x_1 \in X_1 = x_2 \in X_2)}{2} \\ &\quad = 1 - P_{2\mathrm{AFC}}(X_2, X_1) + \frac{P(x_1 \in X_1 = x_2 \in X_2)}{2} \\ &\quad = 1 - \frac{\tilde{U}}{e'e} = \tilde{A}, \end{aligned} \tag{20}$$

i.e. the ROC area is equivalent to a 2AFC test comparing the FP for an event with that for a non-event, plus half the probability that the FPs are the same. In most cases, $P(x \in X = y \in Y)$ should be negligible.

In exactly the same manner, the ROL area can be interpreted as a 2AFC test. In this instance the two samples are warnings and non-warnings, and the selections are made according to event intensities. Let X_1 represent the set of event intensities

when a warning was issued, and X_2 the set of intensities when no warning was issued. From Eqs. (15)–(18), Eq. (19) becomes:

$$P_{2\mathrm{AFC}}(X_1, X_2) = 1 - P_{2\mathrm{AFC}}(X_2, X_1) = 1 - \frac{M}{e'e} = B, \tag{21}$$

i.e. the ROL area is equivalent to a 2AFC test comparing the intensities of events conditioned on whether or not a warning has been issued; it is the probability that an event accompanied by a warning will be more intense than an event not accompanied by a warning.

Finally, the non-parametric nature of the Mann–Whitney U-test confirms the fact that the ROC area is insensitive to conditional and unconditional biases (Mason 1982; Mason and Graham 1999; Wilks 2001); the absolute values of the forecast probabilities are irrelevant, rather it is only their ranking that is important. The ROC diagram and associated area are unaffected by the reliability of the forecast system (reliability is a measure of the degree to which FP is equal to the actual probability of an event; see Wilks (1995)). In Table 2, for example, the FP for 1987 could be varied between $>3.2\%$ and $<28.0\%$ before the ordering of the forecasts is affected, and between $>3.2\%$ and $<58.4\%$ before the ROC area is affected.

6. Comparing ROC or ROL areas

The relationships between the ROC or ROL areas and the U-statistic, and their relationships to the 2AFC test can be exploited for purposes of comparing the ROC or ROL areas of two or more sets of forecasts, as well as for testing the statistical significance of an individual area (e.g. Hanley and McNeil 1983; Metz *et al.* 1984; McClish 1987; DeLong *et al.* 1988; Campbell 1994; Swaving *et al.* 1996). The comparison of ROC areas can be informative when considering the dependence of forecast skill on the season or lead-time, for example (e.g. Mullen and Buizza 2001), or when comparing the areas from competing forecast strategies (e.g. Winston 1988; Buizza *et al.* 1998). The recommended approach for testing the statistical significance of differences in the ROC areas is different in these two cases, and so they are considered separately below. Similarly, the comparison of ROL areas should be informative when comparing the predictive skill of alternative models or for different periods.

The primary consideration when comparing the ROC or ROL areas for weather or climate forecasting systems is whether or not the areas were calculated using independent sets of forecasts. The procedure is simplest in the case of independent sets of forecasts which would, for example, be applicable when comparing forecast performance for two different seasons or (sets of) years using the same model. The comparison of ROC or ROL areas for forecasts at different lead-times will involve independent sets of forecasts only if the difference in lead-times is sufficiently large. If the sets of forecasts can be considered independent, a two-sample Kolmogorov–Smirnov test can be used to compare two ROCs (Gail and Green 1976; Campbell 1994), and makes no parametric assumptions about the distributions of the FPs (Sheskin 2000). Parametric alternatives have been suggested, based primarily on the t-test and χ^2 test (Swets *et al.* 1961; Dorfman and Alf 1969; Metz 1978; Metz and Kronman 1980; Hanley and McNeil 1982), and have been applied in a weather forecasting context (Winston 1988). These techniques could similarly be applied for the comparison of ROL areas.

When the sets of forecasts are not independent, which is the case when comparing the performances of competing forecasting systems over identical time periods, an allowance has to be made for the correlations between the ROC or ROL areas. The tests

proposed for comparing dependent ROC areas differ primarily in the methods of estimating the variances and covariances of the areas (Hanley and McNeil 1983; Metz *et al.* 1984; Nelson 1984; McClish 1987; DeLong *et al.* 1988; Swaving*et al.* 1996). The non-parametric method of DeLong *et al.* (1988) is probably the most widely used and, in the case of a comparison of two ROC areas, A and B, can be expressed as:

$$z = \frac{A - B}{\sqrt{(s_A^2 + s_B^2 - 2r_{AB}s_A s_B)}}, \tag{22}$$

where s_A is the standard error of area A, r_{AB} is the correlation between the two areas, and z is a test statistic whose properties are described below. In practice, the denominator can be more easily calculated by pre- and post-multiplying the variance–covariance matrix of the areas, by a contrast matrix (see DeLong *et al.* (1988), who provide additional details on the comparison of multiple ROC areas), but in the form of Eq. (22), an equivalence to the t-test for paired samples (Conover 1999; Sheskin 2000) becomes evident (Hanley and McNeil 1983). Therefore, one interpretation of z in Eq. (22) is that it is a Student's t-statistic. If the samples used to calculate the ROC areas A and B are assumed to be large, then because Student's t-distribution becomes normal with large samples, z can be interpreted as a standard normal variate, i.e. $z \sim N(0, 1)$ (DeLong *et al.* 1988). The additional implicit assumption that in calculating the ROC areas sampling errors are asymptotically normally distributed is justified by Hoeffding (1948), and is related to the validity of the normal approximation of the Mann–Whitney U-distribution, as discussed in section 3.

The variance–covariance matrix of the ROC areas is calculated from the probabilities that the FPs for the events are greater than for each of the non-events, and from the probabilities that the FPs for the non-events are greater than for each of the events (full details are provided by DeLong *et al.* (1988)). The procedure can be applied to compare the skill of the set of probabilistic predictions for above-median March–May precipitation over north-east Brazil with a similar set of five simulations forced with observed SSTs (the 'AMIP probabilities' in Table 1). Any difference in the ROC areas would then represent the loss of predictability resulting from the forcing of the ECHAM* model with persisted SST anomalies compared to forcing with perfect SST forecasts. The ROC area for the AMIP probabilities is 0.884 ($p = 0.004$), compared to the area of 0.839 ($p = 0.011$) for the forecasts, in the case of ties, from persisted SST anomalies. The difference in area of 0.045 suggests there is some loss of predictive skill resulting from the imperfect SST forecasts, but the standard error of the difference is 0.145 giving a standardized difference of only 0.308 which is not significantly large†. Therefore, by using persisted February SSTs, rather than a perfect set of SST forecasts, for predicting March–May rainfall over north-east Brazil there is only a small loss of skill.

Because of the equivalent relationships between the ROL area and the Mann–Whitney U-statistic and 2AFC test compared to the ROC, Eq. (22) can be used in an identical manner to test for the significance of differences in ROL areas. The ROL area for the AMIP probabilities is 0.857 ($p = 0.010$), compared to an area of 0.661 ($p = 0.168$) for the forecasts, in the case of ties, from persisted SST anomalies. The difference of 0.196 is considerably larger than the difference in the ROC areas, and suggests there is a loss of ability to distinguish between the varying intensities of seasonal precipitation

* European Centre for Medium-Range Weather Forecasts model, HAMburg version (see appendix).

† Note that a one-tailed test is appropriate in this instance: the question is whether the ROC area for the forecasts is less than that for the AMIP simulations.

resulting from the imperfect SST forecasts. The standard error of the difference is 0.198, giving a standardized difference of only 0.839 which is not quite statistically significant.

7. Conclusions

The area beneath the relative (or receiver) operating characteristics (ROC) curve is seeing more frequent use in the atmospheric sciences as a measure of forecast quality, but significance levels for the ROC area are rarely calculated, apparently because of the perceived lack of a suitable significance test. From a review of previous research in the medical and psychology literature, and an application of this research to the relative operating levels (ROL) area, the following key findings can be highlighted.

- The ROC and ROL areas can be interpreted as re-parametrized forms of the Mann–Whitney U-statistic (Bamber 1975).
- Because the Mann–Whitney U-distribution can be specified exactly, this distribution can be used to calculate the statistical significance of the ROC and ROL areas, and gives results that are equivalent to a permutation test.
- A normal approximation provides an accurate estimate of the significance of the areas given large samples.
- Areas can be compared using a test based on the t-test or the paired t-test, depending on whether the forecast sets are independent.
- The equivalence of the Mann–Whitney U-test and the two alternative forced choice (2AFC) test, and their respective relationships to the ROC and ROL areas (Green and Swets 1966), provide useful means of interpreting the areas, and may stimulate extension of such measures to other forecast problems.
- The ROC area defines the probability that the forecast probability issued for when an event occurs is greater than for when there is no event.
- The ROL area defines the probability that the event intensity is greater when a warning is issued than when there is no warning.

Examples have been shown of ROC and ROL curves for predictions of March–May precipitation over north-east Brazil by the ECHAM3.6 model forced with persisted February SST anomalies. Using the principles defined above, it can be concluded that there is a probability of 87.5% that more ensemble members will indicate above-median precipitation when above-median precipitation verifies, compared to when precipitation is below median. This probability is significantly greater than 50% at a confidence level of greater than 99.5%. Similarly, it can be concluded that there is a 77.8% probability that precipitation will be greater when more than 50% of the ensemble members indicate above-median precipitation compared to when less than 50% indicate above-median precipitation. This probability is significantly greater than 50% at a confidence level of greater than 95%. There is thus strong reason to believe that the model provides skilful predictions of precipitation over north-east Brazil, with only a weak loss of predictive skill resulting from the errors in predicted SST anomalies.

Acknowledgements

This research was funded by a grant/co-operative agreement number NA17RJ0453 from the National Oceanic and Atmospheric Administration (NOAA). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies. Helpful suggestions by K. Georgakakos, R. Livezey, and anonymous referees are gratefully acknowledged.

APPENDIX

Climate forecast methodology

The retrospective climate forecasts for March–May rainfall in north-east Brazil (Graham 1994) were prepared using the Max Planck Institute for Meteorology ECHAM3.5 atmospheric general circulation model (AGCM) configured at T-42 resolution (approximately 2.8° latitude/longitude resolution) with 19 layers in the vertical. The prescribed SST fields used the anomalies present in February added to the monthly SST climatologies for March–May. The simulations consisted of an ensemble of five runs beginning in February (an equilibration month) and running through May of each year (1970–96). The values used in this paper are the average of the March–May anomalies from six model grid points centred between 4.2 and 6.9°S latitude and 36.6 and 42.2°W longitude. Medians for the AGCM data were defined with respect to the 1951–80 climatology from an ensemble of ten simulations forced with observed SSTs with the same model. Medians for the observations were defined using data over the same period. For the examples in this paper only the forecasts for 1981–95 are used. In addition, five of the ten ensemble members forced with observed SSTs for 1981–95 were randomly selected. This AMIP-style ensemble was used to estimate the potential predictive skill given perfect SST forecasts. For some of the analyses described here the seasonal forecast ensembles have been artificially expanded using a re-sampling technique (ELVIS—Ensemble Likelihood Values from Inferred Statistics) based on the (generally satisfied) assumption that beyond the effects of prescribed boundary conditions (e.g. SSTs) the month-to-month precipitation anomalies from any given ensemble member are largely uncorrelated (Graham and Mason, personal communication). For the application here to a 3-month season with a 5-member ensemble this technique provides an effective ensemble size of approximately 15.

REFERENCES

Bamber, D. 1975 The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *J. Math. Psychol.*, **12,** 387–415

Begg, C. B. 1991 Advances in statistical methodology for diagnostic medicine in the 1980s. *Stat. Med.*, **10,** 1887–1895

Buckle, N., Kraft, C. H. and van Eeden, C. 1969 An approximation to the Wilcoxon–Mann–Whitney distribution. *J. Am. Stat. Soc.*, **64,** 591–599

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 The 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, E. and Ghelli, A. 1999 Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Weather and Forecasting*, **14,** 168–189

Campbell, G. 1994 Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Stat. Med.*, **13,** 499–508

Centor, R. M. 1991 Signal detectability: the use of ROC curves and their analyses. *Med. Decision Making*, **11,** 102–106

Centor, R. M. and Schwartz, J. S. 1985 An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. *Med. Decision Making*, **15,** 276–282

Conover, W. J. 1973 Rank tests for one sample, two samples, and k samples without the assumption of a continuous distribution function. *Ann. Stat.*, **1,** 1105–1125

1999 *Practical nonparametric statistics*. Wiley, Chichester, UK

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. 1988 Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44,** 837–845

Dineen, L. C. and Blakesley, B. C. 1973 Algorithm AS 62. A generator for the sampling distribution of the Mann–Whitney U statistic. *Appl. Stat.*, **22,** 268–273

Dodge, H. and Romig, H. 1929 A method of sampling inspection. *Bell Systems Tech. J.*, **8,** 613–631

Dooley, K. 2000 The paradigms of quality: evolution and revolution in the history of the discipline. *Adv. Manage. Organ. Qual.*, **5,** 1–28

Dorfman, D. D. and Alf, E. 1969 Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals–rating-method data. *J. Math. Psychol.*, **6,** 487–496

Egan, J. P. 1975 *Signal detection theory and ROC analysis*. Academic Press, New York, USA

Falmagne, J.-C. 1985 *Elements of psychophysical theory*. Oxford University Press, Oxford, UK

Frogner, I.-L. and Iversen, T. 2001 Targeted ensemble prediction for northern Europe and parts of the North Atlantic Ocean. *Tellus*, **53A,** 35–55

Gail, M. H. and Green, S. B. 1976 A generalization of the one-sided two-sample Kolmogorov–Smirnov statistic for evaluating diagnostic tests. *Biometrics*, **32,** 561–570

Graham, N. E. 1994 'Experimental predictions of wet season precipitation in north-eastern Brazil'. Pp. 378–381 in Proceedings of the 18th Annual Climate Diagnostics Workshop, 1–5 November 1993, Boulder, Colorado, USA

Graham, R. J., Evans, A. D. L., Mylne, K. R., Harrison, M. S. J. and Robertson, K. B. 2000 An assessment of seasonal predictability using atmospheric general circulation models. *Q. J. R. Meteorol. Soc.*, **126,** 2211–2240

Green, D. M. and Swets, J. A. 1966 *Signal detection theory and psychophysics*. Peninsula Publishing, Los Altos, California, USA

Grey, D. R. and Morgan, B. J. T. 1972 Some aspects of ROC curve-fitting: normal and logistic models. *J. Math. Psychol.*, **9,** 128–139

Hanley, J. A. 1988 The robustness of the 'binormal' assumptions used in fitting ROC curves. *Med. Decision Making*, **8,** 197–203

Hanley, J. A. and McNeil, B. J. 1982 The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, **143,** 29–36

1983 A method of comparing the areas under receiver operating characteristic curves from the same cases. *Radiology*, **148,** 839–843

Harding, D. E. 1984 An efficient, minimal-storage procedure for calculating the Mann–Whitney U, generalized U and similar distributions. *Appl. Stat.*, **33,** 1–6

Harvey, L. O., 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

Hilden, J. 1991 The area under the ROC curve and its competitors. *Med. Decision Making*, **11,** 95–101

Hoeffding, W. 1948 A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.*, **19,** 293–325

Hsieh, F. and Turnbull, B. W. 1996 Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Stat.*, **24,** 25–40

Kendall, M. G. and Stuart, A. 1977 *The advanced theory of statistics*. Griffin, London, UK

Klotz, J. 1966 The Wilcoxon, ties, and the computer. *Ann. Math. Stat.*, **61,** 772–787

Knuth, D. E. 1998 *The art of computer programming. Volume 3: Sorting and searching*. Addison–Wesley, Reading, Massachusetts, USA

Lovell, D. R., Dance, C. R., Niranjan, M., Prager, R. W. and Dalton, K. J. 1996 'Ranking the effect of different features on the classification of discrete valued data'. Pp. 487–494 in Proceedings of the second international conference on engineering applications of neural networks, 17–19 June 1996, Kingston upon Thames, London, UK

McClish, D. K. 1987 Comparing the areas under more than two independent ROC curves. *Med. Decision Making*, **7,** 149–155

Mann, H. B. and Whitney, D. R. 1947 On a test of whether one or two random variables is stochastically larger than the other. *Ann. Math. Stat.*, **18,** 50–60

Mann, F. A., Hildebolt, C. F. and Wilson, A. J. 1992 Statistical analysis with receiver operating characteristic curves. *Radiology*, **184,** 37–38

Mason, I. 1979 On reducing probability forecasts to yes/no forecasts. *Mon. Weather Rev.*, **107,** 207–211

1982 A model for assessment of weather forecasts. *Aust. Meteorol. Mag.*, **30,** 291–303

Mason, S. J. and Graham, N. E. 1999 Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, **14,** 713–725

Mason, S. J., Goddard, L., Graham, N. E., Yulaeva, E., Sun, L. and Arkin, P. A. 1999 The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bull. Am. Meteorol. Soc.*, **80,** 1853–1873

Metz, C. E. 1978 Basic principles of ROC analysis. *Semin. Nucl. Med.*, **8,** 283–298

Metz, C. E. and Kronman, H. B. 1980 Statistical significance tests for binormal ROC curves. *J. Math. Psychol.*, **22,** 218–243

Metz, C. E., Wang, P. -L. and Kronman, H. B. 1984 A new approach for testing the significance of differences between ROC curves measured from correlated data. In *Information processing in medical imaging*. Ed. F. Deconinck. Nijhof, The Hague, the Netherlands

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

Murphy, A. H. and Winkler, R. L. 1987 A general framework for forecast verification. *Mon. Weather Rev.*, **115,** 1330–1338

Nelson, T. O. 1984 A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychol. Bull.*, **95,** 109–133

1986 ROC curves and measures of discrimination accuracy: a reply to Swets. *Psychol. Bull.*, **100,** 128–132

Neumann, N. 1988 Some procedures for calculating the distributions of nonparametric test statistics. *Stat. Software Newsl.*, **14,** 120–126

Neyman J. and Pearson, E. S. 1933 On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. London*, **A231,** 289–337

Odeh, R. E. 1972 Algorithm AS 55. The generalized Mann–Whitney U-statistic. *Appl. Stat.*, **21,** 348–351

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

Peterson, W. W. and Birdsall, T. G. 1953 'The theory of signal detectability: Part I. The general theory'. Electronic Defense Group, Technical Report 13, June 1953. Available from EECS Systems Office, University of Michigan, 1301 Beal Avenue, Ann Arbor, MI 48109-2122 USA

Peterson, W. W., Birdsall, T. G. and Fox, W. C. 1954 The theory of signal detectability. *Trans. IRE Prof. Group Inf. Theory, PGIT*, **2–4,** 171–212

Preisendorfer, R. 1988 *Principal component analysis in meteorology and oceanography*. Elsevier, New York, USA

Richardson, D. S. 2000 Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.*, **126,** 649–667

Sheskin, D. J. 2000 *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall, Boca Raton, Florida, USA

Shewhart, W. 1931 *Economic control of quality of manufactured products*. D. van Norstand, New York, USA

Simpson, A. J. and Fitter, M. J. 1973 What is the best index of detectability? *Psychol. Bull.*, **80,** 481–488

Sokal, R. R. and Rohlf, F. J. 1973 *Introduction to biostatistics*. Freeman, San Francisco, California, USA

Stanski, H. R., Wilson, L. J. and Burrows, W. R. 1989 'Survey of common verification methods in meteorology'. Research Report No. 89–5, Atmospheric Environment Service, Forecast Research Division, 4905 Dufferin Street, Downsview, Ontario, Canada

Swaving, M., Van Houwelingen, H., Ottes, F. P. and Steerneman, T. 1996 Statistical comparison of ROC curves from multiple readers. *Med. Decision Making*, **16,** 143–153

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

1979 ROC analysis applied to the evaluation of medical imaging techniques. *Invest. Radiol.*, **14,** 109–121

Swets, J. A. 1986 Indices of discrimination of diagnostic accuracy: their ROCs and implied models. *Psychol. Bull.*, **99,** 100–117

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

1995 *Signal detection theory and ROC analysis in psychology and diagnostics: collected papers*. Lawrence Erlbaum Associates, Mahwah, New Jersey, USA

Swets, J. A. and Birdsall, T. G. 1967 Deferred decision in human signal detection: a preliminary experiment. *Perception and Psychophysics*, **2,** 15–28

Swets, J. A. and Pickett, R. M. 1982 *Evaluation of diagnostic systems: methods from signal detection theory*. Academic Press, New York, USA

Swets, J. A., Tanner, W. P. and Birdsall, T. G. 1961 Decision processes in perception. *Psychol. Rev.*, **68,** 301–340

Swets, J. A., Dawes, R. M. and Monahan, J. 2000 Better decisions through science. *Sci. Am.*, **283,** (4), 70–75

Thorncroft, C. and Pytharoulis, I. 2001 A dynamical approach to seasonal prediction of Atlantic tropical cyclone activity. *Weather and Forecasting*, **16,** 725–734

von Storch, H. and Zwiers, F. W. 1999 *Statistical analysis in climate research*. Cambridge University Press, Cambridge, UK

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. 1995 *Statistical methods in the atmospheric sciences*. Academic Press, San Diego, California, USA

2001 A skill score based on economic value for probability forecasts. *Meteorol. Appl.*, **8,** 209–219

Wilson, L. J. 2000 Comments on 'Probabilistic predictions of precipitation using the ECMWF ensemble prediction system'. *Weather and Forecasting*, **15,** 361–364

Winston, H. A. 1988 A comparison of three radar-based severe-storm-detection algorithms on Colorado high plains thunderstorms. *Weather and Forecasting*, **3,** 131–140

WMO 2000 *Standardized verification system (SVS) for long-range forecasts (LRF)*. World Meteorological Organization, Geneva, Switzerland http://www.wmo.ch/web/www/DPS/SVS-for-LRF.html

Zhang, H. and Casey, T. 2000 Verification of categorical probability forecasts. *Weather and Forecasting*, **15,** 80–89

Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K. 2002 The economic value of ensemble-based forecasts. *Bull. Am. Meteorol. Soc.*, **83,** 73–83