A Generic Forecast Verification Framework for Administrative Purposes

SIMON J. MASON

International Research Institute for Climate and Society, Columbia University, Palisades, New York

ANDREAS P. WEIGEL

Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland

(Manuscript received 11 February 2008, in final form 30 May 2008)

ABSTRACT

There are numerous reasons for calculating forecast verification scores, and considerable attention has been given to designing and analyzing the properties of scores that can be used for scientific purposes. Much less attention has been given to scores that may be useful for administrative reasons, such as communicating changes in forecast quality to bureaucrats and providing indications of forecast quality to the general public. The two-alternative forced choice (2AFC) test is proposed as a scoring procedure that is sufficiently generic to be usable on forecasts ranging from simple yes–no forecasts of dichotomous outcomes to forecasts of continuous variables, and can be used with deterministic or probabilistic forecasts without seriously reducing the more complex information when available. Although, as with any single verification score, the proposed test has limitations, it does have broad intuitive appeal in that the expected score of an unskilled set of forecasts (random guessing or perpetually identical forecasts) is 50%, and is interpretable as an indication of how often the forecasts are correct, even when the forecasts are expressed probabilistically and/or the observations are not discrete.

1. Introduction

Few weather or seasonal climate forecasters with any experience can have escaped the question: How often are your forecasts correct? Even when the forecaster acknowledges that the question itself is a reasonable one, only rarely is an answer ready to hand, and then it usually has to be accompanied by various apologies and explanations to help in the interpretation of the quoted score. The naïve expectation is that the forecasts should be correct at least 50% of the time; otherwise, one may just as well guess. However, against this expectation is Finley's accuracy rate of almost 97% (Murphy 1996) on the one hand, which is arguably unskillful, and, on the other hand, the Climate Prediction Center's skillful long-lead 3-month forecasts with accuracy rates of around 40% or less (Livezey and Timofeyeva 2008), all of which makes most simple measures of accuracy paradoxically complicated metrics of performance. In addition, when it comes to probabilistic forecasts, the idea of forecasts being "correct" or "incorrect" is considered a badly formulated question to start with. The forecaster is therefore tempted to present an array of apologies before presenting any verification statistic. However, the danger then arises of leaving the impression that forecast verification practitioners purposely obfuscate the whole problem to hide the fact that weather and climate forecasts are truly as bad as popularly believed.

Unfortunately, as forecasters, we cannot simply withdraw into our scientific community where we do not have to explain concepts like base rates, and where we can work with metrics like Brier and ignorance scores, and can be equitable and effective, or local and proper whenever it suits our particular purposes to be so. Three broad classes of reasons for performing verification analyses have been widely recognized: administrative, scientific, and economic (Brier and Allen 1951; Jolliffe and Stephenson 2003). Because of the needs of the forecasters themselves to understand the strengths and weaknesses of their forecasting algorithms, by far the most attention in the atmospheric sciences literature has been devoted to scientific evaluations of fore-

Corresponding author address: Dr. Simon J. Mason, International Research Institute for Climate and Society, 61 Route 9W, P.O. Box 1000, Columbia University, Palisades, NY 10964-8000.
E-mail: simon@iri.columbia.edu

DOI: 10.1175/2008MWR2553.1

casts, and to devising and diagnosing various scores and procedures for such ends. More recently, concerns over demonstrating the economic value of weather and climate forecasts have been attracting efforts to address this third type of forecast "goodness" in Murphy's (1993) taxonomy. However, the question of how best to service the administrative reasons for forecast verification has received little attention, and requires addressing the apparently competing needs for mathematical rigor and for understandability referred to above.

Although the "administrative" category incorporates a wide range of possible motives for conducting a verification analysis, a common need is for a single score that can be used to summarize forecast quality. It can be added that when the purpose is to communicate to nonspecialists (whether they be bureaucrats who may be making decisions about resource allocations for ongoing forecasting activities and improvements, or potential clients, or the simply curious) the score should be as simple as possible (in the sense of being easy to understand, and not in the sense of capturing only limited aspects of the scientific quality of the forecasts). Of course, it has to be accepted upfront that as soon as one decides to summarize forecast quality into any single score, information is going to be lost; forecast quality is multifaceted (Murphy 1991) and so any single number will have its limitations. Acknowledging that there can be no perfect score, the purpose in this article is to recommend a score that does have some excellent intuitive properties and compromises on questions of information loss as little as possible. The score itself (or scores themselves; since a generic framework is proposed, the precise formulation of the score depends on the nature of the forecasts and observations) is not new, but to the authors' knowledge the framework presented in this article has not previously been proposed in the atmospheric science literature as suitable for the purposes of administrative forecast verification. The proposed scoring framework is sufficiently flexible to be used with forecasts and observations that cover a wide range of complexities, but without reducing the more complex cases to the simpler. Specifically, where the forecasts contain probabilistic information on a continuous target variable, and the target variable itself is measured on the same continuous scale, the score should not discretize the forecasts and observations. But at the same time, where the forecasts and observations are discretized, the score should remain meaningful.

In section 3 of this paper, the proposed framework for evaluating forecasts is presented, and the specific verification scores that are defined by the framework are introduced. The limitations of the scores are then discussed in section 4. The paper closes with some concluding remarks in section 5. First, however, a set of example data to illustrate the calculation of the scores is described.

2. Example data

Forecasts for the Niño-3.4 index from the coupled ocean–atmosphere model of the Centre National de Recherches Météorologiques (CNRM) of Météo-France were compared with observed values. The data were generated as part of the Development of a European Multimodel Ensemble System for Seasonal to Interannual Prediction (DEMETER) project (Palmer et al. 2004), and only the 40 forecasts for January 1961–2000 from model runs initialized using data for the preceding August 1960–99 were considered. There are nine ensemble members. No attempt was made to calibrate the forecasts, but the mean bias is less than 0.3°C, and the model variance is about 80% that of the observed data. The ensemble-mean forecasts account for about 88% of the variance of the January Niño-3.4 index for the 40-yr period, indicating that the model is highly skillful.

The Finley tornado forecasts mentioned in the introduction are considered also, but only within the context of dichotomous forecasts of dichotomous observations.

3. A general framework for evaluating forecasts

The standard way of evaluating forecasts is to take each forecast–observation pair and to compare each in turn, asking the question of how well the forecast corresponds with its respective observation. Given n forecast–observation pairs, this question is addressed n times, and a summary measure is calculated. An alternative procedure is proposed by which a set of two forecast–observation pairs is considered, and the question then becomes whether the forecasts can be used successfully to distinguish between the observations. For example, given a day with rain and one without, can we successfully identify the wet day from the forecasts? Or can we identify the warmer of two days, again from the forecasts? Assuming that the observations are distinguishable, the probability of picking the correct year given unskillful forecasts is 50%. The aim is to compare all possible sets of two forecast–observation pairs, asking the same question each time, and calculating the proportion of times that the question is answered correctly. This proportion is known as the probability of a correct decision, and the question is known as a two-alternative forced choice (2AFC) test (Green and Swets 1989; Mason and Graham 2002). Each time the question is asked, there is a 50% chance of picking the correct observation in the absence of any useful infor-

mation, but if the forecasts are skillful, the proportion of correctly picked observations will exceed 50%, and the better the forecasts are the closer the proportion will be to 100%.

In selecting the two pairs, the observations have to be different in some way, and so it may be the case that not all sets of two pairs can be assessed (e.g., it is not a useful question to ask which of two days was wet if neither of them had observed rainfall). If all the observations differ from one another, a total of $\binom{n}{2} = n(n-1)/2$ sets can be evaluated, although not all the results will be independent[1] (if day 1 is warmer than day 2, and day 2 is warmer than day 3, then it is known that day 1 is warmer than day 3). If some of the observations cannot be distinguished, then the number of sets of two forecast–observation pairs that can be meaningfully compared depends on the numbers of classes of different observations and the numbers of observations in each class, as discussed in detail later.

Technically, a 2AFC test involves asking whether the forecasts can be used to successfully *discriminate* between the observations (Murphy 1991), and so, in the general framework for forecast verification introduced by Murphy and Winkler (1987), the 2AFC procedure proposed in this paper involves a likelihood-base rate factorization. The precise formulation of the question in the 2AFC test depends on the nature of the forecasts and observations, and in some cases the resultant test reduces to a verification score that is already in wide use, and/or is more widely known under a different name. The following sections describe the 2AFC score under these different formulations, starting with the simplest cases of yes–no forecasts for dichotomous outcomes, and proceeding in complexity to probabilistic forecasts of outcomes measured on a continuous scale. A brief discussion of forecasts of observations for which the latter are probability distributions closes this section.

a. Dichotomous observations

The simplest situation is when there is one of only two possible outcomes; one of the possible outcomes is labeled an event, and so an event can either occur or not occur. Let $x_{1,j}$ represent the jth forecast issued when an event did occur, and let there be n_1 events; let $x_{0,i}$ represent the ith forecast issued when an event did not occur, and let there be n_0 nonevents ($n_1 + n_0 = n$). A 2AFC score for dichotomous observations can be defined as the proportion of correctly answered 2AFC tests out of all possible such tests:

$$p_{2\mathrm{AFC}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(x_{0,i}, x_{1,j}), \tag{1a}$$

where the scoring rule $I(x_{0,i}, x_{1,j})$ is defined as

$$I(x_{0,i}, x_{1,j}) = \begin{cases} 0.0 & \text{if} \quad x_{1,j} < x_{0,i} \\ 0.5 & \text{if} \quad x_{1,j} = x_{0,i} \\ 1.0 & \text{if} \quad x_{1,j} > x_{0,i} \end{cases}. \tag{1b}$$

Note that $2 \times I(x_{0,i}, x_{1,j}) - 1 = \mathrm{sgn}(x_{1,j} - x_{0,i})$. The precise calculation of Eq. (1) depends on the nature of the forecasts. Its formulation given different forecast types is detailed in the following subsections.

1) DICHOTOMOUS FORECASTS

Dichotomous forecasts are expressed as a yes or a no for an event to occur. An example is forecasts of rainfall occurrence: Rainfall either occurs or it does not occur, and the forecast indicated either that it would or would not. Such forecasts and observations are most commonly summarized in a 2×2 contingency table, and there are numerous scores that have been applied for summarizing the quality of such forecasts (Mason 2003). Finley's tornado forecasts (Murphy 1996) are a well-known example, and the debate around the quality of these forecasts effectively illustrates some of the difficulties in communicating the quality of this type of forecast.

In a 2AFC test, the procedure would be to select one day on which a tornado occurred, and one on which a tornado did not occur, and then to decide on which day the tornado occurred based on the two corresponding forecasts. This test would be repeated for each possible pairing of a tornado with a nontornado day. For some of the pairings, because some of the forecasts may be incorrect, it may not be possible to decide on which day a tornado occurred: Sometimes both forecasts may indicate no tornado, and other times both may indicate a tornado. For these cases a score of one-half should be recorded. Using the contingency table shown in Table 1 in which a = number of hits, b = number of false alarms, c = number of misses, and d = number of correct rejections, the 2AFC score can be calculated as

$$p_{2\mathrm{AFC}} = \frac{ad + 0.5(ab + cd)}{(a + c)(b + d)}. \tag{2}$$

The denominator defines the total number of forecast–observation pairs to consider and is simply the product of the number of events and nonevents; the numerator is calculated as the number of correctly discriminated observations (ad) plus half the number of observations

[1] This aspect of dependency does not invalidate the test and should be considered distinct from the dependency that arises from spatial and temporal autocorrelation. See section 4 for further discussion.

TABLE 1. 2 × 2 contingency table.

Observations	Forecasts: Yes	Forecasts: No
Yes	*a*	*c*
No	*b*	*d*

that cannot be discriminated either way. The term *bc*, which is obtained if the denominator is expanded, represents the number of incorrectly discriminated observations.

Equation (2) can be shown to be a special case of Eq. (1). From Table 1, *a* is the number of times $x_{1,j} = 1$, *b* is the number of times $x_{0,i} = 1$, *c* is the number of times $x_{1,j} = 0$, and *d* is the number of times $x_{0,i} = 0$, while n_1 is the number of events ($a + c$) and n_2 is the number of nonevents ($b + d$). The denominator of Eq. (2) is therefore the same as in Eq. (1a), while in the numerator *ad* represents all the cases where $x_{1,j} > x_{0,i}$ and $ab + cd$ represents all the cases where $x_{1,j} = x_{0,i}$.

The score defined in Eq. (2) is a version of the area beneath the relative operating characteristics (ROC) curve (Mason 2003). Since the forecasts and observations are both dichotomous, the ROC curve is defined by a single point connected to the corners. Using the trapezium rule, the area beneath the curve can be calculated as the sum of the triangular area to the left of this point, and the trapezoidal area to the right:

$$\mathrm{ROC}_{\mathrm{area}} = 0.5 \times \left[\left(\frac{a}{a+c} \right) \left(\frac{b}{b+d} \right) + \left(\frac{2a+c}{a+c} \right) \left(\frac{d}{b+d} \right) \right]. \quad (3)$$

Equation (3) reduces to Eq. (2), indicating the equivalence of the 2AFC score and the ROC area in this context. The area beneath the ROC graph (when it is calculated using the trapezium rule) can therefore be interpreted as the probability of successfully discriminating the observations (Mason and Graham 2002).

For Finley's forecasts, the 2AFC score is approximately 76%, indicating that his forecasts successfully discriminated tornado from nontornado days over three-quarters of the time. This score is consistent with the recognition that against a strategy of random guessing—as opposed to Gilbert's commonly quoted one of perpetual no tornadoes—Finley's forecasts are skillful (Murphy 1996; Mason 2003; Mason 2008). If Gilbert's strategy of perpetual no-tornado forecasts can outperform Finley's forecasts on the percent correct score, how do they fare on the 2AFC score? In this case *a* and *b* are 0, and so Eq. (2) reduces to

TABLE 2. Contingency table for January 1961–2000 observed values and forecasts of the Niño-3.4 index above or below 27.0°C. Forecasts initialized using data for August 1960–99 are from the coupled ocean–atmosphere model of CNRM.

Observations	Forecasts: >27.0°C	Forecasts: <27.0°C
>27.0°C	14	1
<27.0°C	2	23

$$p_{2\mathrm{AFC}} = \frac{0.5cd}{cd} = 50\%, \quad (4)$$

which is the same as for random guessing. At least in the context of dichotomous forecasts and observations, therefore, the 2AFC score is equitable.

Given that the 2AFC score has been presented as an attempt to address the question of how often the forecasts are correct, how does the 2AFC score of 76% relate to the percent correct score of 97% for Finley's forecasts (Murphy 1996)? The difference is that the percent correct score is a simple count of how often the forecast matched the corresponding observation, and tells us at least as much about the base rate as it does about the quality of the forecasts. The 2AFC score, however, tells us how often the forecasts successfully distinguish a tornado day from a nontornado day and is independent of the base rate.

The CNRM ensemble-mean forecasts and observed values of the Niño-3.4 index were classified as, respectively, predicting and verifying "warm" events if the index exceeded 27.0°C. The resulting contingency table is shown in Table 2. From Eq. (2), the probability of correctly discriminating a warm event from a nonevent ("cool") is approximately 93%.

2) POLYCHOTOMOUS FORECASTS

Polychotomous forecasts of dichotomous outcomes are perfectly reasonable if the forecast categories are viewed as a set of ordered warning levels, which may or may not have explicit probability thresholds assigned to them.[2] Setting the dot subscript to represent any outcome, in terms of Eq. (1), $x_{\cdot,i}$ can take on one of m_f values, where m_f is the number of forecast categories.

[2] It would seem unreasonable for the numbers of observed and forecast categories to differ if the categories were not ordinal: Either at least some of the forecast categories would be incommensurable with the observational categories, or there would be forecast categories for events that never occur and so there would be no point in forecasting them. The case of nominal forecast categories and dichotomous observations is therefore not considered further.

Equation (1b) is independent of the number of possible values for $x_{\cdot,i}$ and so the 2AFC score retains the same format as for the case of the dichotomous forecasts. Let $n_{1,k}$ be the number of forecasts of category k when an event occurred (i.e., the number of times $x_{1,i} = k$), and $n_{0,k}$ the number of forecasts for the same category when an event did not occur (i.e., the number of times $x_{0,j} =$ k).[3] Then,

$$p_{2\mathrm{AFC}} = \frac{\sum_{i=1}^{m_f-1} \sum_{j=i+1}^{m_f} n_{0,i} n_{1,j} + 0.5 \sum_{k=1}^{m_f} n_{0,k} n_{1,k}}{n_0 n_1}. \quad (5)$$

Equation (5) is related to Eq. (1) because the first term in the numerator represents all the cases where $x_{1,j} > x_{0,i}$, and the second term represents all the cases where $x_{1,j} = x_{0,i}$. In the case of perpetually identical forecasts, either $n_{0,i} = 0$ and/or $n_{1,j} = 0$, and either $n_{0,k} = n_0$ and $n_{1,k} = n_1$, or $n_{0,k} = 0$ and $n_{1,k} = 0$. Therefore, just as Eq. (2) reduces to Eq. (4), so also Eq. (5) reduces to

$$p_{2\mathrm{AFC}} = \frac{0.5 n_0 n_1}{n_0 n_1} = 50\%, \quad (6)$$

and the score is again equitable.

The equivalency of the 2AFC score and the area beneath the ROC graph in the case of dichotomous observations and forecasts was noted in section 3a(1). The same is true in the case of ordinal polychotomous forecasts, but, because the forecasts can be ranked, more than one point on the ROC graph can now be constructed. This case is identical to the way that a trapezoidal ROC area is calculated given forecasts expressed as probabilities for categories [section 3a(3)]. Since the probabilities themselves are ignored when performing a ROC analysis (Mason and Graham 2002; Glahn 2004; Wilks 2006), probabilistic forecasts are reduced to forecasts of ordinal polychotomous categories. In the current context, therefore, the 2AFC score is equivalent to the standard way in which the ROC technique is performed in forecast verification, except that the probabilities associated with each point on the curve are undefined (Mason and Graham 2002).

Retaining the definition of observed warm events as a Niño-3.4 index of greater than 27.0°C used in the previous section, the CNRM ensemble-mean forecasts were classified as predicting warm events with very high confidence if the forecast exceeded 28.0°C, high confidence if the forecast exceeded 27.0°C, low confidence if the forecast exceeded 26.0°C, and very low confidence otherwise. The resulting 2 × 4 contingency table is shown in Table 3. From Eq. (5), the probability of correctly discriminating a warm from a cool event is approximately 95%. This value is a slight improvement on the 93% from the 2 × 2 contingency table because of the greater resolution in the forecasts available in the 2 × 4 table, resulting in fewer ties in the forecasts. However, it should be noted that the results are somewhat sensitive to the categorization of the forecasts.

TABLE 3. Contingency table for January 1961–2000 observed values of the Niño-3.4 index above or below 27.0°C from the coupled ocean–atmosphere model of CNRM initialized in August 1960–99.

	Forecasts			
Observations	Very high	High	Low	Very low
>27.0°C	5	9	1	0
<27.0°C	0	2	14	9

3) DISCRETE PROBABILISTIC FORECASTS

The 2AFC score for discrete probabilistic forecasts is essentially the same as for the polychotomous forecasts: The probabilities define a set of ordered categories, with the number of forecast categories being equal to the number of discrete forecast probabilities. Consistent with the relationship shown in Eq. (3), the 2AFC score for discrete probabilistic forecasts is then equivalent to calculating the trapezoidal area under the ROC curve (Mason and Graham 2002). Let $p_{1,j}$ represent the jth forecast probability for an event when an event occurred, and let $p_{0,i}$ represent the ith forecast probability for an event when an event did not occur. Effectively, the forecasts $x_{0,i}$ and $x_{1,j}$ in Eq. (1) are replaced by these probabilities, giving

$$p_{2\mathrm{AFC}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(p_{0,i}, p_{1,j}), \quad (7a)$$

where the scoring rule $I(p_{0,i}, p_{1,j})$ is defined as

$$I(p_{0,i}, p_{1,j}) = \begin{cases} 0.0 & \text{if} \quad p_{1,j} < p_{0,i} \\ 0.5 & \text{if} \quad p_{1,j} = p_{0,i} \\ 1.0 & \text{if} \quad p_{1,j} > p_{0,i} \end{cases} \quad (7b)$$

(see appendix A for further details).

[3] In the case of dichotomous forecasts, a forecast of $x_{\cdot,i} = 1$ implies a forecast of an event, whereas in the case of polychotomous forecasts, a forecast of $x_{\cdot,i} = 1$ implies a forecast of category 1. In the dichotomous case the forecast categories are labeled 0 and 1 to correspond with the observed values of 0 for nonevents and 1 for events. In the polychotomous case the forecast categories are numbered from 1 to m_f, and it is implied that the higher the forecast category, the more likely it is that an event is expected to occur.

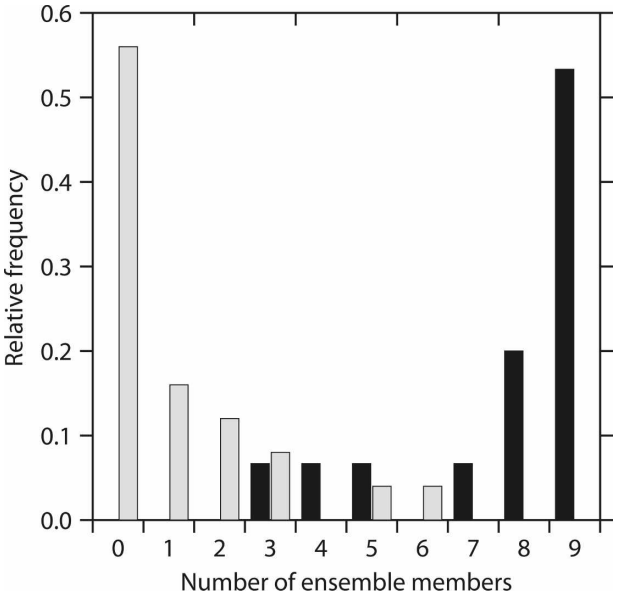

FIG. 1. Relative frequencies of numbers of ensemble members forecasting the Niño-3.4 index to exceed 27.0°C when the index did exceed 27.0°C (black) and when it did not (gray).

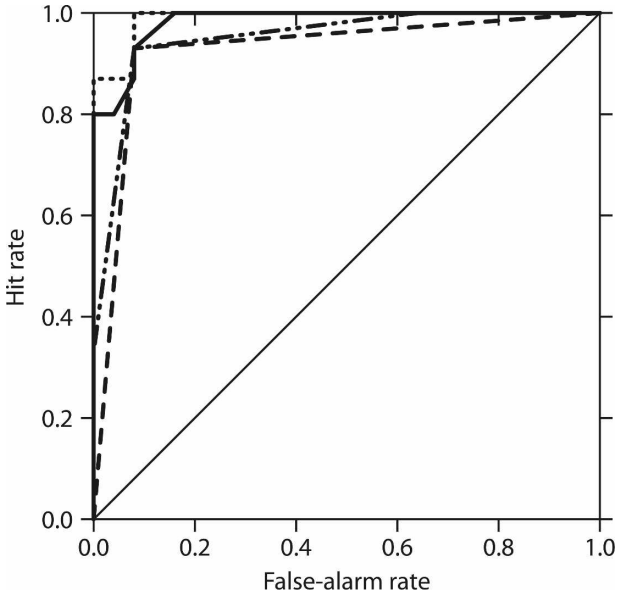

FIG. 2. ROC curve for CNRM forecasts for the Niño-3.4 index to exceed 27.0°C. The ROC curves for dichotomous (dashed), polychotomous (dashed–dotted), discrete probabilistic (solid), and continuous forecasts (dotted) are shown.

One limitation of Eq. (7) is that the score considers only the ordering of the probabilities, ignoring the actual probability values themselves, and thus is insensitive to any monotonic transformation of the probabilities. This insensitivity is the same problem as the insensitivity of the ROC to calibration (Glahn 2004). Although 2AFC scores can be defined that explicitly consider the actual probabilities (appendix B), these scores are not proper (appendix C).

For the CNRM data, forecast probabilities were defined as the proportion of the ensemble members forecasting an index of greater than 27.0°C.[4] With nine ensemble members, the number of probability bins is therefore 10. A histogram of the forecast probabilities for warm and cool events is shown in Fig. 1, indicating that the warm and cool events are well discriminated by the forecasts. The corresponding ROC curve for these forecasts is shown in Fig. 2. The area beneath the curve is about 0.98 (i.e., the 2AFC score is 98%) and is larger than for the dichotomous and polychotomous forecasts because of fewer ties. The ROC curves for the dichotomous and polychotomous forecasts are shown as gray solid and dashed lines, respectively, and it is clear that the poorer resolution in these forecasts compared to the discrete probabilities is responsible for the slightly smaller area.

4) CONTINUOUS FORECASTS

If the forecasts are continuous, the 2AFC score remains essentially the same as for the polychotomous forecasts, except that there will be as many categories as there are distinct forecast values ($m_f = n$; unless there are ties, which may occur if there is a zero bound, e.g., in which case $m_f < n$). Let the forecasts for events and nonevents be pooled, and then ranked in ascending order; if the forecasts are good, the ranks for the forecasts when an event occurred should be higher than for when no event occurred. Setting $r_{1,j}$ as the rank of the forecast for the jth event, the 2AFC score is derived from the equation for the Mann–Whitney U statistic (Conover 1999; Sheskin 2007):

$$p_{2\mathrm{AFC}} = \frac{\sum_{j=1}^{n_1} r_{1,j} - \dfrac{n_1(n_1+1)}{2}}{n_0 n_1}. \tag{8}$$

The second term in the numerator represents the sum of the ranks for the worst possible set of forecasts for the events (i.e., the forecasts for the events are all ranked first), and so the numerator as a whole calculates the number of times a rank for the forecasts for

[4] There are more reliable methods of calculating forecast probabilities (e.g., Kharin and Zwiers 2003), but this method is used purely for the sake of simplifying the example.

when an event occurs is greater than for forecasts for when a nonevent occurs (i.e., the number of correct 2AFC tests). Equation (8) can be shown to be a special case of Eq. (1), because

$$r_{1,j} = 0.5 + \sum_{i=1}^{n_1} I(x_{1,i}, x_{1,j}) + \sum_{i=1}^{n_0} I(x_{0,i}, x_{1,j}) \qquad (9)$$

and

$$\sum_{j=1}^{n_1} \left[0.5 + \sum_{i=1}^{n_1} I(x_{1,i}, x_{1,j}) \right] = \frac{n_1(n_1 + 1)}{2}. \qquad (10)$$

Equation (8) is therefore equal to Eq. (5) not just when $m_f = n$, but also because of the incorporation of forecast ties into the definition of the ranks in Eq. (9), when $m_f < n$.

To illustrate, consider the ensemble-mean CNRM forecasts for the Niño-3.4 index. These forecasts are to be used to predict whether the observed value will exceed 27.0°C (defined as a warm event). To perform an individual 2AFC test, the forecasts are ranked on the basis of the forecast temperature, and the forecast with the higher temperature, for example, is selected as the indicator of the warm event.[5] To calculate the 2AFC across all warm–cool event pairings, an ROC graph could be constructed with the forecast with the highest temperature being assigned as the year most likely to have experienced a warm event (Fig. 2). In essence, the procedure is to assign an arbitrary, but monotonically increasing, probability for a warm event to each forecast as the forecast temperature increases and then to perform a standard ROC as if the forecasts were probabilistic. This procedure was used by Thomson et al. (2006) and is a more comprehensive way of evaluating the quality of continuous forecasts than reducing them to a probabilistic forecast of 0% or 100% depending upon whether the forecast exceeds 27.0°C [as in section 3a(1)] because the conversion to probabilistic forecasts involves considerable loss of information. The 2AFC score for the Niño-3.4 forecasts is almost 99% and is a notably higher score than for the 2 × 2 or 2 × 4 contingency tables because of the elimination of ties.

[5] Note that both or neither of the forecasts in a specific 2AFC test may be above or below the event threshold of 27.0°C, but that a warm event can still successfully be discriminated from a nonevent: if forecast 1 is for 26.5°C and forecast 2 is for 26.0°C, and observation 1 is the warm event, the 2AFC test will be passed correctly. This calibration problem is discussed further in section 4.

5) CONTINUOUS PROBABILISTIC FORECASTS

If a probability distribution is provided as the forecast, that is, a density function that is defined over a continuous scale of values, the 2AFC score can be generalized from that for the discrete probabilistic forecasts [section 3a(3)]. Given forecast distributions, the probability that a value drawn from the one forecast distribution exceeds that from the other can be calculated and conditioned on the prior knowledge that the two observations can be distinguished (appendix A). This conditional probability, $F(p_{0,i}, p_{1,j})$, can then be used as the basis for selecting which of the two corresponding observations represents the event of interest.

If $F(p_{0,i}, p_{1,j}) > 0.5$, then it would seem reasonable to select (in this case correctly) the observation corresponding to forecast $p_{1,j}$ as the event, but to select (in this case incorrectly) the observation corresponding to forecast $p_{0,i}$ if $F(p_{0,i}, p_{1,j}) < 0.5$. Therefore, if $p_{0,i}$ and $p_{1,j}$ in Eq. (7a) are distributions rather than discrete probabilities, Eq. (7b) can be modified to

$$I(p_{0,i}, p_{1,j}) = \begin{cases} 0.0 & \text{if} \quad F(p_{0,i}, p_{1,j}) < 0.5 \\ 0.5 & \text{if} \quad F(p_{0,i}, p_{1,j}) = 0.5 \\ 1.0 & \text{if} \quad F(p_{0,i}, p_{1,j}) > 0.5 \end{cases}. \qquad (11)$$

Although it would seem reasonable to use $F(p_{0,i}, p_{1,j})$ as the scoring rule in place of $I(p_{0,i}, p_{1,j})$ in Eq. (7a), the resulting score is improper (appendix C). Note that the scoring rule for discrete probability forecasts, Eq. (7b), can be derived as a special case from Eq. (11) (appendix A, section b). Also note that if the probability distributions for both forecasts are Gaussian, $F(p_{0,i}, p_{1,j})$ becomes equivalent to a Student's t test for the difference in the means of the two forecast distributions [or a Welch's t test if the variances of the two distributions differ; Sheskin (2007)]. Similarly, if the distributions are similar but non-Gaussian, $F(p_{0,i}, p_{1,j})$ is equivalent to a Mann–Whitney U test. The 2AFC score could therefore be interpreted as being based on a series of Student's t or Mann–Whitney U tests (appendix A, section a).

As an example, the CNRM forecasts for the Niño-3.4 index were converted to Gaussian forecast distributions by using the ensemble mean and standard deviation as parameters for each forecast. The 2AFC score is almost 99%, indicating excellent discrimination between warm and cool events, and is consistent with the score for the continuous forecasts [section 3a(4)]. In fact, if the ensemble distributions are symmetric, then Eq. (11) becomes equivalent to Eq. (8) because $F(p_{0,i}, p_{1,j})$ will always be greater than 0.5 if the mean of $p_{1,j}$ is greater than the mean of $p_{0,i}$, and so it is only the rankings of the ensemble means that determine the score.

This score of 99% for the continuous probabilistic forecasts is larger than that for the discrete probabilistic forecasts. This result is consistent with the observation of Doblas-Reyes et al. (2008) that forecast probabilities should be indicated with as much precision as the forecast system allows since the binning of probabilities tends to result in a degeneration of skill.

b. Polychotomous observations (nominal and ordinal)

If there are more than two possible outcomes, the 2AFC test can still be applied. Let the number of observed categories be m_v, and let the categories be ordinal, with category 1 representing the lowest values and category m_v the highest; the 2AFC score defined in Eq. (1) can then be generalized to

$$p_{2\mathrm{AFC}} = \frac{\sum_{k=1}^{m_v-1} \sum_{l=k+1}^{m_v} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} I(x_{k,i}, x_{l,j})}{\sum_{k=1}^{m_v-1} \sum_{l=k+1}^{m_v} n_k n_l}, \quad (12a)$$

where $I(x_{k,i}, x_{l,j})$ is defined in an analogous way to Eq. (1b):

$$I(x_{k,i}, x_{l,j}) = \begin{cases} 0.0 & \text{if} \quad x_{l,j} < x_{k,i} \\ 0.5 & \text{if} \quad x_{l,j} = x_{k,i} \\ 1.0 & \text{if} \quad x_{l,j} > x_{k,i} \end{cases}. \quad (12b)$$

As with Eq. (1), the objective is to calculate the proportion of correct 2AFC tests for all possible distinguishable pairs of observations. In the case of continuous forecasts, Eq. (12) is equivalent to a simple rescaling of Somer's δ (Agresti 1984), as discussed further in section 3c(4).

If the observed categories are not ordinal, then the less than (<) and greater than (>) symbols in Eq. (12b) do not make sense, and so the score has to be modified to

$$p_{2\mathrm{AFC}} = \frac{\sum_{l=1}^{m_v} \sum_{k \neq l}^{m_v} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} J(x_{k,i}, x_{l,j})}{\sum_{k=1}^{m_v} \sum_{k \neq l}^{m_v} n_k n_l}, \quad (13a)$$

where the scoring rule $J(x_{k,i}, x_{l,j})$ is defined as

$$J(x_{k,i}, x_{l,j}) = \begin{cases} 0.0 & \text{if} \quad x_{l,j} = k \quad \& \quad x_{k,i} \neq k \\ 0.5 & \text{if} \quad x_{l,j} = x_{k,i} \\ 0.5 & \text{if} \quad x_{l,j} \neq k \quad \& \quad x_{k,i} \neq k \\ 1.0 & \text{if} \quad x_{l,j} \neq k \quad \& \quad x_{k,i} = k \end{cases}. \quad (13b)$$

TABLE 4. Contingency table for January 1961–2000 observed values of the Niño-3.4 index in the ranges <26.0° (cold), 26.0°–27° (cool), 27.0°–28° (warm), and >28.0°C (hot) from the coupled ocean–atmosphere model of CNRM initialized in August 1960–99.

	Forecasts			
Observations	Hot	Warm	Cool	Cold
Hot	4	0	0	0
Warm	1	9	1	0
Cool	0	2	7	1
Cold	0	0	7	8

In most cases in the atmospheric sciences the categories will be ordinal, but the nominal case is considered for completeness.

1) DICHOTOMOUS FORECASTS

This case is not considered since some of the outcomes can never be forecast, and so the case is unrealistic.

2) POLYCHOTOMOUS FORECASTS

When there are more than two categories for the observations and the forecasts, the 2AFC score for dichotomous observations and forecasts [section 3a(1)] can be generalized to

$$p_{2\mathrm{AFC}} = \frac{\sum_{k=1}^{m_v-1} \sum_{l=k+1}^{m_v} \left(\sum_{i=1}^{m_f-1} \sum_{j=i+1}^{m_f} n_{k,i} n_{l,j} + 0.5 \sum_{i=1}^{m_f} n_{k,i} n_{l,i} \right)}{\sum_{k=1}^{m_v-1} \sum_{l=k+1}^{m_v} n_k n_l}, \quad (14)$$

where m_f is the number of forecast categories. As for Eq. (5), perpetual forecasts of the same category will score 0.5.

If the observed and forecast Niño-3.4 indices are classified as hot, warm, cool, and cold using ranges of <28.0°, 28.0°–27.0°, 27.0°–26.0°, and <26.0°C, respectively, the resulting contingency table suggests a highly skillful set of forecasts with most of the cases near the diagonal (Table 4). However, primarily because of the large number of forecasts for cool conditions when cold were observed, the 2AFC score drops to about 90%, somewhat less than the scores recorded for the dichotomous observations. The model has a slight warm bias, which is more clearly evident in Table 4 (the model forecasts cold conditions 9 times whereas they are observed 15 times) than in Table 2, and is affecting the

score for the model. Although the 2AFC score for continuous forecasts is insensitive to monotonic transformations of the forecasts (see further discussion in section 4), the score for polychotomous forecasts can be partially affected by biases, depending on how the categories are defined. This sensitivity to the categorization is not unique to the 2AFC score described here and should be considered an argument against the categorization of forecasts rather than as a weakness of the score. The flexibility of the 2AFC test is therefore an advantage: The score can be applied regardless of the data format.

If the categories are nominal, then any distinguishable set of forecast–observation pairs having at least one pair drawn from the diagonal will be marked as correct (i.e., at least one of the forecasts has to be for the correct category). The numbers of observed and forecast categories need to be identical because otherwise categories are forecast that cannot be observed or vice versa. The 2AFC score becomes

$$p_{2\mathrm{AFC}} = \frac{\sum_{k=1}^{m}\sum_{l\neq k}^{m}\left[\sum_{i\neq k}^{m} n_{k,k}n_{l,i} + 0.5\left(\sum_{i\neq k}^{m}\sum_{j\notin\{k,l\}}^{m} n_{k,l}n_{i,j} + \sum_{i=l}^{m} n_{k,i}n_{l,j}\right)\right]}{\sum_{k=1}^{m}\sum_{l\neq k}^{m} n_k n_l}, \tag{15}$$

where $m = m_v = m_f$. If the hot, warm, cool, and cold Niño-3.4 categories defined above are considered as nominal categories, the contingency table is the same as shown in Table 4, but the 2AFC score drops to about 80% because of the loss of information about the ordering of the categories. Given that the categories in this example are ordinal but are scored as if they are nominal, the difference between this score and that from Eq. (14) can be explained as follows. In cases where the forecasts are ranked correctly, but the highest ranking forecast is incorrect, the forecaster is no longer able to make a selection since the requested category is not forecast; in these cases the forecaster scores 0.5 instead of 1. In cases where the forecasts are ranked incorrectly, and neither forecast indicates the requested category, the forecaster is again unable to make a selection; in these cases the forecaster scores 0.5 instead of 0. In some cases, therefore, the forecaster gains, but in most cases, if the forecasts are skillful, the forecaster will lose when using the nominal version of the score, as in the CNRM example.

3) Discrete probabilistic forecasts

For discrete probabilistic forecasts of polychotomous categories, the 2AFC score is a generalized version of that for the dichotomous observations [section 3a(3)]. For ordinal categories, the score assesses the ability to identify the observation in the higher category, just as for the polychotomous forecasts, but makes the selection on the basis of the forecast probabilities, just as for the dichotomous observations. The score is therefore a generalization of Eqs. (7) and (11), and uses the probability that a value drawn from the one forecast distribution exceeds a value drawn from the other, conditioned on the prior knowledge that the two observations can be distinguished (see Appendix A). The 2AFC score is defined as

$$p_{2\mathrm{AFC}} = \frac{\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} I(\mathbf{p}_{k,i}, \mathbf{p}_{l,j})}{\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v} n_k n_l}, \tag{16a}$$

where $\mathbf{p}_{k,i}$ is the vector of forecast probabilities for the ith forecast given category k, and

$$I(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = \begin{cases} 0.0 & \text{if} \quad F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) < 0.5 \\ 0.5 & \text{if} \quad F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = 0.5 \\ 1.0 & \text{if} \quad F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) > 0.5 \end{cases}. \tag{16b}$$

The scoring rule $F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j})$ is defined as

$$F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = \frac{\sum_{r=1}^{m-1}\sum_{s=r+1}^{m} p_{k,i}(r)p_{l,j}(s)}{1 - \sum_{r=1}^{m} p_{k,i}(r)p_{l,j}(r)}. \tag{16c}$$

where $p_{k,i}(r)$ is the forecast probability for the rth category, and for the ith observation in category k. It cannot be used in place of $I(\mathbf{p}_{0,i}, \mathbf{p}_{1,j})$ in Eq. (16a) because the resulting score is improper (see appendix C).

For the CNRM data, using the same four categories for observations and forecasts as for the polychotomous forecasts [section 3b(2)], and using the simple counting procedure for assigning forecast probabilities as was used for the dichotomous forecasts [section 3a(3)], the 2AFC score is about 92%. Although this score contin-

ues to indicate a high degree of skill in forecasting the Niño-3.4 index, it remains lower than the scores for the dichotomous observations, which is attributable to the poor estimates of forecast probabilities using the counting procedure, and to the relatively poor ability to distinguish between the cool and cold conditions.

If the categories are nominal, Eq. (13) is simply applied to the probabilities, just as Eq. (7) was adapted from Eq. (1), except that each pairing has to be tested twice to determine whether the correct choice is made for both outcomes. The 2AFC score is therefore

$$p_{2\mathrm{AFC}} = \frac{\sum_{l=1}^{m_v}\sum_{k\neq 1}^{m_v}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} I[p_{k,i}(l), p_{l,j}(l)]}{\sum_{l=1}^{m_v}\sum_{k\neq l}^{m_v} n_k n_l}, \tag{17a}$$

where

$$I[p_{k,i}(l), p_{l,j}(l)] = \begin{cases} 0.0 & \text{if} \quad p_{l,j}(l) < p_{k,i}(l) \\ 0.5 & \text{if} \quad p_{l,j}(l) = p_{k,i}(l) \\ 1.0 & \text{if} \quad p_{l,j}(l) > p_{k,i}(l) \end{cases}. \tag{17b}$$

The score for the CNRM data is about 86%. As for the dichotomous observed categories, the score for the polychotomous forecasts treated as nominal categories is less than if they are treated as ordinal. However, if the categories are considered individually, the score for being able to identify the hot category is about 99%, while that for the warm is about 93%. The scores for the cold and cool categories are notably less (about 78% and 80%, respectively); the scores for these categories are negatively affected by the previously mentioned warm bias in the model.

4) Continuous forecasts

For continuous forecasts of ordinal polychotomous categories (nominal categories would not make sense), the 2AFC score is adapted in the same way as for the dichotomous categories [section 3a(4)]: For n forecasts (and no ties), there are n categories. The question addressed by the 2AFC test remains identical: Which is the observation with the higher value? However, since there are fewer ties in the observations now that there are more categories, a larger number of tests can be conducted. Let the forecasts for categories k and l only be pooled and then ranked in ascending order. Setting $r_{l,j}$ as the rank of the forecast for the jth event (where an "event" is an occurrence of category l, and the ranks are calculated using only those for forecasts for when an event l occurred), the 2AFC score is generalized from Eq. (8):

$$p_{2\mathrm{AFC}} = \frac{\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v}\left[\sum_{j=1}^{n_l} r_{l,j} - \dfrac{n_l(n_l+1)}{2}\right]}{\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v} n_k n_l}. \tag{18}$$

The 2AFC score for the CNRM data is about 92% and is most negatively affected by a relatively poor ability to discriminate cool from cold conditions (73%). The only other partial scores that were imperfect were between the warm and cool categories (97%), and between warm and hot (98%); that is, the model could not perfectly discriminate between observations in neighboring categories. However, the model's ability to discriminate between observations in neighboring categories is better for the warmer compared to the cooler categories. A similar ability to forecast the warmest categories most successfully was noted when discussing the discrete probabilistic forecasts [section 3b(3)].

5) Continuous probabilistic forecasts

For the dichotomous observed categories, the 2AFC score was based on tests for the difference between the forecast probability distribution functions for events compared to those for nonevents. A similar principle applies in the case of polychotomous observed categories except that the forecast probability distributions are compared to see whether they can be used to identify the observation in the highest category. Equation (11) is therefore generalized to

$$I(p_{k,i}, p_{l,j}) = \begin{cases} 0.0 & \text{if} \quad F(p_{k,i}, p_{l,j}) < 0.5 \\ 0.5 & \text{if} \quad F(p_{k,i}, p_{l,j}) = 0.5 \\ 1.0 & \text{if} \quad F(p_{k,i}, p_{l,j}) > 0.5 \end{cases}, \tag{19a}$$

and the 2AFC score becomes

$$p_{2\mathrm{AFC}} = \frac{\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} I(p_{k,i}, p_{l,j})}{\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v} n_k n_l}, \tag{19b}$$

which is the same as Eq. (16). Note that again the use of $F(p_{k,i}, p_{l,j})$ as the scoring rule in place of $I(p_{k,i}, p_{l,j})$ in Eq. (19b) renders the score improper.

Using the same Gaussian assumption for the CNRM forecasts for the Niño-3.4 index as used for the dichotomous forecasts, the 2AFC score is about 95%, which is a little worse than for the dichotomous forecasts. This reduction is primarily because of a relatively weak abil-

ity to distinguish cold from cool conditions, which has been noted a number of times in previous sections. When the observations were categorized only into warmer or colder than 27.0°C, the cold and cool categories were grouped together and so the forecasts did not have to discriminate between them. The 2AFC score for discriminating cold from cool using the Gaussian forecast distributions is only about 70%. However, the scores for discriminating all other categories exceed 98%.

c. Continuous observations

If the observations are measured on a continuous scale,[6] the 2AFC score can be generalized in the same way as for the difference between continuous and discrete probabilistic forecasts: The n observations are treated as if they were a set of n polychotomous ordinal categories. Assuming that there are no ties in the observations, the 2AFC score defined in Eqs. (1) and (12) can then be generalized to

$$p_{2\mathrm{AFC}} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(x_i, x_j), \tag{20a}$$

where

$$I(x_i, x_j) = \begin{cases} 0.0 & \text{if} \quad x_i < x_j \\ 0.5 & \text{if} \quad x_i = x_j \\ 1.0 & \text{if} \quad x_i > x_j \end{cases}. \tag{20b}$$

If there are ties in the observations, then the scores described in section 3b can be applied.

1) Dichotomous forecasts

This case is not considered because the forecasts cannot be compared with the observations without reducing the observations to categories.

2) Polychotomous forecasts

Similarly, this case is not considered because the forecasts cannot be compared with the observations without reducing the observations to categories.

3) Discrete probabilistic forecasts

As for the dichotomous and polychotomous forecasts, this case is not considered because the forecasts cannot be compared with the observations without reducing the observations to categories.

4) Continuous forecasts

For continuous forecasts, the 2AFC score could be calculated using Eq. (18), setting the number of categories to n. However, it is considered conceptually simpler to use Eqs. (20a) and (20b). Note that Eqs. (20a) and (20b) are a simple transformation of Kendall's correlation coefficient, τ. Kendall's τ is defined as

$$\tau = \frac{4P}{n(n-1)} - 1, \tag{21}$$

where P is the number of so-called concordant pairs (Sheskin 2007). A concordant pair is defined as a pair of bivariate observations, $\mathbf{x} = \{x_1, x_2\}$ and $\mathbf{y} = \{y_1, y_2\}$, in which $\mathrm{sgn}(x_1 - x_2) = \mathrm{sgn}(y_1 - y_2)$, that is, in which the ranking of the two values in $\mathbf{x}$ is the same as in $\mathbf{y}$. Recalling from section 3a that $\mathrm{sgn}(x_1 - x_2) = 2 \times I(x_2, x_1) - 1$, it is easy to show that

$$p_{2\mathrm{AFC}} = \frac{1}{2}(\tau + 1). \tag{22}$$

As mentioned, if there are ties in the observations, the tests in section 3b can be used instead, and the rescaling indicated in Eq. (22) can be applied to Eqs. (18) or (12), defining a simple rescaling of Somer's δ (Agresti 1984). Somer's δ is an alternative to Kendall's τ for situations with tied observational values (Sheskin 2007). Compared to Somer's δ, Goodman and Kruskal's γ is a more commonly used alternative test to Kendall's τ for situations with ties (Sheskin 2007). However, Goodman and Kruskal's γ does not draw a distinction between tied observations and tied forecasts. This distinction is important in the 2AFC scores: Tied observations cannot be discriminated and so are not considered in the score, whereas tied forecasts score 0.5 because of an inability to discriminate observations that are different.

The CNRM forecasts score 87% using Eq. (22), which is less than for the continuous forecasts scored against the polychotomous observations, and less still than against the dichotomous observations. The added precision required to correctly rank all the observed values, rather than just to discriminate successfully between categories, is responsible for the decreased score. This decrease in the score as the information content of the observations increases is not a characteristic peculiar to the scoring procedures proposed in this paper, but is evident with more traditional scoring procedures

[6] The tests in this section apply whether the data are unbounded, single bounded, or double bounded, and so can apply to meteorological variables that can take any value; that have an absolute zero, for example; and to proportions, which have lower limits of zero and upper limits of one. However, the tests may not be easily adaptable to data on a circular scale, such as directions or calendar dates.

also. It should not be taken as an argument in favor of categorizing, but should rather be seen as evidence of a trade-off between precision and skill: The more precise the forecasts have to be, the greater the chance that random errors, whether because of inherent unpredictability or perhaps resulting from observational errors (Bowler 2008), will adversely affect the score.

5) CONTINUOUS PROBABILISTIC FORECASTS

The simple extension of the polychotomous observations tests for continuous forecasts to the case of continuous observations by defining n categories can also be applied for forecasts expressed as probability distribution functions. Equation (19b) therefore becomes

$$p_{2\mathrm{AFC}} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(p_i, p_j), \qquad (23a)$$

where

$$I(p_i, p_j) = \begin{cases} 0.0 & \text{if} \quad p_i < p_j \\ 0.5 & \text{if} \quad p_i = p_j \\ 1.0 & \text{if} \quad p_i > p_j \end{cases}. \qquad (23b)$$

The CNRM forecasts score 87% using Eqs. (23a) and (23b), which is the same as for the continuous forecasts because of the symmetry assumption for the forecast probability distributions. Both these scores for the continuous observations are the lowest of all the scores reported except for the nominal polychotomous forecasts. The added precision required to forecast successfully the ranking of all the observations is responsible for the drop in the scores compared to those for the categorized observations. For the latter, the forecaster is not expected to discriminate between observations within categories.

d. Observations as probability distributions

The observations themselves can be presented as probability distribution functions (pdfs) to represent the observational errors, for example. It is possible to extend the 2AFC scores to apply to such cases. The observational pdf's would be ranked in exactly the same way as the forecast pdf's in section 3a(5), and the test would proceed as for the continuous observations (section 3c) using these ranks rather than those based on the best guesses of the observations. If the observational errors are symmetric, then the ranking will be identical to that from the best guesses.

4. Discussion

A number of potential criticisms that could be targeted at the 2AFC score are addressed in this section. Questions relating to the versions of the score for probabilistic forecasts (discrete and continuous) are discussed first; specifically, the question of its sensitivity to the calibration of the probabilities is considered. Then, in section 4b, some problems with the treatment of continuous observations are discussed. In section 4c the effects of spatial and temporal autocorrelation are discussed, and suggestions for addressing these issues are considered.

a. Propriety and calibration

It was noted in section 3a(3) that Eq. (7) is equivalent to the calculation of the trapezoidal area beneath the ROC curve and that, in the context of discrete probabilistic forecasts, the 2AFC score therefore considers only the ranking of the probabilities, and is insensitive to monotonic transformations of these probabilities. This criticism has to be taken seriously within the context of an administrative verification score because a good score may be misinterpreted as an indication of well-calibrated forecasts, and because improvements in the reliability of probabilistic forecasts would not be registered. Versions of the 2AFC score that do consider the probabilities are discussed in appendix B, where they are rejected as being effectively improper. Strictly speaking, the standard definition of propriety (see Bröcker and Smith 2007) cannot be applied to the 2AFC because the 2AFC is defined in terms of two forecast and observation pairs rather than a single pairing. Nevertheless, from idealized examples, it can be shown that the probabilistic 2AFC scores are not optimized when the forecaster issues probabilities that are consistent with his or her best judgment (see appendix C). The lack of propriety of these probabilistic 2AFC scores may be one of the costs of their equitability, which is considered a fundamental component of its simplicity of interpretation. At least in the case of more traditional scores, which consider single forecast–observation pairs, equitability precludes propriety (Jolliffe and Stephenson 2008).

Given the problems with the lack of propriety with the probabilistic versions of the 2AFC score, the authors recommend the scores presented in the sections above that are insensitive to the calibration of the probabilities. The objective of the proposed scoring framework is to provide a simple metric for indicating the potential usefulness of the forecasts; a score that measures reliability and resolution (or discrimination) cannot easily distinguish a reliable, but poorly resolved, set of forecasts, from an unreliable, but well-resolved, set of forecasts. The potential usefulness of forecasts with negative Brier skill scores, because of the strict requirement for good reliability, is a case in point (Mason

2004). Given that the 2AFC considers only the ability to discriminate observations, may enable a party interested in using forecasts to identify which of two forecast systems, for example, is likely to be the more useful. Thus, the 2AFC score can be considered a measure of the potential usefulness of the forecasts rather than as a comprehensive measure of all aspects of forecast quality. Any serious use of forecasts should consider how the forecasts may need to be calibrated, and so the 2AFC score is able to indicate the forecasts that will prove most useful only after proper calibration. Of course, this argument is not meant to denigrate the importance of issuing well-calibrated forecasts in the first place, and so some consideration of reliability needs to be made, even within administrative contexts. The authors submit that reliability could be measured separately, as a more detailed diagnostic, and we recommend that it be emphasized that a good 2AFC score does not necessarily indicate that the forecasts can be taken at face value.

b. Continuous observations

A second criticism of the 2AFC score is that continuous observations are reduced to an ordinal scale. In this respect, many of the arguments above on the calibration of probabilistic forecasts apply again, but some additional comments are warranted. Whereas Pearson's correlation coefficient is widely used for verification of forecasts of continuous variables, and does not reduce the observations to an ordinal scale, this score is sensitive to distributional assumptions, and so-called nonparametric alternatives are often to be preferred especially since they are almost as powerful (Sheskin 2007). Of the best-known nonparametric measures of association, Spearman's rank-order correlation coefficient is much more widely used than Kendall's τ, partly because of its relative computational efficiency, but also because of its close relationship to Pearson's coefficient. However, in addition to its affinity to the 2AFC score, Kendall's τ has additional advantages over Spearman's coefficient: The sample correlation is an unbiased estimate of the population parameter τ, and the sampling distribution of τ closely follows the normal distribution even for small sample sizes (Lindeman et al. 1980).

Although the conversion of continuous data to ranks may well be a disadvantage, one positive implication is that the 2AFC score can be used effectively on predictands that are measured on a range of different scales. It can be used, for example, on unbounded interval data (for most practical purposes temperature forecasts could be considered an example, although strictly temperatures do have a lower bound), ratio data (e.g., quantitative precipitation forecasts), and proportions (e.g., cloud cover). However, it is not clear that the score would make much sense on circular data (e.g., wind directions) because of the inability to rank data on this scale. Calendar dates are also measured on a circular scale, but in many cases (specifically when the dates apply to only a part of the year) a ranking of dates may be meaningful. Monsoon onset dates, for example, can be ranked from early to late onset because they do not span the whole calendar year. In such cases the 2AFC scores can be applied meaningfully as long as the forecasts display similar properties.[7]

c. Spatial and temporal autocorrelation

In the introduction to section 3 it was mentioned that the individual 2AFC tests that compose the 2AFC scores are not independent. It is often mentioned that most verification scores assume that each forecast–observation pair is independent of all the other pairs (e.g., Wilks 2006). Dependency arises from spatial and/or temporal correlation and invalidates the standard tests for statistical significance of verification scores. However, the authors recommend against calculating the statistical significance of the 2AFC scores because difficulties in their interpretation (e.g., Nicholls 2001; Jolliffe 2004; Mason 2008) are likely to introduce unnecessary confusions when the results are being communicated to nonexperts. Nevertheless, some indication of the uncertainty in the 2AFC score is desirable because of potentially large sampling errors when sample sizes are small. Although they have a close affinity to p values, confidence limits are recommended instead (Jolliffe 2004; Mason 2008), since specifying a range of 2AFC scores is likely to cause much less confusion than trying to communicate two probabilities (the score itself and the p value), each of which has a very different meaning. For some of the versions of the 2AFC score, analytical procedures can be used to obtain confidence intervals [e.g., for the version that is equivalent to the ROC; Mason and Graham (2002)], but these procedures will be invalidated given spatial

[7] One can imagine a pathologically bad model that forecasts onset dates that are randomly distributed throughout the year, making a distinction between very early and very late onsets a somewhat arbitrary distinction. In some cases, therefore, it would be difficult to make a selection in a 2AFC test even though the two forecasts may differ. Ad hoc adjustments to the score could be made to handle cases when the interpretation of the forecast itself is unclear (whether the forecast is indicating an early or a late onset), but this topic is beyond the scope of this paper.

and/or temporal dependency. Bootstrap procedures could be used in these instances.

5. Summary

A framework for forecast verification has been described that is designed to be sufficiently general to be applicable to a wide variety of observation and forecast data types. Most verification situations in the atmospheric sciences can be addressed by this framework, except for data measured on a circular scale, examples of which include forecasts of wind directions, or some forecasts for which the predictand is a calendar date. The framework is based on the two-alternative forced choice (2AFC) test that the authors believe is sufficiently intuitive to make it a suitable procedure for communication of forecast quality to nonspecialists. Although the actual computation of the score and some of the equations may appear fairly complex, the basic concept remains simple: Given any two differing observations, what is the probability that the forecaster can successfully discriminate the observations using the corresponding forecasts? For quantitative precipitation forecasts, for example, the question may be: What is the probability that the forecaster will successfully identify the wettest day? For binary outcomes, for example, the question might be: What is the probability that the forecaster will pick the day on which a tornado occurred? In all cases the score could be loosely interpreted as an attempt to answer the question: How often are the forecasts correct? However, it avoids the numerous interpretive pitfalls of the percent correct score, and various attempts to normalize this score by calculating a skill score, and retains the intuitive property of having an expected value of 50% for forecasts without skill. Because the score is based upon the ability of the forecasts to discriminate observations, the authors propose the name "discrimination score" as a more accessible name than the "two-alternative forced choice score" or its acronym.

Depending upon the nature of the observations and forecasts, the 2AFC score takes different formats because of an attempt to use as much of the information in the observations and forecasts as possible. The aim is to minimize the reduction of the more complex to the simple, although differences in the various formulations of the score using CNRM forecasts of the Niño-3.4 index as examples have indicated some notable conditioning of the quality of the forecasts upon the outcome.

Rather than constituting a suite of entirely new scores, in some cases the 2AFC score has been shown to be equivalent to, or closely related to, statistical tests known under different names. Some of these tests are already widely used in forecast verification (e.g., the trapezoidal area beneath the ROC curve), others are based on tests widely used for purposes other than forecast verification (Student's t test), while still others are not yet widely used in the atmospheric sciences (e.g., Somer's δ and Kendall's τ).

Unavoidably, the 2AFC score has a number of limitations that preclude its use as a generic score for all verification analyses, and that present opportunities for misinterpretation. Its primary weakness is that in some of its formulations the score is insensitive to the calibration of the forecasts. As a result, the score should be interpreted as a measure only of potential usefulness and should not be seen as an infallible indication of whether the forecasts can be taken at face value. These limitations, however, are inevitable products of any attempt to summarize forecast quality in a single number. It is proposed only that the 2AFC score may be a good starting point for beginning a more in-depth discussion with forecast user communities on the quality of forecast sets.

Acknowledgments. This article was funded by Cooperative Agreement AN07GP0213 from the National Oceanic and Atmospheric Administration (NOAA), by EU FP 6 Contract GOCE-CT-2003-505539 from the ENSEMBLES project, and by the Swiss National Science Foundation through the National Centre for Competence in Research (NCCR) Climate. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies. The authors thank the NCCR Climate for financing Weigel's research visit to the IRI, which facilitated this collaboration. Special thanks are due to M. S. J. Harrison for his role in initiating and encouraging this work. Helpful comments from A. G. Barnston, P. Della-Marta, and two anonymous referees are gratefully acknowledged. Computer code for most of the procedures presented in this article, and written in Fortran and R, is available from the authors.

APPENDIX A

The 2AFC for Probabilistic Forecasts

Let $p_{0,i}(x)$ represent the probability density of the ith forecast for an event when an event did not occur, and let $p_{1,j}(x)$ be the jth forecast probability density when an event did occur. The quantities $P_{0,i}(x)$ and $P_{1,j}(x)$ are the corresponding cumulative distributions. Let $X_{0,i}$ be a random sample drawn from $p_{0,i}(x)$, and let $X_{1,j}$ be a random sample drawn from $p_{1,j}(x)$. The probability that $X_{1,j} > X_{0,j}$ $[p(X_{1,j} > X_{0,i})]$ is given by

$$p(X_{1,j} > X_{0,i}) = 1 - p(X_{1,j} \le X_{0,i})$$
$$= 1 - \int_{-\infty}^{\infty} p_{0,i}(x) \int_{-\infty}^{x} p_{1,j}(y)\, dy\, dx$$
$$= 1 - \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx. \quad \text{(A1)}$$

Thus, $p(X_{1,j} > X_{0,i})$ is an obvious basis for a 2AFC score. However, one must consider that, by the nature of the 2AFC test, it is known a priori that the two observations can be discriminated. Therefore, $p(X_{1,j} > X_{0,i})$ needs to be conditioned on the prior knowledge that $X_{1,j} \neq X_{0,i}$. Using Bayes's theorem,

$$p(X_{1,j} > X_{0,i} | X_{1,j} \neq X_{0,i}) = \frac{p(X_{1,j} > X_{0,i})}{p(X_{1,j} \neq X_{0,i})} = \frac{p(X_{1,j} > X_{0,i})}{p(X_{0,i} > X_{1,j}) + p(X_{1,j} > X_{0,i})}. \quad \text{(A2)}$$

Setting $F(p_{0,i}, p_{1,j}) = P(X_{1,j} > X_{0,i} | X_{1,j} \neq X_{0,i})$, and using Eq. (A2) in Eq. (A1), one obtains

$$F(p_{0,i}, p_{1,j}) = \frac{1 - \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx}{2 - \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx - \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx}. \quad \text{(A3)}$$

If $F(p_{0,i}, p_{1,j}) > 0.5$, then it would seem reasonable to select (in this case correctly) the observation corresponding to forecast $p_{1,j}$, but to select (in this incorrectly) the observation corresponding to forecast $p_{0,i}$ if $F(p_{0,i}, p_{1,j}) < 0.5$. It can be shown that

$$F(p_{0,i}, p_{1,j}) > 0.5 \Leftrightarrow \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx < \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx. \quad \text{(A4)}$$

Therefore, the scoring rule formulated in Eq. (11) is equivalent to

$$I(p_{0,i}, p_{1,j}) = \begin{cases} 0.0 & \text{if } \int_{-\infty}^{\infty} p_{1,,j}(x) P_{0,i}(x)\, dx < \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx \\ 0.5 & \text{if } \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx = \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx. \\ 1.0 & \text{if } \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx > \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx \end{cases} \quad \text{(A5)}$$

Now consider two special cases: first, the case that the probability distributions for both forecasts are truly continuous functions, and, second, that only discrete probabilities for the event to happen are issued.

a. Probability distributions that are continuous functions

If both probability distributions are *continuous* functions, then $p(X_{1,j} > X_{0,i}) = p(X_{1,j} \ge X_{0,i})$ and $p(X_{1,j} \neq X_{0,i}) = 1$. Equation (A3) then simplifies to

$$F(p_{0,i}, p_{1,j}) = 1 - \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx = \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx, \quad \text{(A6)}$$

and Eq. (A5) becomes

$$I(p_{0,i}, p_{1,j}) = \begin{cases} 0.0 & \text{if } \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx < 0.5 \\ 0.5 & \text{if } \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx = 0.5. \\ 1.0 & \text{if } \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx > 0.5 \end{cases} \quad \text{(A7)}$$

Note that if the probability distributions for both forecasts are symmetric, Eq. (A7) is equivalent to comparing the ensemble means. If the two distributions are Gaussian, the test becomes equivalent to a Student's t test for the difference in the means of the two forecast distributions, or a Welch's t test if the variances of the

two distributions differ. Otherwise, the test is equivalent to a Mann–Whitney U test.

If the probability distributions reveal delta peaks (e.g., precipitation forecasts that allow the explicit value 0), then the assumption of a continuous distribution is violated and the more general Eqs. (A3) and (A5), respectively, must be applied.

b. Discrete probability forecasts

Equation (7b) for discrete probability forecasts can be derived as a special case from Eq. (A5) above. Consider two discrete probabilities, $p_{0,i}$ and $p_{1,j}$, with $p_{0,i}$ representing the ith forecast probability for an event when an event did not occur and $p_{1,j}$ representing the jth forecast probability when an event did occur. Let the event have a value of 1, and the nonevent a value of 0. Using Dirac's delta distribution, $p_{0,i}$ and $p_{1,j}$ can be formulated in distributional form as follows:

$$p_{0,i}(x) = (1 - p_{0,i})\delta(x) + p_{0,i}\delta(x - 1)$$
$$p_{1,j}(x) = (1 - p_{1,j})\delta(x) + p_{1,j}\delta(x - 1). \quad \text{(A8)}$$

The corresponding cumulative distributions are given by

$$P_{0,i}(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 - p_{0,i} & \text{if} \quad 0 \le x < 1 \\ 1 & \text{if} \quad x \ge 1 \end{cases} \quad \text{(A9a)}$$

and

$$P_{1,j}(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1 - p_{1,j} & \text{if} \quad 0 \le x < 1. \\ 1 & \text{if} \quad x \ge 1 \end{cases} \quad \text{(A9b)}$$

Using the central property of Dirac's delta distribution, namely that $\int_{-\infty}^{\infty} f(x)\delta(x - a) = f(a)$ for any function f, one obtains the following with Eqs. (A8) and (A9) in Eq. (A3):

$$F(p_{0,i}, p_{1,j}) = \frac{1 - (1 - p_{0,i})P_{1,j}(0) - p_{0,i}P_{1,j}(1)}{2 - (1 - p_{0,i})P_{1,j}(0) - p_{0,i}P_{1,j}(1) - (1 - p_{1,j})P_{0,i}(0) - p_{1,j}P_{0i}(1)}$$
$$= \frac{p_{1,j}(1 - p_{0,i})}{p_{1,j}(1 - p_{0,i}) + p_{0,i}(1 - p_{1,j})}. \quad \text{(A10)}$$

From Eq. (A10) it can be shown that $F(p_{0,i}, p_{1,j}) > 0.5$ if and only if $p_{1,j} > p_{0,i}$. Thus, for this special case, Eq. (A5) simplifies to

$$I(p_{0,i}, p_{1,j}) = \begin{cases} 0.0 & \text{if} \quad p_{1,j} < p_{0,i} \\ 0.5 & \text{if} \quad p_{1,j} = p_{0,i} \\ 1.0 & \text{if} \quad p_{1,j} > p_{0,i} \end{cases}; \quad \text{(A11)}$$

that is, Eq. (7b) is retained.

APPENDIX B

A Probabilistic Version of the 2AFC

In Eq. (7) the 2AFC score is calculated by selecting the forecast with the higher probability as corresponding to the event. If this selection is correct, the forecaster scores 1, if incorrect 0, and if the two forecasts are identical, the forecaster scores 0.5. However, given that the forecasts are probabilities, the forecaster can indicate a degree of belief that a specific selection will be correct. This degree of belief can be expressed as a probability that is not the same as the forecast probability for the event: By the nature of the 2AFC test, it is known a priori that one and only one of the two observations is an event, but when making these forecasts, no such prior knowledge was available. It would seem reasonable to score the forecasts using this probability (see appendix A) rather than using the simple scoring metric defined above. These probabilistic versions of the 2AFC score are described in the following sections, using the same breakdown of observations into dichotomous, polychotomous, and continuous scalings, and of forecasts into discrete and continuous probability distributions.

a. Dichotomous observations

Let $p_{1,i}$ represent the ith forecast probability for an event when an event occurred, and let $p_{0,i}$ represent the ith forecast probability for an event when an event did not occur. The appropriate level of belief in a correct selection of the event, $F(p_{0,i}, p_{1,j})$, would replace $I(x_{0,i}, x_{1,j})$ in Eq. (1), which would then become

$$p_{2\text{AFC}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} F(p_{0,i}, p_{1,j}). \quad \text{(B1)}$$

The calculation of Eq. (B1) depends on whether $p_{0,i}$ and $p_{1,j}$ represent discrete probabilities or continuous probability distributions.

1) DISCRETE PROBABILISTIC FORECASTS

The appropriate level of belief in a correct selection of the event, $F(p_{0,i}, p_{1,j})$, has been derived in Eq. (A10) and is given by

$$F(p_{0,i}, p_{1,j}) = \frac{p_{1,j}(1 - p_{0,i})}{p_{1,j}(1 - p_{0,i}) + p_{0,i}(1 - p_{1,j})}. \quad \text{(B2)}$$

Equation (A2) is greater than 50% if $p_{1,j} > p_{0,i}$, and $F(p_{0,i}, p_{1,j}) = 100\%$ only if $p_{1,j} > p_{0,i}$ and $p_{1,j} = 100\%$ or $p_{0,I} = 0\%$, and so for deterministic forecasts, in which the implied probabilities are 0% or 100%, Eq. (A2) reduces to Eq. (1b). The 2AFC score for discrete probabilistic forecasts becomes

$$p_{2\text{AFC}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_j} \frac{p_{1,j}(1 - p_{0,i})}{p_{1,j}(1 - p_{0,i}) + p_{0,i}(1 - p_{1,j})}. \quad \text{(B3)}$$

Note that if $p_{1,j} = p_{0,i}$, the fraction on the right-hand side of Eq. (B3) reduces to

$$p_{2\text{AFC}} = 0.5. \quad \text{(B4)}$$

This result is appropriate because the equal probabilities do not enable the forecaster to decide which of the two cases was an event and is consistent with Eq. (1b), which scores 0.5 if $x_{1,j} = x_{0,i}$.

2) CONTINUOUS PROBABILISTIC FORECASTS

If a probability distribution is provided as the forecast, that is, a density function that is defined over a continuous scale of values, the 2AFC score can be generalized from that for the discrete probabilistic forecasts [Eq. (B3)]. Given forecast distributions, the probability that a value drawn from the one forecast distribution exceeds that drawn from the other can be calculated, and is used in place of Eq. (1b) again. Let $p_{\cdot,i}(x)$ and $P_{\cdot,i}(x)$ represent, respectively, the probability density and the cumulative probability at x. Based on the derivation in Eq. (A3), the 2AFC score is

$$p_{2\text{AFC}} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left[\frac{1 - \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx}{2 - \int_{-\infty}^{\infty} p_{1,j}(x) P_{0,i}(x)\, dx - \int_{-\infty}^{\infty} p_{0,i}(x) P_{1,j}(x)\, dx} \right]. \quad \text{(B5)}$$

Note that if the shape of the density functions is mathematically continuous (i.e., it does not reveal steps or delta peaks), then the 2AFC score of Eq. (B5) can be based on the much simpler Eq. (A6) rather than (A3).

b. Polychotomous observations

If there are m_v possible outcomes, Eq. (B1) can be generalized to

$$p_{2\text{AFC}} = \frac{\sum_{k=1}^{m_v - 1} \sum_{l=k+1}^{m_v} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} F(p_{k,i}, p_{l,j})}{\sum_{k=1}^{m_v - 1} \sum_{l=k+1}^{m_v} n_k n_l}. \quad \text{(B6)}$$

As with Eq. (B1), the exact calculation of Eq. (B6) depends on whether $p_{0,i}$ and $p_{1,j}$ represent discrete probabilities or continuous probability distributions.

1) DISCRETE PROBABILISTIC FORECASTS

For ordinal categories, the score assesses the ability to identify the observation in the higher category, just as for the polychotomous forecast, but makes the selection on the basis of the forecast probabilities. The score is a generalization of Eq. (B3) but uses the cumulative probabilities of the first k categories for the ith forecast given that category l occurred, $P_{l,i}(k)$. The 2AFC score is defined as

$$p_{2\text{AFC}} = \frac{\sum_{k=1}^{m_v - 1} \sum_{l=k+1}^{m_v} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} \frac{[1 - P_{l,j}(k)] P_{k,i}(k)}{[1 - P_{l,j}(k)] P_{k,i}(k) + P_{l,j}(k)[1 - P_{k,i}(k)]}}{\sum_{k=1}^{m_v - 1} \sum_{l=k+1}^{m_v} n_k n_l}. \quad \text{(B7)}$$

If the categories are nominal, Eq. (B6) is simply applied across each of them. The 2AFC score is therefore

$$p_{2\mathrm{AFC}} = \frac{\displaystyle\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l}\frac{p_{l,j}(l)[1-p_{k,i}(l)]}{p_{l,j}(l)[1-p_{k,i}(l)]+[1-p_{l,j}(l)]p_{k,i}(l)}}{\displaystyle\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v} n_k n_l}. \tag{B8}$$

The summation over j is across the forecasts for all the observations not in category k. In contrast to Eq. (B7), Eq. (B8) is based on the probability assigned to the category in question only (category l).

2) CONTINUOUS PROBABILISTIC FORECASTS

For the dichotomous observed categories, the 2AFC score on continuous probabilistic forecasts used the probability that a value drawn from the forecast distribution for the event exceeds that drawn from the one for the nonevent. A similar principle applies in the case of polychotomous observed categories except that the forecast probability distributions are compared to see whether they can be used to identify the observation in the highest category. Equation (B5) is therefore generalized to

$$p_{2\mathrm{AFC}} = \frac{\displaystyle\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l}\left[\frac{1-\displaystyle\int_{-\infty}^{\infty}p_{k,i}(x)P_{l,j}(x)\,dx}{2-\displaystyle\int_{-\infty}^{\infty}p_{k,i}(x)P_{l,j}(x)\,dx-\int_{-\infty}^{\infty}p_{l,j}(x)P_{k,i}(x)\,dx}\right]}{\displaystyle\sum_{k=1}^{m_v-1}\sum_{l=k+1}^{m_v} n_k n_l}. \tag{B9}$$

Again, this expression simplifies significantly if the distributions are continuous functions, that is, if the 2AFC scores can then be based on Eq. (A6) rather than Eq. (A3).

c. Continuous observations

For continuous observations, only continuous probabilistic forecasts are considered. The number of categories, m_v, in Eq. (B6) is equal to the number of cases, n, and so the equation becomes

$$p_{2\mathrm{AFC}} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\frac{1-\displaystyle\int_{-\infty}^{\infty}p_i(x)P_j(x)\,dx}{2-\displaystyle\int_{-\infty}^{\infty}p_i(x)P_j(x)\,dx-\int_{-\infty}^{\infty}p_j(x)P_i(x)\,dx}\right], \tag{B10}$$

APPENDIX C

Is the Probabilistic Version of the 2AFC Proper?

As mentioned in section 3a(3) and elsewhere, the standard definition of propriety (see Bröcker and Smith 2007) cannot be applied to the probabilistic version of the 2AFC in appendix B because the 2AFC is defined in terms of two forecast and observation pairs rather than a single pairing. This makes a theoretical proof of the 2AFC score's impropriety problematic. Nevertheless, from a simple constructed example it can be shown that the 2AFC scores for probabilistic forecasts are not optimized when the forecaster issues probabilities that are consistent with his or her best judgment.

Consider the situation of issuing discrete probabilistic forecasts for an event E to happen. Assume that $2n$ forecast–observation pairs are available for the verification. Further assume that n of these pairs (group A) were sampled under a climatologic regime under which E occurs with a probability $p_A = 0.8$, while the remaining n pairs (group B) were sampled under a regime which allows E to happen only with, say, $p_B = 0.2$.

Finally, assume a forecaster based his or her forecasts on distinguishing between the two climatologic regimes and issued the same probability forecast p_A^* (p_B^*) for all samples of group A (group B). Applying Eq. (B1), the expectation of the *probabilistic version* of the 2AFC score can be calculated:

$$\langle p_{2\mathrm{AFC}} \rangle = \frac{n^2[p_A(1-p_B)F(p_B^*, p_A^*) + p_A(1-p_A)F(p_A^*, p_A^*) + (1-p_A)p_BF(p_A^*, p_B^*) + p_B(1-p_B)F(p_B^*, p_B^*)]}{n^2[p_A(1-p_B) + p_A(1-p_A) + (1-p_A)p_B + p_B(1-p_B)]}$$
$$= 0.64F(p_B^*, p_A^*) + 0.16F(p_A^*, p_A^*) + 0.04F(p_A^*, p_B^*) + 0.16F(p_B^*, p_B^*). \quad \text{(C1)}$$

The first summand on the r.h.s. of Eq. (C1) is the expected contribution of those 2AFC tests that compare forecasts corresponding to events in group A with forecasts corresponding to nonevents in group B. (The second summand is for events in group A with nonevents in group A, third summand for events in group B with nonevents in group A, and fourth summand for events in group B with nonevents in group B.)

Using Eq. (B2) in (C1), and choosing $p_A^* = p_A$ and $p_B^* = p_B$, $\langle p_{2\mathrm{AFC}} \rangle$ becomes 76.5%. However, by choosing $p_A^* = 1$ and $p_B^* = 0$, $\langle p_{2\mathrm{AFC}} \rangle$ becomes 80%. In other words, if the forecaster issues forecasts that are overconfident w.r.t. his or her own true belief, the skill score can be enhanced, implying that the probabilistic version of the $p_{2\mathrm{AFC}}$ as formulated in appendix B is not proper.

REFERENCES

Agresti, A., 1984: *Analysis of Ordinal Categorical Data*. Wiley, 304 pp.

Bowler, N., 2008: Accounting for the effect of observation errors on verification of MOGRPS. *Meteor. Appl.,* **15,** 199–205.

Brier, G. W., and R. A. Allen, 1951: Verification of weather forecasts. *Compendium of Meteorology,* T. F. Malone, Ed., Amer. Meteor. Soc., 841–848.

Bröcker, J., and L. A. Smith, 2007: Scoring probabilistic forecasts: The importance of being proper. *Wea. Forecasting,* **22,** 382–388.

Conover, W. J., 1999: *Practical Nonparametric Statistics*. Wiley, 584 pp.

Doblas-Reyes, F. C., C. A. S. Coelho, and D. B. Stephenson, 2008: How much does simplification of probability forecasts reduce forecast quality? *Meteor. Appl.,* **15,** 155–162.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Glahn, H. R., 2004: Discussion of verification concepts in *Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wea. Forecasting,* **19,** 769–775.

Green, D. M., and J. A. Swets, 1989: *Signal Detection Theory and Psychophysics*. Peninsula Publishing, 521 pp.

Jolliffe, I. T., 2004: P stands for . . . . *Weather,* **59,** 77–79.

——, and D. B. Stephenson, 2003: Introduction. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 1–12.

——, and ——, 2008: Proper scores for probability forecasts can never be equitable. *Mon. Wea. Rev.,* **136,** 1505–1510.

Kharin, V. V., and F. W. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate,* **16,** 1684–1701.

Lindeman, R. H., P. F. Merando, and R. Z. Gold, 1980: *Introduction to Bivariate and Multivariate Statistics*. Scott Foresman, 444 pp.

Livezey, R. E., and M. Timofeyeva, 2008: The first decade of long-lead U.S. seasonal forecasts: Insights from a skill analysis. *Bull. Amer. Meteor. Soc.,* **89,** 843–854.

Mason, I. T., 2003: Binary events. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 37–76.

Mason, S. J., 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **132,** 1891–1895.

——, 2008: Understanding forecast verification statistics. *Meteor. Appl.,* **15,** 31–40.

——, and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.,* **128,** 2145–2166.

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

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

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

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

Nicholls, N., 2001: The insignificance of significance testing. *Bull. Amer. Meteor. Soc.,* **82,** 981–986.

Palmer, T. N., and Coauthors, 2004: Development of a European ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

Sheskin, D. J., 2007: *Handbook of Parametric and Nonparametric Statistical Procedures.* Chapman and Hall/CRC, 1776 pp.

Thomson, M. C., F. J. Doblas-Reyes, S. J. Mason, R. Hagedorn, S. J. Connor, T. Phindela, A. P. Morse, and T. N. Palmer, 2006: Multi-model ensemble seasonal climate forecasts for malaria early warning. *Nature,* **439,** 576–579.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 648 pp.