Conditional Exceedance Probabilities

SIMON J. MASON

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

JACQUELINE S. GALPIN

School of Statistics and Actuarial Science, University of the Witwatersrand, Johannesburg, South Africa

LISA GODDARD

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

NICHOLAS E. GRAHAM

Hydrologic Research Center, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California

BALAKANAPATHY RAJARTNAM

Department of Statistical Sciences, Cornell University, Ithaca, New York

(Manuscript received 31 October 2005, in final form 19 April 2006)

ABSTRACT

Probabilistic forecasts of variables measured on a categorical or ordinal scale, such as precipitation occurrence or temperatures exceeding a threshold, are typically verified by comparing the relative frequency with which the target event occurs given different levels of forecast confidence. The degree to which this conditional (on the forecast probability) relative frequency of an event corresponds with the actual forecast probabilities is known as reliability, or calibration. Forecast reliability for binary variables can be measured using the Murphy decomposition of the (half) Brier score, and can be presented graphically using reliability and attributes diagrams. For forecasts of variables on continuous scales, however, an alternative measure of reliability is required. The binned probability histogram and the reliability component of the continuous ranked probability score have been proposed as appropriate verification procedures in this context, but are subject to some limitations. A procedure is proposed that is applicable in the context of forecast ensembles and is an extension of the binned probability histogram. Individual ensemble members are treated as estimates of quantiles of the forecast distribution, and the conditional probability that the observed precipitation, for example, exceeds the amount forecast [the conditional exceedance probability (CEP)] is calculated. Generalized linear regression is used to estimate these conditional probabilities. A diagram showing the CEPs for ranked ensemble members is suggested as a useful method for indicating reliability when forecasts are on a continuous scale, and various statistical tests are suggested for quantifying the reliability.

1. Introduction

Because of the inability to forecast the atmosphere with absolute certainty, the forecaster's confidence in a specific forecast provides useful additional information beyond a simple indication of what is considered the best estimate (Murphy 1973, 1997; Murphy and Winkler 1987; Wilks 2006). Confidence in the forecast can be expressed in a number of ways, but whichever form is used, any comprehensive forecast verification system must consider more than just the accuracy of the forecasts: the appropriateness of the forecaster's confidence in the forecasts should also be examined (Murphy and Winkler 1987; Murphy and Wilks 1998). [For detailed

Corresponding author address: Dr. Simon J. Mason, IRI, 61 Route 9W, P.O. Box 1000, Palisades, NY 10964-8000.
E-mail: simon@iri.columbia.edu

DOI: 10.1175/MWR3284.1

discussions of the various aspects of forecast quality and how they can be measured, see Murphy (1993, 1997) and Jolliffe and Stephenson (2003).]

The appropriateness of the forecaster's confidence is usually measured by considering the reliability, or calibration, of the forecasts. Reliability is defined as a consistency between the a priori predicted probabilities of an event and the a posteriori observed relative frequencies of this event (Murphy 1973; Toth et al. 2003). The way reliability is measured depends on how the uncertainty in the forecast is indicated. Perhaps the simplest way of indicating forecast uncertainty is to specify a range of values between which the verification is expected to occur with a predefined level of confidence α (Montgomery and Peck 1992; Seber and Lee 2003). For each forecast, the level of confidence α is kept fixed, but the width of the interval is varied to reflect the varying uncertainty of the forecaster: decreased (increased) uncertainty is indicated by a narrowing (widening) of the interval. The reliability of such prediction intervals can be assessed by comparing the capture rate for the respective confidence intervals (for the specified α); a forecaster is judged to be overconfident (or underconfident) if the verification falls too infrequently (or frequently) within the range defined by the prediction intervals. The forecaster's confidence is appropriate when the capture rate (the proportion of times the verification is contained within the prediction interval) corresponds with the confidence level.

Given problems in user interpretation of intervals (see, e.g., Teigen and Jørgensen 2005), an attractive alternative method to the specification of prediction intervals for indicating forecast uncertainty involves fixing the interval and allowing the forecaster's level of confidence to vary instead. Examples include defining the probability that daily precipitation will exceed a trace amount, or that seasonally averaged maximum temperatures will be warmer than 30°C. Forecasts that involve assigning a variable probability to an observed value falling within a predefined range (or taking a specific discrete value) are widely referred to as probabilistic forecasts. Although the measurement of reliability in such forecasts becomes more complicated than in the context of prediction intervals, the principle involved is identical: Does the verification fall within the predefined interval or category more or less frequently than the forecaster anticipates? Because the forecast confidence varies, the observed relative frequency of the verifying event is calculated for forecast probabilities within predefined ranges (e.g., >0.05, 0.05–0.10, . . .). As a result, the dimensionality of the verification problem can become much larger than in the case of verification of confidence intervals (Murphy and Wilks 1998).

Observed relative frequencies conditional on the forecast probability are often plotted as reliability or attributes diagrams, which are simple and effective methods of illustrating reliability graphically (Hsu and Murphy 1986; Atger 2004). Various summary measures of the reliability diagram have been proposed. The Murphy (1973) decomposition of the (half) Brier score provides a useful quantitative measure of forecast reliability; one of the components of the decomposition represents the sum of the squared distances between the empirical reliability curve and the diagonal line of perfect reliability, weighted by the frequency with which forecasts of each probability (or in each probability bin) are issued (Hsu and Murphy 1986; Mason 2004). Alternative measures considered by Murphy and Wilks (1998) involve fitting a weighted regression line to the empirical reliability curve. For good reliability, the slope of the regression line should be close to 1.0 and the intercept close to 0.0 (and should fit the empirical curve well). For forecasts that contain no useful information, the empirical reliability curve is a horizontal line, indicating that the verification is independent of the forecast.

Reliability diagrams have been criticized for their arbitrary categorization of the target variable and binning of the forecast probabilities, and for the large sampling errors that occur given small samples and/or when applied to probabilities of rare events (Atger 2004). Ideally, when forecasts are expressed, and the observations are measured, on a continuous scale, the reliability of the forecasts (and their verification more generally) should be measured without discretization. One option is the continuous ranked probability score, which is the integral of the Brier score for all possible threshold values of the target variable (Gneiting et al. 2005). Like the Brier score, the continuous ranked probability score can be decomposed to yield a reliability component (Hersbach 2000). This reliability component is closely related to the rank histogram (Anderson 1996; Hamill and Colucci 1998; Hamill 2001) in that it measures whether, on average, the cumulative forecast distribution correctly indicates the probability that the observed value is less than each ranked ensemble member (Hersbach 2000). In this paper the basic concepts of the binned histogram and the reliability component of the continuous ranked probability score are extended with the aim of addressing some of the limitations of these verification techniques (Hamill 2001). Specifically, the reliability of an ensemble forecast is examined by considering the probability that the observed value is more than the forecast from each ensemble member. If this

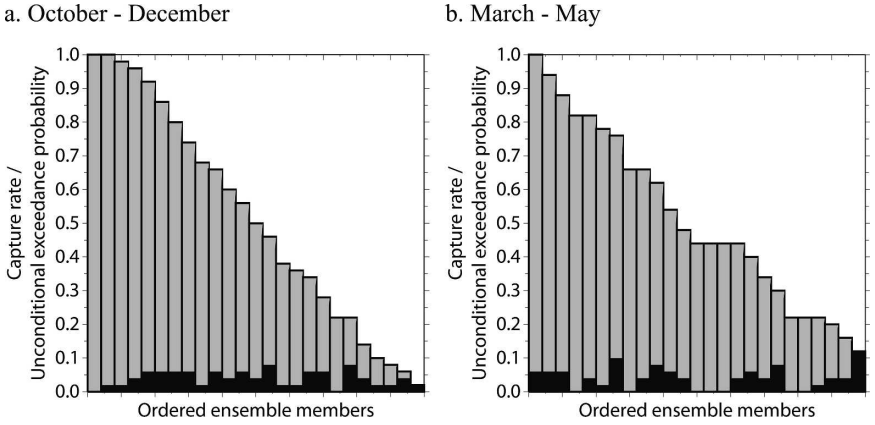

FIG. 1. Capture rates (black) and UEPs (gray) for 24 ordered ensemble member simulations of area-averaged precipitation over eastern Africa (10°N–10°S, 30°–50°E) for (a) October–December and (b) March–May 1951–2000.

probability is conditional upon the actual values forecast, then it is argued that the forecasts from the ensembles are unreliable.

2. Unconditional exceedance probabilities

Given an ensemble of forecasts, errors in the central tendency, spread, and shape of the distribution of the ensemble members constitute possible sources of error in their representation of the probability distribution of possible outcomes. Consistent errors in the ensemble distribution can be estimated using the binned probability ensemble (Anderson 1996; Talagrand et al. 1998), also known as the rank histogram (Hamill and Colucci 1998; Hamill 2001). Given a history of ensemble forecasts, the verification should fall between ordered ensemble-member forecasts an equal number of times. Histograms of these capture rates are a useful means of determining whether there are errors in the ensemble distribution. Since the proportion of observations in each bin should follow a uniform distribution, a Kolmogorov–Smirnov (Wilks 2006) or Cramér–von Mises test (Elmore 2005) could be used to test for systematic errors (Sheskin 2003).

Rank histograms often are U shaped, indicating that the verification falls outside of the ensemble's range too frequently. A U shape can be indicative of an ensemble spread that is consistently too small (Anderson 1996), so that the probabilities that the observation will fall within either of the outer bins is inflated for each forecast. A conditional bias in the forecasts can result in similarly shaped histograms because the probability that the observation will fall within one of the outer bins is inflated at each forecast (Hamill 2001). Because some forms of conditional bias and some forms of unconditional bias can result in U-shaped rank histograms, the histograms can be difficult to interpret. Similarly, although the rank histogram is uniform when the ensemble distribution reliably reproduces the distribution of possible outcomes, a uniform histogram is no guarantee that this distribution is being represented by the ensemble. Hence, if two sets of forecasts generate similar rank histograms, it cannot automatically be concluded that the sets of forecasts are equally good.

To illustrate this problem of forecasts of notably different quality generating similarly shaped rank histograms, examples for two sets of simulations of precipitation are presented in Fig. 1. The figure shows capture rates for a 24-member ensemble of simulations of area-averaged precipitation over eastern Africa (10°N–10°S, 30°–50°E) for October–December (Fig. 1a) and March–May (Fig. 1b) 1951–2000 using the ECHAM4.5 atmospheric general circulation model (Roeckner et al. 1996). The model was forced using observed sea surface temperatures, and forecasts were verified against station-based observed data (Mitchell et al. 2003) averaged over the same area. After the spatial averaging, the observed and simulated precipitation estimates were standardized to correct for mean and variance errors in the simulated precipitation. The examples presented in Fig. 1, and Kolmogorov–Smirnov test statistics for a uniform distribution of the observations in each bin, suggest that the ensembles are reasonably well calibrated for both seasons ($p = 0.438$ for both seasons). The rank histograms therefore provide no clear indication that the simulations for one season are superior to those for the other.

Capture rates can be expressed alternatively as exceedance probabilities, defining the probability that the observed precipitation exceeds the amount forecast by

the kth ranked ensemble member. This probability is termed the unconditional exceedance probability (UEP), and in a well-calibrated model the expected value of the UEP is given by

$$P(X_0 > X_k) = 1 - \frac{k}{m+1}, \quad (1)$$

where X_0 is the observed precipitation, X_k is the forecast precipitation from the kth ensemble member sorted in ascending order, and m is the number of ensemble members. In a model in which the ensemble variance is consistently too small, $P(X_0 > X_1) < 1 - 1/(m+1)$ and $P(X_0 > X_m) > 1 - m/(m+1)$. The UEPs for the eastern African precipitation simulations are shown as the gray bars in Fig. 1. For observations uniformly distributed amongst the bins, the graph of the exceedance probabilities will decrease evenly from the top-left corner of the plot to the bottom-right corner.

Hersbach (2000) derives a measure of reliability from a decomposition of the continuous ranked probability score that is closely related to the UEPs, but explicitly considers the average width of each bin. The measure is based in part on the difference between a measure of the relative frequency with which the observed value is less than the central point of each bin and the cumulative probability of the corresponding bin. As with the reliability score from Murphy's (1973) decomposition of the Brier score, good reliability is indicated by values close to zero. For the ECHAM4.5 simulations, the score for the October–December season (0.012) is similar to that for the March–May season (0.052), thus supporting the results of the Kolmogorov–Smirnov test for uniformity that reliability is good for both seasons.

3. Conditional exceedance probabilities

If there is an error in the central tendency of the ensemble distribution, the exceedance probabilities will be conditional upon the forecast, not just on the ensemble shape and spread. Given only one ensemble member, there should be a 50% probability that the observed precipitation is more than forecast, regardless of how much precipitation was forecast. A rank histogram would indicate only the proportion of observed values exceeding all the forecasts, but in an imperfect forecast system, if the forecast is for anomalously wet (dry) conditions, then the probability that conditions that are wetter than forecast is likely to be less (more) than 50%. More specifically, in a forecast system that contains no useful information, the probability that conditions are wetter than forecast is equal to the climatological probability of precipitation being more than the forecast. In such a case, because the forecasts are varying randomly, the climatological probabilities of exceeding differing rainfall amounts are the only useful information available. Extending the argument to an imperfect forecast ensemble, if the wettest ensemble member indicates relatively wet (dry) conditions the probability that conditions are wetter than forecast is likely to be less than (exceed) $1 - m/(m+1)$. It is therefore possible that a uniform rank histogram indicates only that the model reproduces the observed climatology well, but does not necessarily imply that the model has any predictive skill (Hamill 2001). A measure of conditional forecast bias would resolve this shortcoming.

The conditional exceedance probability (CEP) is defined here as the probability that the observed precipitation exceeds the amount forecast, conditional on the amount forecast. More generally, the CEP is defined as the probability that the observed value exceeds the forecast value, given the forecast. The CEPs can be calculated using generalized linear models with binomial errors and a logit link function (appendix A; McCullagh and Nelder 1989). Alternative link functions, such as the probit and complementary log–log could be used (Mason and Mimmack 2002), but differences in results are likely to be small (McCullagh and Nelder 1989). The logit link was selected because the parameters can be interpreted meaningfully, as discussed below, but further research is required before any definite recommendations can be made about the most appropriate link function to use.[1] Using the logit link, the CEP is defined as

$$P(X_0 > X_k | X_k) = \frac{\exp(\beta_{0,k} + \beta_{1,k} X_k)}{1 + \exp(\beta_{0,k} + \beta_{1,k} X_k)} = \frac{1}{1 + \exp(-\beta_{0,k} - \beta_{1,k} X_k)}, \quad (2)$$

where $\beta_{0,k}$ and $\beta_{1,k}$ are parameters to be estimated, and X_k is the forecast of the kth driest in an m-member ensemble. Since $\beta_{1,k}$ determines the slope (on the logit scale) of the CEP curve while $\beta_{0,k}$ determines the height, in a perfectly reliable model $\beta_{1,k}$ will be equal to 0, while nonzero values of $\beta_{1,k}$ will provide indications of conditional model biases or poor skill.

The CEPs can be usefully plotted for any forecast value within the range for which historical forecasts are available. Examples are shown in Fig. 2 for the en-

[1] When the observed data are normally distributed, a probit link function may be an appropriate choice, since this would correctly reproduce the climatological cumulative distribution when there is no useful information in the forecast. However, the probit link does not conveniently allow for negatively sloping CEP curves.

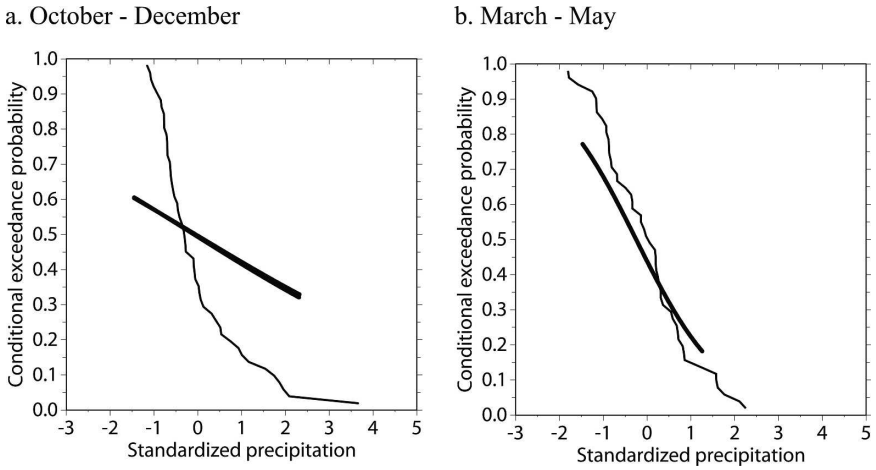

FIG. 2. CEPs for ensemble median simulations of area-averaged (a) October–December and (b) March–May precipitation over eastern Africa (10°N–10°S, 30°–50°E), 1951–2000. The thin line indicates the climatological probabilities of exceedance.

semble medians from the simulations of October–December, and March–May precipitation over East Africa. In a perfectly reliable model, where the exceedance probability is independent of the forecast, the CEPs will form a horizontal line. In most cases, however, where forecast skill is moderate, the CEP will decrease as the forecast precipitation increases, and so the CEP curve is likely to slope downward to the right (see appendix B for further discussion of the slope of the CEP curve). As the CEP curve approaches the line of the observed climatological probability of exceedance curve, the reliability of the forecast decreases. For both seasons, the slope of the CEP curve is negative, but is much more so for the simulations of March–May precipitation, for which it is only marginally flatter than the climatological probability of exceedance. The differences in the slope of the CEP curves for the two seasons are consistent with differences in the correlations of the ranked ensemble members with the observed precipitation (Table 1). The exceptionally high skill in simulating October–December precipitation ($r = 0.723$ for the ensemble median) suggests much smaller errors in the central tendency of the ensemble distribution compared to the simulations for March–May ($r = 0.140$).

The slope of the CEP curve therefore provides useful information about the reliability of the forecasts: if the slope is close to zero, the probability that the observed precipitation exceeds the forecast amount is near constant, and it can be assumed that the exceedance probability for a new forecast value will be equal to this constant. A horizontal CEP curve indicates "complete calibration" since the exceedance probabilities for subsets of the forecasts will be asymptotically equal to this same constant value (Seillier-Moiseiwitsch and Dawid 1993).

The CEPs can be calculated for each ranked ensemble member over their respective ranges of forecast values. As well as being horizontal, the CEP curves should be evenly spaced at values represented by Eq. (1). Examples of CEPs for the 24-member ensemble simulations for eastern African precipitation are shown in Fig. 3. The fact that many of the CEP curves cross each other is a result of sampling errors. There are two sources of sampling errors in estimating the regression parameters: sampling errors arising from an insufficient number of forecasts, and inaccuracies in estimating the quantiles of the ensemble distribution. The first source of error is common to all verification methods, but better estimates of the quantiles could be obtained by increasing the ensemble size or by fitting a distribution to the ensemble members and calculating the quantiles of the fitted distribution (provided that a distribution can be found that estimates the quantiles well). The CEPs are therefore likely to be estimated most accurately given large ensemble sizes, and for those ensemble members close to the median. As in the case of the curves for the ensemble medians (Fig. 2), there is a clear difference between the simulations for October–December (Fig. 3a) and those for March–May (Fig. 3b). For the March–May season, the curves for all the ensemble members follow the climatological exceedance probabilities closely, and suggest that the simulations do not provide reliable indications of the observed variability in seasonal precipitation.

Statistical significance tests for $\beta_{1,k} = 0$ provide a more meaningful indication of the dependence of the exceedance probability on the forecast than visual in-

TABLE 1. Reliability and skill measures for ECHAM4.5 simulated (a) October–December and (b) March–May 1951–2000 area-averaged precipitation over eastern Africa (10°N–10°S, 30°–50°E). The slope parameter β_{1k} of the generalized linear regression model for conditional exceedance probabilities and the p value for $\beta_{1k} = 0$ are given in the last column. The statistics for the ensemble median are shown in the last row.

(a) October–December

Ranked ensemble member k	Correlation with obs precipitation	$\beta_{1,k}$	Probability that the p value for $\beta_{1,k} = 0$
1	0.745	N/A	N/A
2	0.672	−2.029	0.207
3	0.665	−0.830	0.503
4	0.720	−0.640	0.464
5	0.712	−0.355	0.597
6	0.716	−0.291	0.620
7	0.716	−0.038	0.942
8	0.719	−0.042	0.931
9	0.716	−0.102	0.829
10	0.718	−0.461	0.311
11	0.716	−0.399	0.367
12	0.720	−0.313	0.467
13	0.724	−0.168	0.696
14	0.731	0.106	0.811
15	0.747	0.272	0.529
16	0.753	0.540	0.214
17	0.752	0.481	0.271
18	0.749	0.661	0.150
19	0.735	0.658	0.156
20	0.736	0.410	0.457
21	0.715	0.122	0.845
22	0.676	−0.377	0.591
23	0.686	−0.638	0.417
24	0.635	1.459	0.176
Ensemble median	0.723	−0.300	0.486

(b) March–May

Ranked ensemble member k	Correlation with obs precipitation	$\beta_{1,k}$	Probability that the p value for $\beta_{1,k} = 0$
1	0.005	−7.082	0.004
2	0.042	−3.851	0.001
3	0.097	−2.240	0.002
4	0.192	−1.988	<0.001
5	0.216	−2.282	<0.001
6	0.228	−2.857	<0.001
7	0.215	−1.236	0.034
8	0.216	−1.192	0.041
9	0.204	−0.798	0.141
10	0.177	−0.639	0.208
11	0.151	−0.846	0.082
12	0.144	−0.953	0.057
13	0.134	−1.033	0.002
14	0.157	−0.982	0.050
15	0.145	−1.021	0.042
16	0.138	−0.843	0.088
17	0.154	−0.985	0.058
18	0.151	−1.153	0.033
19	0.104	−1.117	0.069
20	0.082	−1.258	0.045
21	0.074	−1.186	0.059
22	0.073	−1.189	0.059
23	0.022	−1.875	0.010
24	0.007	−1.425	0.033
Ensemble median	0.140	−0.997	0.049

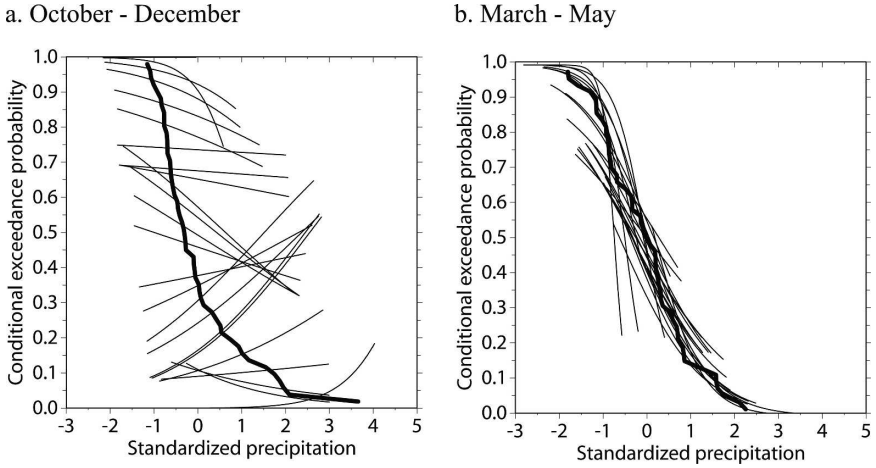

FIG. 3. CEPs for 24 ordered ensemble member simulations of area-averaged (a) October–December and (b) March–May precipitation over eastern Africa (10°N–10°S, 30°–50°E), 1951–2000. The thick line indicates the climatological probabilities of exceedance.

spection of the curves. Approximate significance tests for $\beta_{1,k} = 0$ are based on the reduction in the deviance compared to a model with no $\beta_{1,k}$ term (McCullagh and Nelder 1989). A model with no $\beta_{1,k}$ term assumes that the exceedance probability is not conditioned by the actual forecast, and thus equivalence to the UEP can be upheld. If D_{k0} is the deviance associated with a model calculating the UEP for the kth ranked ensemble member, and D_{k1} is the deviance associated with a model calculating the CEPs, then the reduction in deviance can be calculated over each forecast, j, using

$$D_{k0} - D_{k1} = -2\sum_j \left[y_{kj} \log\left(\frac{u_k}{\mu_{kj}}\right) + (1 - y_{kj}) \log\left(\frac{1 - u_k}{1 - \mu_{kj}}\right)\right], \quad (3)$$

where y_{kj} is equal to 1 if the observation exceeds the forecast and is equal to 0 otherwise, μ_{kj} is the CEP, and u_k is the UEP. This statistic is approximately distributed as χ^2_1. If the forecasts are reliable, the $\beta_{1,k}$ term will not substantially reduce the deviance, and so the value of Eq. (3) will be small. Conversely, large reductions in the deviance indicate a conditioning of the exceedance probability, and hence poor reliability in the forecasts.

The values of $\beta_{1,k}$ and the probabilities that $\beta_{1,k} = 0$ for the individual ranked ensemble members are provided in Table 1. For October–December (Table 1a), the observed rainfall exceeded the forecast from the driest ensemble member for every year, and so the CEP for this member is one for all forecast values. Although the CEP is thus independent of the forecast value, in this case good reliability cannot be claimed because the unconditional exceedance probability should be equal to $1 - 1/(m + 1)$ [96%, Eq. (1)]. The slope of the CEP curve does not provide a complete indication of reliability: it reflects only the conditional bias in the forecasts, but there is an unconditional bias in the example, which could be diagnosed by considering the intercept term $\beta_{0,k}$.

Apart from the driest ensemble member for October–December, the slopes of the CEP curves for this season are reasonably close to zero for all ensemble members (as indicated in the last two columns of Table 1a). However, there is some suggestion that reliability weakens toward the tails of the forecast distribution, which would be a natural result of the increase in sampling errors in the percentiles of the distribution here. The larger sampling errors in the tails are reflected by the weakening of the correlations with the observed precipitation for the highest and lowest ranked ensemble members.

The conditioning of the exceedence probabilities on the simulated precipitation for March–May is strong (Table 1b). The CEPs follow the climatological exceedance probability curve much more closely than for October–December, and are indicative of the much weaker skill in simulating boreal spring rainfall over eastern Africa (Mason and Graham 1999). All but two of the ensemble members have CEPs with slope parameters that differ significantly from zero at a 10% level of significance (Table 1b). The poor reliability indicated by the CEPs is consistent with the weak level of skill as measured by the correlations with the observed precipitation, both for individual ensemble members and for the ensemble median.

The differences in skill of the model in simulating rainfall for the two seasons is a reflection of differences in the potential predictability: the October–December period has a higher potential predictability than March–May (Indeje et al. 2000). Ideally, the model should be able to identify the poor potential predictability of the March–May season, but the model's signal is strong for both seasons, as discussed below. The CEPs for cases of low potential predictability should be evenly spaced, but should have very short horizontal extents since if the ensemble distribution is consistently reproducing the climatological distribution there will be minimal variance in the quantiles of the ensemble distribution. In the limiting case of no signal, for such climatological forecasts, although perfectly reliable (Jolliffe and Stephenson 2003; Wilks 2006), the CEP curves would be representable only by a single point because the variance of the percentiles of the climatological distribution would be zero. Even in a very large ensemble forecast system, however, sampling errors inevitably generate some differences from forecast to forecast, including situations where there is no signal. It is therefore informative to consider the sampling distribution of the CEPs in the context of no signal. Given m ensemble members and no signal in the model, the CEP for the kth ordered member will be the same as for randomly generated order statistics:

$$P(x \geq X_k | X_k) = \sum_{i=k}^{m} \binom{m}{i} [F(x)]^i [1 - F(x)]^{m-i}, \quad (4)$$

where $F(x)$ is the climatological probability of observing (or, more strictly, of forecasting) x or more (Balakrishnan and Cohen 1991; Arnold et al. 1992). Equation (4) represents the right-tail area of the binomial distribution with m trials, and $F(x)$ is probability of success.

Although the joint distribution of the ensemble members could be used to test for a signal in the model, a simpler approach can be used by considering the numbers of ensemble members forecasting above-median conditions [applying Eq. (4) only to the case of the median ensemble member]. If the ensemble members are independent, these numbers should follow a binomial distribution with parameters m and 0.5. The Kolmogorov–Smirnov statistic (Wilks 2006) can be used to test whether the proportions follow this distribution, and thus acts as an alternative to mean interensemble correlations (Dix and Hunt 1995) or analysis of variance (Rowell 1998) for detecting model signals. Based on this test, the ECHAM4.5 ensemble distributions differ significantly from the binomial distribution, indicating that the model has a strong signal for both seasons.

4. Summary

The reliability, or calibration, of probabilistic forecasts is an important component of forecast skill. The most commonly used procedures for indicating reliability require an arbitrary categorization when the target variable is measured on a continuous scale. The rank histogram has been proposed for use when a continuous scale is preferred, but interpretation of the histograms, and hence of the reliability component of the continuous ranked probability score, requires caution (Hamill 2001). It has been argued in this paper that the binned probability histogram is of restricted value in verification because it is based on the assumption that the probability of the observed value falling in each of the bins is constant, irrespective of the forecasts. A test has been proposed to assess, in effect, whether this probability is conditional upon the forecast. The test is based on calculating the probability that the observed value will exceed the value forecast by a ranked ensemble member. If a model generates reliable forecast distributions, the exceedance probability will not depend on the forecast.

The conditional exceedance probabilities can be estimated using generalized linear models with binomial errors. The models can be used to diagnose conditional and unconditional biases in the forecasts. The slope term indicates conditional bias by indicating whether the exceedance probability is conditional upon the forecast, while the intercept term can be used to indicate unconditional bias.

Acknowledgments. This paper was funded in part by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA), Contract NA17RJ1231, with the University of California, San Diego, and Contract NA050AR4311004 with the trustees of Columbia University. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies. The anonymous comments of referees, and helpful discussions with A. G. Barnston, L. Clarke, I. T. Jolliffe, G. V. Kass, B. Rajagopalan, L. A. Smith, R. L. Smith, U. Lall, and D. S. Wilks are gratefully acknowledged.

APPENDIX A

Generalized Linear Models

A CEP curve is designed to compare values of forecasts with the probabilities that the observations exceed these values. Given a set of forecasts from a single ranked ensemble member, together with a set of observations that are represented as 1s if the observation

exceeds the forecast, and 0s otherwise, the aim is to fit a model with the forecasts as the independent variable, and the observations as the dependent variable y. Rather than fitting a standard regression model to the data, it is more appropriate to fit an S-shaped curve that is bounded by 0 and 1, and which can represent the probability that the observed value will exceed the forecast (Wilks 2006). Certain forms of generalized linear regression models (McCullagh and Nelder 1989) are ideally suited to modeling of probabilities. Generalized linear regression involves a model that is linear in its parameters, but introduces a link function to transform the values of the predictand, and allows for model errors that are not normally distributed. The generalized linear model used to represent the CEPs [Eq. (2)] has two parameters, both of which are analogous to the parameters of a linear regression model: $\beta_{0,k}$ is a regression constant that defines the height of the curve for the kth ranked ensemble member, and thus the exceedance probability when the forecast value is zero; $\beta_{1,k}$ defines the slope of the curve, and indicates whether the exceedance probability is conditional upon the forecast.

APPENDIX B

Interpretation of the CEP Curve

Some idealized CEP curves are indicated in Fig. B1. A curve is shown only for the ensemble median. For a completely reliable set of forecasts the CEP curve for the ensemble median should be horizontal and should indicate an exceedance probability of 0.5, as shown in Fig. B1a. In Fig. B1b there is a similarly horizontal curve showing that the exceedance probability does not depend on the forecast, but the shorter length of the curve indicates a model with a weaker signal, and the horizontal displacement of the curve indicates an unconditional bias. Specifically, if the curves are for forecasts of precipitation, the forecast is consistently too dry.

Positively sloping CEP curves (Fig. B1c) occur when the model has positive skill but a signal that is consistently too weak. In this case, when an ensemble member indicates anomalously wet (dry) conditions, wet (dry) conditions are more likely to occur, but their probability will be underestimated, and so the probability of exceeding the forecast rainfall will be increased (decreased). However, if the signal is consistently too strong (but the model still has positive skill), the CEP curve will tend to slope negatively (Fig. B1d): the probability of exceeding the forecast if it is for anomalously wet (dry) conditions will be lower (higher) than indicated.

If the model has negative skill (Fig. B1e), then when an ensemble member indicates anomalously wet (dry) conditions, the opposite is more likely to occur, and so the exceedance probability will be larger (smaller) than indicated by the climatological exceedance probability. The CEP curve will thus be steeper than the climatological exceedance probability curve. When the model has no skill (Fig. B1f) the CEP curve will follow the climatological exceedance probability curve.

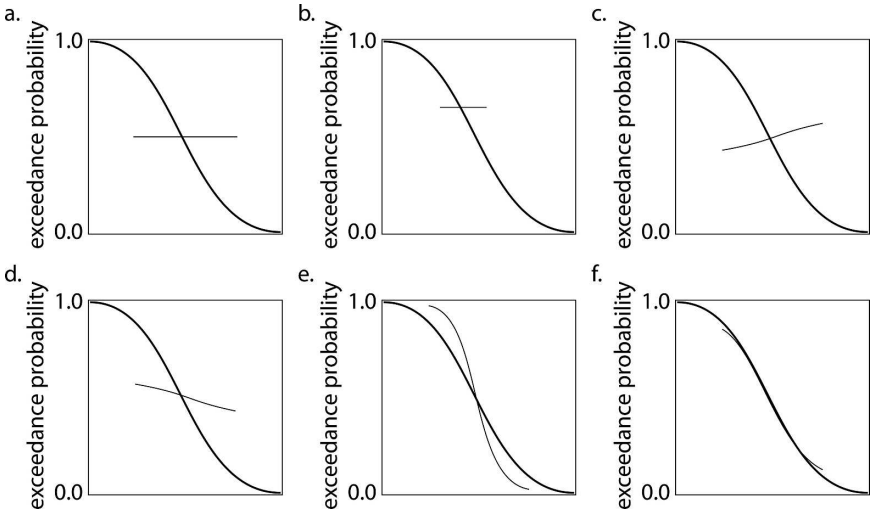

FIG. B1. Idealized CEPs for an ensemble median forecast with (a) perfect reliability and a strong signal; (b) no conditional bias, but an unconditional bias and a weak signal; (c) positive skill, but a signal that is too weak; (d) positive skill, but a signal that is too strong; (e) negative skill; and (f) no skill.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Arnold, B. C., N. Balakrishnan, and H. N. Nagaraja, 1992: *A First Course in Order Statistics*. Wiley, 279 pp.

Atger, F., 2004: Estimation of the reliability of ensemble-based probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.,* **130,** 627–646.

Balakrishnan, N., and A. C. Cohen, 1991: *Order Statistics and Inference*. Academic Press, 377 pp.

Dix, M. R., and B. G. Hunt, 1995: Chaotic influences and the problem of deterministic seasonal predictions. *Int. J. Climatol.,* **15,** 729–752.

Elmore, K. L., 2005: Alternatives to the chi-squared test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting,* **20,** 789–795.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1113.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

——, and S. J. Colucci, 1998: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical frame work for assessing the quality of probability forecasts. *Int. J. Forecasting,* **2,** 285–293.

Indeje, M., F. H. M. Semazzi, and L. Ogallo, 2000: ENSO signals in East African rainfall seasons. *Int. J. Climatol.,* **20,** 19–46.

Jolliffe, I., and D. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 240 pp.

Mason, S. J., 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.,* **137,** 1891–1895.

——, and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics and relative operating levels. *Wea. Forecasting,* **14,** 713–725.

——, and G. M. Mimmack, 2002: A comparison of statistical methods of probabilistic seasonal climate forecasting. *J. Climate,* **15,** 8–29.

McCullagh, P., and J. A. Nelder, 1989: *Generalized Linear Models*. Chapman and Hall, 511 pp.

Mitchell, T. D., T. R. Carter, P. D. Jones, M. Hulme, and M. New, 2003: A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: The observed record (1901–2000) and 16 scenarios (2001–2100). Working Paper 55, Tyndall Centre for Climate Change Research, Norwich, United Kingdom, 30 pp.

Montgomery, D. C., and E. A. Peck, 1992: *Introduction to Linear Regression Analysis.* Wiley, 527 pp.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

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

——, 1997: Forecast verification. *Economic Value of Weather and Climate Forecasts,* R. W. Katz and A. H. Murphy, Eds., Cambridge University Press, 19–74.

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

——, and D. S. Wilks, 1998: A case study for the use of statistical models in forecast verification: Precipitation probability forecasts. *Wea. Forecasting,* **13,** 795–810.

Roeckner, E., and Coauthors, 1996: The atmospheric circulation model ECHAM-4: Model description and simulation of present-day climate. MPI-Rep. 218, Max-Planck-Institut für Meteorologie, Hamburg, Germany, 90 pp.

Rowell, D. P., 1998: Assessing potential seasonal predictability using an ensemble of multidecadal GCM simulations. *J. Climate,* **11,** 109–120.

Seber, G. A. F., and A. J. Lee, 2003: *Linear Regression Analysis*. Wiley-Interscience, 582 pp.

Seillier-Moiseiwitsch, F., and A. P. Dawid, 1993: On testing the validity of sequential probability forecasts. *J. Amer. Stat. Assoc.,* **88,** 355–359.

Sheskin, D. J., 2003: *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall, 1232 pp.

Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, ECMWF, 17–28.

Teigen, K. H., and M. Jørgensen, 2005: When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Appl. Cognit. Psychol.,* **19,** 455–475.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. Jolliffe and D. Stephenson, Eds., Wiley, 137–163.

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