NOTES AND CORRESPONDENCE

On the ROC Score of Probability Forecasts

VIATCHESLAV V. KHARIN AND FRANCIS W. ZWIERS

Canadian Centre for Climate Modelling and Analysis, Meteorological Service of Canada, Victoria, British Columbia, Canada

22 October 2002 and 19 May 2003

ABSTRACT

The relative operating characteristic (ROC) is a measure of the quality of probability forecasts that relates the hit rate to the corresponding false-alarm rate. This paper examines some aspects of the ROC curve for probability forecasts of three equiprobable categories (below normal, near normal, and above normal) in the framework of a simple analog of a climate forecasting system. The insensitivity of the ROC score to some types of forecast biases is discussed and the link to deterministic potential predictability is established in the context of the simple forecasting system. The findings are illustrated with a collection of 24-member ensemble hindcasts of seasonal mean 700-hPa temperature produced with the second-generation general circulation model of the Canadian Centre for Climate Modelling and Analysis.

1. Introduction

The relative (or receiver) operating characteristic (ROC) is a representation of the skill of a forecasting system in which the hit rate and the false-alarm rate are compared (Swets 1973; Mason 1982). The related ROC score is often used to evaluate the quality of probability forecasts (Stanski et al. 1989; Buizza and Palmer 1998; Mason and Graham 1999).

The ROC has its origins in the signal detection theory [e.g., see the review by Swets (1973)]. Typically, the ROC is described in terms of the parameters of two hypothetical, often Gaussian, probability distributions (e.g., Mason 1982; Harvey et al. 1992). One distribution represents the evidence strength associated with the occurrence of the event, and the other represents the nonevent distribution. The two distributions are used to obtain the hit rate and the false-alarm rate as a function of the decision criterion. In the present study, the ROC is derived in the framework of a simple analog of a climate forecasting system (Kharin and Zwiers 2003, hereinafter KZ2003) and the link between the ROC score and deterministic potential predictability (Zwiers 1996; Rowell 1998) is established.

The ROC score is relatively independent of forecast calibration, that is, the correspondence between the forecast probability and observed relative frequency. It is qualitatively similar to resolution, that is, the ability of the forecast system to discriminate between event occurrences and nonoccurences [see, e.g., Mason (1982) and, more recently, Wilks (2001) and Mullen and Buizza (2001)]. The insensitivity to some types of forecast biases is demonstrated here in the context of a simple analog of a climate forecasting system. In particular, it is shown that under certain conditions the ROC score is a function of the properties of the observed system only.

The outline of this note is as follows. The ROC curve and score are derived and discussed in section 2. Some findings are illustrated with a collection of 24-member ensemble hindcasts of seasonal mean 700-hPa temperature in section 3, followed by a summary in section 4.

2. Relative operating characteristic

In this section, a simple analog of a climate forecasting system is introduced and expressions for the hit rate and false-alarm rate, the essential ingredients in the ROC curve definition, are derived. Examples of ROC curves are given for a ''perfect'' forecasting system in a Gaussian setting, and the relationship between the ROC score and potential predictability is illustrated.

a. Climate forecasting system analog

The simple climate forecasting system considered here has the form

$$X = \beta + \epsilon \quad \text{and} \quad F = \beta' + \epsilon', \tag{1}$$

Corresponding author address: Viatcheslav V. Kharin, Canadian Centre for Climate Modelling and Analysis, P.O. Box 1700, STN CSC, Victoria, BC V8W 2Y2, Canada.
E-mail: slava.kharin@ec.gc.ca

TABLE 1. The two-by-two contingency table for verification of a dichotomous forecasting system.

	Forecasts		
Obs	Warning	No warning	Total
Event	H	M	O
Nonevent	FA	CR	O'
Total	W	W'	N

where X is a scalar predictand and F is its forecast. The predictand X is assumed to be a sum of two independent components: a potentially predictable signal β and an unpredictable component that is represented by stochastic Gaussian noise ϵ. On seasonal and longer time scales, the primary predictability source is thought to be the atmospheric response to slowly varying lower boundary conditions such as sea surface temperature. The effect of the initial conditions is felt primarily in the first few forecast weeks. The noise term represents the unpredictable effects of day-to-day weather. The properties of the Gaussian noise are fully characterized by a single parameter σ_ϵ^2, the noise variance. Similarly, the forecast F is assumed to be a sum of the model-simulated signal β' and Gaussian noise ϵ'. In this simplified world, a perfect forecasting system has $\beta' = \beta$ and $\sigma_{\epsilon'}^2 = \sigma_\epsilon^2$.

b. The hit rate and false-alarm rate

We begin by considering a deterministic forecasting system in which a warning is issued when an event is predicted to occur. The operation of such a forecasting system can be summarized in a 2×2 contingency table (Table 1). Assume that there are a total of N forecasts and verifying observations. Let the total number of warnings issued be W and the number of nonwarnings be $W' = N - W$. Similarly, let O be the number of events that occurred, and let $O' = N - O$ be the number of nonevents. Also let H be the number of hits, for which an event occurred and a warning was issued; let FA be the number of false alarms, for which a warning was issued but an event did not occur; let M be the number of misses, for which an event occurred but a warning was not issued; and let CR be the number of correct rejections, for which an event did not occur and a warning was not issued.

The hit rate (HR) and the false-alarm rate (FAR) are defined as (e.g., Mason 1982)

$$\text{HR} = \frac{H}{O} \quad \text{and} \quad \text{FAR} = \frac{\text{FA}}{O'}. \tag{2}$$

Let the random variable E denote a dichotomous predictand, defined as $E = 1$ when the event occurs and $E = 0$ otherwise. Asymptotically, for $N \to \infty$, the ratios in (2) converge to conditional probabilities

$$\text{HR} = \Pr\{\text{warning} \,|\, E = 1\} \quad \text{and}$$
$$\text{FAR} = \Pr\{\text{warning} \,|\, E = 0\}. \tag{3}$$

The HRs and FARs are equal in a forecasting system with no skill. When the forecasting system has some skill, the hit rate exceeds the false-alarm rate. For perfect forecasts, HR = 1 and FAR = 0. Both HR and FAR are equal to 1 in a forecasting system that always issues warnings, and both are equal to 0 in a system that never issues a warning.

For probabilistic forecasts, a warning may be issued when the forecast probability P exceeds some critical threshold P_{cr}. Thus, the HR and FAR can be expressed as a function of P_{cr} as follows:

$$\text{HR}(P_{\text{cr}}) = \Pr\{\Omega_P \,|\, E = 1\} = \int_{\Omega_P} f(P \,|\, E = 1)\, dP$$

and

$$\text{FAR}(P_{\text{cr}}) = \Pr\{\Omega_P \,|\, E = 0\} = \int_{\Omega_P} f(P \,|\, E = 0)\, dP, \tag{4}$$

where Ω_P denotes forecast probabilities $P > P_{\text{cr}}$ and $f(P \mid E)$ is the conditional probability density function of probability forecasts. By using the calibration–refinement and likelihood–base rate factorizations of the join distribution of forecasts and predictands (Murphy and Winkler 1987), $f(P \mid E)$ in (4) can be expressed as

$$f(P \,|\, E) = \frac{f(E \,|\, P) f_P(P)}{f_E(E)}, \tag{5}$$

where $f(E \mid P)$ is the expectation of the predictand E conditional on the forecast probability P and $f_P(P)$ and $f_E(E)$ are the marginal probability density functions of probability forecasts and predictands, respectively. By substituting (5) into (4), the HR and FAR can be written as

$$\text{HR}(P_{\text{cr}}) = \frac{1}{\Pr(E = 1)} \int_{\Omega_P} f(E = 1 \,|\, P) f_P(P)\, dP \quad \text{and}$$

$$\text{FAR}(P_{\text{cr}}) = \frac{1}{\Pr(E = 0)} \int_{\Omega_P} f(E = 0 \,|\, P) f_P(P)\, dP. \tag{6}$$

If sampling variability can be neglected, for example, when the ensemble size of independent forecasts produced with the system (1) is sufficiently large, the forecast probability P may be assumed to be a function of the forecast signal β' only. In this case, the condition $P \geq P_{\text{cr}}$ for issuing a warning can be expressed equivalently as a condition on β'. Given the probability density function $f_{\beta'}$ of β', expressions in (6) may be rewritten as

$$\text{HR}(P_{\text{cr}}) = \frac{1}{\Pr(E = 1)} \int_{\Omega_{\beta'}} f(E = 1 \,|\, \beta') f_{\beta'}(\beta')\, d\beta'$$

and

$$\text{FAR}(P_{\text{cr}}) = \frac{1}{\Pr(E = 0)} \int_{\Omega_{\beta'}} f(E = 0 \,|\, \beta') f_{\beta'}(\beta')\, d\beta', \tag{7}$$

where $\Omega_{\beta'}$ denotes all values of β' for which $P > P_{\text{cr}}$.

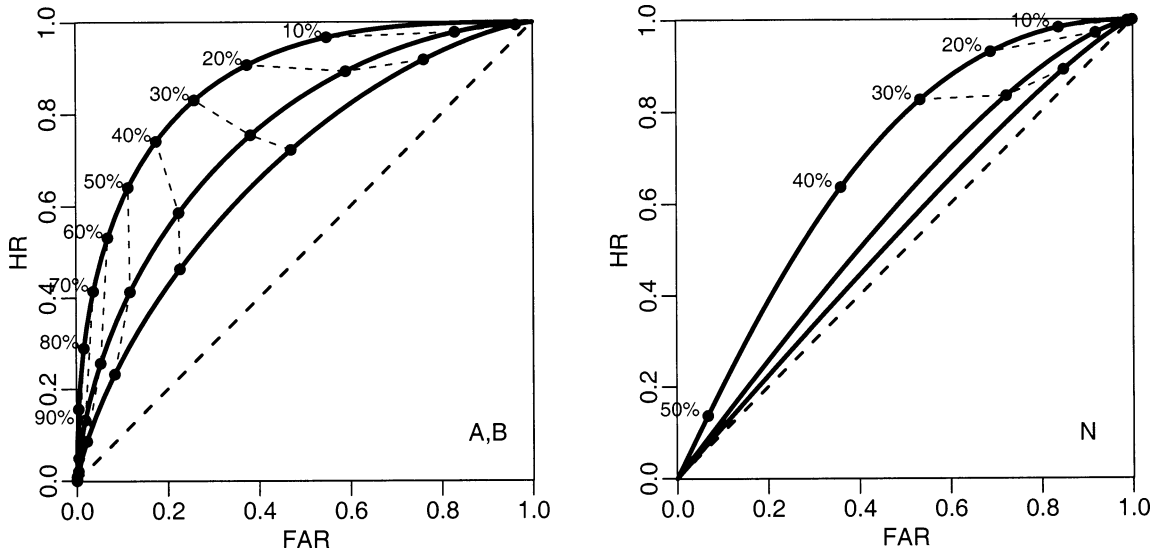

FIG. 1. ROC curves for probability forecasts of the (left) below- or above-normal category and (right) near-normal category from a perfect forecasting system (1) for $\rho^2_{\text{pot}} = 0.15$ (the closest curve to the diagonal), 0.30, and 0.60 (the farthest curve from the diagonal) in the Gaussian setting. The probability thresholds are indicated at the dots on the curves.

The conditional expectation $f(E \mid \beta')$ depends explicitly on the forecast signal β' and the properties of the observed system. Therefore, HR and FAR generally depend on the properties of both the observed and forecasting systems.

The expressions under the integrals in (7) become functions of the observed system only when there is a one-to-one correspondence between β' and β. In this case, HR and FAR take the form

$$\text{HR}(P_{\text{cr}}) = \frac{1}{\Pr(E=1)} \int_{\Omega_\beta} f_\beta(\beta) f(E=1 \mid \beta)\, d\beta \quad \text{and}$$

$$\text{FAR}(P_{\text{cr}}) = \frac{1}{\Pr(E=0)} \int_{\Omega_\beta} f_\beta(\beta) f(E=0 \mid \beta)\, d\beta, \qquad (8)$$

where Ω_β denotes all values of β for which $P > P_{\text{cr}}$. In this case, the HR and FAR depend on the forecast probability P only implicitly through the domain of integration Ω_β.

c. The ROC curve

A new contingency table and the corresponding HR and FAR can be determined for every probability threshold P_{cr}. A ROC curve is obtained by varying P_{cr} and plotting the resulting HRs versus the FARs. The ROC curve for no-skill forecasts coincides with the 45° line from the origin, and that for perfect forecasts connects the points (0, 0), (0, 1), and (1, 1). For deterministic forecasts, a ROC curve can be constructed by plotting the HRs and FARs for the deterministic forecast system and connecting it to the HRs and FARs obtained for a forecast system issuing perpetual warnings (1, 1) and a perpetual no-warnings forecast system (0, 0) (e.g., Mason and Graham 1999). Examples of ROC curves are displayed in Fig. 1. These curves are constructed for probability forecasts of three equiprobable categories (below normal "B," near normal "N," and above normal "A") produced with a perfect forecasting system ($\beta' = \beta$ and $\sigma^2_{\epsilon'} = \sigma^2_\epsilon$) in a Gaussian setting. The HRs and FARs in (8) were estimated by using an algebraic manipulator ("Maple V;" Char et al. 1991). The ROC curves for the A and B categories (left panel of Fig. 1) are identical. The three curves on each diagram correspond to three values of the potential predictability $\rho^2_{\text{pot}} \equiv \sigma^2_\beta / \sigma^2_X$ (0.15, 0.3 and 0.6), the ratio of the variance of the potentially predictable signal σ^2_β to the total observed variance σ^2_X (Zwiers 1996; Rowell 1998). The former two values are typical for seasonal atmospheric variations in northern midlatitudes, and the latter value is typical in the Tropics. The greater the potential predictability ρ^2_{pot} is, the more closely the ROC curves approach the upper-left corner of the diagrams. The numbers near the dots on the curves indicate threshold values P_{cr} for issuing a warning. The ROC curves for the N category (right panel of Fig. 1) are noticeably closer to the "no skill" diagonal than those for the A or B categories for the same potential predictability level. This result is a consequence of the fact that the probability of the N category is relatively insensitive to the amplitude of the potentially predictable signal, resulting in low resolution—that is, the forecasting system is not able to discriminate strongly between event occurrences and nonoccurences of the N category (KZ2003). Note that the near-normal probabilities are bounded from above by a value smaller than 100%. The maximum probability is achieved for $\beta = 0$. The lower the potential predictability is, the closer the maximum possible

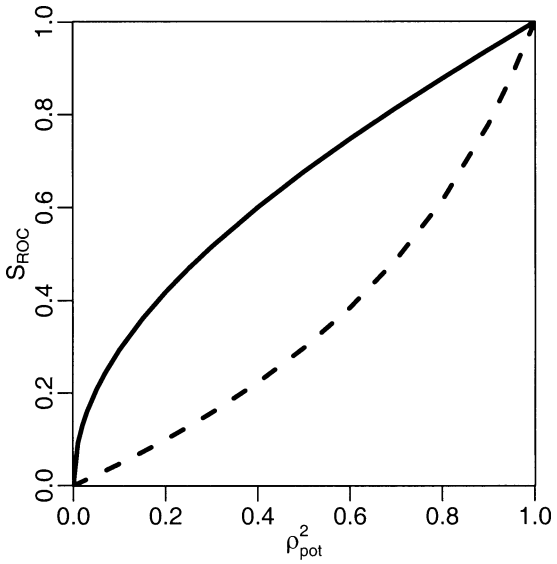

FIG. 2. The ROC skill score S_{ROC} of the probability forecasts from a perfect forecasting system (1) as a function of potential predictability ρ^2_{pot} in the Gaussian setting. The solid curve is for the B and A categories, and the dashed curve is for the N category.

probability of the N category is to the climatological probability.

d. The ROC score

The ROC score A_{ROC} is defined as the area under the ROC curve. No-skill forecasts have $A_{ROC} = 0.5$. The perfect forecast has $A_{ROC} = 1$. The ROC skill score is therefore often defined as

$$S_{ROC} = 2A_{ROC} - 1. \tag{9}$$

This skill score is equal to 1 for the perfect forecast and 0 for a no-skill forecast.

Figure 2 displays the ROC skill score of three-category probability forecasts produced with the perfect forecasting system in the Gaussian setting as a function of the potential predictability ρ^2_{pot}. The skill score for the B and A categories (solid line curve) increases rapidly for small values of the potential predictability. In contrast, the ROC skill score for the N category (dashed line curve) is less sensitive to changes in ρ^2_{pot} when the potential predictability is small. Thus, the same level of potential predictability may imply different ROC scores depending on the ''event'' definition.

Expressions for the HR and FAR [(7), (8)] demonstrate clearly that the ROC score is insensitive to some types of forecast biases, a fact that is already well known to the meteorological community, as mentioned in the introduction. In particular, any two forecasts that are related as $P_2(\beta) = P_1(a\beta)$, $a \neq 0$, have identical ROC curves and equal ROC scores. Rescaling the signal to noise ratio in the forecast system (1) does not change the ROC score. Thus, the ROC score should be regarded as a measure of potential rather than actual skill.

KZ2003 discuss a statistical technique for improving biased probability forecasts. The technique assumes a biased system of the form (1) and adjusts the amplitude of the forecast signal and noise to minimize the corresponding Brier score (i.e., the mean-square error of probability forecasts). It is clear from the above considerations that the ROC score cannot be improved by such a technique. On the contrary, the cross-validated ROC scores of statistically improved forecasts are likely to be degraded by sampling errors in the parameters of the statistical improvement technique.

3. ROC scores of HFP hindcasts

In this section we describe the ROC scores of probability hindcasts of 700-hPa temperature (T_{700}) derived from a collection of 24-member ensemble hindcasts. The hindcasts were produced for 26 northern winters [December–January–February (DJF)] for the 1969–95 period with the second-generation general circulation model of the Canadian Centre for Climate Modelling and Analysis (McFarlane et al. 1992). These integrations were performed as the part of the Canadian Historical Forecast Project (HFP; Derome et al. 2001). Each HFP ensemble member is initialized from reanalyzed fields (Kalnay et al. 1996) lagged at 6-h intervals prior to the forecast season. The monthly mean sea surface temperature anomalies observed in the month prior to the forecast period are ''persisted'' throughout the forecast season. These anomalies are obtained from the Global Sea Ice and Sea Surface Temperature (GISST) dataset (version 2.2; Rayner et al. 1996). Sea ice extent is specified from climatological data. The initial snow line in the Northern Hemisphere is specified from National Centers for Environmental Prediction satellite observations for the week before the forecast period. The soil conditions are initialized from climatological data.

KZ2003 described a number of probability hindcasts of three equiprobable categories of below-normal, near-normal, and above-normal DJF T_{700} in the Tropics (30°S–30°N) and the North American sector (20°–80°N, 150°–45°W). Two of those hindcasts are considered here, the unadjusted Gaussian hindcasts and statistically improved hindcasts. The unadjusted Gaussian probability hindcasts $\hat{P}_{\mathcal{G}}$ are obtained from

$$\hat{P}_{\mathcal{G}}(\mathrm{B}) = \mathcal{F}_{\mathcal{N}}\left(\frac{-\hat{\beta}'_t - x_a}{\hat{\sigma}_{\epsilon'}}\right),$$

$$\hat{P}_{\mathcal{G}}(\mathrm{A}) = \mathcal{F}_{\mathcal{N}}\left(\frac{\hat{\beta}'_t - x_a}{\hat{\sigma}_{\epsilon'}}\right), \quad \text{and}$$

$$\hat{P}_{\mathcal{G}}(\mathrm{N}) = 1 - \hat{P}_{\mathcal{G}}(\mathrm{B}_t) - \hat{P}_{\mathcal{G}}(\mathrm{A}_t), \tag{10}$$

where $\mathcal{F}_{\mathcal{N}}$ is the distribution function of the standard normal distribution, $x_a = \hat{\sigma}_X \mathcal{F}^{-1}_{\mathcal{N}}(2/3) \approx 0.43\hat{\sigma}_X$ is the boundary between the N and A categories, and $\hat{\beta}'_t$ and $\hat{\sigma}_{\epsilon'}$ are estimated from N-member ensembles of seasonal

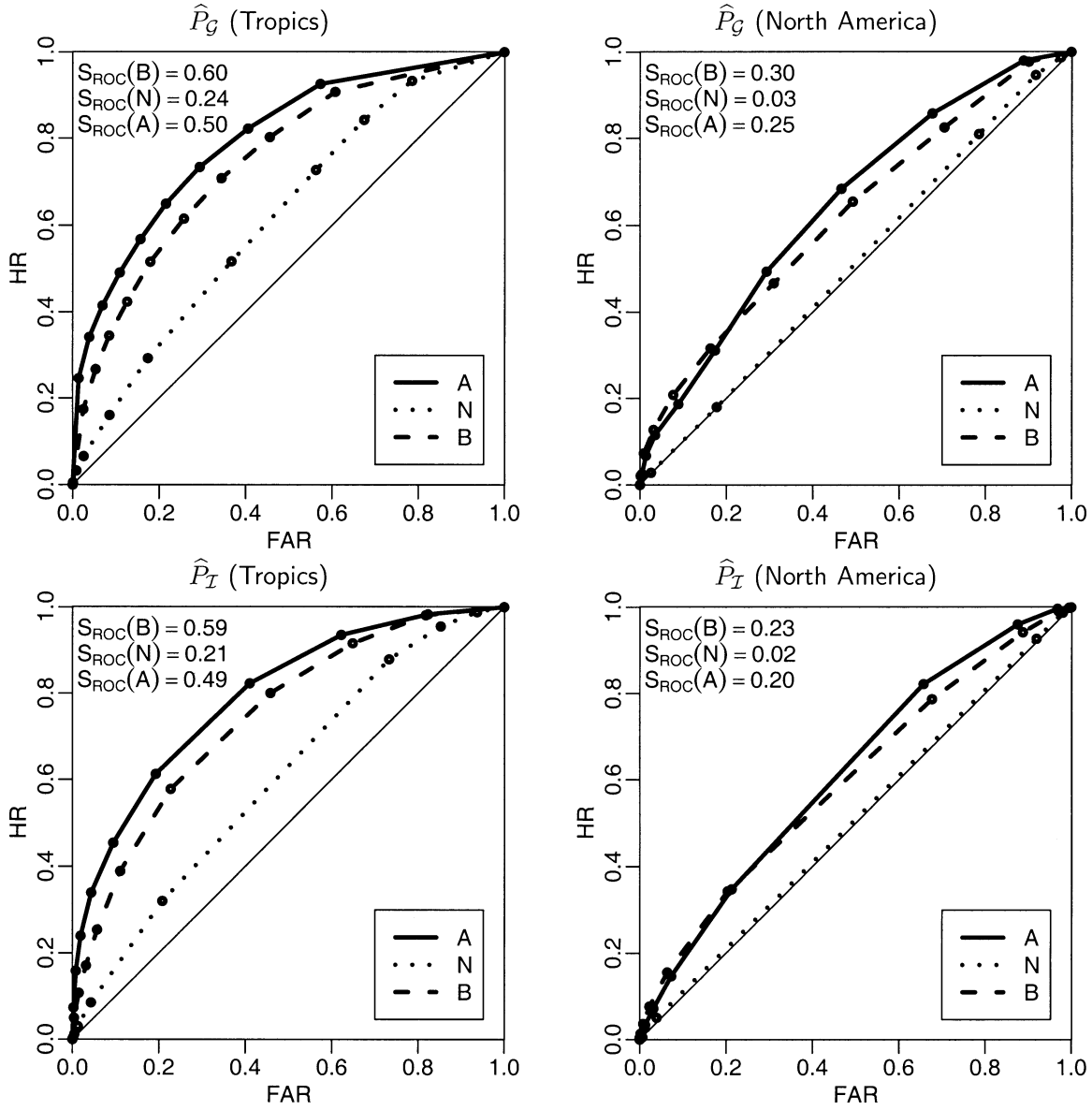

FIG. 3. The ROC curves for the probability hindcasts of the A, N, and B categories derived from the 24-member HFP ensemble DJF T_{700} hindcasts. The HR and FAR are averaged over (left) the Tropics and (right) the North American sector. (top) The unadjusted Gaussian hindcasts $\hat{P}_{\mathcal{G}}$; (bottom) the statistically improved hindcasts $\hat{P}_{\mathcal{I}}$.

anomaly hindcasts for T years $\{F_{nt}, n = 1, \ldots, N; t = 1, \ldots, T\}$ as

$$\hat{\beta}'_t = \frac{1}{N}\sum_{n=1}^{N} F_{nt} \quad \text{and}$$

$$\hat{\sigma}^2_{\epsilon'} = \frac{1}{(N-1)T}\sum_{t=1}^{T}\sum_{n=1}^{N}(F_{nt} - \hat{\beta}'_t)^2. \qquad (11)$$

KZ2003 demonstrated that the probability forecast obtained with (10)–(11) is superior to the standard nonparameteric probability forecast obtained as the relative number of the ensemble members in an event category for T_{700} on seasonal time scales.

The improved hindcasts $\hat{P}_{\mathcal{I}}$ are obtained by adjusting $\hat{\beta}'_t$ and $\hat{\sigma}_{\epsilon'}$ in (10) by factors $\hat{a}$ and $\hat{b}$, respectively, to minimize the sum of the Brier scores for each category. The adjusting factors are chosen to satisfy the relationship $\hat{a}^2\hat{\sigma}^2_{\beta'} + \hat{b}^2\hat{\sigma}^2_{\epsilon'} = \hat{\sigma}^2_{\bar{X}}$, that is, to match the observed variability estimate, and are obtained in a cross-validation procedure in which the data for the forecast year are excluded from the training dataset (see KZ2003 for details). KZ2003 showed that the unadjusted T_{700} probability hindcasts, which are not very reliable (i.e., they are conditionally biased) are substantially improved by this statistical technique.

Figure 3 shows the ROC curves of the unadjusted

Gaussian and improved probability hindcasts in the Tropics and in the North American sector. The ROC skill score for each category is indicated in the upper-left corner of each diagram. The ROC curves for the probability hindcasts in the Tropics deviate farther away from the no-skill diagonal toward the upper-left corner than those for the North American sector, indicating better skill in the Tropics. The hindcasts of the A category are somewhat more skillful than those of the B category, and the hindcasts of the N category are substantially less skillful than those of the other two categories. These results are consistent with the findings in KZ2003.

The statistically improved hindcasts have slightly smaller ROC skill scores than unadjusted hindcasts, a result that is not unexpected given the discussion of the ROC score properties in section 2 in which it was argued that the ROC score cannot be improved by adjusting the amplitude of the predictable signal. The degradation of the ROC skill scores of the ''improved'' hindcasts is apparently the result of sampling errors introduced by the statistical technique.

4. Summary

The properties of the ROC curve and ROC score have been discussed in the context of a simple analog of a seasonal forecasting system in which the total variability may be decomposed into two components, one that is associated with the potentially predictable signal and another that is unpredictable. The relationship between the ROC skill score and the potential predictability, defined as the ratio of the variance of the potentially predictable signal variance to the total variance, is established in the Gaussian setting. It is shown that there is a one-to-one correspondence between the deterministic potential predictability and the ROC scores of optimal probability forecasts. This correspondence depends on the definition of the forecast event of interest. In particular, the same level of potential predictability results in a smaller ROC score for the N category than that for the A or B category.

The insensitivity of the ROC score to certain types of forecast biases is demonstrated in the framework of the simple forecasting system. In an idealized setting in which the model-simulated signal is a function of the observed potentially predictable signal only, the ROC score is shown to be a function of the properties of the observed system only. If sampling variability can be neglected, probability forecasts derived from a forecast system that systematically under- or oversimulates the amplitude of the potentially predictable signal have the same ROC score as that of the perfectly reliable probability forecasts. That is, the ROC score does not penalize forecasting biases. Thus, care is needed when interpreting and intercomparing the performance of model forecasts based on the ROC score.

The findings are illustrated with a collection of 24-member ensemble seasonal hindcasts of 700-hPa temperature. It was shown that although statistically improved temperature hindcasts are more reliable (less biased) than the unadjusted hindcasts, the ROC scores of the statistically improved hindcasts are slightly degraded as compared with those of unadjusted hindcasts.

REFERENCES

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.,* **126,** 2503–2518.

Char, B. W., K. O. Geddes, G. H. Gonnet, B. L. Leong, M. B. Monagan, and S. M. Watt, 1991: *Maple V Library Reference Manual*. Springer, 698 pp.

Derome, J., G. Brunet, A. Plante, N. Gagnon, G. J. Boer, F. W. Zwiers, S. Lambert, and H. Ritchie, 2001: Seasonal predictions based on two dynamical models. *Atmos.–Ocean,* **39,** 485–501.

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

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

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

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

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

McFarlane, N. A., G. J. Boer, J.-P. Blanchet, and M. Lazare, 1992: The Canadian Climate Centre second-generation general circulation model and its equilibrium climate. *J. Climate,* **5,** 1013–1044.

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF ensemble prediction system. *Mon. Wea. Rev.,* **129,** 638–663.

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

Rayner, N. A., E. B. Horton, D. E. Parker, C. K. Folland, and R. B. Hackett, 1996: Version 2.2 of the Global Sea-Ice and Sea Surface Temperature Data Set, 1903–1994. Hadley Centre Climate Research Tech. Note CRTN 74, 21 pp.

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

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. WMO World Weather Watch Tech. Rep. 8, WMO TD 358, 114 pp.

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

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.,* **8,** 209–219.

Zwiers, F. W., 1996: Interannual variability and predictability in an ensemble of AMIP climate simulations conducted with the CCC GCM2. *Climate Dyn.,* **12,** 825–848.