Verification Techniques and Simple Theoretical Forecast Models

ROBERT FAWCETT

National Climate Centre, Australian Bureau of Meteorology, Melbourne, Victoria, Australia

(Manuscript received 9 November 2007, in final form 23 May 2008)

ABSTRACT

This paper investigates the performance of some skill measures [e.g., linear error in probability space, (LEPS), relative operating characteristics score (ROCS), Brier scores, and proportion correct rates], commonly used in the validation and verification of seasonal climate forecasts, within the context of some simple theoretical forecast models. The models considered include linear regression and linear discriminant analysis types, where the forecasts are presented in the form of above/below median probabilities and tercile probabilities. Above and below the median categories are also explored within the context of stratified climatology models, while tail categories are explored within the context of the linear regression type. The skill scores for the models are calculated in each case as functions of a parameter that expresses the strength of the relationship between the predictor and predictand. The skill scores investigated are found to exhibit different dependencies on the model parameter, implying that a given skill score value (0.1 say) can imply a range of strengths in the relationship between predictor and predictand, depending on which skill score is being considered. On the other hand, interrelationships between pairs of skill scores are found to be similar across the different types of models, provided model reliability is preserved. The two-category and three-category LEPS skill scores are found to be on approximately the same scale for the linear regression–type model, thereby enabling a direct comparison.

1. Introduction

In this paper, some standard forecast verification techniques commonly used in the practice of seasonal climate forecasting are explored within the context of some simple theoretical forecast models, with a view toward calculating their skill levels analytically. This approach permits the various verification techniques to be compared and contrasted within a controlled environment, without having to resort to very large Monte Carlo simulations. Skill scores are found to exhibit a range of dependencies on the underlying model parameters, implying that a knowledge of these differences is important when assessing the results obtained from these skill scores in real forecasting examples. The existence of analytic results is also helpful, because they permit checks on the results of Monte Carlo simulations.

The first theoretical model to be considered is a simple linear model (section 2), familiar within the context of standard linear regression theory. This model supposes a linear relationship between the predictor and predictand, together with a noise component. All three variables (predictor, predictand, and noise) are normally distributed. The model could be applied within the context of a slightly idealized forecasting of seasonal temperature anomalies (which are approximately normally distributed) from a single linear predictor such as the Southern Oscillation index (SOI) or a Niño-region sea surface temperature (SST) index. Two-category (section 2a) and three-category (section 2b) forecast probabilities are derived for the model. The model is by construction perfectly reliable (Wilks 2006; Hartmann et al. 2002), so measures of forecast assessment that are designed to detect departures from reliability are irrelevant to this particular example. A variant example exhibiting a pronounced departure from reliability is also discussed (section 2c). Forecasting the tails of the distribution, something (arguably) of more relevance to weather forecasting, is explored in section 2d.

The second theoretical model (section 3) arises within the context of linear discriminant analysis (LDA), a statistical technique embedded within the Australian Bureau of Meteorology's operational sea-

Corresponding author address: Robert Fawcett, National Climate Centre, Australian Bureau of Meteorology, GPO Box 1289, Melbourne, VIC 3001, Australia.
E-mail: r.fawcett@bom.gov.au

DOI: 10.1175/2008WAF2007091.1

sonal forecasting system (Jones 1998; Drosdowsky and Chambers 2001). Two-category (section 3a) and three-category (section 3b) versions are considered, with the predictand being treated as being essentially discrete or categorical.

Finally, a simple stratified climatology model is presented (section 4), illustrating yet another statistical technique used in Australian seasonal (e.g., Stone et al. 1996) and weather (e.g., Stern 1980; Dahni and Stern 1995) forecasting. This has the opposite property of the predictand being continuous but the predictor discrete or categorical.

Comparisons are made between the results of the three different model types, and some interesting similarities are found (section 5), along with an application of the results obtained in section 2 to some actual forecast verification data. Concluding remarks are given in section 6, while some of the longer calculations are relegated to appendixes A and B.

The forecasts for the various model types are presented for the most part in the form of probabilities above (below) the median outcomes and tercile probabilities, consistent with current operational practice in seasonal forecasting at the Australian Bureau of Meteorology (e.g., Fawcett et al. 2005). In the linear model, the two- and three-category forecast probabilities are derived in a statistically consistent fashion from the same underlying model. In the LDA case, this is not done. Rather, as previously mentioned, separate versions are used in the two- and three-category cases. For the stratified climatology model, detailed results for just the two-category case are presented, for reasons of space. The construction of the analogous three-category case is sketched briefly.

The skill scores considered include the linear error in probability space (LEPS) score (Potts et al. 1996), the Brier score (Brier 1950; Wilks 2006), the relative operating characteristics (ROC) score (ROCS; Mason 1982; Mason and Graham 1999), and some scores directly associated with 2×2 contingency tables (Mason 2003), including the proportion correct (the verification statistic given the most prominence by the Australian Bureau of Meteorology in its seasonal forecasting efforts, but called in that context the percent consistent). While the use of the proportion correct is sometimes considered inadvisable [see the discussion in Mason (2003)], its use in the present context is justified because the various forecast categories are climatologically equally likely and the forecast models preserve reliability. Verification statistics are converted into skill scores in the usual way, using climatological and perfect forecasts as the reference forecasts.

Numerical (and some algebraic) evaluations of integrals have generally been performed using *Mathematica*. Details of some calculations have been given in situations (in particular, the ROCS calculations) where more involved transformations of the various integrals have been required.

Scores, rates, and skill scores in section 2 for the limiting parameter values $a = 0$ and 1 have typically been calculated in terms of the limits $a \downarrow 0$ and $a \uparrow 1$, rather than by direct substitution. For example, direct substitution of $a = 1$ usually leads to a division by zero. Similar considerations hold for the subsequent sections. [Here, $a \uparrow b$ has its usual mathematical meaning. "a approaches b from below." Likewise, "$a \downarrow b$" means "a approaches b from above," while "$a \rightarrow b$" means "a approaches b" (without regard to which side of the limiting process is taken).]

2. The linear regression model

Let X, Y, and Z be standard normally distributed random variables, X and Z independently so, such that $Y = aX + bZ$, where a and b are constants ($0 < a < 1$ and $b = \sqrt{1 - a^2}$). Here, X is interpreted as the predictor, Y as the predictand, and Z as a noise variable. The parameter a is the model correlation coefficient between the predictor X and predictand Y, and measures the strength of the relationship between them. Hence, the fraction of the variance of Y explained by X is a^2. The conditional distribution $Y|X = x$ is then a normal distribution with mean ax and variance b^2. It represents the conditional (or forecast) distribution within the seasonal forecasting context, in contrast to the unconditional (or climatological) distribution for Y, which as previously mentioned is the standard normal distribution.

For $a = 0$, the model is completely unskilled, with its forecasts always equal to the unconditional (or climatological) forecast. Forecast skill is constructed to increase as the model correlation coefficient a increases, and perfect skill is attained as $a \uparrow 1$. Note that negative values of a do not result in a negatively skilled model, but rather a positively skilled model with a negative correlation between the predictor and predictand. Accordingly, the discussion will be restricted to cases with $a \geq 0$.

For convenience, let $\rho_Z(z) = (1/\sqrt{2\pi}) \exp(-z^2/2)$ denote the probability density function (p.d.f.) of the standard normal distribution, $P_Z(z) = \int_{-\infty}^{z} \rho_Z(u)du$ the corresponding cumulative distribution function, and $Q_Z(z) = 1 - P_Z(z) = \int_{z}^{\infty} \rho_Z(u)du$ the corresponding probability of exceedance function. These designations will be used throughout the paper.

Some more general types of linear models can be reduced to the form described above. For example, sup-

pose $Y' = a'X' + b'Z' + c'$, where X' is normally distributed with mean $\mu_{X'}$ and variance $\sigma^2_{X'}(\sigma_{X'} > 0)$, Z' is normally distributed with mean $\mu_{Z'}$ and variance $\sigma^2_{Z'}(\sigma_{Z'} > 0)$, X' and Z' independently so, and a', b', and c' are nonzero constants. Then, Y' is normally distributed with mean $\mu_{Y'} = a'\mu_{X'} + b'\mu_{Z'} + c'$ and variance $\sigma^2_{Y'} = a'^2\sigma^2_{X'} + b'^2\sigma^2_{Z'}$. As before, X' and Y' are the predictor and predictand, respectively, while Z' is the noise variable. To reduce this more general model to the original form $Y = aX + bZ$, we set $X = \text{sign}(a')(X' - \mu_{X'})/\sigma_{X'}$, $Z = \text{sign}(b')(Z' - \mu_{Z'})/\sigma_{Z'}$, and

$$a = |a'|\sigma_{X'}/\sqrt{a'^2\sigma^2_{X'} + b'^2\sigma^2_{z'}},$$

$$b = |b'|\sigma_{Z'}/\sqrt{a'^2\sigma^2_{X'} + b'^2\sigma^2_{Z'}}, \quad \text{and}$$

$$Y = (Y' - \mu_{Y'})/\sqrt{a'^2\sigma^2_{X'} + b'^2\sigma^2_{Z'}}.$$

As before, X, Y, and Z are all standard normally distributed, X and Z independently so; a and b are both positive, with $a^2 + b^2 = 1$.

a. Above the median forecasts

With a and b both positive, the conditional probability of an above median outcome is

$$q = \Pr(Y > 0|X = x) = Q_Z\left(-\frac{ax}{b}\right),$$

with the corresponding probability of a below median outcome as

$$p = 1 - q = \Pr(Y < 0|X = x) = P_Z\left(-\frac{ax}{b}\right).$$

It is expedient at this point to calculate the actual distribution of the forecast probabilities generated by the model, under the assumption that the predictor X is standard normally distributed. The p.d.f. is

$$\rho_Q(q) = \rho_Z(x)\frac{dx}{dq} = \frac{b}{a}\exp\left\{\frac{1}{2}\left(1 - \frac{b^2}{a^2}\right)[Q_Z^{-1}(q)]^2\right\}.$$

This distribution is symmetric about $q = 1/2$, resulting in the mean and median probabilities being $q = 1/2$. In the special case $a = 1/\sqrt{2}$, it is the uniform distribution on [0, 1]. Here, $[a, b]$ has its usual mathematical meaning of the closed interval $a \le x \le b$. Likewise, (a, b) will be used to denote the open interval $a < x < b$. As $a \downarrow 0$, the p.d.f. approaches the Dirac delta function, $\delta[q - (1/2)]$, and for small values of a, the forecast probabilities are very narrowly distributed around $q = 1/2$. For $0 < a < 1/\sqrt{2}$, the value $q = 1/2$ is the modal value, but for $1/\sqrt{2} < a < 1$, the p.d.f. becomes U-shaped, with $q = 1/2$ the least likely forecast value. The symmetry in the distribution leads to the result $q(x) + q(-x) \equiv 1$, using the functional notation for q as a function of x.

1) LEPS2 SCORES

The linear error in probability space methods of scoring forecasts are described in Potts et al. (1996), but for the present purpose, a summary of how LEPS2 scores (the particular version used here) and skill scores are computed in the specific above/below median case is given in Fawcett et al. (2005). If the outcome is above the median, the forecast attracts a LEPS2 score of $(1/6)q - (1/6)p$, whereas if the outcome is below the median, the forecast attracts a LEPS2 score of $(1/6)p - (1/6)q$. For the model, these two outcomes occur with probabilities q and $1 - q$, respectively. Thus, the expected value of the LEPS2 score for the given value $X = x$ of the predictor is

$$s_1(q) = q\left(\frac{1}{6}q - \frac{1}{6}p\right) + (1 - q)\left(\frac{1}{6}p - \frac{1}{6}q\right)$$

$$= \frac{1}{6}(1 - 2q)^2. \tag{1}$$

The mean LEPS2 score for the model is then

$$\overline{\text{LEPS2}} = \int_{-\infty}^{\infty}\rho_Z(x)s_1(q)\,dx = \frac{1}{6\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2}x^2\right)\left[1 - 2Q_Z\left(-\frac{ax}{b}\right)\right]^2 dx,$$

with the LEPS2 skill score (LEPS2SS) being 6 times this quantity (LEPS2SS = $6\overline{\text{LEPS2}}$). LEPS2SS ranges from 0 to 1 as a ranges from 0 to 1. While an explicit form of this integral is not easily obtainable, some exact values for the skill score are 0 when $a = 0$, 1/3 when $a = 1/\sqrt{2}$, and 1 when $a = 1$. For $a = 0.2$, 0.3, 0.4, and 0.5, under the linear model the predictor explains 4%, 9%, 16%, and 25% of the predictand variance, respectively, and achieves LEPS2 skill scores of approximately 2.5%, 5.7%, 10.2%, and 16.1%, respectively. A Taylor series expansion of about 0 for $Q_Z(z)$ permits a calculation of the behavior of the LEPS2 skill score for small a; LEPS2SS $\sim (2/\pi)[a^2/(1 - a^2)]$. Higher-order terms are also readily obtainable. The LEPS2 skill

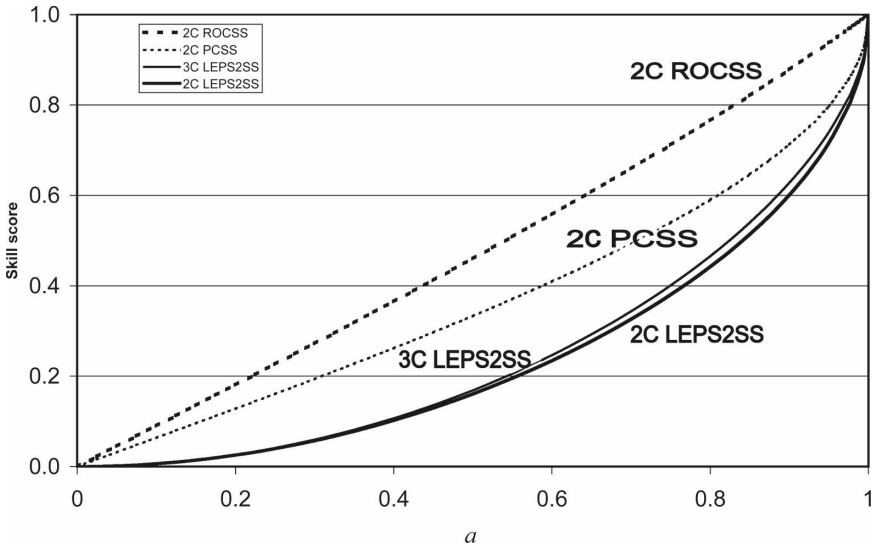

FIG. 1. Plot of skill scores (vertical axis) for the linear regression model against the model correlation coefficient a (horizontal axis): two-category LEPS2SS (thick solid line), three-category LEPS2SS (thin solid line), two-category PCSS (thin dashed line), and two-category ROCSS (thick dashed line).

score for this model as a function of a is shown in Fig. 1.

2) BRIER SCORES

A similar type of score is the Brier score (Brier 1950; Wilks 2006). For a sequence of probabilistic forecasts for a particular category, it is defined as $\mathrm{BS} = (1/n)\Sigma_{i=1}^{n}(p_i - o_i)^2$, where the p_i are the forecast probabilities and the o_i are the outcomes (1 if the category outcome occurs and 0 if it does not). [The form of the Brier score used here is the so-called half-Brier score, not the original form $\mathrm{OBS} = (2/n)\Sigma_{i=1}^{n}(p_i - o_i)^2$.] Within the present context, the particular category outcome is of course an above median outcome. If the outcome is above median, the forecast attracts a Brier score contribution of $(q - 1)^2$, while if the outcome is below median, the score contribution is $(q - 0)^2$. As these occur with probabilities q and $1 - q$ respectively, the expected Brier score given, $X = x$, is therefore

$$s_2(q) = (q - 1)^2 q + q^2(1 - q) = q(1 - q), \qquad (2)$$

and the mean Brier score for the above median category is

$$\overline{\mathrm{BS}} = \int_{-\infty}^{\infty} \rho_Z(x) s_2(q)\, dx.$$

$\overline{\mathrm{BS}}$ ranges from 1/4 down to 0 as a ranges from 0 to 1. It can be converted into a Brier skill score by setting $\mathrm{BSS} = 1 - 4\overline{\mathrm{BS}}$. For this model, the Brier skill score for the above median forecasts is identical to the LEPS2 skill score given above. [This result is actually somewhat more general than the current context, as

$$\begin{aligned}\mathrm{BSS} &= 1 - 4\int_{-\infty}^{\infty} \rho_Z(x) q(1 - q)\, dx \\ &= \int_{-\infty}^{\infty} \rho_Z(x)(1 - 2q)^2\, dx \\ &= \mathrm{LEPS2SS}, \end{aligned} \qquad (3)$$

independently of the precise functional dependence of q on x and also independently of the precise functional form of the predictor p.d.f. $\rho_Z(x)$.] Also, while the Brier skill score for the above median category has been computed above, the Brier skill score for the below median category is the same, so (like the LEPS2 skill score) the Brier skill score is effectively a skill score for the model rather than for the category.

3) CONTINGENCY TABLE SCORES

We now consider various verification statistics that arise from reinterpreting the probabilistic forecast as categorical.

The "proportion correct" scoring approach[1] gives a score of 1 to a forecast if $q > 1/2$ and $Y > 0$, and also if $q < 1/2$ and $Y < 0$, but 0 otherwise. Hence, if the outcome is above the median, the forecast attracts a

[1] This is $(a + d)/n$ in the 2×2 standard contingency table formulation of Mason (2003).

proportion correct contribution of $H[q - (1/2)]$, where $H(x)$ is the unit Heaviside step function. This has $H(x) = 1$ if $x > 0$, and 0 otherwise; although for the calculations presented in this paper, it is desirable to assume that $H(0) = 1/2$. Likewise, if the outcome is below the median, the forecast attracts a proportion correct contribution of $H[(1/2) - q]$. As these two outcomes occur with probabilities of q and $1 - q$, respectively, the expected proportion correct given $X = x$ is

$$s_3(q) = qH\left(q - \frac{1}{2}\right) + (1 - q)H\left(\frac{1}{2} - q\right) = \frac{1}{2} + \left|q - \frac{1}{2}\right|. \quad (4)$$

The proportion correct for the model as a whole is then

$$\overline{\text{PC}} = \int_{-\infty}^{\infty} \rho_Z(x) s_3(q)\, dx = \frac{1}{2} + \int_{-\infty}^{\infty} \rho_Z(x) \left|q - \frac{1}{2}\right| dx = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{a}{\sqrt{1 - a^2}}\right). \quad (5)$$

The value of $\overline{\text{PC}}$ ranges from 1/2 to 1 as a ranges from 0 to 1. Some special values include proportions correct of 1/2, 2/3, and 3/4 for $a = 0$, 1/2 and $1/\sqrt{2}$, respectively. The proportion correct can be converted into a skill score via

$$\text{PCSS} = 2\overline{\text{PC}} - 1 = \frac{2}{\pi} \arctan\left(\frac{a}{\sqrt{1 - a^2}}\right).$$

The PCSS ranges from 0 to 1 as a ranges from 0 to 1. The proportion correct skill score for this model is shown as a function of a in Fig. 1. Note that for small levels of forecast skill ($a \ll 1$), the PCSS shows an approximately linear dependence on the model correlation coefficient a, unlike the approximately quadratic dependence on a for the LEPS2SS when $a \ll 1$. This means, for example, that an LEPS2SS of 0.1 (10%) implies a much stronger relationship between the predictor and predictand than does a PCSS of 0.1.

The "hit rate" scoring approach[2] for the above median category calculates the fraction of times the conditional probability of an above median outcome exceeds the climatological value given that the above median outcome occurs. This is the conditional probability

$$\overline{H} = \Pr\left(q > \frac{1}{2} \,\middle|\, Y > 0\right) = \frac{\Pr\left(q > \frac{1}{2} \text{ and } Y > 0\right)}{\Pr(Y > 0)} = 2\Pr\left(q > \frac{1}{2} \text{ and } Y > 0\right).$$

To calculate this, we integrate the joint probability distribution function of X and Z over the region defined by $x > 0$ (arising from $q > 1/2$) and $ax + bz > 0$ (arising from $Y > 0$). The joint p.d.f. is $\rho_{XZ}(x, z) = \rho_Z(x)\rho_Z(z)$, and the required mean hit rate is

$$\overline{H} = 2\int_{x=0}^{\infty} \int_{z=-ax/b}^{\infty} \rho_Z(z)\rho_Z(x)\, dz\, dx = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{a}{\sqrt{1 - a^2}}\right).$$

This is the same as $\overline{\text{PC}}$ in Eq. (5), and is converted into a hit rate skill score in the same way. Specifically,

$$\text{HSS} = \frac{2}{\pi} \arctan\left(\frac{a}{\sqrt{1 - a^2}}\right) = \text{PCSS}.$$

In a similar manner, the "false alarm rate" scoring approach[3] for the above median category calculates the fraction of times the conditional probability of an above median outcome exceeds the climatological probability given that the above median outcome does not occur. This is the conditional probability:

$$\overline{F} = \Pr\left(q > \frac{1}{2} \,\middle|\, Y < 0\right) = 2\Pr\left(q > \frac{1}{2} \text{ and } Y > 0\right).$$

To calculate this, we integrate $\rho_{XZ}(x, z)$ over the area defined by $x > 0$ (arising from $q < 1/2$) and $ax + bz < 0$ (arising from $Y < 0$), and obtain

$$\overline{F} = 2\int_{x=0}^{\infty} \int_{z=-\infty}^{-ax/b} \rho_Z(z)\rho_Z(x)\, dz\, dx = \frac{1}{2} - \frac{1}{\pi} \arctan\left(\frac{a}{\sqrt{1 - a^2}}\right).$$

Unlike the mean proportion correct and mean hit rate, this runs from 1/2 down to 0 as a runs from 0 to 1. Conversion to a skill score via $\text{FSS} = 1 - 2\overline{F}$, however, yields the same result as PCSS and HSS.

4) ROC SCORES

The hit rate and false alarm rate calculations given above, treated more generally, lead to the calculation of the ROC score (Mason 1982; Mason and Graham 1999). We assume that the probabilistic forecast of an

[2] This is $a/(a + c)$ in the standard contingency table formulation of Mason (2003).

[3] This is $b/(b + d)$ in the standard contingency table formulation of Mason (2003).

above median outcome is reinterpreted as categorical, if q exceeds a threshold probability q_0. Then,

$$\overline{H}(q_0) = \Pr(q > q_0 | Y > 0) = 2\Pr(q > q_0 \quad \text{and} \quad Y > 0)$$

and

$$\overline{F}(q_0) = 2\Pr(q > q_0 \quad \text{and} \quad Y < 0).$$

The ROC score used here is the area under the curve, $\{[\overline{F}(q_0), \overline{H}(q_0)] | 0 \le q_0 \le 1\}$, calculated via

$$\text{ROCS} = \int_{q_0=1}^{0} \overline{H}(q_0) \frac{d}{dq_0} [\overline{F}(q_0)] \, dq_0.$$

The integral runs from $q_0 = 1$ to 0, because these values correspond to the first and last points, respectively, on the ROC curve. It simplifies eventually to

$$\text{ROCS} = \frac{3}{2} - 2\int_{-\infty}^{\infty} \rho_Z(u) Q_Z^2\left(-\frac{b}{a}u\right) du,$$

but the calculation is somewhat involved. The details are given in appendix A. The ROC score varies almost linearly with a, running from 1/2 to 1 as a runs from 0 to 1. A ROC skill score can be calculated as ROCSS = 2ROCS − 1 and is plotted in Fig. 1. An exact result is obtainable when $a = 1/\sqrt{2}$, which leads to ROCS = 5/6 and ROCSS = 2/3. The near-linear dependence of the ROC skill score on a means, for example, that a ROCSS of 0.2 (20%) implies a much weaker relationship between predictor and predictand than does a LEPS2SS of 0.2.

b. Terciles forecasts

The climatological terciles for the predictor in this model are the intervals $(-\infty, z_{12})$, (z_{12}, z_{23}), and (z_{23}, ∞), where $z_{23} = -z_{12} \approx 0.4307$ satisfies $Q_Z(z_{23}) = 1/3$. The conditional forecast probabilities for the model, given $X = x$, are $p_1 = 1 - Q_Z[(1/b)(z_{12} - ax)]$, $p_2 = 1 - p_1 - p_3$, and $p_3 = Q_Z[(1/b)(z_{23} - ax)]$ for terciles 1, 2, and 3, respectively. The symmetry of the model results in the relationships $p_2(x) \equiv p_2(-x)$ and $p_1(x) \equiv p_3(-x)$. The tercile 2 probability attains its maximum value at $x = 0$.

1) LEPS2 scores

A summary of how LEPS2 scores and skill scores are computed in the terciles case is given in Fawcett et al. (2005). If the outcome is in tercile 1 ($Y < z_{12}$), the forecast attracts a LEPS2 score of $(8/27)p_1 - (1/27)p_2 - (7/27)p_3$; if it is in tercile 2 ($z_{12} < Y < z_{23}$), the score is $-(1/27)p_1 + (2/27)p_2 - (1/27)p_3$; and if in it is in tercile 3 ($Y > z_{23}$), the score is $-(7/27)p_1 - (1/27)p_2 + (8/27)p_3$. Since for the model these three outcomes occur with the probabilities p_1, p_2, and p_3, respectively, the expected value of the LEPS2 score given $X = x$ is

$$\begin{aligned} s_4(p_1, p_2, p_3) &= p_1\left(\frac{8}{27}p_1 - \frac{1}{27}p_2 - \frac{7}{27}p_3\right) + p_2\left(-\frac{1}{27}p_1 + \frac{2}{27}p_2 - \frac{1}{27}p_3\right) + p_3\left(-\frac{7}{27}p_1 - \frac{1}{27}p_2 + \frac{8}{27}p_3\right) \\ &= \frac{2}{27}(6p_1^2 + 6p_3^2 - 3p_1p_3 - 3p_1 - 3p_3 + 1). \end{aligned} \tag{6}$$

The mean value of the LEPS2 score is then

$$\overline{\text{LEPS2}} = \int_{-\infty}^{\infty} \rho_Z(x) s_4(p_1, p_2, p_3) \, dx.$$

As the model is positively skilled for $a > 0$, the LEPS2 skill score is calculated as LEPS2SS = $(9/2)\overline{\text{LEPS2}}$. LEPS2SS ranges from 0 to 1 as a ranges from 0 to 1. The LEPS2 skill score for this model as a function of a is shown in Fig. 1. The three-category LEPS2 skill score is slightly higher than the corresponding two-category LEPS2 skill score of section 2a over the range $0 < a < 1$, but the graph shows that the two scores are basically comparable. This suggests that it is quite reasonable to interpret the two as being on the same scale, something that could not be said (for example) for the two-category skill scores LEPS2SS and PCSS taken together. The reason for the three-category LEPS2 skill scores being slightly higher than the corresponding two-category skill scores is that there is slightly more information about the underlying linear model present in the three-category forecasts. Further, it would be reasonable to expect seasonal forecasting systems producing two- and three-category forecasts in a statistically consistent way (e.g., by counting outcomes from an ensemble of coupled general circulation model runs) to generate similar LEPS2 skill scores across large sets of parallel forecasts, and that it would be prudent to investigate any substantial discrepancies between the two- and three-category LEPS2 skill scores.

2) Brier scores

Regarding the calculation of Brier scores, if the outcome is in tercile 1, then the forecast attracts a tercile 1

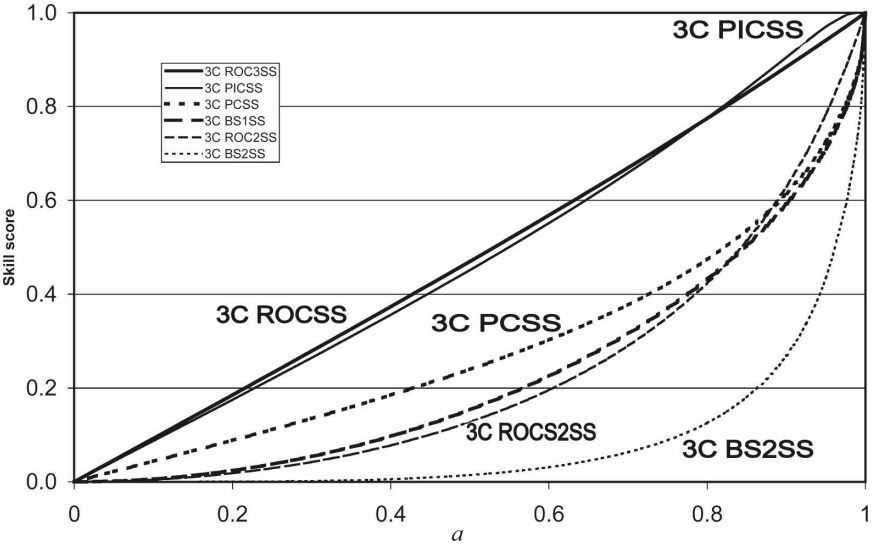

FIG. 2. Plot of skill scores (vertical axis) for the linear regression model against the model correlation coefficient a (horizontal axis): three-category PCSS (thick dotted line), PICSS (thin solid line line), BS1SS (thick dashed line), BS2SS (thin dotted line), ROC3SS (thick solid line), and ROC2SS (thin dashed line).

Brier score contribution of $(p_1 - 1)^2$, whereas if the outcome is in tercile 2 or tercile 3, the tercile 1 Brier score contribution is $(p_1 - 0)^2$. Since these outcomes occur with probabilities p_1 and $1 - p_1$, respectively, the expected tercile 1 Brier score given $X = x$ is

$$s_2(p_1) = (p_1 - 1)^2 p_1 + p_1^2(1 - p_1) = p_1(1 - p_1),$$

where $s_2(\cdot)$ is as given in Eq. (2). The mean tercile 1 Brier score is

$$\overline{\text{BS1}} = \int_{-\infty}^{\infty} \rho_Z(x) s_2(p_1)\, dx.$$

It ranges from 2/9 down to 0 as a ranges from 0 to 1. Hence, a skill score may be calculated as $\text{BS1SS} = 1 - (9/2)\overline{\text{BS1}}$. Values of BS1SS are plotted in Fig. 2. They are very similar to those of the two- and three-category LEPS2 skill scores (Fig. 1), but slightly smaller than both. The mean Brier scores for the other two terciles are analogously calculated:

$$\overline{\text{BS2}} = \int_{-\infty}^{\infty} \rho_Z(x) s_2(p_2)\, dx$$

and

$$\overline{\text{BS3}} = \int_{-\infty}^{\infty} \rho_Z(x) s_2(p_3)\, dx.$$

The tercile 3 Brier scores are, not surprisingly, the same as the tercile 1 scores, given the underlying symmetry of the model and, hence, so are the corresponding skill scores. The tercile 2 Brier scores, which are also shown in Fig. 2, are converted into skill scores in the same way. They are however much lower than those for terciles 1 and 3, indicating that there is much less skill available in predicting the middle tercile than the outer terciles. Skill at predicting the middle tercile occurs here only with a strong relationship between the predictor and predictand.

3) CONTINGENCY TABLE SCORES

As in section 2a, various verification statistics can be obtained by reinterpreting the probabilistic forecast as categorical and constructing a standard 3×3 contingency table.

To calculate the proportion correct for the model, we note that when the outcome is in tercile 1, the forecast attracts a proportion correct contribution of 1 if $p_1 > p_2$ and $p_1 > p_3$, and 0 otherwise. This can be represented as $H(p_1 - p_2)H(p_1 - p_3)$, in terms of the Heaviside step function. The other two categories are treated similarly. Given that the outcome occurs in the three categories with probabilities p_1, p_2, and p_3, respectively, the expected contribution to the proportion correct, given $X = x$, is

$$\begin{aligned} s_5(p_1, p_2, p_3) = {} & p_1 H(p_1 - p_2) H(p_1 - p_3) \\ & + p_2 H(p_2 - p_1) H(p_2 - p_3) \\ & + p_3 H(p_3 - p_1) H(p_3 - p_2). \end{aligned} \quad (7)$$

The mean proportion correct for the model as a whole is therefore

$$\overline{\mathrm{PC}} = \int_{-\infty}^{\infty} \rho_Z(x) s_5(p_1, p_2, p_3)\, dx.$$

We find that $\overline{\mathrm{PC}} \downarrow 1/3$ as $a \downarrow 0$, and $\overline{\mathrm{PC}} \uparrow 1$ as $a \uparrow 1$. Hence, a skill score for the three-category proportion correct can be constructed as $\mathrm{PCSS} = (3/2)\overline{\mathrm{PC}} - 1/2$. It ranges from 0 to 1 as a ranges from 0 to 1. A comparison of numerically computed three-category PCSS values with the two-category PCSS (for which an exact result was obtained in section 2a) leads to the observation that

$$\mathrm{PCSS} \approx \frac{2}{\pi} \arctan\left(\frac{0.7a}{\sqrt{1-a^2}}\right)$$

in the three-category case.

To complement the proportion correct statistic, a "proportion incorrect" is also considered. While the proportion correct counts the fraction of times the *most probable* tercile is the one subsequently observed, the proportion incorrect counts the fraction of times the *least probable* tercile is the one subsequently observed. The expected contribution to the proportion incorrect given $X = x$ is

$$\begin{aligned} s_6(p_1, p_2, p_3) = {} & p_1 H(p_2 - p_1) H(p_3 - p_1) \\ & + p_2 H(p_1 - p_2) H(p_3 - p_2) \\ & + p_3 H(p_1 - p_3) H(p_2 - p_3). \end{aligned} \quad (8)$$

The mean proportion incorrect for the model as a whole is therefore

$$\overline{\mathrm{PIC}} = \int_{-\infty}^{\infty} \rho_Z(x) s_6(p_1, p_2, p_3)\, dx.$$

We find that $\overline{\mathrm{PIC}} \uparrow 1/3$ as $a \downarrow 0$, and $\overline{\mathrm{PIC}} \downarrow 0$ as $a \uparrow 1$. Hence, a skill score for the three-category proportion incorrect can be constructed as $\mathrm{PICSS} = 1 - 3\overline{\mathrm{PIC}}$. It ranges from 0 to 1 as a ranges from 0 to 1, in an almost linear fashion, and is higher than PCSS. Both are shown in Fig. 2. As before, a given value (0.2 say) for the PCSS and PICSS means two quite different things for the strength of the relationship between the predictor and predictand.

4) ROC SCORES

A ROC score for tercile 3 can be calculated in a manner analogous to that of the above median category. We reinterpret the probabilistic forecast of tercile 3 as categorical if p_3 exceeds a threshold probability p_0. Then, the hit rate is

$$\begin{aligned} \overline{H}(p_0) &= \Pr(p_3 > p_0 | Y > z_{23}) \\ &= 3\Pr(p_3 > p_0 \quad \text{and} \quad Y > z_{23}). \end{aligned}$$

In addition, the false alarm rate is

$$\overline{F}(p_0) = \frac{3}{2} \Pr(p_3 > 0 \quad \text{and} \quad Y < z_{23}).$$

The area under the ROC curve is, as before,

$$\mathrm{ROCS} = \int_{p_0 = 1}^{0} \overline{H}(p_0) \frac{d}{dp_0} [\overline{F}(p_0)]\, dp_0,$$

which here simplifies to

$$\mathrm{ROCS} = \frac{5}{4} - \frac{9}{2} \int_{-\infty}^{\infty} Q_Z(v) p_3(v) \rho_Z(v)\, dv\,.$$

The details of this calculation are given in appendix B. The ROC score for tercile 3 can be converted into a skill score via $\mathrm{ROC3SS} = 2\ \mathrm{ROCS} - 1$. Again, by virtue of the basic symmetry of the model, the ROC scores and skill scores for tercile 1 are the same as those for tercile 3 (ROC1SS = ROC3SS). These values are plotted in Fig. 2.

For tercile 2, we have

$$\overline{H}(p_0) = 3\Pr(p_2 > p_0 \quad \text{and} \quad z_{12} < Y < z_{23})$$

and

$$\overline{F}(p_0) = \frac{3}{2} \Pr[p_2 > p_0 \quad \text{and} \quad (Y < z_{12} \quad \text{or} \quad Y > z_{23})],$$

giving

$$\mathrm{ROCS} = \frac{5}{4} - \int_{0}^{\infty} 9 p_2(u) [Q_Z(-u) - Q_Z(u)] \rho_Z(u)\, du.$$

Again, the details of the calculation are given in appendix B. As before, this is converted into a skill score via $\mathrm{ROC2SS} = 2\ \mathrm{ROCS} - 1$. The skill scores are shown in Fig. 2. They are considerably smaller than the tercile 3 ROC skill scores, indicating the much reduced predictability of tercile 2, compared to terciles 1 and 3, which is consistent with the Brier score results.

c. *Reliability*

As previously mentioned, the linear regression model presented at the start of this section (and its interpretation in section 2a in the cases of the above and below median categories) is by construction perfectly reliable (Wilks 2006; Hartmann et al. 2002). That is, for a given type of forecast category and forecast probability, the observed frequency of the outcome is the same as the forecast probability of it. To explore the issue of reliability within this context, it is necessary to degrade the model so as to make it less than perfectly reliable.

This can be done in many ways, but for illustrative

purposes we construct a variant model in which the forecast probability departure from climatology of above median outcomes is half the observed one. [We cannot do it the other way around (forecast departure twice the observed departure) in this example, because the observed frequencies run from 0 to 1, which would cause the forecast frequencies to run from −1/2 to 3/2.] In other words, if $q_o(x)$ and $q_f(x)$ are the observed and forecast probabilities, respectively, of an above the median outcome given $X = x$, then we suppose

$$\left[q_f(x) - \frac{1}{2}\right] = \frac{1}{2}\left[q_o(x) - \frac{1}{2}\right] \quad \text{or} \quad q_f(x) = \frac{1}{4} + \frac{1}{2}q_o(x).$$

As the observed frequencies run from 0 to 1, the forecast frequencies run from 1/4 to 3/4. As before, $q_o(x) = Q_Z(-ax/b)$. The corresponding below median probabilities are of course $p_o(x) = 1 - q_o(x)$ and $p_f(x) = 1 - q_f(x)$. [In the original formulation of section 2a, we had $q_f(x) \equiv q_o(x)$. It is possible to write down explicitly a variant model that generates the probability

$$\Pr(\tilde{Y} > 0 | X = x) = \frac{1}{4} + \frac{1}{2}q_o(x),$$

although it is rather artificial. It has

$$\tilde{Y} = af(X) + bZ, \quad \text{where}$$

$$f(X) = -\frac{b}{a}Q_Z^{-1}\left[\frac{1}{4} + \frac{1}{2}Q_Z\left(-\frac{a}{b}X\right)\right],$$

with a, b, X, and Z as before. This amounts to a nonlinear transformation of the predictor variable, although for small values of a it is approximately linear (with $f'(0) = 1/2$) over the region of most interest (i.e., roughly $-5 \le X \le 5$) for the standard normal distribution. For larger values of a, it becomes increasingly sigmoid shaped, retaining the property $f(X) = -f(-X)$. For the present purpose of illustrating the effect of the departure from reliability, the variant model is verified against "observations" from the original model, as previously indicated, rather than against its own "observations" with respect to which it would by construction be perfectly reliable.]

Following section 2a, the expected value of the LEPS2 score given $X = x$ is

$$q_o\left(\frac{1}{6}q_f - \frac{1}{6}p_f\right) + (1 - q_o)\left(\frac{1}{6}p_f - \frac{1}{6}q_f\right) = \frac{1}{6}(1 - 2q_o) \times (1 - 2q_f).$$

The LEPS2 skill scores are therefore

$$\text{LEPS2SS} = \int_{-\infty}^{\infty} \rho_Z(x)(1 - 2q_f)(1 - 2q_o)\, dx.$$

The expected value of the Brier score for the above median category given $X = x$ is $(q_f - 1)^2 q_o + q_f^2(1 - q_o)$. The Brier skill score is therefore

$$\text{BSS} = 1 - 4\int_{-\infty}^{\infty} \rho_Z(x)[(q_f - 1)^2 q_o + q_f^2(1 - q_o)]\, dx$$

$$= \int_{-\infty}^{\infty} \rho_Z(x)(1 - 2q_f)(1 + 2q_f - 4q_o)\, dx.$$

Because q_f is different from q_o, the Brier skill score is not equal to the LEPS2 skill score in this case. By analogy with Eq. (4), the proportion correct skill score is

$$\text{PCSS} = -1 + 2\int_{-\infty}^{\infty} \rho_Z(x)\left[q_o H\left(q_f - \frac{1}{2}\right) + (1 - q_o)H\left(\frac{1}{2} - q_f\right)\right] dx.$$

Because $\text{sign}(q_f - 1/2) = \text{sign}(q_o - 1/2)$, the proportion correct skill score is not degraded at all by this particular form of departure from perfect reliability, unlike the other two skill scores, which are reduced somewhat, with perfect skill no longer obtainable in the limit $a \uparrow 1$. The LEPS2SS and BSS scores are plotted in Fig. 3, along with the LEPS2SS and PCSS for the perfectly reliable case.

In fact, since $(1 - 2q_f) = (1/2)(1 - 2q_o)$, the LEPS2 skill scores for this example are exactly half those of the original (perfectly reliable) model, and since further $(1 + 2q_f - 4q_o) = (3/2)(1 - 2q_o)$, the Brier skill scores are exactly three-quarters those of the original model. In the LEPS2 case, the result arises out of the fact that the LEPS2 scores are proportional to the departures from climatology of the forecast probabilities. Halving the departures halves the scores. On the other hand, a parallel calculation (not shown) to that given in appendix A for the perfectly reliable model shows that the ROC skill score is not affected by degrading the reliability in this way.

d. Forecasting the tails

The two (above and below median) and three (terciles) categories considered in sections 2a and 2b arise frequently within the context of climate forecasting, whereas categories representing the tails of a distribution are arguably more relevant in weather forecasting.

Suppose the category of interest for the predictand is now (y_0, ∞), where the base (climatological) rate of its occurrence is $q_0 = Q_Z(y_0)$. [Situations corresponding to $(y_0,\ q_0) = (0,\ 1/2)$ for the above median category and $(y_0,\ q_0) = (z_{23},\ 1/3)$ for the tercile 3 category have

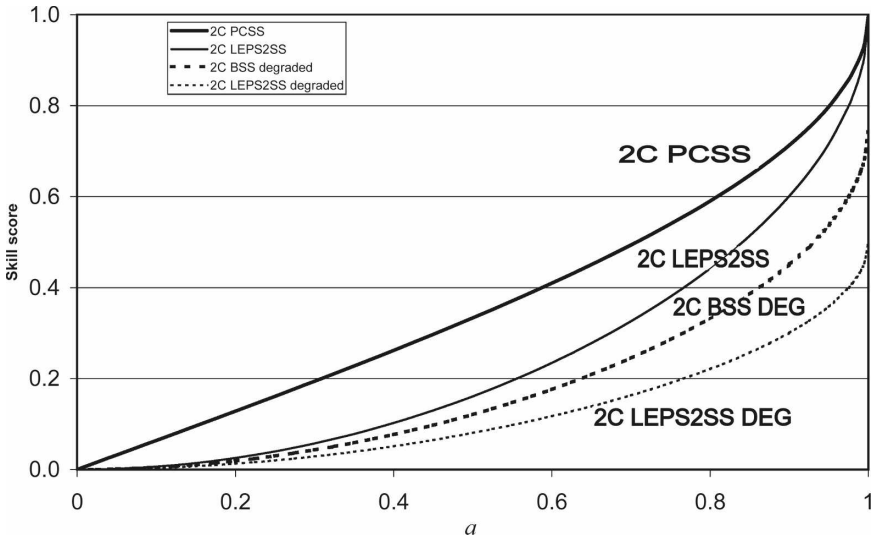

FIG. 3. Plot of skill scores (vertical axis) for the linear regression model against the model correlation coefficient a (horizontal axis): two-category LEPS2 skill score (thin solid line) and proportion correct skill score (thick solid line) for the perfectly reliable model, together with the LEPS2 skill score (thin dotted line) and Brier skill score (thick dotted line) for the degraded unreliable model. The PCSS for the degraded model is the same as that for the original model (thick solid line), while the BSS for the original model is the same as LEPS2SS for the original model (thin solid line).

already been considered in section 2.] The forecast probability for an outcome in this category is

$$q = \Pr(Y > y_0 \,|\, X = x) = Q_Z\left[\frac{1}{b}(y_0 - ax)\right],$$

given the predictor value $X = x$.

From Potts et al. (1996), if the outcome is in this category, then the forecast attracts a LEPS2 score of

$$\frac{2}{3}(1 - q_0)^2 q - \frac{2}{3} q_0(1 - q_0)(1 - q) = \frac{2}{3}(1 - q_0)(q - q_0),$$

while if the outcome is not in this category, then the forecast attracts a LEPS2 score of

$$-\frac{2}{3} q_0(1 - q_0)q + \frac{2}{3} q_0^2(1 - q) = \frac{2}{3} q_0(q_0 - q).$$

For the model, these outcomes occur with probabilities q and $1 - q$, respectively, so the expected value of the LEPS2 score for the given value $X = x$ of the predictor is

$$q\frac{2}{3}(1 - q_0)(q - q_0) + (1 - q)\frac{2}{3} q_0(q_0 - q) = \frac{2}{3}(q_0 - q)^2.$$

The mean LEPS2 score for the model as a whole is then

$$\overline{\text{LEPS2}} = \int_{-\infty}^{\infty} \rho_Z(x)\frac{2}{3}\left\{q_0 - Q_Z\left[\frac{1}{b}(y_0 - ax)\right]\right\}^2 dx,$$

with the LEPS2 skill score being given by

$$\text{LEPS2SS} = \frac{3}{2q_0(1 - q_0)}\,\overline{\text{LEPS2}}.$$

These skill scores are plotted in Fig. 4 as functions of a for the values $q_0 = n^{-1}$, $n = 2$, 4, 8, 16, and 32. The $n = 2$ curve in Fig. 4 is the same as the "2C LEPS2SS" curve in Fig. 1, since they both represent the LEPS2 skill score associated with predicting the above median category. As the tail category gets smaller ($q_0 \downarrow 0$, $y_0 \to \infty$), arising from the prediction of more extreme values, the LEPS2 skill scores for a given strength of the underlying relationship between the predictor and predictand decreases. Alternatively, as the tail category gets smaller, better and better predictors must be found in order to preserve the LEPS2 skill score. In short, the smaller the tail category, the harder it is to predict it.

3. The LDA model

a. Above the median forecasts

We now consider a simple linear discriminant analysis (LDA) model composed of two categories for the predictand variable (effectively, Y above and below the median), which are climatologically equally likely. We suppose that the conditional distribution $X|Y > Y_{\text{median}}$ is a normal distribution with mean $\alpha\sigma$ and variance σ^2,

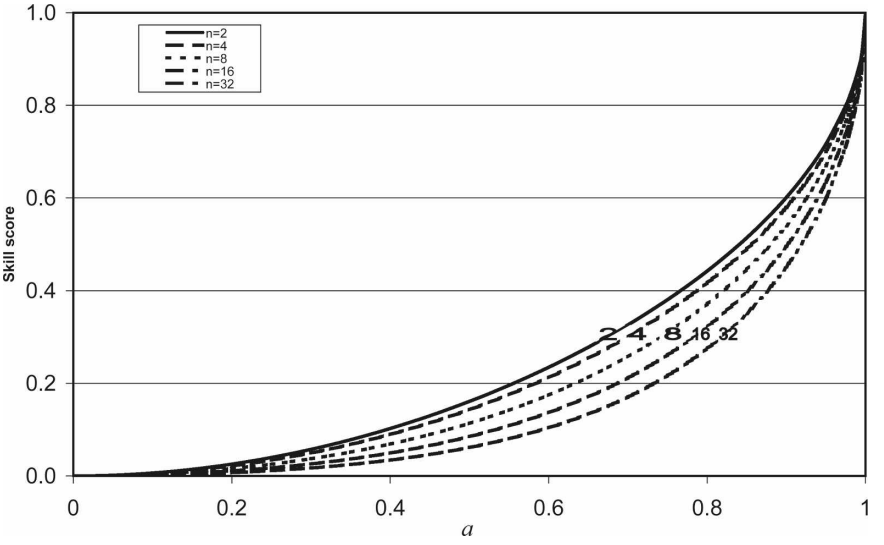

FIG. 4. LEPS2 skill scores for predicting the tails of the linear regression model, as functions of the model correlation coefficient a. The tail categories are defined as (y_0, ∞), where $\Pr(Y > y_0) = n^{-1}$ for $n = 2$, 4, 8, 16, and 32.

and that the conditional distribution $X|Y < Y_{\text{median}}$ is a normal distribution with mean $-\alpha\sigma$ and variance σ^2. Both α and σ are positive parameters of the distributions (although we will consider the limit $\alpha \downarrow 0$). The standard deviation σ has the same physical dimensions as the predictor variable X, but α is nondimensional. The p.d.f.'s for these two distributions are therefore

$$\rho_{X|Y>Y_{\text{median}}}(x) = \frac{1}{\sigma}\rho_Z\left(\frac{x-\alpha\sigma}{\sigma}\right)$$

and

$$\rho_{X|Y<Y_{\text{median}}}(x) = \frac{1}{\sigma}\rho_Z\left(\frac{x+\alpha\sigma}{\sigma}\right)$$

The means of these two distributions are $2\alpha\sigma$ (or 2α standard deviations) apart. The unconditional (climatological) distribution for the predictor X therefore has p.d.f.

$$\rho_X(x) = \frac{1}{2\sigma}\left[\rho_Z\left(\frac{x-\alpha\sigma}{\sigma}\right) + \rho_Z\left(\frac{x+\alpha\sigma}{\sigma}\right)\right].$$

This distribution is symmetric about $x = 0$ (hence zero mean and median) and has one mode ($x = 0$) for $0 \leq \alpha \leq 1$, but is bimodal for $\alpha > 1$. This bimodality of the climatological predictor distribution, which is particularly pronounced for large values of α, makes the LDA model less plausible than the linear regression model in the seasonal forecasting context, because the commonly used predictors, such as El Niño–Southern Oscillation (ENSO) indices, tend to unimodality.

Small values of α imply low skill [i.e., low discrimination; with $\alpha = 0$ corresponding to the no-skill (i.e., climatological) forecast] case, while large values of α imply high skill (i.e., high discrimination). Perfect skill is attained in the limit $\alpha \to \infty$. Negative values of α do not imply negative skill, rather positive skill with an oppositely signed relationship between the predictor and predictand.

Within the context of LDA forecasting, it is not necessary to suppose a functional dependence of the predictor Y on the predictand x. Indeed, Y does not even have to be a quantitative/continuous variable. We could simply designate $Y = 1$ for a "below median" outcome and $Y = 2$ for an "above median" outcome, treating Y as a categorical/discrete variable.

An application of Bayes's theorem leads to the conditional probability of an above median outcome for the predictand as

$$q = \Pr(Y > Y_{\text{median}}|X = x) = \frac{\rho_Z(x/\sigma - \alpha)}{\rho_Z(x/\sigma - \alpha) + \rho_Z(x/\sigma + \alpha)}.$$

By construction, the model has a symmetry: $q(x) + q(-x) \equiv 1$. It will be expedient in what follows to simplify matters by setting $\sigma = 1$; this can be done without loss of generality. Hence, from now on

$$\rho_X(x) = \frac{1}{2}[\rho_Z(x-\alpha) + \rho_Z(x+\alpha)]$$

and

$$q = \frac{\rho_Z(x-\alpha)}{\rho_Z(x-\alpha) + \rho_Z(x+\alpha)}.$$

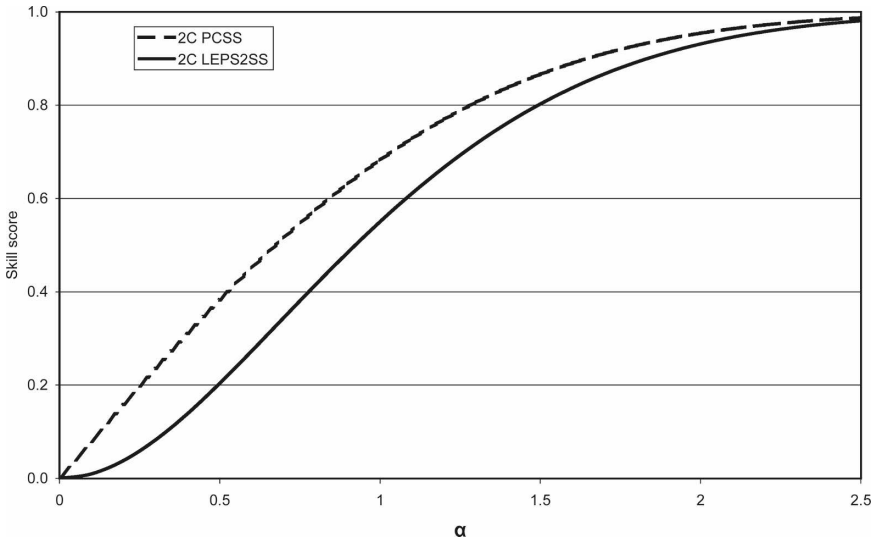

FIG. 5. Skill scores (LEPS2SS and PCSS) for the two-category LDA model, graphed against the conditional predictor distribution separation parameter α.

This implies that the probability of a below median outcome is

$$p = 1 - q = \frac{\rho_Z(x - \alpha)}{\rho_Z(x - \alpha) + \rho_Z(x + \alpha)}.$$

As in the linear regression model of section 2a, the LEPS2 skill score is calculated using Eq. (1) as

$$\text{LEPS2SS} = 6\int_{-\infty}^{\infty} \rho_X(x) s_1(q)\, dx$$

$$= \int_{\infty}^{\infty} \rho_X(x)(1 - 2q)^2\, dx.$$

Brier skill scores may also be calculated, along the lines of section 2a and using Eq. (2):

$$\text{BSS} = 1 - 4\int_{-\infty}^{\infty} \rho_X(x) s_2(q)\, dx.$$

They turn out to equal the LEPS2 skill scores [see Eq. (3)].

The mean proportion correct for the model is also calculated in a similar way to the linear regression model using Eq. (4):

$$\overline{\text{PC}} = \int_{-\infty}^{\infty} \rho_X(x) s_3(q)\, dx$$

$$= \int_{-\infty}^{\infty} \rho_X(x)\left(\frac{1}{2} + \left|q - \frac{1}{2}\right|\right) dx.$$

This score increases from 1/2 as a increases from 0. It is converted into a skill score in the usual way: PCSS = $2\overline{\text{PC}} - 1$. The proportion correct skill score is plotted in Fig. 5, along with the LEPS2 skill score, for a range of values of the separation parameter α. If we plot LEPS2SS against PCSS for this model (Fig. 7) and compare it against the corresponding results from the linear regression model of section 2a, the two curves are almost the same.

b. Tercile forecasts

An analogous three-category forecasting model is easily constructed. Suppose $X|Y \in$ tercile 1 is normally distributed with mean $-\alpha\sigma$ and variance σ^2, $X|Y \in$ tercile 2 is normally distributed with mean 0 and variance σ^2, and $X|Y \in$ tercile 3 is normally distributed with mean $\alpha\sigma$ and variance σ^2. Here, α and σ have the same meanings as in the two-category case (section 3a), with the same eventual consequences for the skill as α increases from 0 to ∞. Once again, it is supposed that the predictand terciles are climatologically equally likely. Then,

$$\rho_X(x) = \frac{1}{3\sigma}\left[\rho_Z\left(\frac{x - \alpha\sigma}{\sigma}\right) + \rho_Z\left(\frac{x}{\sigma}\right) + \rho_Z\left(\frac{x + \alpha\sigma}{\sigma}\right)\right]$$

and

$$p_1 = \Pr(Y \in \text{ tercile } 1 | X = x)$$

$$= \frac{\rho_Z(x/\sigma + \alpha)}{\rho_Z(x/\sigma - \alpha) + \rho_Z(x/\sigma + \alpha)},$$

$$p_2 = \Pr(Y \in \text{ tercile } 2 | X = x)$$

$$= \frac{\rho_Z(x/\sigma)}{\rho_Z(x/\sigma - \alpha) + \rho_Z(x/\sigma) + \rho_Z(x/\sigma + \alpha)},$$

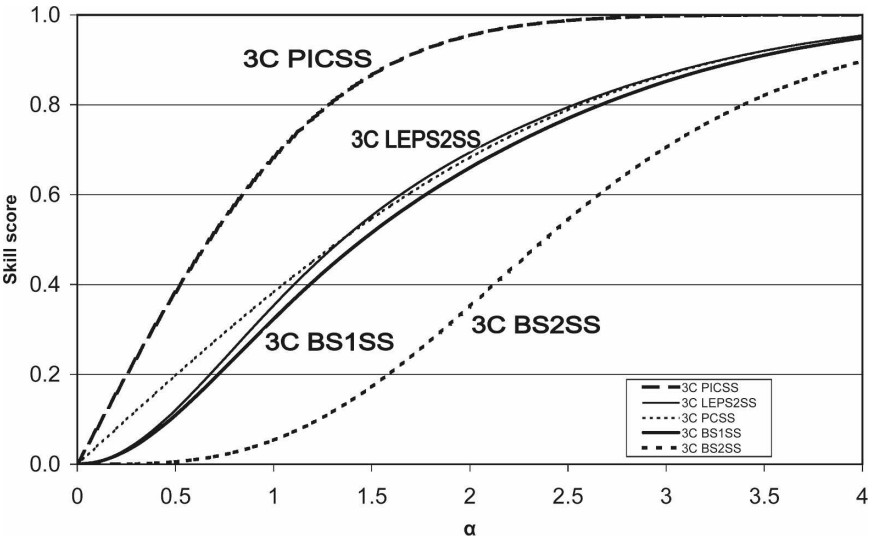

FIG. 6. Skill scores (LEPS2SS, BS1SS, BS2SS, PCSS, and PICSS) for the three-category LDA model, graphed against the conditional predictor distribution separation parameter α.

and

$$p_3 = \Pr(Y \in \text{tercile } 3 | X = x) = \frac{\rho_Z(x/\sigma - \alpha)}{\rho_Z(x/\sigma - \alpha) + \rho_Z(x/\sigma) + \rho_Z(x/\sigma + \alpha)}.$$

Obviously, $p_1 + p_2 + p_3 = 1$. By construction, this model has a symmetry with respect to x; $p_2(x) \equiv p_2(-x)$ and $p_1(x) \equiv p_3(-x)$. Asymmetric choices in the means of the conditional predictor distributions could easily remove this symmetry. Likewise, different variances in the three conditional predictor distributions could remove the symmetry.

As in the two-category case, it will be expedient to set $\sigma = 1$; this can be done without loss of generality. Hence,

$$p_1 = \frac{\rho_Z(x + \alpha)}{\rho_Z(x - \alpha) + \rho_Z(x) + \rho_Z(x + \alpha)},$$

$$p_2 = \frac{\rho_Z(x)}{\rho_Z(x - \alpha) + \rho_Z(x) + \rho_Z(x + \alpha)}, \quad \text{and}$$

$$p_3 = \frac{\rho_Z(x - \alpha)}{\rho_Z(x - \alpha) + \rho_Z(x) + \rho_Z(x + \alpha)}.$$

Following section 2b, the LEPS2 skill score is calculated using Eq. (6) as

$$\text{LEPS2SS} = \frac{9}{2} \int_{-\infty}^{\infty} \rho_X(x) s_4(p_1, p_2, p_3)\, dx,$$

the tercile 1 Brier skill score [using Eq. (2)] as

$$\text{BS1SS} = 1 - \frac{9}{2} \int_{-\infty}^{\infty} \rho_X(x) s_2(p_1)\, dx,$$

the tercile 2 Brier skill score as

$$\text{BS2SS} = 1 - \frac{9}{2} \int_{-\infty}^{\infty} \rho_X(x) s_2(p_2)\, dx,$$

the proportion correct skill score [using Eq. (7)] as

$$\text{PCSS} = -\frac{1}{2} + \frac{3}{2} \int_{-\infty}^{\infty} \rho_X(x) s_5(p_1, p_2, p_3)\, dx,$$

and the proportion incorrect skill score [using Eq. (8)] as

$$\text{PICSS} = 1 - 3 \int_{-\infty}^{\infty} \rho_X(x) s_6(p_1, p_2, p_3)\, dx.$$

These are plotted in Fig. 6, as functions of the separation parameter α. Once again, the skill score for tercile 2 (BS2SS) is quite a bit lower than the corresponding score for tercile 1 (BS1SS), with substantial skill for the middle tercile only obtainable where there is strong discrimination in the model. Symmetry considerations lead to the conclusion that the Brier skill score for tercile 3 is the same as BS1SS for this model.

4. A simple stratified climatology model

Another statistical forecasting technique used in the context of Australian weather (e.g., Stern 1980; Dahni and Stern 1995) and seasonal climate (e.g., Stone et al. 1996) prediction is the stratified climatology technique. For simplicity, we suppose just two predictor catego-

ries,[4] which are climatologically equally likely. Again, we suppose the predictand Y is standard normally distributed, but now we suppose that the predictor X is a discrete random variable, $X \in \{1, 2\}$, such that $\Pr(X = 1) = \Pr(X = 2) = 1/2$. [This situation of having Y continuous and X discrete is the opposite of the LDA case in section 3.] We could think of this in the context of an ENSO predictor such as the SOI being stratified into a positive phase and a negative phase.

We model $Y|X = 1$ as having the p.d.f. $\rho_{Y|X=1}(y) = \rho_Z(y)g_1(y)$, and $Y|X = 2$ the p.d.f. $\rho_{Y|X=2}(y) = \rho_Z(y)g_2(y)$. The functions $g_j(y)(j = 1, 2)$ must satisfy (i) $g_j(y) \geq 0$, so that the conditional p.d.f.'s are nonnegative ($\rho_{Y|X=j}(y) \geq 0$); (ii) the normalization condition $\int_{-\infty}^{\infty} \rho_Z(y)g_j(y)\,dy = 1$; and (iii) $g_1(y) + g_2(y) \equiv 2$ so that consistency with the climatological predictand distribution is preserved {i.e., $1/2[\rho_{Y|X=1}(y) + \rho_{Y|X=2}(y)] \equiv \rho_Z(y)$}.

The no-skill case comes from $g_1(y) \equiv g_2(y) \equiv 1$, for then $\rho_{Y|X=1}(y) \equiv \rho_{Y|X=2}(y) \equiv \rho_Z(y)$. That is, the two conditional distributions are identical to the climatological (unconditional) distribution, with X showing no predictability for Y. If we are interested in forecasts of the form $\Pr(Y > 0)$ or $\Pr(Y < 0)$, then perfect skill comes from having $g_1(y) \equiv 2H(-y)$ and $g_2(y) \equiv 2H(y)$ (or the converse for an oppositely signed relationship between the predictor and predictand). As before, $H(y)$ is the unit Heaviside step function. Perfect skill here means that X correctly predicts whether Y is above or below median in all cases.

To proceed further, it is necessary to specify the functions $g_1(y)$ and $g_2(y)$. So, let $g_1(y) = 2Q_Z(ay)$ and $g_2(y) = 2P_Z(ay) = 2[1 - Q_Z(ay)]$ for $a \geq 0$. The no-skill case is $a = 0$, leading to $g_1(y) \equiv g_2(y) \equiv 1$, as noted above. Perfect skill arises in the limit $a \to \infty$, with $g_1(y) \to 2H(-y)$ and $g_2(y) \to 2H(y)$. The four relevant probabilities are

$$q_1 = \Pr(Y > 0|X = 1) = \int_0^{\infty} \rho_Z(y)2Q_Z(ay)\,dy = \frac{1}{2} - \frac{1}{\pi}\arctan(a),$$

$$q_2 = \Pr(Y > 0|X = 2) = \int_0^{\infty} \rho_Z(y)2P_Z(ay)\,dy = \frac{1}{2} + \frac{1}{\pi}\arctan(a),$$

$p_1 = Pr(Y < 0|X = 1) = 1 - q_1$, and $p_2 = Pr(Y < 0|X = 2) = 1 - q_2$. The particular choice above for $g_1(y)$ and $g_2(y)$ imparts a symmetry to the model, whose consequence is $q_2 = 1 - q_1 = p_1$.

a. LEPS2 scores

If $X = 1$, the forecast attracts a LEPS2 score [from section 2a and using Eq. (1)] of $s_1(q_1) = 1/6(1 - 2q_1)^2$, while if $X = 2$, the forecast attracts a LEPS2 score of $s_1(q_2) = 1/6(1 - 2q_2)^2$. The mean LEPS2 score for the model is therefore

$$\overline{\text{LEPS2}} = \frac{1}{2}[s_1(q_1) + s_1(q_2)],$$

with a LEPS2 skill score of

$$\text{LEPS2SS} = 6\overline{\text{LEPS2}} = (1 - 2q_2)^2.$$

As a increases from 0 to ∞, q_2 increases from 1/2 to 1, and LEPS2SS increases from 0 to 1.

b. Proportion correct

In a similar way, the mean proportion correct can be calculated using Eq. (4) as

$$\overline{\text{PC}} = \frac{1}{2}[s_3(q_1) + s_3(q_2)],$$

with the proportion correct skill score as

$$\text{PCSS} = 2\overline{\text{PC}} - 1 = |1 - 2q_2|.$$

Likewise, as a increases from 0 to ∞, $\overline{\text{PC}}$ increases from 1/2 to 1 and PCSS increases from 0 to 1, with LEPS2SS = $(\text{PCSS})^2$. This relationship is plotted in Fig. 7 where it is compared with the analogous relationships of the models in sections 2a and 3a. This comparison is considered further in the following section.

An analogous three-category forecast model can be constructed with conditional predictand distributions $\rho_{Y|X=j}(y) = \rho_Z(y)g_j(y)$ $(j = 1, 2, 3)$, where $g_1(y) = 3cP_Z[a(z_{12} - y) + z_{12}]$, $g_2(y) = 3 - g_1(y) - g_3(y)$, and $g_3(y) = 3cP_Z[a(y - z_{23}) - z_{23}]$. Here, z_{12} and z_{23} are as in section 2b, a has the same interpretation as in the two-category case above, and c (which has a weak dependence on a) is the normalization constant required to satisfy condition (ii) above. Further details of this three-category model will be omitted for reasons of space.

5. Discussion

The linear regression–type (LR) model of section 2 is parameterized in terms of the correlation coefficient a, while the LDA model of section 3 is parameterized in terms of α, which dictates the degree of separation of

[4] The SOI-phase system of Stone et al. (1996) has five predictor categories.

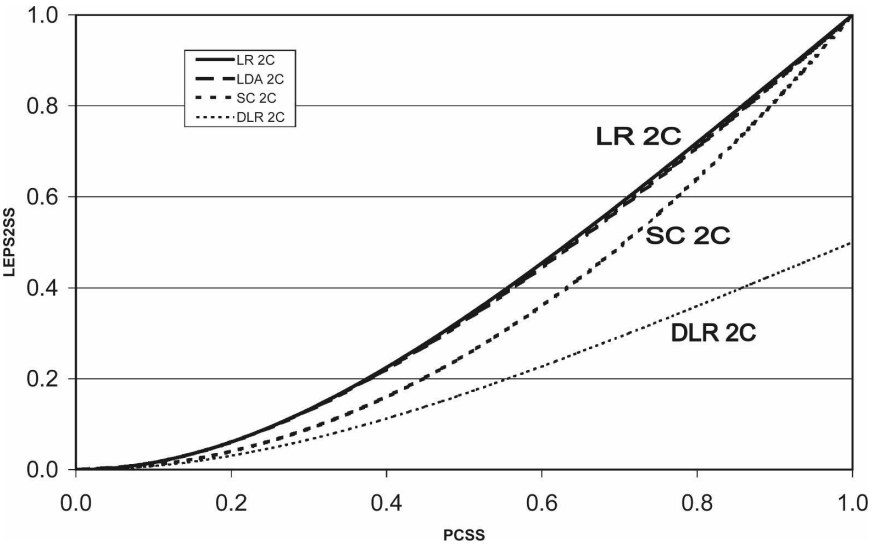

FIG. 7. Plot of LEPS2SS (vertical axis) against PCSS (horizontal axis) for the two-category linear regression model (thick solid line), the two-category linear discriminant analysis model (thick dashed line) and the degraded unreliable version (thin dotted line), and the two-category stratified climatology model (thick dashed line).

the underlying conditional predictor distributions. There is no particular connection between these two parameters, so a direct comparison between Figs. 1 and 5 (and likewise a comparison between Figs. 2 and 6) is not especially revealing. Instead, Fig. 7 plots the relationship between the two-category LEPS2SS and PCSS scores for the two models, along with the corresponding results for the degraded LR model and the stratified climatology (SC) model. In spite of the very different construction principles of the LR and LDA models, the LEPS2SS–PCSS relationships are very closely matched. The LEPS2SS–PCSS relationship for the SC model is somewhat similar, lying slightly below the corresponding LR and LDA curves, but the analogous relationship for the degraded LR model is quite different because only one of the two scores (LEPS2SS) is affected by the imposed loss of reliability. The extent to which departures from perfect reliability manifest more generally in changed relationships between the various skill scores is however beyond the scope of this paper.

The close connection between the LR and LDA model results is also seen in the three-category forecasts, again in spite of the very different construction principles. Figure 8 plots the relationships between PCSS and each of LEPS2SS, PICSS, BS1SS, and BS2SS for the two models, with very close matches between the four pairs of curves. The corresponding four curves for the three-category SC model (not shown) generally have the same relationship to those of the LR and LDA models as in the two-category case (Fig. 7), in that each of the four SC curves lies slightly below the corresponding LR and LDA curves over most of the range. The largest differences are for high values of the PIC skill score.

The results obtained here can be applied to the question of choosing suitable contours for skill maps. As has been previously noted, skill scores can show a range of different responses to the underlying strength of the relationship between the predictor and predictand, and failure to take account of this when choosing contour values can result in misleading impressions of overall skill patterns.

Suppose for example that within an above/below median forecasting context, LEPS2, proportion correct, and ROC skill scores are generated, and the set of contour values {0.05, 0.1, 0.2, 0.3} is chosen for mapping the LEPS2 skill scores. From Fig. 1, it may be concluded that if the same contour values are used for the PCSS, a stronger impression of the overall skill pattern would be obtained, with an even worse outcome if the same contour values were applied to the ROCSS. Instead, we could use Fig. 1 to select the contour choices {0.2, 0.3, 0.4, 0.5} as approximately equivalent contour values for the PCSS and {0.25, 0.35, 0.5, 0.65} as approximately equivalent contour values for the ROCSS. These three sets of contour values correspond (subject to a moderate amount of rounding to obtain "conventional" values) to approximately the same underlying correlation coefficient a of the LR model in section 2a. (The results shown in Figs. 7 and 8 indicate that the use of the LR model in this way has a more general application.) The use of these variant sets of contours would seem to be

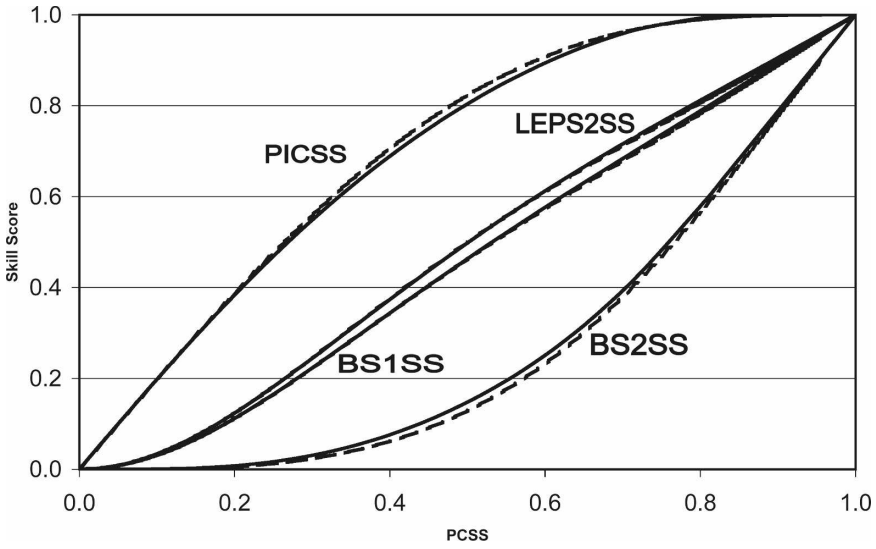

FIG. 8. Comparison of the relationships between PCSS and each of LEPS2SS, PICSS, BS1SS, and BS2SS for the three-category LR (solid lines) and LDA (dashed lines) models.

a significant improvement over the use of the same contour set for all three skill scores, together with the added benefit of a theoretical justification for the adoption of the variant sets. This idea will now be illustrated within the context of an actual forecasting example.

The Australian Bureau of Meteorology issues each month seasonal (3 month) rainfall outlooks for the following season. For example, seasonal outlooks for February–April are issued in January. The primary forecast format is the probability of the seasonal total exceeding the climatological median, but terciles probabilities are also issued, as are seasonal outlooks for seasonal mean maximum and minimum temperatures. These forecasts are verified in near-real time. [See Fawcett et al. (2005) for a description of the verification arrangements and a summary of the forecast model, which principally uses the statistical techniques of LDA and principal component analysis in its construction. Further information is available in Jones (1998) and Drosdowsky and Chambers (2001).] LEPS2 and proportion correct verification skill scores for a sequence of official seasonal rainfall outlooks covering the overlapping seasons from June–August (JJA) 2000 to November–January (NDJ) 2007/08 are shown in Figs. 9 and 10, respectively. These are mapped in correlation coefficient–equivalent units for the LR model of section 2a. That is, the data used to generate Fig. 1 were used to choose LEPS2SS and PCSS contour values, which map onto the values a = 0.0, 0.1, 0.2, . . . , 1.0 of the correlation coefficient of section 2a, with the contours in Figs. 9 and 10 being labeled with the a values. [The results shown in Figs. 7 and 8 suggest that the use of the section 2a (LR model) results, rather than the use of the less easily interpreted section 3a (LDA model) results, is not a cause for concern here.] The PCSS patterns (Fig. 10) are slightly stronger and spatially somewhat noisier than the LEPS2SS patterns (Fig. 9). The Brier skill score patterns (not shown) are intermediate in strength between the LEPS2SS and PCSS patterns, and spatially similar to the LEPS2SS patterns. Similar observations may be made about the verification results (also not shown) for the seasonal maximum and minimum temperature forecasts over a slightly longer sequence of forecasts [March–May (MAM) 2000 to NDJ 2007/08]. A particular aspect of the way in which the forecast probabilities are generated [again, see Jones (1998) and Drosdowsky and Chambers (2001) for more detail] may explain why the PCSS values are higher in correlation coefficient–equivalent units than the LEPS2SS values. Interestingly, in the rainfall forecasts at least, this discrepancy does not appear to have arisen at the expense of reliability, something that might have been suspected in light of the results of section 2c. Verification statistics in their original units over a shorter sequence of forecasts can be found in Fawcett et al. (2005).

6. Concluding remarks

This paper has compared skill scores using analytic techniques for several different simple seasonal forecast models, across varying strengths of relationships between predictor and predictand. This provides a theoretical framework for exploring how different skill scores relate to one another, and how they vary with the underlying model strength.

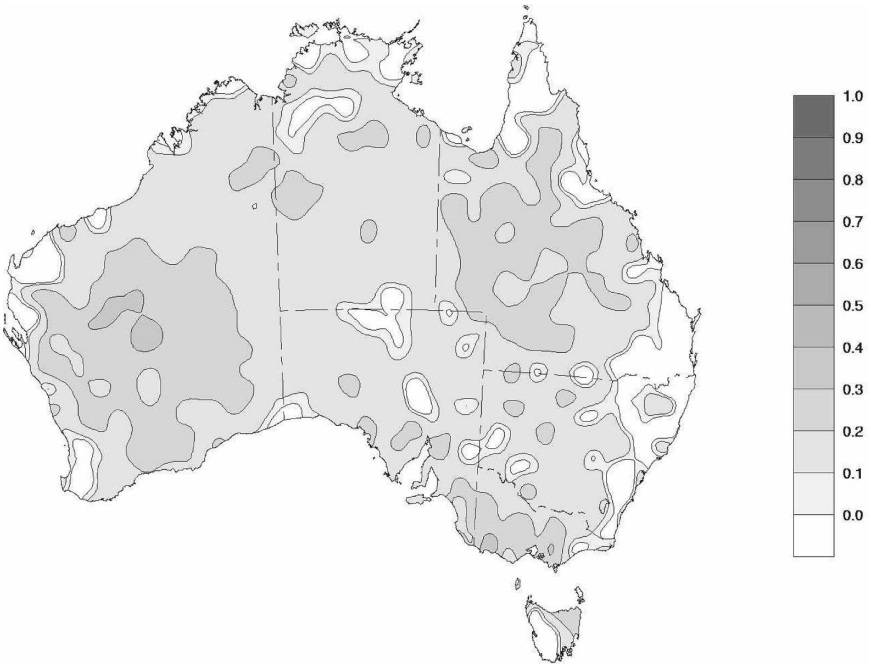

FIG. 9. LEPS2 skill scores for forecasts of Australian seasonal rainfall (all overlapping 3-month seasons from JJA 2000 to NDJ 2007/08), expressed in correlation coefficient–equivalent units. Negative values in the map are associated with negative skill scores.

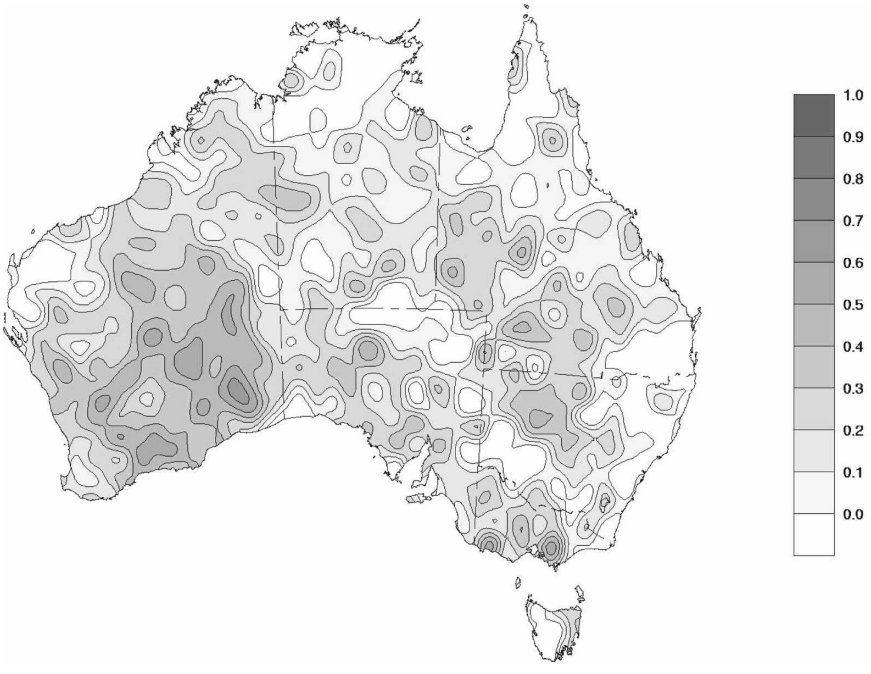

FIG. 10. The PC skill scores for forecasts of Australian seasonal rainfall (all overlapping 3-month seasons from JJA 2000 to NDJ 2007/08), expressed in correlation coefficient–equivalent units. Negative values in the map are associated with negative skill scores.

Theoretical insight is obtained into how skill scores for different types of forecasts (above median forecasts *versus* terciles forecasts) can be intercompared. Likewise, the reasons why tercile 2 forecasts are typically much less skillful than the corresponding tercile 1 and tercile 3 forecasts in real situations become clear. In a similar way, the relative predictability of the inner quantiles in finer-scale categories (e.g., quartiles or quintiles) relative to the outer quantiles could be explored within this framework.

The forecast models explored have been constructed to be symmetric with respect to both the predictor and predictand (i.e., the p.d.f.'s are symmetric about their respective means), as well as to their relationship (leading to the stated symmetry relations among the various category probabilities), in part to simplify the calculations and to permit more analytic results to be obtained. This makes the results more applicable to temperature forecasting, than to rainfall forecasting, where the asymmetries can be much more pronounced. Nevertheless, the techniques described here can be extended into asymmetric examples.

Acknowledgments. The author thanks Dr David Jones for his encouragement of the exploration of the ideas presented here, and Prof. Roger Stone for organizing the meeting at which they were first presented. The helpful suggestions from three anonymous referees are also acknowledged.

APPENDIX A

ROCS Calculation—Two-Category Cases

This appendix gives the details of the ROCS calculation from section 2a. We assume that the probabilistic forecast of an above median outcome is reinterpreted as categorical, if q exceeds a threshold probability q_0. Then,

$$\overline{H}(q_0) = \Pr(q > q_0 | Y > 0) = 2\Pr(q > q_0 \quad \text{and} \quad Y > 0).$$

Now $q > q_0$ if $x > -(b/a)Q_Z^{-1}(q_0)$, so

$$\overline{H}(q_0) = 2\int_{x=-(b/a)Q_Z^{-1}(q_0)}^{\infty}\int_{z=-ax/b}^{\infty}\rho_Z(z)\rho_Z(x)\,dz\,dx = 2\int_{-(b/a)Q_Z^{-1}(q_0)}^{\infty}Q_Z\left(-\frac{ax}{b}\right)\rho_Z(x)\,dx.$$

Similarly,

$$\overline{F}(q_0) = 2\Pr(q > q_0 \quad \text{and} \quad Y < 0),$$

which becomes

$$\overline{F}(q_0) = 2\int_{x=-(b/a)Q_Z^{-1}(q_0)}^{\infty}\int_{z=-\infty}^{-ax/b}\rho_Z(z)\rho_Z(x)\,dz\,dx = 2\int_{-(b/a)Q_Z^{-1}(q_0)}^{\infty}\left[1 - Q_Z\left(-\frac{ax}{b}\right)\right]\rho_Z(x)\,dx$$
$$= 2Q_Z\left[-\frac{b}{a}Q_Z^{-1}(q_0)\right] - \overline{H}(q_0).$$

The ROC score is the area under the curve, $\{[\overline{F}(q_0), \overline{H}(q_0)] | 0 \le q_0 \le 1\}$, which is computed as

$$\text{ROCS} = \int_{q_0=1}^{0}\overline{H}(q_0)\frac{d}{dq_0}[\overline{F}(q_0)]\,dq_0 = \int_{1}^{0}\overline{H}(q_0)\frac{d}{dq_0}\left\{2Q_Z\left[-\frac{b}{a}Q_Z^{-1}(q_0)\right] - \overline{H}(q_0)\right\}dq_0$$
$$= -\frac{1}{2} + \int_{1}^{0}\overline{H}(q_0)\frac{d}{dq_0}\left\{2Q_Z\left[-\frac{b}{a}Q_Z^{-1}(q_0)\right]\right\}dq_0.$$

The integral runs from 1 to 0 rather than from 0 to 1 because the point (0, 0) on the ROC curve corresponds to $q_0 = 1$, while the point (1, 1) corresponds to $q_0 = 0$. Application of integration by parts gives

$$\text{ROCS} = -\frac{1}{2} + 2 - \int_{1}^{0}\left\{2Q_Z\left[-\frac{b}{a}Q_Z^{-1}(q_0)\right]\right\}\frac{d}{dq_0}[\overline{H}(q_0)]\,dq_0$$
$$= \frac{3}{2} + \int_{1}^{0}\left\{2Q_Z\left[-\frac{b}{a}Q_Z^{-1}(q_0)\right]\right\}\left\{\frac{2q_0b}{a}\frac{\rho_Z\left[-\frac{b}{a}Q_Z^{-1}(q_0)\right]}{\rho_Z[Q_Z^{-1}(q_0)]}\right\}dq_0$$
$$= \frac{3}{2} - \frac{4b}{a}\int_{-\infty}^{\infty}Q_Z\left(-\frac{b}{a}u\right)Q_Z(u)\rho_Z\left(-\frac{b}{a}u\right)du,$$

using the substitution $u = Q_Z^{-1}(q_0)$. This integral is now simple enough to evaluate numerically, but an additional application of integration by parts allows this last result to be transformed into

$$\text{ROCS} = \frac{3}{2} - 2\int_{-\infty}^{\infty} \rho_Z(u) Q_Z^2\left(-\frac{b}{a}u\right) du,$$

which is marginally simpler.

APPENDIX B

ROCS Calculation—Three-Category Cases

This appendix gives the details of the ROCS calculations from section 2b. A ROC score for tercile 3 can be calculated in a manner analogous to that of the above median category (see the details in appendix A). We reinterpret the probabilistic forecast of tercile 3 as categorical if p_3 exceeds a threshold probability p_0. Then,

$$\begin{aligned}\overline{H}(p_0) &= \Pr(p_3 > p_0 | Y > z_{23}) \\ &= 3\Pr(p_3 > p_0 \quad \text{and} \quad Y > z_{23}).\end{aligned}$$

Now, $p_3 > p_0$ if $x > (z_{23}/a) - (b/a)Q_Z^{-1}(p_0)$, and $Y > z_{23}$ if $z > (1/b)(z_{23} - ax)$, so

$$\overline{H}(p_0) = 3\int_{x=\frac{z_{23}}{a}-\frac{b}{a}Q_Z^{-1}(p_0)}^{\infty} \int_{z=\frac{1}{b}(z_{23}-ax)}^{\infty} \rho_Z(z)\rho_Z(x)\, dz\, dx = 3\int_{\frac{z_{23}}{a}-\frac{b}{a}Q_Z^{-1}(p_0)}^{\infty} \rho_Z(x) Q_Z\left[\frac{1}{b}(z_{23} - ax)\right] dx.$$

Similarly,

$$\overline{F}(p_0) = \frac{3}{2}\Pr(p_3 > 0 \quad \text{and} \quad Y < z_{23}),$$

which becomes

$$\begin{aligned}\overline{F}(p_0) &= \frac{3}{2}\int_{x=\frac{z_{23}}{a}-\frac{b}{a}Q_Z^{-1}(p_0)}^{\infty} \int_{z=-\infty}^{\frac{1}{b}(z_{23}-ax)} \rho_Z(z)\rho_Z(x)\, dz\ \ dx = \frac{3}{2}\int_{\frac{z_{23}}{a}-\frac{b}{a}Q_Z^{-1}(p_0)}^{\infty} \rho_Z(x) \\ &\quad \times \left\{1 - Q_Z\left[\frac{1}{b}(z_{23} - ax)\right]\right\} dx = \frac{3}{2}Q_Z\left[\frac{z_{23}}{a} - \frac{b}{a}Q_Z^{-1}(p_0)\right] - \frac{1}{2}\overline{H}(p_0).\end{aligned}$$

The area under the ROC curve is, as before,

$$\begin{aligned}\text{ROCS} &= \int_{p_0=1}^{0} \overline{H}(p_0)\frac{d}{dp_0}[\overline{F}(p_0)]\, dp_0 = \frac{1}{2}\int_{p_0=1}^{0} \overline{H}(p_0)\frac{d}{dp_0}\left\{3Q_Z\left[\frac{z_{23}}{a} - \frac{b}{a}Q_Z^{-1}(p_0)\right] - \overline{H}(p_0)\right\} dp_0 \\ &= -\frac{1}{4} + \frac{1}{2}\int_{p_0=1}^{0} \overline{H}(p_0)\frac{d}{dp_0}\left\{3Q_Z\left[\frac{z_{23}}{a} - \frac{b}{a}Q_Z^{-1}(p_0)\right]\right\} dp_0.\end{aligned}$$

We apply integration by parts to obtain

$$\begin{aligned}\text{ROCS} &= -\frac{1}{4} + \frac{3}{2} - \frac{1}{2}\int_{p_0=1}^{0} 3Q_Z\left[\frac{z_{23}}{a} - \frac{b}{a}Q_Z^{-1}(p_0)\right]\frac{d}{dp_0}[\overline{H}(p_0)]\, dp_0 \\ &= \frac{5}{4} - \frac{1}{2}\int_{p_0=1}^{0} \left\{3Q_Z\left[\frac{z_{23}}{a} - \frac{b}{a}Q_Z^{-1}(p_0)\right]\right\}\left\{\frac{-3bp_0}{a}\frac{\rho_Z\left[\frac{z_{23}}{a} - \frac{b}{a}Q_Z^{-1}(p_0)\right]}{\rho_Z[Q_Z^{-1}(p_0)]}\right\} dp_0.\end{aligned}$$

Setting $u = Q_Z^{-1}(p_0)$ gives

$$\text{ROCS} = \frac{5}{4} - \int_{-\infty}^{\infty} \frac{9b}{2a} Q_Z\left(\frac{z_{23}}{a} - \frac{bu}{a}\right) Q_Z(u)\rho_Z\left(\frac{z_{23}}{a} - \frac{bu}{a}\right) du,$$

which is simple enough to evaluate numerically. A further (linear) change of variable, however, makes for something more numerically tractable:

$$\mathrm{ROCS} = \frac{5}{4} - \frac{9}{2}\int_{-\infty}^{\infty} Q_Z(v)p_3(v)\rho_Z(v)\,dv\,.$$

When it comes to calculating ROC scores for tercile 2, things are somewhat more complicated. We have

$$\overline{H}(p_0) = 3\mathrm{Pr}(p_2 > p_0 \quad \text{and} \quad z_{12} < Y < z_{23}).$$

The second of these conditions, $z_{12} < Y < z_{23}$, is equivalent to

$$\frac{z_{12}}{a} - \frac{ax}{b} < z < \frac{z_{23}}{a} - \frac{ax}{b}\,.$$

The first of the conditions involves

$$p_2(x) = Q_Z\left(\frac{z_{12}}{a} - \frac{ax}{b}\right) - Q_Z\left(\frac{z_{23}}{a} - \frac{ax}{b}\right) = p_2(-x),$$

which cannot readily be inverted, so let $p_2^{-1}(p_0)$ denote the functional inverse of the positive branch ($0 < p_0 < p_2(0)$, since $\overline{H}(p_0) \equiv 0$ for $p_0 > p_2(0)$. Then,

$$\begin{aligned}\overline{H}(p_0) &= 3\int_{x=-p_2^{-1}(p_0)}^{p_2^{-1}(p_0)} \int_{z=\frac{z_{12}}{a}-\frac{ax}{b}}^{\frac{z_{23}}{a}-\frac{ax}{b}} \rho_Z(z)\rho_Z(x)\,dz\,dx \\ &= 3\int_{-p_2^{-1}(p_0)}^{p_2^{-1}(p_0)} p_2(x)\rho_Z(x)\,dx.\end{aligned}$$

Similarly,

$$\begin{aligned}\overline{F}(p_0) &= \frac{3}{2}\mathrm{Pr}(p_2 > p_0 \quad \text{and} \quad (Y < z_{12} \quad \text{or} \quad Y > z_{23})) \\ &= \frac{3}{2}\int_{-p_2^{-1}(p_0)}^{p_2^{-1}(p_0)} [1 - p_2(x)]\rho_Z(x)\,dx \\ &= \frac{3}{2}\{Q_Z[-p_2^{-1}(p_0)] - Q_Z[p_2^{-1}(p_0)]\} - \frac{1}{2}\overline{H}(p_0).\end{aligned}$$

The area under the ROC curve is

$$\begin{aligned}\mathrm{ROCS} &= \int_{p_2(0)}^{0} \overline{H}(p_0)\frac{d}{dp_0}[\overline{F}(p_0)]\,dp_0 = -\frac{1}{4} + \frac{1}{2}\int_{p_2(0)}^{0} \overline{H}(p_0)\frac{d}{dp_0}\langle 3\{Q_Z[-p_2^{-1}(p_0)] - Q_Z[p_2^{-1}(p_0)]\}\rangle\,dp_0 \\ &= -\frac{1}{4} + \frac{3}{2} - \frac{1}{2}\int_{p_2(0)}^{0} \langle 3\{Q_Z[-p_2^{-1}(p_0)] - Q_Z[p_2^{-1}(p_0)]\}\rangle\frac{d}{dp_0}[\overline{H}(p_0)]\,dp_0,\end{aligned}$$

making use of integration by parts. Evaluation of the derivative gives

$$\begin{aligned}\mathrm{ROCS} = \frac{5}{4} - \frac{1}{2}\int_{p_2(0)}^{0} &\langle 3\{Q_Z[-p_2^{-1}(p_0)] \\ &- Q_Z[p_2^{-1}(p_0)]\}\rangle\left\{6p_0\frac{\rho_Z[p_2^{-1}(p_0)]}{p_2'[p_2^{-1}(p_0)]}\right\}dp_0,\end{aligned}$$

and setting $u = p_2^{-1}(p_0)$ gives

$$\mathrm{ROCS} = \frac{5}{4} - \int_0^{\infty} 9p_2(u)[Q_Z(-u) - Q_Z(u)]\rho_Z(u)\,du.$$

REFERENCES

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

Dahni, R. R., and H. Stern, 1995: The development of a generalised UNIX version of the Victorian Regional Office's operational analogue statistics model. BMRC Research Rep. 47, Melbourne, VIC, Australia, 34 pp.

Drosdowsky, W., and L. E. Chambers, 2001: Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *J. Climate,* **14,** 1677–1687.

Fawcett, R. J. B., D. A. Jones, and G. S. Beard, 2005: A verification of publicly issued seasonal forecasts issued by the Australian Bureau of Meteorology: 1998–2003. *Aust. Meteor. Mag.,* **54,** 1–13.

Hartmann, H. C., T. C. Pagano, S. Sorooshian, and R. Bales, 2002: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.,* **83,** 683–698.

Jones, D. A., 1998: The prediction of Australian land surface temperatures using near global sea surface temperature patterns. BMRC Research Rep. 70, Melbourne, VIC, Australia, 44 pp.

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

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

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

Potts, J. M., C. K. Folland, I. T. Jolliffe, and D. Sexton, 1996: Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *J. Climate,* **9,** 34–53.

Stern, H., 1980: A system for automated forecasting guidance. *Aust. Meteor. Mag.,* **28,** 141–154.

Stone, R. C., G. L. Hammer, and T. Marcussen, 1996: Prediction of global rainfall probabilities using phases of the Southern Oscillation index. *Nature,* **384,** 252–255.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 2nd ed. Academic Press, 467 pp.