Do statistical models trade resolution for reliability?

S. J. Mason

International Research Institute for Climate and Society, The Earth Institute of Columbia University Palisades, NY, USA

Statistical models are widely used to calibrate and recalibrate predictions from dynamical model. In this paper “calibration” refers to the correction of errors in the mean and/or the variance of the predictions, whereas “recalibration” involves correction for the skill of the predictions by adjusting the signal (Mason, 2008). The extent to which statistical post-processing schemes successfully improve the quality of dynamical model predictions is considered. Quality is defined here in terms of reliability, resolution and discrimination (Stephenson, 2012). Formally, a forecast system has resolution when the marginal distribution of the outcomes is conditioned on the forecast; it has discrimination when the marginal distribution of the forecasts is conditioned on the outcome. It can be argued that resolution and discrimination are the fundamental properties of good forecasts, since in their absence there is no useable information. Reliability is achieved when the expected value of the observations, conditioned upon the forecast, is equal to the forecast for all forecast values. Most statistical correction procedures that are implemented to calibrate or recalibrate the model outputs on a gridbox-by-gridbox basis are actually designed primarily to address reliability. The extent to which they do achieve reliability, and what the effects might be on resolution and discrimination, are the subjects of this paper.

Let Y be a variable we wish to forecast, and let X be a forecast of Y. Let X and Y be bivariate normal: $X, Y \sim MN(\mu, \Sigma)$, where $\mu = \{\mu_X, \mu_Y\}, \Sigma = \begin{Bmatrix} \varsigma_X & c \\ c & \varsigma_Y \end{Bmatrix}$, and $c = \mathbf{cov}(Y, X)$. An example of bivariate normally distributed data is provided in Figure 1, where $\mu = 0$ and $\Sigma = \begin{Bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{Bmatrix}$. If the parameters of the bivariate normal distribution describing X and Y are known, then the values of various category-based verification scores can be defined as a function of Pearson’s product-moment correlation, ρ (where $\rho = c/\varsigma_X \varsigma_Y$), between the forecasts and the actual values. For example, consider the case where there are two equi-probable categories, so an event occurs when $Y > \mu_Y$ and a warning is issued when $X > \mu_X$. Define a hit when $X > \mu_X$ & $Y > \mu_Y$, a correct rejection when $X < \mu_X$ & $Y < \mu_Y$, a miss when $X < \mu_X$ & $Y > \mu_Y$, and a false-alarm when $X > \mu_X$ & $Y < \mu_Y$. Since the bivariate normality assumption implies $\Pr(X > \mu_X) = \Pr(Y > \mu_Y) = 0.5$, the probability of a hit is the same as the probability of a correct-rejection, while the probability of a miss is the same as the probability of a false-alarm. In fact, each of these probabilities can be calculated from the corresponding tail areas of the bivariate-normal distribution. The probability of a hit, for example, is the right tail area of the distribution and is calculated as:

$$\Pr\left(X > \mu_x, Y > \mu_y\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \times$$
$$\int_{\mu_Y}^{\infty}\int_{\mu_X}^{\infty} \exp\left(-\frac{1}{2\left(1-\rho^2\right)}\left(\frac{\left(X \sqrt{\mu_X}\right)^2}{\varsigma_X^2} - 2\rho\frac{\left(X-\mu_X\right)\left(Y-\mu_Y\right)}{\varsigma_X\varsigma_Y} + \frac{\left(Y-\mu_Y\right)}{\varsigma_Y^2}\right)\right) dXdY \tag{1}$$

which simplifies to

$$\Pr\left(X > \mu_X, Y > \mu_Y\right) = \frac{1}{2} \times \left(\frac{1}{2} + \frac{\sin^{-1}\left(\rho\right)}{\pi}\right) \tag{2}$$

(Kotz et al. 2000). Similarly, the false-alarm rate simplifies to

$$\Pr\left(X > \mu_X, Y > \mu_Y\right) = \frac{1}{2} \times \left(\frac{1}{2} - \frac{\sin^{-1}\left(\rho\right)}{\pi}\right). \tag{3}$$

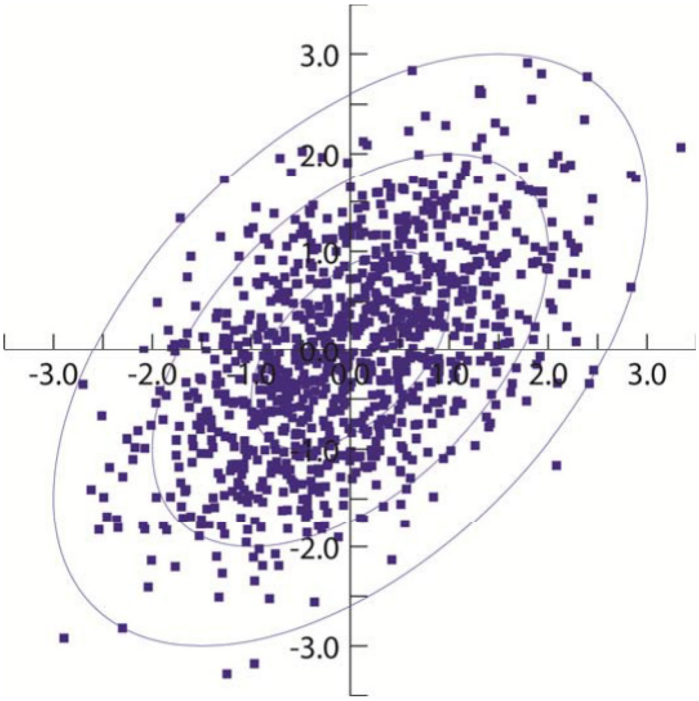

Figure 1: Example of bivariate normally distributed data given ρ = 0.5. The thin blue lines are ellipses of equal density.

Given Eqs (2) and (3), the two-category scores listed in Table 1 (and selected from Table 3.3 of Hogan and Mason (2012)) can be defined purely as a function of ρ (note that some of the scores are identical because of the constraints imposed by the assumptions). A corresponding version of Eq. (1) for categories that are not defined by the mean and are not necessarily unbounded (as is the case when three equi-probable categories are used, for example) does not simplify because the integrals cannot be defined in closed form (Divgi, 1979). However, polynomial approximations to Eq. (1) allow it to be calculated with a high degree of accuracy. For example, the hit rates for three equi-probable categories are shown in Figure 2 as a function of the correlation. The hit rates are the same for the two outer categories, but the score for the middle ("near-normal") category is lower when $0 < \rho < 1$, and remains near its minimum except when the correlation is very strong. The effect is that the values of scores for the "near-normal" category are inevitably weak unless the correlation between the forecasts and the observations is very strong. This result is purely an effect of the shape of the bivariate normal density, and provides a mathematical reason for the low skill in predicting the "near-normal" category (van den Dool and Toth, 1991).

Table 1: Values of two-category verification scores as a function of the correlation, ρ, for cases when the probability of a warning and the probability of an event are both 0.5, and the predictand and predictor are bivariate-normal.

Score	Value
Hit Rate, H Proportion Correct, PC ROC area	$\dfrac{\sin^{-1}(\rho)}{\pi}+\dfrac{1}{2}$
False-Alarm Rate, F	$\dfrac{1}{2}-\dfrac{\sin^{-1}(\rho)}{\pi}$
Critical Success Index, CSI	$\dfrac{\pi+2\sin^{-1}(\rho)}{3\pi-2\sin^{-1}(\rho)}$
Gilbert Skill Score, GSS	$\dfrac{\sin^{-1}(\rho)}{\pi-\sin^{-1}(\rho)}$
Heidke Skill Score, HSS Pierce Skill Score, PSS Clayton Skill Score, CSS Doolittle Skill Score, DSS	$\dfrac{2}{\pi}\sin^{-1}(\rho)$
Odds Ratio Skill Score, ORSS	$\dfrac{4\pi+\sin^{-1}(\rho)}{\pi^2+4\left(\sin^{-1}(\rho)\right)^2}$

The values of the scores indicated in Table 1 all assume that the parameters μ_X and μ_Y are known (ς_X and ς_Y do not affect the scores in the case of two-category forecasts). If either parameter is estimated incorrectly then the assumption that Pr $(X > \mu_X)$ = Pr$(Y > \mu_Y)$ = 0.5 is no longer valid, and the scores are affected. Errors in estimating μ_X and μ_Y introduce a mean-bias into the predictions, which translate into errors in the base-rate and/or the forecast rate in the two-category set-up. The effects on the scores of varying errors in the forecast rate when ρ = 0.5 are shown in Figure 3. The hit rate (dark blue line) and false-alarm rate (pink line) both increase from 0.0 to 1.0 as the forecast rate increases, but when the base rate is 0.5 (Figure 3a) the difference between the two (which is measured by the Pierce Skill Score (grey line)) is maximised when the bias is zero. All the other scores in Figure 3a decrease as the bias increases (away from the vertical dotted line), except the Clayton and Odds Ratio Skill Scores (green line and orange lines) as discussed by Hogan and Mason (2012), and the Critical Success Index (yellow line), which is inequitable and can be hedged (Mason, 1989). Of greatest importance in the current context is the fact that the scores that measure discrimination, namely the Pierce Skill Score and the ROC area (grey and red lines, respectively), are optimised in unbiased forecasts. However, these scores are not optimized (in fact, none of the scores are optimized) if the base rate is not 0.5 (Figure 3b), implying that calibrating dynamical model predictions for errors in the mean could actually result in a deterioration in discrimination, depending on the underlying correlation and how the data are categorized.

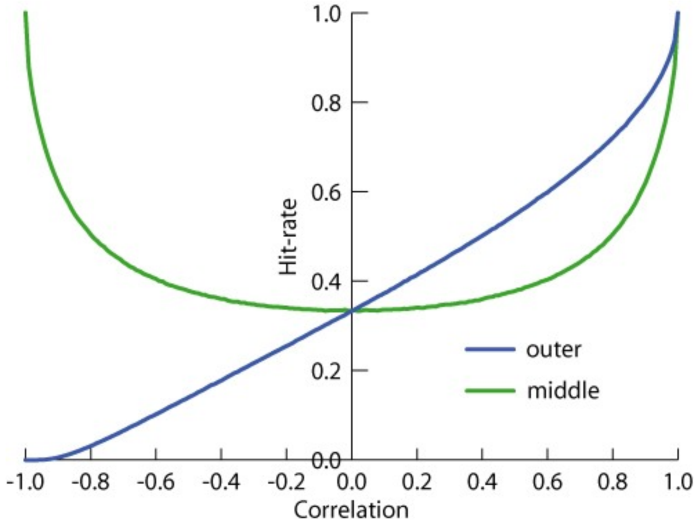

Figure 2: Hit rates for equi-probable three-category forecasts and observations as a function of the correlation for bivariate normally distributed data.

Instead of converting the forecasts to deterministic categorical predictions, probabilistic estimates of $Y > t$, where t is a threshold of interest, can be obtained from least-squares estimates of Y. Forecast probabilities can be calculated using

$$\rho = \Pr\left(Y > t; \hat{y}\right) = \frac{1}{\varsigma_Y \sqrt{2\pi\left(1-\rho^2\right)}} \int_t^\infty \exp\left(\frac{\left(u - \hat{Y}\right)^2}{2\varsigma_Y^2\left(1-\rho^2\right)} \right) du \tag{4}$$

where $\hat{Y}$ is a least squares estimate of Y:

$$\hat{Y} = \mu_Y - \rho \frac{\varsigma_Y}{\varsigma_X}\left(\mu_X - X\right) \tag{5}$$

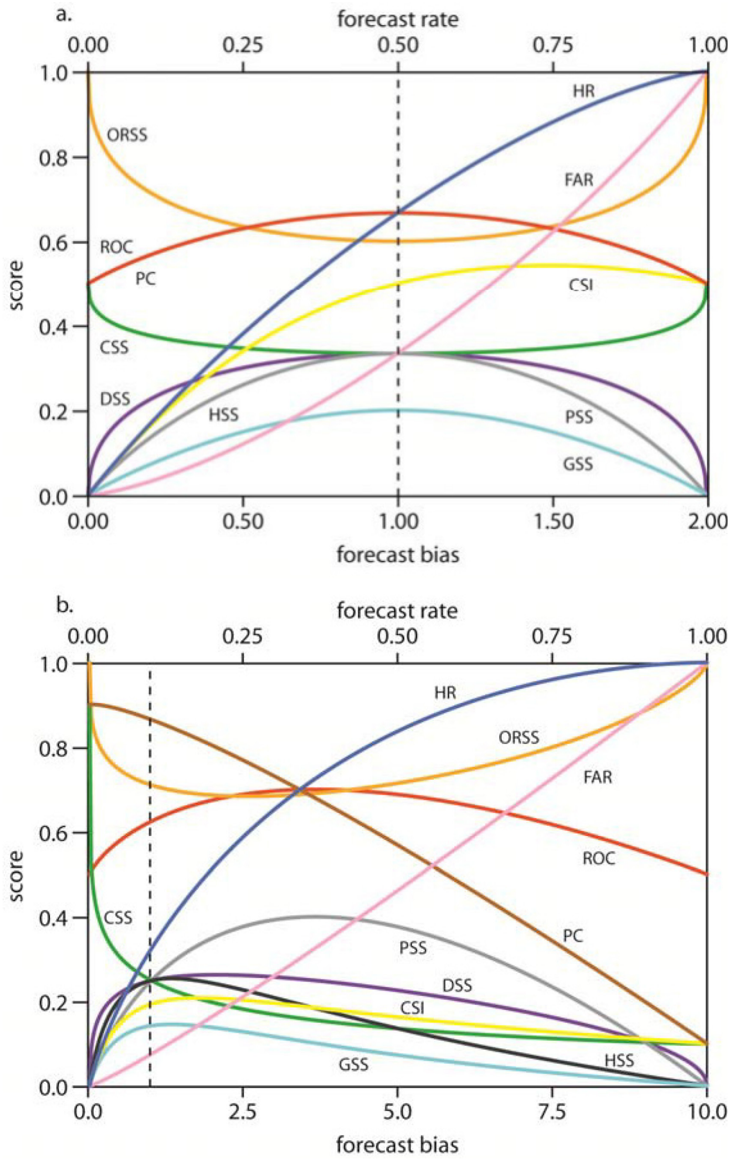

Figure 3: Values of two-category verification scores as a function of the forecast bias or forecast rate when the correlation is 0.5, and the base rate is (a) 0.5 and (b) 0.1, assuming the predictand and predictor are bivariate-normal. The dashed vertical line indicates unbiased forecasts (forecast bias is calculated as the forecast rate divided by the base rate). The score abbreviations are indicated in Table 1. Note that the curve for the ROC Area in (a) is the same as that for Percentage Correct (red line), and the curve for the Pierce Skill Score is the same as that for the Heidke Skill Score (grey line).

If $\rho = 0, \hat{Y} \mu_Y$ regardless of the value of *X*, and *p* is the climatological probability for all *X*, so the centred forecasts have no resolution, but do have perfect reliability. If $|\rho|$=1, Eq. (4) is not strictly defined, but $\hat{Y} = Y$ regardless of the value of *X*, and *p* = 0.0 when $\hat{Y} < t$ and *p* = 1.0 when $\hat{Y} > t$, so the forecasts have maximum resolution and perfect reliability.

In practice, 0 < $|\rho|$ < 1, and the distribution of *p* approximates a beta distribution (Richardson 2001). For $t = \mu_Y$, *p* has a symmetric distribution. Some examples of the distribution of *p* given different values of ρ are shown in Figure 4. In the special case of

$\rho = 1/\sqrt{2}$ (orange line), p has a uniform distribution. If $\rho > 1/\sqrt{2}$ then the distribution of the probabilities is U-shaped, but is unimodal (with mode 0.5) otherwise.

Because the distribution is symmetric, the mean forecast probability is 0.5, and the sharpness of the forecasts can be measured by the variance or standard deviation (Figure 5). The variance is

$$\operatorname{var}[p] = \frac{1}{4} - \frac{1}{\pi}\tan^{-1}\sqrt{\frac{1-\rho^2}{1+\rho^2}}. \tag{6}$$

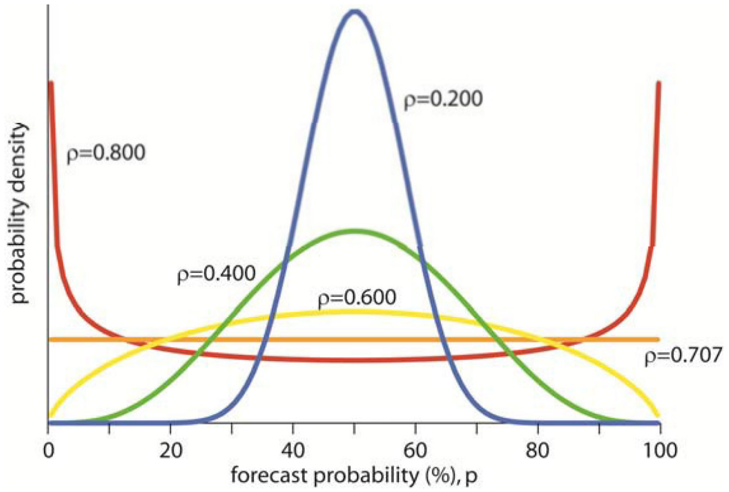

Figure 4: Distributions of forecast probabilities given least squares predictions, where the predictions and observations are bivariate normally distributed, with correlation *ρ*.

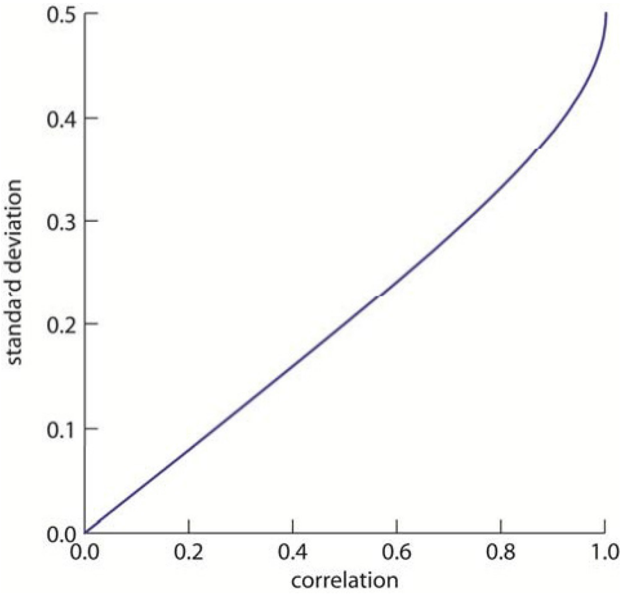

Figure 5: Standard deviation of forecast probabilities derived from least squares predictands, given different values of the correlation.

Let $p_{\hat{Y}}$ be the forecast probability for an event given least squares forecast $\hat{Y}$. The (half-) Brier score (Broecker 2012), *s*, is then defined as

$$s = \int_{-\infty}^{t} f_{\hat{Y}} p_{\hat{Y}}^{2} d\hat{Y} + \int_{t}^{\infty} f_{\hat{Y}} \left(1 - p_{\hat{Y}}\right)^{2} d\hat{Y}, \tag{7}$$

where *d* is the density of the forecasts at $\hat{Y}$. In the case of $t = \mu_Y$, Eq. (7) simplifies to

$$s = \frac{1}{\pi} \tan^{-1} \sqrt{\frac{1-\rho^2}{1+\rho^2}} \tag{8}$$

This relationship between the Brier score and the correlation is shown in Figure 6.

As with the deterministic scores, if the parameters μ_X and μ_Y are not known exactly then errors in calibration or recalibration will affect the probabilistic scores. Errors in estimating μ_X and μ_Y introduce a mean-bias into the predictions, which translate into a systematic increase or decrease in forecast probabilities leading to over- or under-forecasting (Wilks 2011. The distribution of the forecast probabilities is no longer symmetrical (Fig. 7). The effects on the Brier skill score are shown in Figure 8, which indicates a rapidly increasing loss of skill with increasing correlation. The loss of skill is entirely because of poor reliability in the forecasts (the resolution is unchanged).

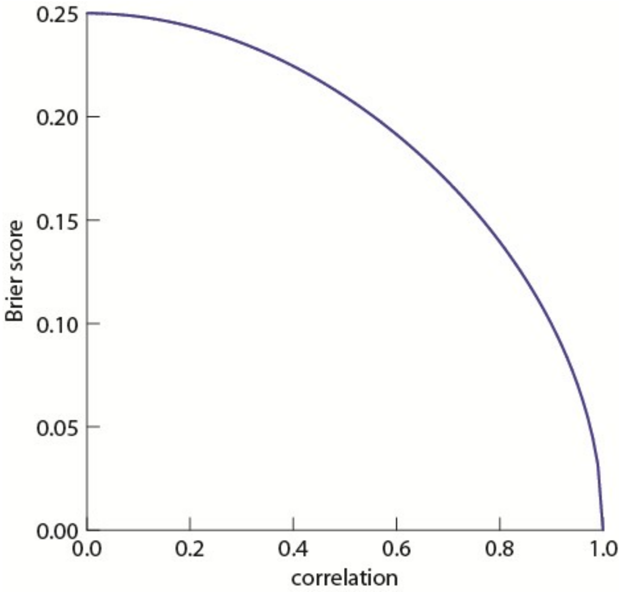

Figure 6: Brier scores given forecast probabilities for a positive anomaly where the predictor and predictand are bivariate normal. The forecast probabilities are derived from least squares predictions given different values of the correlation.

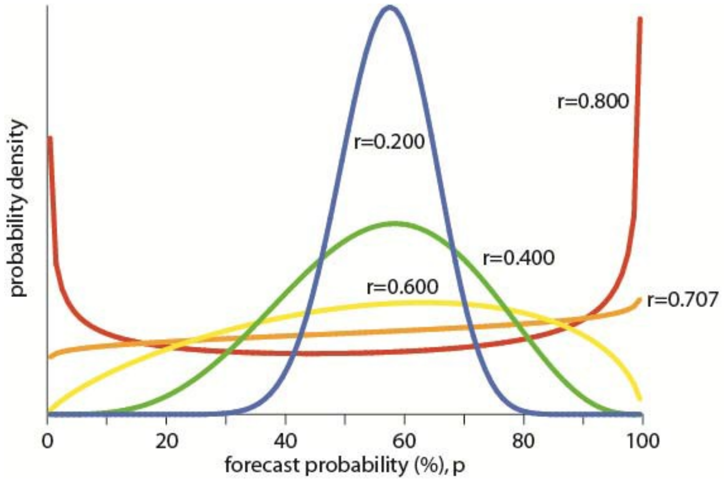

Figure 7: Distributions of forecast probabilities given least squares predictions, where the predictions and observations are bivariate normally distributed, with correlation ρ, and where the mean forecast is biased by one standard error assuming a sample size of 30.

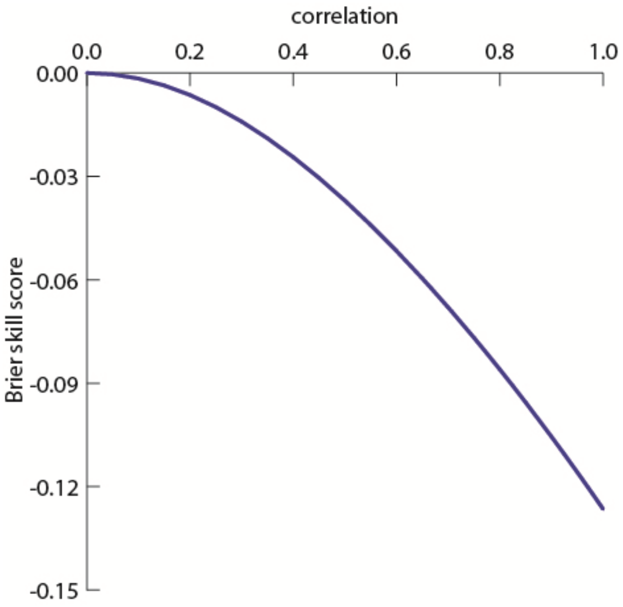

Figure 8: Brier skill scores given forecast probabilities for a positive anomaly where the predictor and predictand are bivariate normal, with correlation ρ, and where the mean forecast is biased by one standard error assuming a sample size of 30. The forecast probabilities are derived from least squares predictions.

If the parameters ς_X and ς_Y and/or the covariance, c , are not known exactly then probabilistic scores will again be affected. Consequent errors in estimating ρ introduce a conditional-bias into the predictions, which translate into a systematic increase or decrease in the sharpness of the forecasts leading to over- or under-confidence (Wilks 2011), but the distribution of the forecast probabilities remains symmetrical. The effects on the Brier score are shown in Figure 9, which again is a reflection of a loss of reliability since the resolution of the forecasts is unchanged. The reliability deteriorates given strong correlations because of the artificially high sharpness of the forecasts; and, for the same reason, the deterioration is most marked for cases where the strength of the correlation is over-estimated (blue line).

In summary, sampling errors in estimating the parameters of a model for calibrating or recalibrating forecasts result in a deterioration of forecast quality for most binary deterministic verification scores when the categories are equiprobable, but can have complicated effects on forecasts of relatively rare events. Similarly, sampling errors introduce reliability and resolution errors into probabilistic forecasts that are reflections of over- or under-confidence and over- or under-forecasting. All these effects can be quantified if the forecasts and observations are bivariate normal.

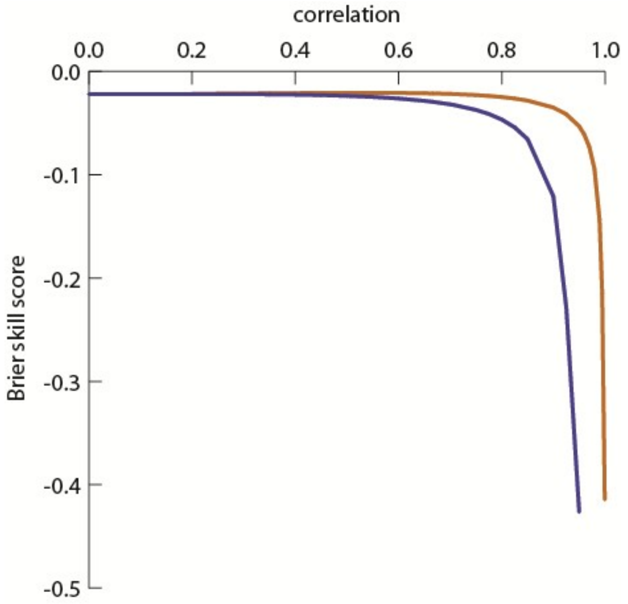

Figure 9: Brier skill scores given forecast probabilities for a positive anomaly where the predictor and predictand are bivariate normal, with correlation ρ, and where the strength of the association is positively (blue) and negatively (red) biased by one standard error assuming a sample size of 30. The forecast probabilities are derived from least squares predictions.

ACKNOWLEDGMENTS: Considerable assistance from C. A. T. Ferro is gratefully acknowledged, particularly with regard to Eqs. (6) and (8). This work was funded by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA12OAR4310086). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies.

References

Broecker, J., 2012: Probability forecasts. In I.T. Jolliffe and D.B. Stephenson (Eds), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, Chichester, 120–139.

Divgi, D.R., 1979: Calculation of univariate and bivariate normal probability functions. *Ann. Statist.*, **7**, 903–910.

Hogan, R.J., and I.B. Mason, 2012: Deterministic forecasts of binary events. In I.T. Jolliffe and D.B. Stephenson (Eds), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, Chichester, 31–59.

Kotz, S., N. Balakrishnan, and N.L. Johnson, 2000: *Continuous Multivariate Distributions: Volume 1, Models and Applications*, Wiley, Chichester, 752 pp.

Mason, I.B., 1989: Dependence of the Critical Success Index on sample climate and threshold probability. *Austr. Met. Mag.*, **37**, 75–81.

Mason, S.J., 2008: From dynamical predictions to seasonal forecasts. In Troccoli, A., M.S.J. Harrison, D.L.T. Anderson, and S.J. Mason (Eds), *Seasonal Climate Variability: Forecasting and Managing Risk*, Springer Academic Publishers, Dordrecht, 205–234.

Richardson, D.S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Met. Soc.*, **127**, 2473–2489.

Stephenson, D.B., 2012: Glossary. In I.T. Jolliffe and D.B. Stephenson (Eds), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, Chichester, 241–249.

Van den Dool, H.M., and Z. Toth, 1991: Why do forecasts for "near normal" often fail? *Wea. Forecasting*, **6**, 76–85.

Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Sciences*, Wiley, Chichester, 704 pp.