Forecasts of “normal”

Simon J. Mason[1], Christopher A. T. Ferro[2], and Willem A. Landman[3]

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

2. Department of Mathematics, University of Exeter, Exeter, United Kingdom

3. Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Pretoria, South Africa

Submitted: *Quarterly Journal of the Royal Meteorological Society*

Date: 13 October 2020

Corresponding author address: Willem A. Landman, Department of Geography, Geoinformatics and Meteorology, University of Pretoria, Private bag X20, Hatfield, 0028

South Africa

E-mail: Willem.Landman@up.ac.za

ABSTRACT

The difficulty of forecasting "normal" climate conditions is demonstrated in the context of bivariate normally distributed forecasts and observations. Deterministic and probabilistic skill scores for the normal category are less than for the outer category for all-but-perfect models. There are two important mathematical properties of the normal category in a three-category climatologically equiprobable forecast system that affect the scores for this category. First, the normal category can achieve the highest probability less frequently than the outer categories, and far less frequently in contexts of weak to moderate skill. Second, there are upper limits to the probability the normal category can reach. These mathematical constraints suggest that summary measures of skill may underestimate the predictability and forecast-skill of extreme events, and that subjective inputs to probabilistic forecasts may need to take greater account of limitations to the predictability of normal conditions.

***Key Words*:** seasonal climate forecasting; probabilistic forecasts; deterministic forecasts; normal; bivariate normal

1. Introduction

The difficulty of forecasting “normal” climate conditions at virtually any temporal or spatial scale has been long-recognised (Namias, 1964; Epstein, 1988; Toth, 1989; Livezey, 1990; Livezey *et al.*, 1990; van den Dool and Toth, 1991; Mason and Mimmack, 2002; Kharin and Zwiers 2003; Landman *et al.*, 2005, 2012; Johansson, 2007; van den Dool 2007; Kleeman, 2008; Arribas *et al.*, 2011; Manzanas *et al.*, 2014). Much of the difficulty is a result of the narrowness and boundedness of the normal category (here “normal” is taken to mean within the inter-tercile range in a three-category forecast system) compared to the outer categories (except in the case of the below-normal category for precipitation and for other random variables with absolute limits), and because skill usually is defined using reference forecast strategies that perform best when forecasting near-normal (van den Dool and Toth, 1991; van den Dool 2007).

Despite these difficulties, there may be an understandable tendency to hedge seasonal forecasts towards normal in order to avoid possible negative implications of issuing a forecast for one extreme when the opposite extreme verifies (Roulston and Smith, 2002; Dahal and Hagelmann, 2012). This hedging may be partly a function of the implicit use of inequitable verification scores (Mason, 2012), but regardless of the underlying reasons, over-forecasting of normal is prevalent in many subjectively based seasonal forecasts. For example, in the African Regional Climate Outlook Forum (RCOF) forecasts, 70 to 80% of the forecasts have highest probabilities on the normal category, and around 90% of the forecasts have probabilities on normal exceeding the climatological probability (Mason and Chidzambwa, 2008; Walker *et al.*, 2019). In the context of these RCOFs, the observed relative frequencies of normal conditions have been notably less than the forecasts have implied, which suggests that the methods used for setting probabilities in the RCOF forecasts are sub-optimal. Although there is an appropriate lack of sharpness in the probabilities on normal from the RCOFs (the probabilities are almost always between 35% and 45%), consistent with the weak skill for this category (Wilks, 2000a; Wilks and Godfrey, 2002), the over-forecasting for this category should ideally be addressed.

Over-forecasting of normal is less of an issue when objective systems are used to make real-time forecasts (e.g., the Climate Prediction Center[1], the European Centre for Medium Range Weather Forecasts[2], the International Research Institute for Climate and Society (IRI)[3]): the normal category is seldom indicated as the most likely outcome in such forecasts. Nevertheless, these systems experience problems in achieving reliability in probabilistic forecasts of normal, and IRI, at least, shows some tendency to over-forecast this category (Barnston *et al.*, 2010).

In this note, we expand on the theory of van den Dool and Toth (1991) and Kharin and Zwiers (2003) for explaining the poor predictability of the normal category. We indicate mathematically how probabilities for the normal category are constrained by the skill of the forecast system and the strength of the forecast signal, and derive mathematical limits for the sharpness of probabilistic forecasts of normal. This paper primarily addresses forecasts of the "normal" category as widely used in seasonal forecasting (Ogallo *et al.*, 2008; Mason, 2012), although the results are applicable to tercile-category forecasts at any timescale. We also derive functional relationships between Pearson's correlation as a measure of deterministic forecast skill and many category-based and probabilistic forecast verification scores, under the assumption of bivariate-normality.

[1] http://www.cpc.ncep.noaa.gov/products/NMME/prob/usPROBprate.html

[2] https://www.ecmwf.int/en/forecasts/charts/catalogue/seasonal_system5_public_standard_rain

[3] https://iri.columbia.edu/our-expertise/climate/forecasts/seasonal-climate-forecasts/

2. Idealised Forecasts

Let Y be a variable we wish to forecast, and let X be a forecast of Y. Let $X, Y \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \{\mu_X, \mu_Y\}$, $\boldsymbol{\Sigma} = \begin{Bmatrix} \varsigma_X^2 & c \\ c & \varsigma_Y^2 \end{Bmatrix}$, and $c = \operatorname{cov}(Y, X) = \rho\varsigma_X\varsigma_Y$ [i.e., X and Y are bivariate normal with expectation (μ_X, μ_Y), standard deviations ς_X and ς_Y, and correlation ρ (Forbes *et al.*, 2010)]. The parameter, ρ, is used as a measure of skill of the forecasts, but unless specified otherwise, it is not assumed that the forecasts are least squares estimates of the observations. An example of bivariate normally distributed data is provided in Figure 1, where $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \begin{Bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{Bmatrix}$. In the idealised case it is assumed that the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known.

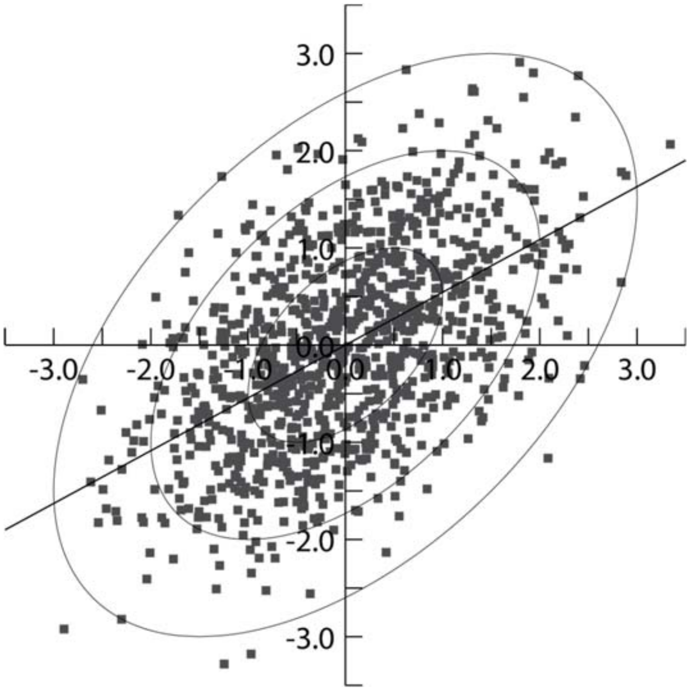

Figure 1. Example of bivariate normally distributed data given $\rho = 0.5$. The thin lines are ellipses of equal density, while the diagonal line is the least squares regression line. Note that the ellipses are oriented at 45° for all $\rho > 0.0$, whereas the regression line is oriented at $\tan^{-1}(\rho^{-1})$.

2.1 Two-Category Deterministic Forecasts

In the simplest case, consider a binary deterministic forecast system: "yes / no" forecasts are issued to indicate whether or not an event is expected to occur. An event occurs when $Y > t$, where t is a threshold of interest. If forecasts are unbiased, a warning is issued when $\frac{X - \mu_X}{\varsigma_X} > \frac{t - \mu_Y}{\varsigma_Y}$ (assuming $\rho > 0$). In the special case of $t = \mu_Y$, 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 that $\Pr\left(X > \mu_X\right) = \Pr\left(Y > \mu_Y\right) = 0.5$, 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. Each of these probabilities can be calculated from the corresponding tails of the bivariate-normal distribution. The probability of a hit (and of a correct rejection), for example, is the right tail area of the distribution and is calculated as:

$$\begin{aligned}\Pr\left(\text{hit}\right) &= \Pr\left(X > \mu_X \,\&\, Y > \mu_Y\right) \\ &= \frac{1}{2\pi\varsigma_X\varsigma_Y\sqrt{1-\rho^2}} \int_{\mu_Y}^{\infty}\int_{\mu_X}^{\infty} \exp\left(-\frac{\zeta}{2\left(1-\rho^2\right)}\right) dX\ dY\end{aligned} \tag{1a}$$

where

$$\zeta = \frac{\left(X-\mu_X\right)^2}{\varsigma_X^2} + \frac{\left(Y-\mu_Y\right)^2}{\varsigma_Y^2} - 2\rho\frac{\left(X-\mu_X\right)\left(Y-\mu_Y\right)}{\varsigma_X\varsigma_Y}. \tag{1b}$$

Eq. (1) simplifies (thankfully) 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 (and miss rate) simplifies to

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

Given Eqs (2) and (3), the two-category verification 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 (cf. Tippett et al. 2010). The forms of these relationships are illustrated in Figure 2. In all cases the scores improve monotonically as the correlation increases (the false-alarm rate is a negatively-oriented score, while all the others are positively oriented). Some scores (such as hit and false alarm rates, percent correct, ROC area, and the critical success index) are most sensitive to changes in correlation when the correlation is close to 1, whereas others are most sensitive when the correlation is close to zero (such as the odds ratio skill score). However, when compared to the Fisher-z transform of the correlation, all the scores are most sensitive when the skill is near zero.

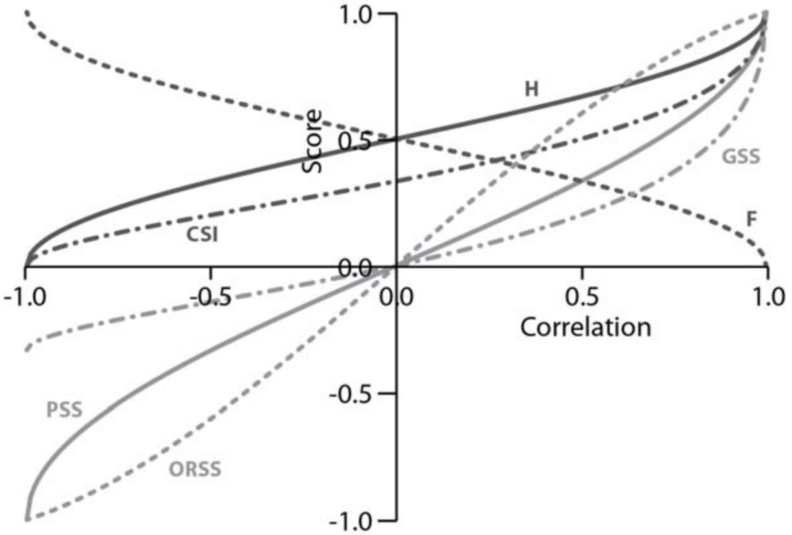

Figure 2. Verification scores for equi-probable two-category forecasts and observations as a function of the correlation for bivariate normally distributed data. The scores shown are for: the hit rate, percent correct, and ROC area (dark solid line marked H); the false alarm rate (dark dashed line marked F); the critical success index (dark dash-dotted line marked CSI); the Gilbert skill score (light dash-dotted line marked GSS); the Peirce skill score, (light solid line marked PSS); and the odds ratio skill score (light dashed line marked ORSS).

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. Some of the scores are identical because of the constraints imposed by the bivariate normality assumptions.

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

2.2 Three-Category Deterministic Forecasts

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 3 as a function of the correlation. The hit rates are the same for the two outer categories, but the score for the middle ("normal") category is lower whenever $0 < \rho < 1$, and remains near its minimum except when the correlation is very strong. The effect is that the values of scores for the normal category are inevitably weak unless the correlation between the forecasts and the observations is very strong. This result provides a mathematical reason for the low skill in predicting the normal category (van den Dool and Toth, 1991; Kharin and Zwiers 2003).

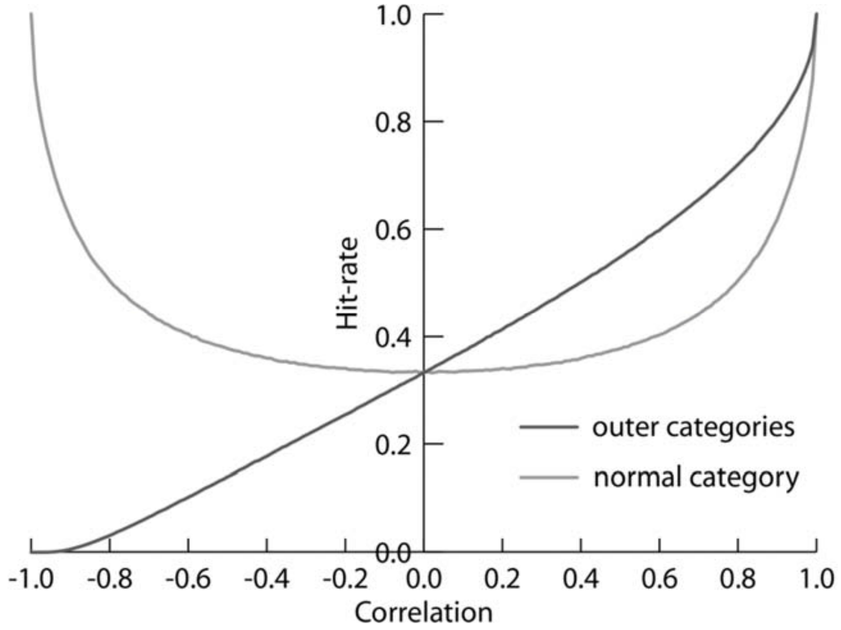

Figure 3. Hit rates for unbiased equi-probable three-category forecasts and observations as a function of the correlation for bivariate normally distributed data.

This difference in the scores between outer categories and the normal category is not specific to the hit rate. When $0 < \rho < 1$, the value for the normal category is lower than for the outer categories for all of the scores in Table 1, except for the false-alarm rate, which is the only negatively oriented score listed, i.e., the normal category scores worst on all the scores.

2.3 Two-Category Probabilistic Forecasts

Instead of converting the forecasts to deterministic categorical predictions, probabilistic forecasts can be derived from the error variance of least-squares estimates of Y. Let $\hat{Y}$ be a least squares estimate of Y:

$$\hat{Y} = \mu_Y + \rho \frac{\varsigma_Y}{\varsigma_X} \left(X - \mu_X \right). \tag{4}$$

Forecast probabilities can then be calculated using

$$P_t = \Pr\left(Y > t \mid X \right) = \frac{1}{\varsigma_Y \sqrt{2\pi\left(1-\rho^2\right)}} \int_t^\infty \exp\left(-\frac{\left(u - \hat{Y}\right)^2}{2\varsigma_Y^2\left(1-\rho^2\right)} \right) du \tag{5}$$

(Montgomery *et al.*, 2012). If $\rho = 0$, then $\hat{Y} = \mu_Y$ regardless of the value of X, and P_t is the climatological probability for all X, in which case the forecasts have no resolution, but do have perfect reliability (Wilks, 2020). If $\left|\rho\right| = 1$, Eq. (5) is not strictly defined, but $\hat{Y} = Y$ regardless of the value of X, and $P_t = 0.0$ when $\hat{Y} < t$, and $P_t = 1.0$ when $\hat{Y} > t$, so the forecasts have maximum resolution and perfect reliability.

If $0 < \left|\rho\right| < 1$, the distribution of P_t approximates a beta distribution (Richardson, 2001). For $t = \mu_Y$, P_t has a symmetric distribution. Some examples of the distribution of P_t given different values of ρ are shown in Figure 4, which indicates how frequently different forecast probabilities would be indicated for forecast models with different levels of skill. In the special case of $\rho = 1/\sqrt{2}$ (horizontal grey line), P_t 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. If $\rho = 0$, the forecast probability is always 0.5.

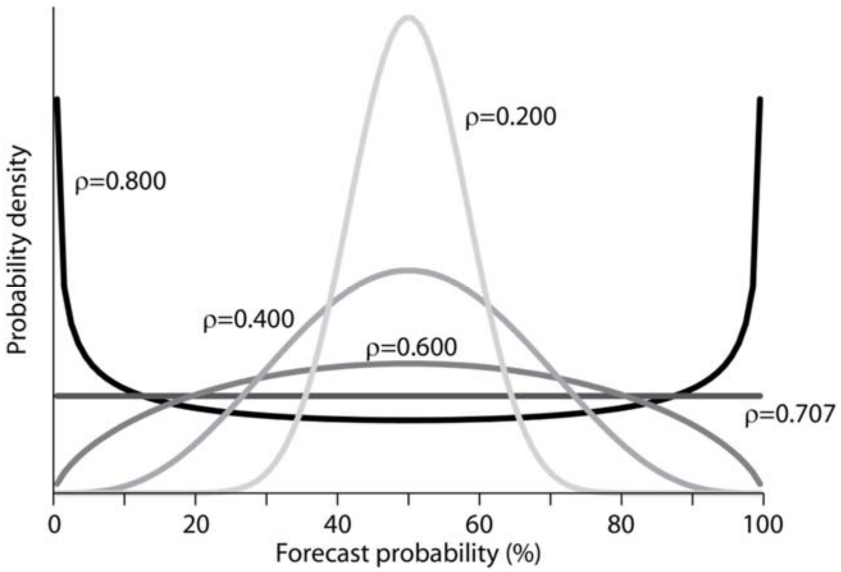

Figure 4. Distributions of forecast probabilities of the above- (or below-) median category. The predictions and observations are bivariate normally distributed, with correlation ρ defining the skill, and hence the variance, of the corresponding least-squares deterministic forecasts. The different curves are an indication of the sharpness of the forecasts from models with the indicated levels of skill, which indicates how frequently different forecast probabilities would be indicated for forecast models with different levels of skill.

The graph indicates that the sharpness of the forecasts increases as the skill of the forecasts improves (as one might expect). If the skill of the forecasts is known and the forecasts are properly calibrated, it is possible to determine how frequently certain probability thresholds will be exceeded that might be important for triggering action (Carbone and Dow, 2005; McInerny and Keller, 2008; Vizard and Anderson, 2009).

To determine the skill of the probabilistic forecasts, the (half-) Brier score (Broecker, 2012), S , can be calculated:

$$S = \left[P_t - I\left(Y > t\right) \right]^2 . \tag{6}$$

In the case of $t = \mu_Y$ and $\rho \geq 0$, Eq. (6) simplifies to

$$S = \frac{1}{\pi} \tan^{-1} \sqrt{\frac{1-\rho^2}{1+\rho^2}} \tag{7}$$

(Appendix). This relationship between the Brier score and the correlation is shown in Figure 5. The score decreases monotonically (the score is negatively oriented, and so small scores are better than large scores) from a maximum of 0.25 when the correlation is zero, to a minimum when the correlation is 1.0. This result is consistent with those for the deterministic verification scores that all indicated an improvement in the forecasts as the correlation increases. Like most of the deterministic scores shown in Figure 2, the Brier score is most sensitive to changes in correlation when the correlation is close to 1.

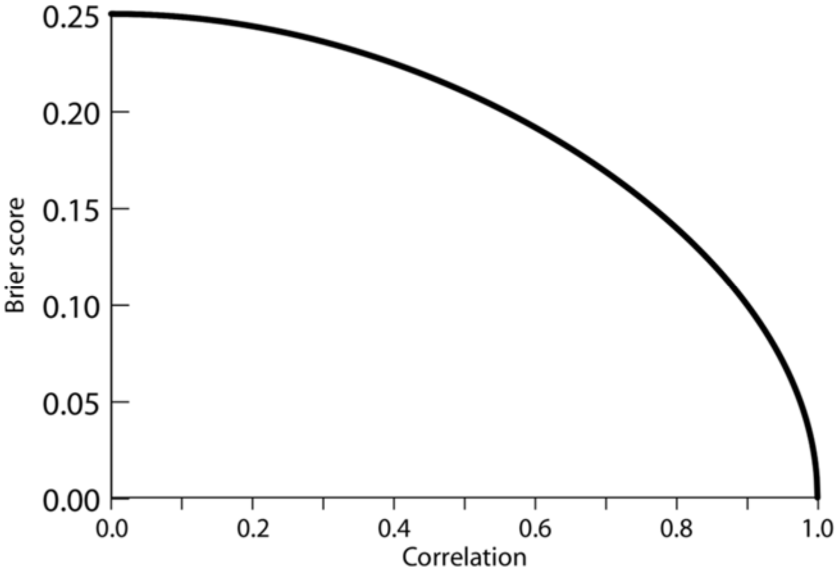

Figure 5. Brier scores given forecast probabilities for a positive anomaly where the predictor and predictand are bivariate normal. The forecast probabilities are derived from least squares predictions given different values of the correlation.

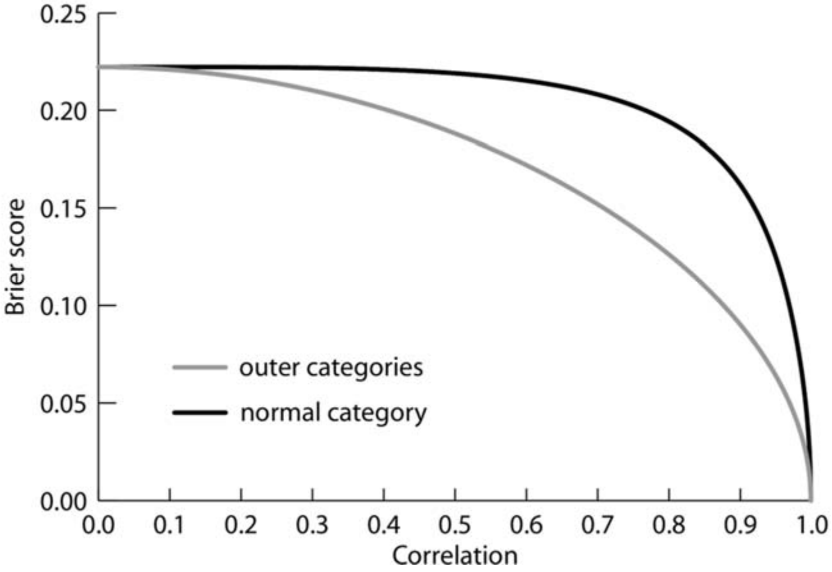

Figure 6. Brier scores given forecast probabilities for a positive anomaly where the predictor and predictand are bivariate normal. The forecast probabilities are derived from least squares predictions given different values of the correlation.

2.4 Three-Category Probabilistic Forecasts

When there are three equi-probable categories the Brier scores for the outer categories are notably less (i.e., better) than the score for the normal category for all positive correlations (Figure 6), which provides further evidence for the difficulty of forecasting the normal category. The differences in the Brier scores are an effect of substantial differences in the frequency distributions of the forecast probabilities (i.e., sharpness) for the outer categories compared to the normal category. For the below- and above-normal categories (Figure 7a) the mode of the forecast probabilities flattens progressively from 33.3% (the vertical dashed line) and the sharpness of the probabilities increases as the skill increases. However, what is most striking is the distribution of forecast probabilities for the normal category (Figure 7b): the lack of sharpness even for high-skill forecasts, is clearly apparent (Kharin and Zwiers 2003), and, perhaps most importantly, there is an upper bound to the forecast probability. This upper bound occurs when $\hat{Y} = \mu_Y$, and the probability is constrained by the skill of the forecasts, ρ. Figure

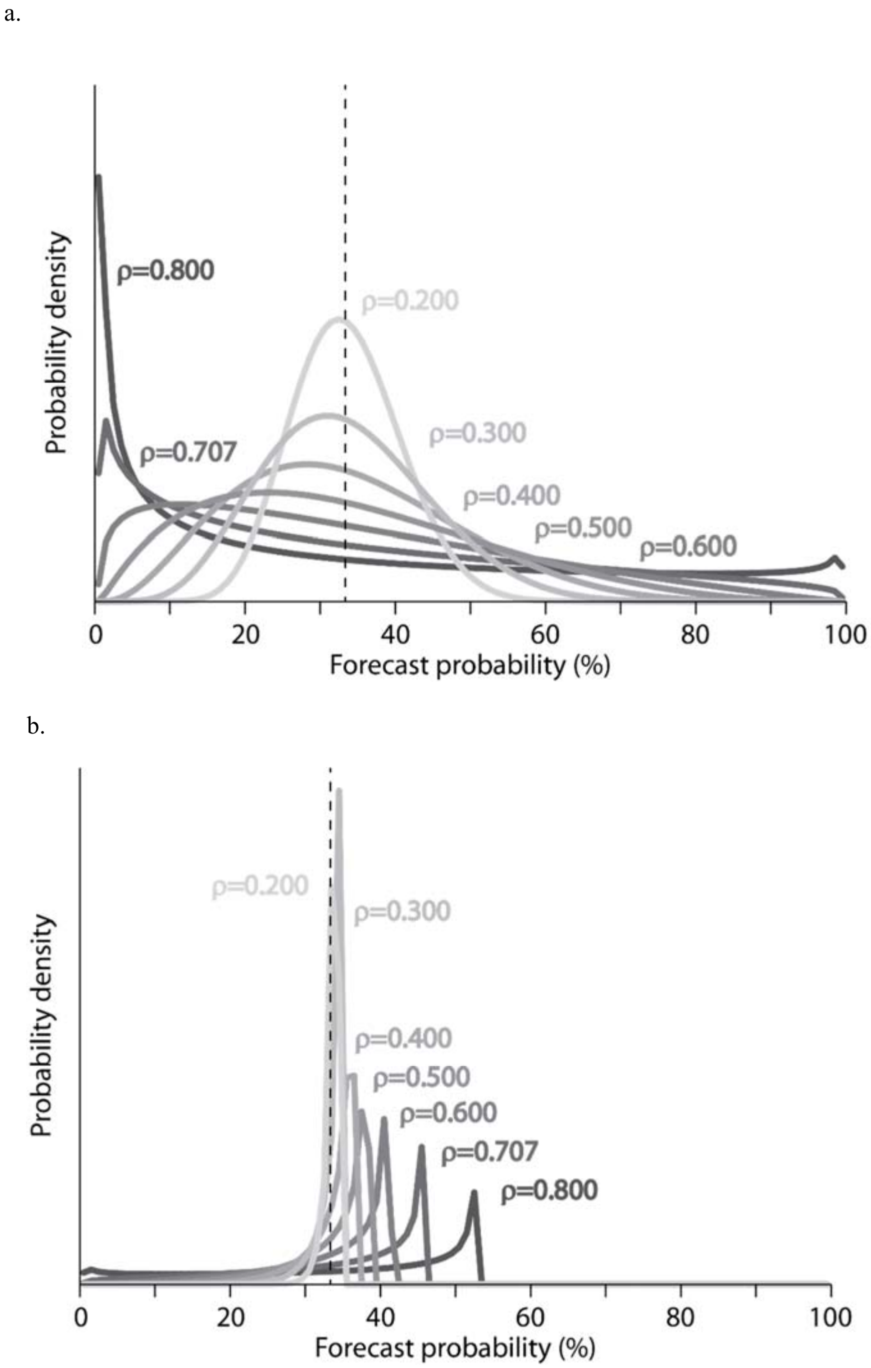

Figure 7. Distributions of forecast probabilities for the (a) outer categories, and (b) the normal category, given least squares predictions, where the predictions and observations are bivariate normally distributed, with correlation ρ . (Note that the vertical scaling on the two graphs differs.) The vertical dashed lines indicate the climatological probability.

8 shows the maximum possible forecast probability as a function of the skill. The maximum probability is less than 35% for skill levels up to about 0.47; probabilities of 40% cannot be exceeded for skill levels up to about 0.57, while a probability of 45% would require a skill above 0.64. However, these upper limits to the normal probability are only reached when the mean forecast is very close to the climatological mean.

Given the lack of sharpness and the upper bound to the normal probability shown in Figure 8 and 9, under what conditions could the normal category have the highest probability? The normal category can only be the most likely outcome when the mean deterministic forecast lies well within the range of the normal category. When the mean deterministic forecast corresponds exactly with the lower- (or upper-) tercile, the probability for below-normal (above-normal) will be 50% exactly, and the remaining probability will be divided between normal and the opposite category. But the below-normal category will have the highest probability even before its probability reaches 50%. Similarly, the above-normal category will have the highest probability when the mean forecast approaches the upper tercile, but is still within the normal category. Therefore, the normal category does not necessarily have the highest probability even when the mean forecast is within the normal category; the normal category will have the highest probability only in the less frequent case of the mean forecast being close to average.

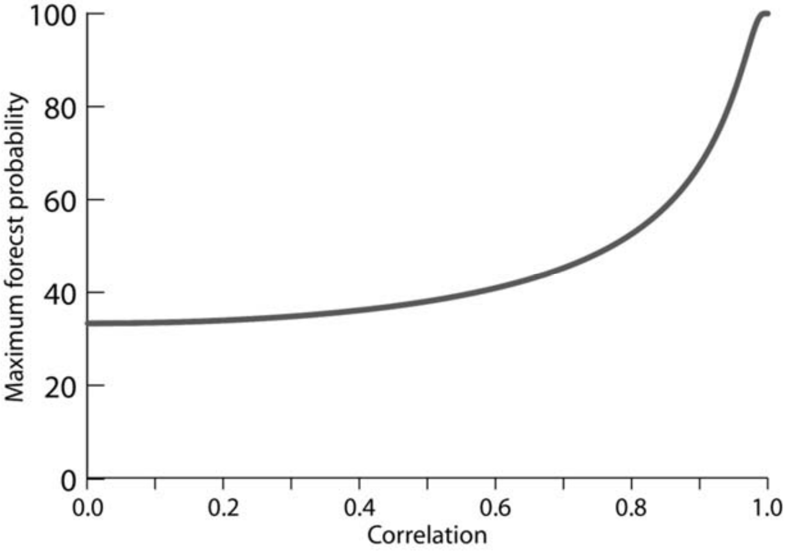

Figure 8. Maximum possible forecast probability for the normal category as a function of the correlation skill for bivariate normally distributed predictions and observations.

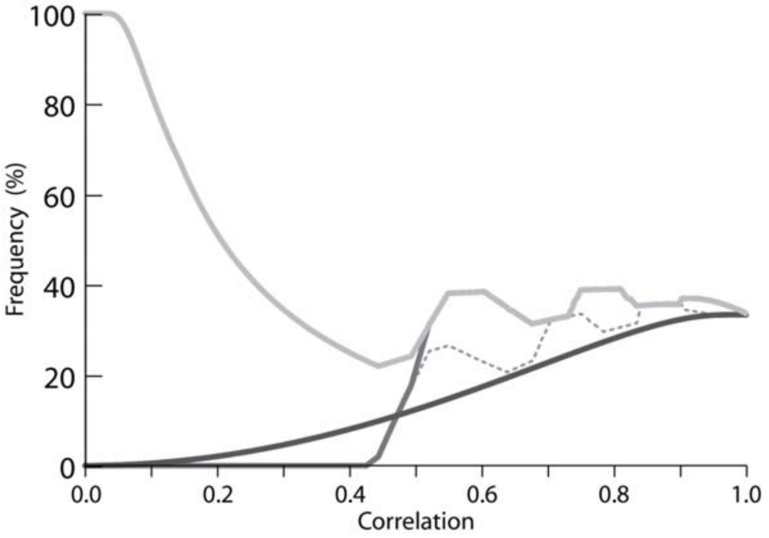

Figure 9. Relative frequency with which the normal category has the highest or equal highest probability in an equi-probable three category forecast system as a function of the correlation skill, assuming bivariate normally distributed predictions and observations, and assuming: no rounding of the forecast probabilities (dark line); rounding to the nearest 5% (light line); and rounding to the nearest 5% and resetting to climatological probabilities (in which case an equal highest probability is not counted) if the highest rounded probability is only 35% (medium line). Relative frequency with which the normal category has a higher probability than both outer categories after rounding of the forecast probabilities to the nearest 5% (dashed line).

How close to average does the mean forecast need to be for normal to have the highest probability? That depends on the skill. Assuming $\mu_Y = 0$, define $\hat{Y} = z_0$ as the point at which the normal and the above-normal categories have identical probability; at $\hat{Y} = -z_0$ the normal and the below-normal category will have identical probabilities. The normal category will then have the highest probability if the absolute value of the mean forecast, $\left|\hat{Y}\right|$, is less than z_0. Given that the variance of the mean forecast is $\rho^2 \varsigma_Y^2$, the frequency with which the mean forecast will be within this range can be calculated; the results are shown by the dark line in Figure 9. The frequency reaches a maximum of 33.3% when the forecasts are perfect, but is otherwise always less than the climatological probability. The implication is that for all imperfect forecast models, the normal category has the highest probability less frequently than the outer categories, and for models with moderate to weak skill, the normal category has the highest probability only infrequently. For example, if the correlation skill is weaker than about 0.45, the normal category has the highest probability in less than 10% of the forecasts.

If the normal category is the most likely outcome less frequently than the outer categories does that not mean that on average it is less likely to occur? Not at all! Whereas the normal category has highest probability only infrequently, whenever skill is positive it can never be the least likely category (except in special cases resulting from sampling errors when skill is weak, as discussed below); the normal category will always have higher probability than at least one of the outer categories. Figure 8 indicates that the normal category not only has high probabilities less frequently than the outer categories, but also has low probabilities less frequently. The expected probability is 33% for all three categories regardless of the skill, but for the normal category the probabilities lack sharpness.

2.5 Rounded Three-Category Probabilistic Forecasts

The results shown by the dark line in Figure 9 are based on the assumption that the forecast probability is a continuous value so that the probability for normal is never equal to that of either of the outer categories. In practice, forecast probabilities are often rounded (typically to the nearest 5% in the case of seasonal forecasts). In that case, the normal category may have the highest or equal highest probability more frequently than indicated by the dark line. In fact, after rounding[4], when the skill is weak the normal category has the highest or equal highest probability all the time (light line on Figure 9). However, in many of these cases the forecast probabilities for the normal category are tied with one of the outer categories. After rounding the probabilities to the nearest 5%, the normal category can never have higher probabilities than both outer categories unless $\rho > 0.426$ (dashed line on Figure 9). At higher skill levels, because of probability rounding, the normal category does have the highest probability slightly more than one third of the time: the maximum relative frequency for normal reaches 36%, and occurs at skill levels of about $\rho = 0.886$.

The light line on Figure 9 indicates that the normal category can have the highest or equal highest probability frequently when skill is low, but the dashed line indicates that these cases are virtually always cases of tied equal-highest probability. In fact, most of these cases are instances where the normal probability is tied at only 35%, with the other outer category having a probability of 30%. There is a widespread practice in seasonal forecasting to leave such a small shift in probability as a climatological forecast, and to only indicate a shift if the highest probability is at least 40%. In that case the normal category can have the highest or equal

[4] When rounding probabilities of three or more categories care has to be taken to ensure that the total probability is unchanged. For example, a simple rounding of 43%, 33%, and 24% would round all the values up to the nearest 5%, and the total would then be 105%. In this paper, to decide how to round these probabilities, the rounded values that lead to the least (i.e., best) expected ignorance score (Broecker, 2012) are selected (in the case of the example, 40%, 35%, 25%). Apart from being strictly proper, using the ignorance score prevents any probabilities close to zero from being rounded down to zero.

highest probability (not counting climatological probabilities as equal highest) only when $\rho > 0.426$ (medium line on Figure 9).

3. Sample Forecasts

In practice, the parameters of the bivariate normal model typically are unknown, and have to be estimated from a sample of data, quite possibly of very limited size. In this case, the forecast probabilities from Eq. (5) are no longer valid, and instead have to be derived from sample estimates of the model parameters, and the Student's *t*-distribution in place of the Gaussian distribution. The forecast probabilities become

$$P_t = \Pr\left(Y > t \mid X\right) = \frac{\Gamma\left[\left(n-1\right)/2\right]}{\Gamma\left[\left(n-2\right)/2\right]\sqrt{\left(n-2\right)\pi}} \int_t^{\infty} \left(1 + \frac{\left(u - \hat{Y}\right)^2}{\left(n-2\right) s_Y^2 \left(1 - r^2\right)}\right)^{\left(n-1\right)/2} du \qquad (8)$$

where Γ is the gamma function, n is the size of the sample, s_Y^2 is the sample variance of the observations, and r is the sample correlation (Montgomery *et al.*, 2012). Eq. (8) indicates the probability for exceeding threshold t, so probabilities for the normal category can be obtained by setting t to the lower tercile, and subtracting the probability for the above-normal category.

The effects of sample size on the maximum forecast probability for the normal category are indicated in Figure 10. The effect is small (reducing the maximum probability by less than 1% for sample sizes larger than 20), and so the effects on other results are not discussed further.

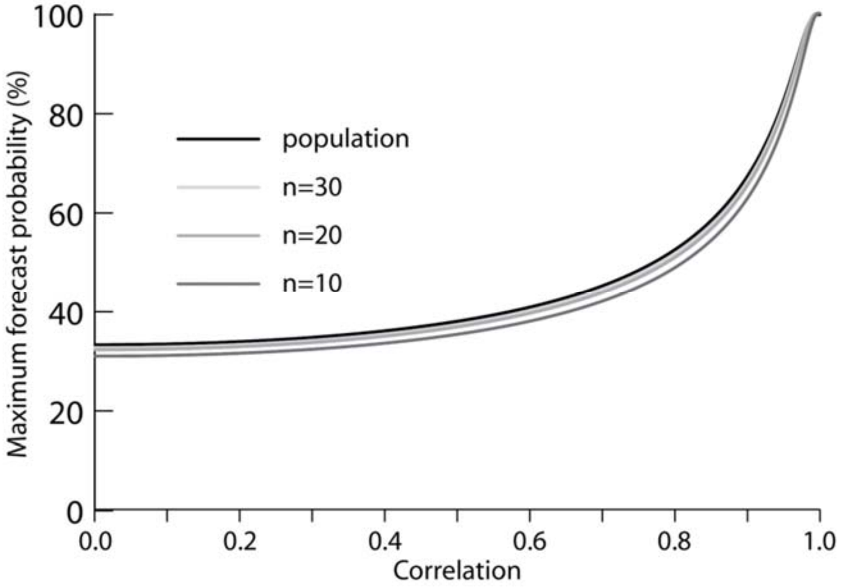

Figure 10. Maximum possible forecast probability for the normal category as a function of the correlation skill for bivariate normally distributed predictions and observations given different sample sizes, n.

4. Discussion

These results indicate that the skill of forecasting the normal category, in the context of bivariate normal forecasts and observations, is weaker than that of the outer categories because of weaker sharpness in the probabilities for normal. High probabilities for normal can occur only if the corresponding deterministic forecast is close to the climatological mean, and if the skill is well-above 0.4 (see section 2.4). This skill threshold exceeds estimates of predictability for many parts of the world (Rowell, 1998; Doblas-Reyes *et al.*, 2013; Landman *et al.*, 2019), and suggests that for at least some forecast systems normal should have the highest probability only occasionally. Objective forecast systems appear to reflect at least some of these limitations on the probability of normal: they rarely indicate highest probability on normal. The National Oceanic and Atmospheric Administration's (NOAA) seasonal forecasts explicitly limit the probabilities on normal in a manner that is essentially consistent with the results in this paper (van den Dool, 2007). Other forecast systems that use either some form of model output statistics [such as the IRI's (Barnston *et al.,* 2010)] or a purely empirical approach, will most likely include appropriate mathematical constraints on probabilities implicitly. Similarly, dynamical model forecasts that do not involve any calibration beyond a correction for the model climatology will partially reflect the limitations on normal if the ensemble distribution provides a reasonably reliable indication of the forecast uncertainty. Although there are some problems with reliability of dynamical seasonal forecasts for some parts of the world (Weisheimer and Palmer 2014), the forecasts from these models are broadly consistent with the results of this paper, perhaps partly because of physical reasons for greater predictability of the outer categories when boundary forcing is strong (Becker et al 2013). However, there are many contexts in which forecast probabilities are issued subjectively, and where these mathematical limits apparently are not respected. It is worth considering whether, and how, the mathematical results presented in this paper might be useful for informing the subjective setting of probabilities.

The primary considerations are that the results in this paper apply strictly only if the bivariate normality assumption is valid, and apply only for cases of three climatologically equiprobable categories. However, the general conclusions are likely to be relevant for other bivariate distributions and other categories bounded on both sides. Climatologically equiprobable categories are used widely in seasonal forecasting, and the results are likely to be of most relevance in that context. More specifically, empirical or recalibrated dynamical model forecasts made using least-squares regression-based procedures (e.g., canonical correlation analysis), which are widely used in research and operations, typically assume bivariate normality (Friederichs and Hense, 2003). Terciles are also used at timescales from sub-seasonal to decadal, but such forecasts and projections are made almost exclusively using dynamical models (Hamill *et al.*, 2004; Saha *et al.*, 2014; Vigaud *et al.*, 2017). While forecasts for the normal category from well-calibrated dynamical model outputs appear to display similar properties to those described in section 2 for a bivariate-normal system, they are not mathematically constrained to do so. Regardless, we might expect similar results if these model outputs are recalibrated in similar ways to the seasonal forecasts.

Strict assumptions of bivariate normality are unlikely to be met in practice, even if the forecasts and the observations are both normally distributed. However, although violations of the assumptions will change the numerical details, similar restrictions on the sharpness of probabilities for the normal category will exist in other distributions simply because of the boundedness of the normal category. Therefore, the general conclusions presented here are likely to be valid, at least as approximate guidelines, in many practical settings, including for some of the inputs used in Regional Climate Outlook Forums. It appears to be appropriate for issued forecast probabilities of the normal category to be less sharp than those of the outer categories, whether or not the relevant data have an approximately normal distribution.

The second specified consideration is that the results only apply to tercile-based categories. Where forecasts are for three categories that are not equi-probable, the most frequent situation is for the middle category to have a climatological probability that is larger than 33%. For example, many El Niño – Southern Oscillation (ENSO)-based forecasts have a middle category ("neutral" ENSO conditions) that has a climatological probability of about 50% (Barnston *et al.*, 2012). In such cases, the general conclusions still apply – the middle category will still have an upper-bound that is less than one, and the sharpness of forecasts for this category will still be restricted. If the climatological probability of the middle category is increased, it may become the category with the highest probability most frequently, but the probabilities for this category will exceed its climatological probability less frequently than for the outer categories, and only when there is positive skill and the deterministic forecast is near the mean.

5. Summary

Given bivariate normally distributed forecasts and observations, many commonly used skill scores based on deterministic and probabilistic forecasts of below- and above-average (two equi-probable categories) can be defined as functions of Pearson's correlation, ρ, between the forecasts and observations. Regardless of the correlation, the expected forecast probability is 50%, but the sharpness of the forecasts is affected by the correlation: forecast probabilities are unimodal if $\rho < 1/\sqrt{2}$, uniformly distributed if $\rho = 1/\sqrt{2}$, and u-shaped for $\rho > 1/\sqrt{2}$.

Given three equi-probable categories in this bivariate-normal setting, deterministic and probabilistic skill scores for the normal category are consistently less than for the outer category for all but perfect models. For probabilistic forecasts, the sharpness of the forecasts for the normal category is less than for the outer categories for all but perfect models; in fact, there is a mathematical upper limit to the probability on the normal category that is a function of ρ. The maximum probability is less than 35% for $\rho < 0.47$; probabilities of 40% cannot be exceeded if $\rho < 0.57$; a probability of 45% would require $\rho > 0.64$. Not only is there an upper limit to the probability on the normal category, but the normal category can have the highest probability only rarely: in less than 10% of the forecasts if $\rho < 0.45$. If probabilities are rounded to the nearest 5% the normal category can never have the highest probability if $\rho < 0.42$. Although the upper bound on the outer forecast probability is 100%, and these categories have the highest probability most of the time, for $\rho = 0.3$ the forecast probability exceeds 40% only about one time in four, and so forecasts will lack sharpness much of the time.

These mathematical results provide an analytical explanation for the widely observed poor skill in forecasting normal, and suggest that verification measures specifically for conditions that are not normal may give higher estimates of skill than do summary measures. The results also may be worthy of consideration for subjective forecasting: forecasters should be aware of limitations on the sharpness of the probabilities for the normal category when those probabilities are derived from a bivariate normal relationship

Acknowledgements

This work was funded in part by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA12OAR4310086). The views expressed are those of the authors and do not necessarily reflect the views of NOAA or its subagencies. Some of the material presented here is based upon work partly supported financially by the National Research Foundation (NRF) of South Africa through the ACyS –ACCESS project (Grant No: 114689). Helpful discussions with A. G. Barnston, L. Goddard, and A. M. Greene are gratefully acknowledged, as are constructive comments from anonymous referees.

APPENDIX

Relationship between the Brier score and Pearson’s correlation

Let $\left(X,Y\right)$ be bivariate normal with expectation $\left(\mu_X,\mu_Y\right)$, standard deviations ς_X and ς_Y, and correlation ρ (as in the main text). Define the ideal forecast for the event $\{Y>t\}$ to be

$$P_t = \Pr\left(Y>t \mid X\right) = E_{Y|X}\left\{I\left(Y>t\right)\right\}, \tag{A.1}$$

where I is the indicator function. Equation A.1 defines the forecast as not only unbiased, but also perfectly reliable; but the resolution of the forecasts, and therefore most measures of skill, are not explicitly specified. However, because of the bivariate normality assumption

$$Y \mid X \square N\left(\mu_Y + \rho {}^{\varsigma_Y}\!/_{\varsigma_X}\left(X-\mu_X\right), \left(1-\rho^2\right)\varsigma_Y^2\right), \tag{A.2}$$

and so

$$P_t = \Pr\left\{Z > \frac{t-\mu_Y-\rho {}^{\varsigma_Y}\!/_{\varsigma_X}\left(X-\mu_X\right)}{\varsigma_Y\sqrt{1-\rho^2}}\right\} = 1-\Phi\left(a+bZ\right), \tag{A.3}$$

where Φ is the distribution function of $Z \sim N\left(0,1\right)$, and where $a = \dfrac{t-\mu_Y}{\varsigma_Y\sqrt{1-\rho^2}}$ and $b = -\dfrac{\rho}{\sqrt{1-\rho^2}}$. Therefore, as in Eq. (5), the forecast probability can be defined partly as a function of the skill ρ.

The skill of probabilistic forecasts can be measured using the Brier score (Broecker, 2012), as defined in Eq. (6). Given a forecast X, the expected value of the Brier score is

$$\begin{aligned} E(S \mid X) &= P_t^2 - 2P_t E\{I(Y>t) \mid X\} + E\{I(Y>t) \mid X\} \\ &= P_t^2 - 2P_t^2 + P_t \\ &= P_t - P_t^2 \end{aligned} . \qquad \text{(A.4)}$$

The expected Brier score is therefore

$$E(S) = E(P_t) - E(P_t^2). \qquad \text{(A.5)}$$

The expectation of P_t, denoted q, is

$$q = E_X\left[E_{Y|X}\{I(Y>t)\}\right] = E_Y\{I(Y>t)\} = \Pr(Y>t) = 1 - \Phi\left(\frac{t-\mu_Y}{\varsigma_Y}\right), \qquad \text{(A.6)}$$

and the variance of P_t is

$$\begin{aligned} \operatorname{var}(P_t) &= E(P_t^2) - \{E(P_t)\}^2 \\ &= E(P_t^2) - q^2 \end{aligned} . \qquad \text{(A.7)}$$

From Eqs (A.5 and A.7):

$$E(S) = q - \operatorname{var}(P_t) - q^2. \qquad \text{(A.8)}$$

For reliable forecasts, therefore, the Brier score is a function of the mean and variance of the forecast probabilities.

Given that $\operatorname{var}(P_t) = \operatorname{var}(1-P_t)$, Eq. (A.7) can be redefined as

$$\operatorname{var}(P_t) = E\left\{(1-P_t)^2\right\} - \{E(1-P_t)\}^2, \qquad \text{(A.9)}$$

and since $\Phi(a+bZ) = 1 - P_t$ [as in Eq. (A.3)], Eq. (A.9) becomes

$$\begin{aligned}\operatorname{var}\left(P_t\right) &= E\left\{\Phi\left(a+bZ\right)^2\right\}-\left\{E\left(1-P_t\right)\right\}^2 \\ &= E\left\{\Phi\left(a+bZ\right)^2\right\}-\left(1-q\right)^2\end{aligned}. \tag{A.10}$$

According to Owen (1980)

$$E\left\{\Phi\left(a+bZ\right)^2\right\}=\Phi\left(t'\right)-2T\left(t',\theta\right), \tag{A.11}$$

where

$$t'=\frac{a}{\sqrt{1+b^2}}=\frac{t-\mu_Y}{\varsigma_Y}, \tag{A.12}$$

$$\theta=\frac{1}{\sqrt{1+2b^2}}=\sqrt{\frac{1-\rho^2}{1+\rho^2}}, \tag{A.13}$$

$$T\left(t',\theta\right)=\phi\left(t'\right)\int_0^\theta \frac{\phi\left(t'x\right)}{1+x^2}dx, \tag{A.14}$$

and ϕ is the density function of Z . As a result Eq. (A.10) can be redefined as

$$\operatorname{var}\left(P_t\right)=\Phi\left(t'\right)-2T\left(t',\theta\right)-\left(1-q\right)^2=q\left(1-q\right)-2T\left(t',\theta\right). \tag{A.15}$$

Substituting Eq. (A.15) into Eq. (A.8), the expected Brier score can be expressed as:

$$\begin{aligned}E\left(S\right) &= q-q\left(1-q\right)+2T\left(t',\theta\right)-q^2 \\ &= 2T\left(t',\theta\right)\end{aligned}. \tag{A.16}$$

In the special case of $t=\mu_Y$, $t'=0$ [Eq. (A.12)] and $\phi\left(t'\right)=1/\sqrt{2\pi}$, so

$$T\left(0,\theta\right)=\phi\left(0\right)\int_0^\theta \frac{\phi\left(0x\right)}{1+x^2}dx=\frac{1}{2\pi}\int_0^\theta \frac{1}{1+x^2}dx=\frac{\tan^{-1}\theta}{2\pi}. \tag{A.17}$$

Combining Eqs (A13), (A16) and (A17):

$$E\left(S\right)=2T\left(t',0\right)=2\frac{\tan^{-1}\sqrt{\frac{1-\rho^{2}}{1+\rho^{2}}}}{2\pi}=\frac{1}{\pi}\tan^{-1}\sqrt{\frac{1-\rho^{2}}{1+\rho^{2}}}\,, \tag{A.18}$$

as given in Eq. (7).

References

Arribas A, Glover M, Maiden A, Peterson K, Gordon M, MacLachlan C, Graham R, Fereday D, Camp J, Scaife AA, Xavier P, McLeann P, Colman A, Cusack S. 2011. The GloSea4 Ensemble Prediction System for seasonal forecasting. *Mon. Wea. Rev.* 139: 1891–1910.

Barnston AG, Li S, Mason SJ, DeWitt DG, Goddard L, Gong X. 2010. Verification of the first 11 years of IRI's seasonal climate forecasts. *J. Appl. Meteorol. Climatol.* 49: 493–520.

Barnston AG, Tippett MK, L'Heureux ML, Li S., DeWitt DG. 2012. Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing?. *Bull. Amer. Met. Soc.*, 93: 631–651.

Becker EJ, Van Den Dool H, Peña M. 2013. Short-term climate extremes: prediction skill and predictability. *J. Climate* 26: 512–531.

Broecker J. 2012. Probability forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB (eds.): 120–139. Wiley: Chichester.

Carbone GJ, Dow K. 2005. Water resource management and drought forecasts in South Carolina. *J. Amer. Water Res. Assoc.* 41: 145–155.

Dahal KR, Hagelman R. 2011. People's risk perception of glacial lake outburst flooding: a case of Tsho Rolpa Lake, Nepal. *Env Hazards – Human Pol. Dim.* 10: 154–170.

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

Doblas-Reyes FJ, García-Serrano J, Lienert F, Biescas AP, Rodrigues LRL. 2013. Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdiscip. Rev. Clim. Change* 4: 245–268.

Epstein ES. 1988. Long-range weather prediction: Limits of predictability and beyond. *Wea. Forecasting* 3: 69–75.

Forbes C, Evans M, Hastings N, Peacock B. 2010. *Statistical Distributions*. Wiley: Chichester; 230 pp.

Friederichs P, Hense A. 2003. Statistical inference in canonical correlation analyses exemplified by the influence of North Atlantic SST on European climate. *J. Climate* 16: 522–534.

Hamill TM, Whitaker JS, Wei X. 2004. Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.* 132: 1434–1447.

Hogan RJ, Mason IB. 2012. Deterministic forecasts of binary events. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB (eds.): 31–59. Wiley: Chichester.

Johansson A. 2007. Prediction skill of the NAO and PNA from daily to seasonal time scales. *J. Climate* 20: 1957–1975.

Kharin, V.V. and Zwiers, F.W., 2003. Improved seasonal probability forecasts. *Journal of Climate*, *16*(11), pp.1684-1701.

Kleeman R. 2008. Limits, variability, and general behavior of statistical predictability of the midlatitude atmosphere. *J. Atmos. Sci.* 65: 263–275.

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

Landman WA, Botes S, Goddard L, Shongwe M, 2005: Assessing the predictability of extreme rainfall seasons over southern Africa. *Geophys. Res. Lett.* 32: L23818.

Landman WA, DeWitt DG, Lee DE, Beraki A, Lötter D. 2012. Seasonal rainfall prediction skill over South Africa: 1- vs 2-tiered forecasting systems. *Wea. Forecasting* 27: 489–601.

Landman WA, Barnston AG, Vogel C, Savy J. 2019. Use of ENSO-related seasonal precipitation predictability in developing regions for potential societal benefit. *Int. J. Climatol.* 1–11. doi.org/10.1002/JOC.6157

Livezey RE. 1990. Variability of skill of long-range forecasts and implications for their use and value. *Bull. Amer. Met. Soc.* 71: 300–309.

Livezey RE, Barnston AG, Neumeister BK. 1990. Mixed analog/persistence prediction of United States seasonal mean temperatures. *Int. J. Climatol.* 10: 329-340.

Manzanas R, Frias MD, Cofiño AS, Gutiérrez JM. 2014. Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. *J. Geophys. Res. Atmos.* 119: 1708 – 1719.

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

Mason SJ. 2012. Seasonal and longer-range forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB (eds.): 203–220. Wiley: Chichester.

Mason SJ, Mimmack GM. 2002. Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate* 15: 8–29.

Mason SJ, Chidzambwa S. 2008. Verification of African Regional Climate Outlook Forum forecasts. doi.org/10.7916/D85T3SB0.

McInerney D, Keller K. 2008. Economically optimal risk reduction strategies in the face of uncertain climate thresholds. *Clim. Change* 91: 29–41.

Montgomery DC, Peck EA, Vinning GG. 2012. *Introduction to Linear Regression Analysis*. Wiley: Chichester; 672 pp.

Murphy AH. 1991. Forecast verification: its complexity and dimensionality. *Mon. Wea. Rev.* 119: 1590–1601.

Namias J. 1964. A 5-year experiment in the preparation of seasonal outlooks. *Mon. Wea. Rev.* 92: 449–464.

Ogallo LJ, Bessemoulin P, Ceron JP, Mason SJ, Connor SJ. 2008. Adapting to climate variability and change: the Climate Outlook Forum process. *J. World Meteor. Org.* 57: 93–102.

Owen DB. 1980. A table of normal integrals. *Commun. Stat.: Simul. Comput.* 9: 389–419.

Richardson DS. 2001. Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.* 127: 2473–2489.

Roulston MS, Smith LA. 2004. The boy who cried wolf revisited: The impact of false alarm intolerance on cost-loss scenarios. *Wea. Forecasting* 19: 391–397.

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

Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, Behringer D, Hou YT, Chuang HY, Iredell M, Ek M. 2014. The NCEP climate forecast system version 2. *J. Climate* 27: 2185–2208.

Stephenson DB. 2012. Glossary. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Jolliffe IT, Stephenson DB (eds.): 241–249. Wiley: Chichester.

Tippett MK, Barnston AG, DelSole T. 2010. Comments on "Finite samples and uncertainty estimates for skill measures for seasonal prediction". *Mon. Wea. Rev.* 138, 1487–1493.

Toth Z. 1989. Long-range weather forecasting using an analog approach. *J. Climate* 2: 594-607.

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

Van den Dool HM. 2007 *Empirical Methods in Short-term Climate Prediction*. Oxford University Press: Oxford; 240 pp.

Vigaud N, Robertson AW, Tippett MK. 2017. Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.* 145, 3913–3928.

Vizard AL, Anderson GA. 2009. The resolution and potential value of Australian seasonal rainfall forecasts based on the five phases of the Southern Oscillation Index. *Crop Pasture Sci.* 60: 230–239.

Walker DP, Birch CE, Marsham JH, Scaife AA, Graham RJ, Segele ZT. 2019. Skill of dynamical and GHACOF consensus seasonal forecasts of East African rainfall. *Climate Dyn.*, https://doi.org/10.1007/s00382-019-04835-9.

Weisheimer A, Palmer TN. 2014. On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface* 11: 20131162.

Wilks DS. 2000a. Diagnostic verification of the Climate Prediction Center long-lead outlooks, 1995–98. *J. Climate* 13: 2389–2403.

Wilks DS. 2000b. On interpretation of probabilistic climate forecasts. *J. Climate* 13: 1965–1971.

Wilks DS. 2020. *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, 840 pp.

Wilks DS, Godfrey CM. 2002. Diagnostic verification of the IRI Net Assessment forecasts, 1997–2000. *J. Climate* 15: 1369–1377.