Reliability of Regression-Corrected Climate Forecasts

MICHAEL K. TIPPETT
Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York, and Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

TIMOTHY DELSOLE
George Mason University, Fairfax, Virginia, and Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland

ANTHONY G. BARNSTON
International Research Institute for Climate and Society, Palisades, New York

(Manuscript received 18 September 2013, in final form 31 December 2013)

ABSTRACT

Regression is often used to calibrate climate model forecasts with observations. Reliability is an aspect of forecast quality that refers to the degree of correspondence between forecast probabilities and observed frequencies of occurrence. While regression-corrected climate forecasts are reliable in principle, the estimated regression parameters used in practice are affected by sampling error. The low skill and small sample sizes typically encountered in climate prediction imply substantial sampling error in the estimated regression parameters. Here the reliability of regression-corrected climate forecasts is analyzed for the case of joint-Gaussian distributed ensemble forecasts and observations with regression parameters estimated by least squares. Hypothesis testing of the regression parameters provides direct information about the skill and reliability of the uncorrected ensemble-based probability forecasts. However, the regression-corrected probability forecasts with estimated parameters are systematically “overconfident” because sampling error causes a positive bias in the regression forecast signal variance, despite the fact that the estimates of the regression parameters are themselves unbiased. An analytical description of the reliability diagram of a generic regression-corrected climate forecast is derived and is shown to depend on sample size and population correlation skill, with small sample size and low skill being factors that increase overconfidence. The analytical reliability estimate is shown to capture the effect of sampling error in synthetic data experiments and in a 29-yr dataset of NOAA Climate Forecast System version 2 predictions of seasonal precipitation totals over the Americas. The impact of sampling error on the reliability of regression-corrected forecast has been previously unrecognized and affects all regression-based forecasts. The use of regression parameters estimated by shrinkage methods such as ridge regression substantially reduces overconfidence.

1. Introduction

Linear regression has long played an important role in weather and climate forecasting, both in empirical prediction models and statistical postprocessing of physics-based prediction model output (e.g., Glahn and Lowry 1972; Penland and Magorian 1993). Here we focus on the use of regression to correct and calibrate climate forecasts, although our findings are generally applicable to all regression-based forecasts. A regression between past model output and observations provides an effective means of combining available forecast model guidance and past model behavior to predict a future observation (e.g., Landman and Goddard 2002; Tippett et al. 2005; Vecchi et al. 2011). Describing a model forecast $f$ and its corresponding observation $o$ as random variables, the regression forecast distribution is the conditional distribution $p(o|f)$, that is, the probability distribution of the observation $o$ conditional on the
forecast $f$ (DelSole 2005). The regression forecast distribution, by definition, contains all the information about the observation that can be extracted from the model forecast. When the model forecast provides no useful information (i.e., the model forecast is independent of the observation), the regression forecast distribution $p(o | f)$ is the same as the unconditional or climatological distribution $p(o)$. The mean of the regression forecast distribution is the best estimate of the observation in the sense that it minimizes the expected squared forecast error, irrespective of the forecast and observation distributions. Additional statistics of the regression forecast distribution (e.g., variance) can be used to produce probabilistic forecasts (Kharin and Zwiers 2003).

The regression forecast distribution generally must be determined empirically from data. When forecasts and observations have a joint-Gaussian distribution, the regression forecast distribution $p(o | f)$ is itself Gaussian and is specified completely by its mean and variance. Moreover, in this case, the regression forecast mean is given by linear regression, and the regression forecast variance is the error variance of the linear regression. Such Gaussian regression forecasts, where the mean and variance are known exactly, are reliable, essentially by construction as we show below (see also Johnson and Bowler 2009).

Reliability is an aspect of forecast quality that measures the extent to which probabilistic forecasts capture the actual likelihood of the event being predicted. A fundamental challenge in assessing forecast reliability is that nature provides only a single outcome for a particular forecast. Consequently, a reliability assessment necessarily involves some pooling of forecasts. There are several ways of quantifying reliability (Gneiting et al. 2007). Probabilistic calibration is assessed via the probability integral transform or rank histogram (Hamill 2001). Marginal calibration refers to the equality of the forecast and observed climate. Exceedance calibration is defined in terms of thresholds; there is no empirical method in the literature for assessing exceedance calibration (Machete 2013).

A commonly used graphical tool for assessing the reliability of categorical probability forecasts is the reliability diagram, also called an attributes diagram (Murphy and Winkler 1977). A reliability diagram consists of the graph (the calibration function) of the observed frequency of an event as a function of its forecast probability along with the frequency of forecast issuance. In particular, suppose that $O$ is the binary variable whose value is one when the event occurs and zero otherwise, and denote by $P$ the forecast probability of the event $O$. For instance, $O$ might be the event that the observation $o$ exceeds a specified value, or falls within a specified interval. Later, we choose $O$ to be the event that the observed seasonal precipitation does not exceed the 33rd percentile of the climatological distribution. The calibration function is the graph of $E[O | P]$, that is, the frequency of occurrence as a function of forecast probability. Ideally, the frequency of occurrence equals the forecast probability, and the ideal calibration function is a line passing through the origin with slope one (i.e., the line $E[O | P] = P$). In the case that the forecast probability comes from the regression forecast distribution, ideal reliability is ensured since $P = E[O | f]$, and


The regression forecast distribution, when known, leads to ideal reliability by construction.

In practice, the regression forecast distribution must be estimated from data; in the case of seasonal climate forecasts, those records are fairly short. Therefore, the impact of sampling error on the regression parameter estimates and in turn on forecast quality is an important consideration. In this paper we analyze the impact of sampling error on the reliability of Gaussian regression forecast distributions. In section 2 we present the Gaussian (linear) regression forecast, demonstrate its reliability with population parameters, and use those parameters to characterize the reliability of the underlying uncorrected forecast ensemble in a procedure that parallels detection and attribution analysis (Allen and Tett 1999). In section 3 we demonstrate that applying regression to seasonal precipitation forecasts improves reliability on dependent data but that sampling errors in the regression parameter estimates lead to unreliable overconfident forecasts on independent data. We use synthetic data and theoretical arguments in section 4 to show that estimated regression forecast distributions are unreliable because of a positive bias in the forecast signal variance and characterize that unreliability as function of sample size and skill level. A summary of the results is given in section 5.

2. Univariate Gaussian forecasts

In this paper we consider the case in which an ensemble of forecasts is available, the forecast ensemble mean and corresponding observation are scalars, and their joint distribution is Gaussian. We show that the resulting regression forecast distribution is reliable when the population regression parameters are known and that the regression parameters characterize the reliability of the underlying uncorrected forecast ensemble. In particular, the calibration function for the forecast ensemble is an explicit function of the regression parameters, and
the requirements for the reliability of ensemble-based probabilities can be stated in terms of the regression parameters. We also show that testing the consistency of the sample regression parameters with the reliability requirements parallels the procedure used in detection and attribution analysis.

a. Reliability of Gaussian regression forecasts

An ensemble of interchangeable (i.e., indistinguishable) forecasts $f_1, f_2, \ldots$ is often available rather than just a single forecast. In this case, the regression forecast distribution $p(o \mid f_1, f_2, \ldots)$ is equivalent to $p(o \mid \bar{f})$, where $\bar{f}$ is the ensemble mean (DelSole 2005). In the more general case in which the distributions are not joint-Gaussian, regression methods that incorporate information regarding ensemble members (e.g., Unger et al. 2009) in addition to the ensemble mean would be required. Additionally, we make the large-ensemble assumption that the true (population) ensemble mean $\bar{f}$ and ensemble variance $\sigma_f^2$ are known; we do not consider the impact of ensemble size on skill and reliability (Richardson 2001; Weigel et al. 2007). The regression forecast distribution $p(o \mid \bar{f})$ is Gaussian and therefore is determined by its mean and variance. The mean (signal) of the regression forecast distribution is

$$E[o \mid \bar{f}] = b\bar{f} + a,$$

where $b = r(\sigma_o/\sigma_f)$ and $a = E[o] - bE[\bar{f}]$; $\sigma_o^2$ and $\sigma_f^2$ are the observation and ensemble mean variances, respectively, and $r$ is the correlation between the ensemble mean and the observations. The parameters $a$ and $b$ are the intercept and slope, respectively, of the regression line fit between $o$ and $\bar{f}$. The variance (noise) of the regression forecast distribution is

$$\text{var}[o \mid \bar{f}] = E[(o - E[o \mid \bar{f}]^2) / \bar{f}] = \sigma_o^2(1 - r^2) = \sigma^2.$$  \hspace{1cm} (3)

We have shown in (1) that probability forecasts based on the regression distribution result in ideal reliability. Now we demonstrate the reliability of the Gaussian regression forecast distribution. Suppose that $O$ is the event that $o < c$ for some constant $c$. Then the forecast probability of $o < c$ based on the regression forecast distribution is

$$P = \Phi\left(\frac{c - b\bar{f} - a}{\sigma}\right),$$

where $\Phi$ is the cumulative distribution function of a Gaussian random variable with mean zero and unit variance. The expected occurrence of $O$ conditional on $P$ is

$$E[O \mid P] = E[O \mid \bar{f}] = E[o < c \mid \bar{f}] = \Phi\left(\frac{c - b\bar{f} - a}{\sigma}\right) = P;$$

the expected occurrence is equal to the forecast probability. The Gaussian regression forecast distribution results in an ideal calibration function.

b. Regression parameters and ensemble reliability

We now characterize the reliability of the underlying uncorrected forecast ensemble in terms of the parameters of the Gaussian regression forecast distribution. The members of an ideal, reliable forecast ensemble are statistically indistinguishable from observations, both in an average sense and conditional on specific forecast signals (Hamill 2001; Mason et al. 2007). This requirement of “exchangeability” or “replaceability” can be stated as

$$p(o \mid \bar{f}) = p(f_i \mid \bar{f}).$$  \hspace{1cm} (4)

In other words, for a given ensemble mean $\bar{f}$, the corresponding observations have the same distribution as an ensemble member $f_i$, and that distribution is the same as the regression forecast distribution. In the Gaussian setting, the equality (4) is equivalent to the two distributions having the same means and variances. Taking the means and variances of (4) and using (2) gives

$$b\bar{f} + a = \bar{f}$$

and

$$\text{var}[o \mid \bar{f}] = \sigma^2 = \text{var}(f_i \mid \bar{f}) = \sigma_e^2, $$

respectively, as the requirements for reliability, where $\sigma_e^2$ is the ensemble variance. Equation (5) implies that the parameters of the linear regression forecast distribution satisfy $b = 1$ and $a = 0$ if $\bar{f}$ is not constant. In the case that $\bar{f}$ is constant, $r = 0$, $b = 0$, $a = E[o]$, and (5) reduces to $\bar{f} = E[o]$; that is, the ensemble mean is equal to the observed mean. These conditions along with $\sigma_e^2 = \sigma^2$ are the requirements for the forecast ensemble to be reliable. These conditions on the regression parameters $b$, $a$, and $\sigma_e^2$ characterize the signal, mean, and variance calibration of the ensemble, respectively, under the assumption of Gaussian distributions.

Only sample estimates of the regression parameters are available in practice. The consistency of the sample values with the hypothesis of a skillful and reliable forecast ensemble can be checked using linear hypothesis testing in a procedure that parallels that employed
in detection and attribution analysis (Allen and Tett 1999). First, the forecast ensemble mean is regressed with the observations and a confidence interval is obtained for the estimated regression coefficient \( \hat{a} \). We note that the uncertainty in estimating the ensemble mean itself must be accounted for in the case of small ensemble size (Allen and Stott 2003). If the interval does not include zero, the hypothesis of skill cannot be rejected (detection). If the interval includes 1, the hypothesis of signal calibration cannot be rejected (attribution). Likewise, the hypothesis of mean calibration cannot be rejected if the interval for the estimate of \( a \) contains zero. Finally, the variance of the linear regression forecast distribution is compared with the ensemble variance (residual consistency). Linear reliability hypothesis testing is parametric and requires fewer data than the usual reliability diagram analysis. Reliability diagrams are typically computed nonparametrically by pooling forecasts from multiple locations and lead times, binning forecast probabilities, and computing the frequency of occurrence for each bin. Linear reliability hypothesis testing avoids pooling forecasts with possibly distinct characteristics and potentially provides more detailed assessment.

c. Regression parameters and the calibration function

We now compute the calibration function for the uncorrected forecast ensemble directly from the regression parameters. The ensemble-based predicted probability \( P_e \) of the event \( o < c \) is

\[
P_e = \Phi \left( \frac{c - \hat{T}}{\sigma} \right).
\]

The calibration function for the ensemble-based forecast is

\[
E[O \mid P_e] = E[O \mid \hat{T}] = \Phi \left( \frac{c - b \hat{T} - a}{\sigma} \right)
= \Phi \left( \frac{c - b (c - \sigma_e \Phi^{-1}(P_e)) - a}{\sigma} \right).
\]  

Again, we see that the ideal relation \( E[O \mid P_e] = P_e \) holds for all values of \( P_e \) if and only if \( b = 1, a = 0, \) and \( \sigma_e = \sigma \) in the case of nonconstant \( \hat{T} \); that is, the ensemble distribution is reliable if its distribution is the same as the linear regression forecast distribution.

Although the expression (7) gives the exact dependence of the calibration function on the linear regression forecast parameters and ensemble variance, it is informative to examine the individual impacts of signal and variance ensemble miscalibration on the calibration function. Linear interpolation between \( b = 0 \) and \( b = 1 \) gives the approximation

\[
\Phi \left( \frac{c - b \hat{T}}{\sigma} \right) \approx b \Phi \left( \frac{c - \hat{T}}{\sigma} \right) + (1 - b) \Phi \left( \frac{c}{\sigma} \right),
\]

which when substituted into (7) yields in the case of signal miscalibration (\( \sigma_e = \sigma, a = 0, \) but \( b \neq 1 \)) that

\[
E[O \mid P_e] = \Phi \left( \frac{c - b \hat{T}}{\sigma} \right) \approx b P_e + (1 - b) \Phi \left( \frac{c}{\sigma} \right).
\]

The slope of the calibration function is approximated by the regression coefficient \( b \). The calibration function intersects the ideal line \( E[O \mid P_e] = P_e \) at approximately \( P_e = \Phi(c/\sigma) \), which for small values of \( r^2 \) is close to \( \Phi \{ (1 - r^2) c/\sigma \} \), the climatological probability of the event \( o < c \). Likewise, linear interpolation approximation

\[
\Phi \left( \frac{c - b \hat{T}}{\sigma} \right) \approx \frac{\sigma}{\sigma_e} \Phi \left( \frac{c - b \hat{T}}{\sigma_e} \right) + \frac{1}{2} \left( 1 - \frac{\sigma}{\sigma_e} \right),
\]

in the case of variance miscalibration (\( b = 1, a = 0, \) but \( \sigma_e \neq \sigma \)) gives

\[
E[O \mid P_e] = \Phi \left( \frac{c - \hat{T}}{\sigma} \right) \approx \frac{\sigma}{\sigma_e} P_e + \frac{1}{2} \left( 1 - \frac{\sigma}{\sigma_e} \right).
\]

The slope of the calibration function is approximated by the ratio of the regression and ensemble standard deviations. In this case, the calibration function intersects the ideal line \( E[O \mid P_e] = P_e \) at approximately \( P_e = \frac{1}{2} \). Figure 1 shows the calibration functions for ensembles with various values of \( b \) and standard deviation ratios for \( r = 0.5 \) and \( c = \Phi^{-1}(1/3) \approx -0.431 \), that is, forecasts of the probability of the observations being in the lower tercile category. The approximations are quite good over most of the range of values of \( P_e \). Overconfidence results when \( b > 1 \) or \( \sigma_e < \sigma \). The two cases can be distinguished by the fact that the intersection of the calibration function with the ideal line is near the climatological probability \( \frac{1}{3} \) in the case of signal miscalibration, whereas the intersection is always near \( \frac{1}{2} \) in the case of variance miscalibration.

3. Example: Seasonal precipitation forecasts

We now examine the reliability of seasonal precipitation forecasts made by a current coupled general
circulation model. The distribution of seasonal precipitation, suitably transformed, has been shown to permit a Gaussian description (Tippett et al. 2007). Figure 2a shows the ranked probability skill score (RPSS) of the National Oceanic and Atmospheric Administration (NOAA) Climate Forecast System version 2 (CFSv2) forecasts of December–February (DJF) precipitation over North and South America made the previous 1 November during the period 1982–2010 (Saha et al. 2014). The forecast data are those provided to the National Multi-Model Ensemble project (Kirtman et al. 2014). Precipitation observations were taken from the Climate Prediction Center (CPC) Unified Gauge-Based Analysis of Global Daily Precipitation averaged onto a $1^\circ \times 1^\circ$ grid (Xie et al. 2007; Chen et al. 2008). Forecast probabilities of the below-normal, normal, and above-normal categories are computed from ensemble frequencies of membership in the model-defined tercile-based categories; the ensemble size is 24. Although positive RPSS values are seen in ENSO teleconnection regions, there are substantial areas with negative RPSS values. The corresponding reliability diagrams for the above-normal and below-normal categories shown in Fig. 2c are computed by pooling all forecasts and locations and indicates that, overall, forecasts are overconfident in the sense that the calibration function has slope less than one. The pooling of forecasts in the reliability diagram means that the observed overconfidence

![Fig. 1. Calibration functions (solid), linear interpolation approximation (dashed), and ideal (dotted) for (a) $b = 0.5$, (b) $b = 2$, (c) $\sigma / \sigma_c = 2$, and (d) $\sigma / \sigma_c = 0.5$.](image-url)
cannot be ascribed to particular locations. We note that the calibration function intersects the ideal line near the climatological probability of 1/3, consistent with signal miscalibration.

To estimate the regression forecast distribution, we first apply the square root transformation to both the observed and forecast precipitation values to reduce the positive skewness and then compute a regression between the transformed ensemble mean forecasts and observations at each grid point. Least squares estimates of the regression parameters are made using all 29 years of data. A map of RPSS values for the linear regression forecast distribution probabilities is shown in Fig. 2b, and values are positive in essentially the same locations as the RPSS of the ensemble frequency-based forecasts. However, where the ensemble frequency-based forecasts have negative RPSS values, the regression forecast distribution forecasts have near-zero RPSS. The corresponding reliability diagram in Fig. 2d indicates that the regression distribution forecasts are well calibrated with the occurrence frequencies being virtually equal to the forecast probabilities. The histograms of forecast probability issuance indicate that the regression distribution forecast probabilities are more tightly clustered around the climatological frequency than those of the ensemble frequency-based forecasts. Linear regression reduces

Fig. 2. RPSS of CFSv2 forecasts of DJF precipitation made 1 November with probabilities computed using (a) ensemble frequencies and (b) linear regression. Map averages are shown in title. Reliability of probabilities computed using (c) ensemble frequencies and (d) linear regression. Inset histograms show the frequency of forecast probability issuance.
forecast probability excursions from climatology and improves reliability.

We can examine the calibration of the forecast ensemble on a spatial basis using the linear reliability hypothesis tests introduced in the previous section. Where the intercept interval includes zero, the calibrated mean hypothesis cannot be rejected. Figure 3a shows that the intercept calibration hypothesis is rejected at the 95% level at relatively few points, mostly over South America. The no-skill hypothesis cannot be rejected where the 95% confidence interval for \( b \) includes zero. Where the 95% confidence interval for \( b \) does not include 1, the uncalibrated signal hypothesis cannot be rejected. Finally, the regression variance is compared to the ensemble variance. Figure 3 shows that locations with positive RPSS closely correspond to those that pass the linear reliability hypothesis tests. There are a few isolated points with apparent skill that do not pass the calibration test, and those in southernmost South America present significant, but negative, correlations. Standard methods can be applied to assess the field significance of the calibrated mean and no-skill test results (Livezey and Chen 1983; DelSole and Yang 2011).

Unfortunately, the quality of these regression forecasts is overestimated because the regression parameters were estimated using the same data that are used to evaluate skill and reliability. Figures 4a and 4c show the RPSS and reliability, respectively, of regression forecasts developed in a leave-one-out cross-validated fashion. There is a modest decrease in RPSS, but a substantial reduction in reliability with a reappearance of overconfidence. The forecast issuance histograms show no marked changes. Some aspect of the cross-validation procedure results in loss of reliability. Figures 4b and 4d show the RPSS and reliability, respectively, of regression forecasts where only the regression coefficient \( b \) is subject to cross-validation. In this case, the reliability is very similar to that of the fully cross-validated forecasts. Conversely, there is little loss of reliability when the regression coefficient \( b \) is exempted from the cross-validation procedure, and the other regression parameters are computed in a cross-validated manner (not shown). This behavior suggests that the loss of reliability is associated with the estimation of the slope parameter and, consequently, with signal miscalibration.

To avoid sources of signal miscalibration other than sampling error, such as the observations or forecasts not satisfying the assumptions of linear regression or some unknown degeneracy associated with the cross-validation procedure (Barnston and Van den Dool 1993), we use synthetic data with specified distributions and theoretical estimates in the following section to analyze the reliability of estimated regression forecast distributions.

4. Reliability of estimated regression forecast distributions

a. Synthetic data

We design a synthetic data experiment that mimics the seasonal climate regression forecasts of the previous section. The use of synthetic data permits the population distributions to be specified and cross-validation to be
avoided. We take the number of “grid points” to be 5000 and the training period to consist of 29 samples. Additional experiments (not shown) indicate that the number of grid points in the synthetic data affects only the calibration function error bars, with more grid points leading to smaller error bars. The population correlation of the forecasts with observations is taken to be uniformly distributed between 0 and a specified value $r_{\text{max}}$. Linear regression parameters are estimated from the 29 training samples using least squares. The resulting regression is used to make probability forecasts of the observations occurring in the lower tercile category for an independent verification dataset with 100 samples. Figure 5 shows the resulting reliability diagrams for $r_{\text{max}} = 0.3$ and 0.5, respectively. While the probabilities are reliable during the training period (blue lines), they are not reliable during the verification period (red lines) and display overconfidence as did the cross-validated CFSv2 regression forecasts. The overconfidence is relatively greater for $r_{\text{max}} = 0.3$. Applying cross-validation to one regression parameter at a time reveals that the overconfidence is almost entirely due to sampling error in the estimate of the regression coefficient $\hat{b}$. These numerical simulations demonstrate that the lack of calibration...
in the cross-validated CFSv2 regression forecasts is not due to any features particular to the data (e.g., not being Gaussian distributed) or to the cross-validation procedure, but rather is the result of sampling error.

b. Theoretical estimates

We now quantify the dependence of the estimated regression forecast miscalibration on sample size and skill level. The observed overconfidence of the estimated regression forecasts is consistent with misspecification of the regression coefficient: in particular, it is consistent with $b > b$, as shown in section 2. However, the least squares estimate $b$ is an unbiased estimate of $b$. Therefore the overconfidence is not due to a biased estimate. Likewise, the estimate of regression forecast variance is unbiased as well. Since overconfidence suggests that the signal component is too large on average, we now examine the estimated regression forecast signal variance:

$$E[(\hat{b} f)^2] = E[(b + (b - b) f)^2] = (b^2 + \sigma^2/n)(\alpha^2_f),$$

where we have used the large-sample approximation that the estimate $\hat{b}$ is Gaussian with mean $b$ and variance $\sigma^2/n$ and is independent of $f$. Since the correct value of the regression forecast variance is $b^2\sigma^2_f$, (8) means that the estimated regression forecast signal variance has a positive bias due to the variance of the regression coefficient estimate. This bias is large when the skill level is low ($\sigma$ is large) and when the sample size $n$ is small.

In section 2c we showed that the miscalibration of an unreliable forecast can be quantified using the regression parameters themselves. We now use this approach to quantify the impact of the signal variance miscalibration on reliability and express the calibration function as a function of sample size and skill level. We treat the estimated regression forecast mean as an uncalibrated forecast and regress it with the observations. That is, we compute

$$E[o \mid \hat{b} f] = E[bf \mid \hat{b} f] = \gamma bf,$$

where the slope $\gamma$ is given by

$$\gamma = \frac{E[b(\hat{b} f)^2]}{E[\hat{b} f^2]} \approx \frac{b^2}{b^2 + \sigma^2/n} = \frac{r^2}{r^2 + 1 - r^2}.$$  

Again we use the large-sample approximation for the distribution of $b$. This expression for $\gamma$ depends on the population parameters $b$ and $\sigma$, and while these quantities are not available from a sample, the expression in (9) is a useful characterization of the dependence of $\gamma$ on sample size and skill level. From (7) we expect the calibration function of the estimated regression forecast distribution to be

$$E(O \mid P) = \Phi \left( \frac{c - \gamma(c - \sigma \Phi^{-1}(P)) - a}{\sigma} \right)$$

$$\approx \gamma P + (1 - \gamma)\Phi \left( \frac{c}{\sigma} \right).$$
The slope of the calibration function is approximately \( \gamma \), and since \( \gamma < 1 \), overconfidence is indicated. In Figs. 4c, 4d, and 5 we have plotted the calibration function from (10) as the gray curve labeled “Analytical.” We use the average (over grid points squared) sample correlation value to estimate \( r \) in (9). The analytical calibration function is a good approximation of the actual calibration function, especially for forecast probabilities near the climatological probability of \( \frac{1}{3} \). A limitation of the analytical function is that it depends on the population correlation.

A property of the conditional mean is that it minimizes the expected squared error. Therefore,

\[
E(o - \gamma \hat{f})^2 \leq E(o - \hat{f})^2.
\]

This result seems paradoxical at first because the regression coefficient is itself designed to minimize squared error. However, it is the population regression coefficient \( b \) that minimizes expected squared error, and the least squares estimate \( \hat{b} \) is the estimate of \( b \) with the smallest variance among all linear unbiased estimates according to the Gauss–Markov theorem (Hastie et al. 2009). The expected squared error contains contributions from the variance of the estimate of \( b \) and the square of the bias of the estimate. The least squares estimate is unbiased and minimizes the variance contribution to the squared error. Biased estimates, such as those produced by ridge regression, attempt to minimize the squared error further by trading an increase in bias for a reduction in variance (Hoerl and Kennard 1988). Ridge regression shrinks the regression coefficient to reduce the squared error, and indeed \( \gamma \) here is strictly less than one. In fact, the form of \( \gamma \) in (9) agrees with a population estimate of the ridge shrinkage parameter [Eq. (3.12) of Copas (1983)]. Returning to the numerical simulations, using the ridge regression estimate gives more reliable probabilities as shown in Fig. 5. The ridge parameter controls the degree of shrinkage, and its selection is crucial (DelSole 2007). Here, we select a single ridge parameter, similar to DelSole et al. (2013), by minimizing the cross-validated error in the 29-yr training data.

5. Summary

For a given model forecast and verifying observations, the regression forecast distribution is the probability distribution of the observations conditional on the model forecast (DelSole 2005). The regression forecast distribution contains all the information about the observation that can be extracted from the model forecast and is the optimal correction of the model forecast. The regression forecast distribution produces reliable calibrated probability forecasts in the sense that the expected occurrence of an event is equal to its forecast probability. The relation between expected occurrence and forecast probability is summarized by the calibration function that appears in the reliability diagram (Murphy 1973). Here we considered joint-Gaussian distributed forecasts and observations. In this case, the regression forecast distribution is also Gaussian. Additionally, we suppose that an ensemble of interchangeable model forecasts is available; by interchangeable forecasts we mean forecasts from a single model rather than from multiple models. In this case, we demonstrated that the parameters of the regression forecast characterize the skill and reliability of the underlying uncorrected forecast ensemble. In fact, the calibration function for the underlying forecast ensemble is a function of the regression parameters. A consequence of this functional dependence is that examination of the calibration function can distinguish between overconfidence due to a too strong signal and overconfidence due to a too small variance. The consistency of estimated regression parameters with the hypothesis of ensemble skill and reliability can be assessed using linear hypothesis testing in a procedure that parallels detection and attribution analysis (Allen and Tett 1999).

Linear reliability hypothesis testing offers some advantages over the reliability diagram approach for Gaussian distributions. Calibration functions are generally computed nonparametrically and often require pooling forecasts from multiple locations and lead times. On the other hand, linear reliability hypothesis testing is parametric, requires less data, and thus permits finer assessments of reliability than do reliability diagrams. For instance, linear reliability hypothesis testing allowed us to characterize, on a spatially varying basis, the reliability of a 29-yr set of CFSv2 ensemble seasonal precipitation forecasts for North and South America.

The regression forecast distribution generally must be estimated from data. Least squares estimates of the regression forecast mean and variance are unbiased. However, sampling error in the estimate of the mean causes a positive bias in the regression forecast signal variance that manifests itself in overconfident forecast probabilities. This result is consistent with that of Richardson (2001), who noted that sampling error due to finite ensemble size introduced unreliability into probability forecasts. No comparable bias is seen in the noise variance. The level of sampling error and the corresponding signal bias depend on the sample size and the underlying level of skill. We analytically estimated the signal bias, as well as its impact on the calibration function, as a function of the sample size and population correlation skill. For the low-skill, small-sample
situations typical in climate forecasting, there is substantial sampling error, leading to a substantial positive bias in signal variance and overconfident probability forecasts. With increasingly larger forecast sample sizes and/or higher average correlation skills, the positive bias of the signal variance lessens until it becomes inconsequential. The reality of the signal bias and the accuracy of the calibration function estimate were demonstrated using idealized numerical simulations with synthetic data and CFSv2 ensemble seasonal precipitation forecasts. The impact of this bias on reliability has not been previously recognized, although its impact on (out of sample) prediction error is known and provides the motivation for shrinkage methods such as ridge regression that reduce prediction error and improve reliability by decreasing the forecast signal variance.

Acknowledgments. The authors thank two anonymous reviewers for their insightful and useful comments and corrections. MKT and AGB are supported by grants from the National Oceanic and Atmospheric Administration (NA05OAR4311004 and NA08OAR4320912) and the Office of Naval Research (N00014-12-1-0911). TD gratefully acknowledges support from grants from the NSF (0830068), the National Oceanic and Atmospheric Administration (NA09OAR4310058), and the National Aeronautics and Space Administration (NNX09AN50G). 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


