Ensemble Regression

DAVID A. UNGER, HUUG VAN DEN DOOL, EDWARD O'LENIC, AND DAN COLLINS

NOAA/NWS/NCEP/Climate Prediction Center, Camp Springs, Maryland

(Manuscript received 25 March 2008, in final form 17 November 2008)

ABSTRACT

A regression model was developed for use with ensemble forecasts. Ensemble members are assumed to represent a set of equally likely solutions, one of which will best fit the observation. If standard linear regression assumptions apply to the best member, then a regression relationship can be derived between the full ensemble and the observation without explicitly identifying the best member for each case. The ensemble regression equation is equivalent to linear regression between the ensemble mean and the observation, but is applied to each member of the ensemble. The ''best member'' error variance is defined in terms of the correlation between the ensemble mean and the observations, their respective variances, and the ensemble spread. A probability density function representing the ensemble prediction is obtained from the normalized sum of the best-member error distribution applied to the regression forecast from each ensemble member. Ensemble regression was applied to National Centers for Environmental Prediction (NCEP) Climate Forecast System (CFS) forecasts of seasonal mean Niño-3.4 SSTs on historical forecasts for the years 1981–2005. The skill of the ensemble regression was about the same as that of the linear regression on the ensemble mean when measured by the continuous ranked probability score (CRPS), and both methods produced reliable probabilities. The CFS spread appears slightly too high for its skill, and the CRPS of the CFS predictions can be slightly improved by reducing its ensemble spread to about 0.8 of its original value prior to regression calibration.

1. Introduction

a. Background

An ensemble forecasting system addresses the chaotic nature of the atmosphere by providing a dynamic estimate of the prediction confidence. Such systems exploit the stochastic nature of the atmosphere by generating many solutions based on slightly perturbed initial states (Toth and Kalnay 1993). The chaotic nature of the predicted system leads the model solutions to diverge from one another with time, resulting in different realizations representing possible future atmospheric states (Epstein 1969a; Leith 1974). For well-calibrated models, closely grouped model realizations (in phase space) are assumed to indicate low uncertainty in the final atmospheric state, while widely scattered solutions indicate higher uncertainty. Ensemble forecasting is also used for climate predictions, where predictive skill may be low or negligible, but where ensembles are hoped to accurately reflect a range of possible climatic anomalies, and identify areas of potential predictability associated with boundary forcing (Barnett 1995; Stern and Miyakoda 1995; Kumar and Hoerling 1995).

Ensemble predictions generally require calibration to remove biases and to ensure that the forecast frequency of events gives a realistic representation of that in the atmosphere. Many methods have been developed to calibrate ensemble predictions to provide reliable probabilistic forecasts. The choice of an appropriate method depends on the characteristics of the ensemble forecasts and on their intended applications. Since a primary mission of the Climate Prediction Center (CPC) is seasonal climate prediction, we require a calibration method suitable for use in global climate models. Skill in climate prediction is low and varies considerably both spatially and temporally (Livezey 1990; Rowell 1998; Phelps et al. 2004; Livezey and Timofeyeva 2008; O'Lenic 2008). This may be reflected in the ensemble spread of GCM predictions, so it is critical to retain as much of this information as possible. The amount of data available to develop relationships for seasonal prediction is very limited.

Corresponding author address: David A. Unger, NOAA/NWS/NCEP/Climate Prediction Center, 5200 Auth Rd., Camp Springs, MD 20746.
E-mail: david.unger@noaa.gov

DOI: 10.1175/2008MWR2605.1

The National Centers for Environmental Prediction's (NCEP) Climate Forecast System (CFS) model (Saha et al. 2006), for example, has hindcast data available only since 1981 so any calibration procedure must do well with very limited sample sizes. An additional requirement is the ability to estimate the full probability density function (PDF) of the forecast element in order to support CPC's Probability of Exceedance product (Barnston et al. 2000).

The small amount of hindcast data available from most GCM predictions eliminates some ensemble calibration methods from serious consideration. Some calibration methods such as binning procedures (Anderson 1996; Hamill and Colucci 1997, 1998; Eckel and Walters 1998) or logistic regression (Hamill et al. 2004) divide the range of the forecast element into a series of categories (bins). The limited data available for seasonal prediction make subdivision of the data sample into multiple categories impractical. The analog approach described by Hamill and Whitaker (2006) is similarly impractical for climate prediction in view of the difficulty in finding good analogs for the limited data available for seasonal prediction (van den Dool 1994).

Regression-based approaches to the problem are appealing because of their ability to optimally "fit" data to minimize errors. When applied to continuous variables, this allows the entire dataset to simultaneously contribute to the regression relationship, enabling maximum use of small samples. Several methods based on regression have been proposed for ensemble calibration. Gneiting et al. (2005) use a pair of regression relationships to calibrate ensemble forecasts, one to correct the ensemble mean forecast and another to correct the ensemble spread. This method was among the most successful of those tested in a side-by-side comparison of a variety of ensemble calibration methods, both on an idealized model (Wilks 2006) and on GFS reforecasts of temperature and precipitation (Wilks and Hamill 2007). The method requires that the ensemble members be fit with a parametric distribution in each case. While this may be beneficial if the distribution of the residuals about the ensemble mean forecast is believed known, some aspects of atmospheric circulation are known to exhibit complex PDFs (Benzi and Speranza 1989), so a calibration method that can retain information from nonparametric distributions forecast by a model is desired for CPC's operations, at least for model diagnostics, if not for prediction.

We propose a regression model specifically designed for use on ensemble forecasts. This "ensemble regression" (EREG) model is formulated for the commonly held assumption that individual ensemble members represent possible solutions for a given initial state (Sivillo et al. 1997). Of the many solutions, one will be "best" and if the ensembles are generated from the same model, it can reasonably be assumed that the probability of each member being best is about equal. We will show that, given this assumption, together with other assumptions usually made for linear regression, a "best member" regression equation can be estimated from the statistics of the ensemble set as a whole, with no need to explicitly identify a best member in each case. The regression model produces a calibrated set of ensemble forecasts, together with an estimated error distribution around each member that can be used to produce a PDF of the predicted variable from the ensemble forecasts.

The EREG estimates, together with their estimated errors, resemble the "dressed ensemble" approach to ensemble calibration (Roulston and Smith 2003; Wang and Bishop 2005; Fortin et al. 2006), except the ensemble members are fully calibrated for a least squares fit to the data (rather than just bias corrected as in the dressed ensemble approach) and the kernel distributions are derived from the regression estimates of the distribution of residuals about the calibrated best member.

The regression coefficients for the best-member equation are identical to those derived from the ensemble mean and are applied to each member of the ensemble. Therefore, EREG can be implemented by developing a regression relationship between the ensemble mean forecast and the observation, and applying the result to the individual members in the ensemble. This approach has recently been pragmatically explored by the National Weather Service's Meteorological Development Laboratory (Glahn et al. 2009) on short-range forecasts. We show theoretical justification for such an approach here.

We will present the mathematical basis for the EREG procedure in section 2. EREG is applied to long-lead seasonal predictions of sea surface temperature (SST) in the Niño-3.4 region of the Pacific Ocean from NCEP's Climate Forecast System to produce calibrated probabilistic predictions and these results are presented in section 3.

b. Terminology

For this discussion, we assume that statistics are accumulated over a sample of forecasts, such as a series of forecasts issued daily or monthly. The time average over the entire sample of M cases is indicated by angle brackets;

$$\langle x \rangle = \frac{1}{M} \sum_{j=1}^{M} x_j.$$

The subscript, j, represents the temporal dimension and will usually not appear in equations for individual forecast cases. For each case, a collection of N ensemble members are available, F_i, $i = 1, N$, and the ensemble

mean is denoted as F_m. Note that over M cases, $\langle F_i \rangle = \langle F_m \rangle$, since the summation over the N ensemble members is implied in averaging.

It is usually appropriate to eliminate the seasonal cycle from the predictions by expressing both forecasts and observations as departures from long-term climatology. Statistics can also be stratified by month and lead time so that data for all forecasts initialized at a given time of year and for a given lead time are pooled together to form a regression relationship.

2. Regression relationships

a. Simple linear regression

Regression has been applied to the output from dynamic numerical prediction models for over 40 yr (Glahn and Lowry 1972; Glahn et al. 2009). Regression analysis usually begins with a tentative assumption of a linear relationship between the predictors (in this case the forecasts from a numerical model) and the predictand (observations), with errors represented by the term, ε. For reasons that will become clear later, this will be illustrated by the relationship between the ensemble mean, F_m, and the observation, Y:

$$Y = \alpha_0 + \alpha_1 F_m + \varepsilon.$$

Linear regression minimizes the quantity $\langle (F_m^{'} - Y)^2 \rangle$ to estimate the α's and obtains the equation, $F_m^{'} = a_0 + a_1 F_m$ (Draper and Smith 1981), where $F_m^{'}$ is the regression estimate, and coefficients are given by

$$a_1 = R_m \frac{S_Y}{S_m}, \quad a_0 = \langle Y \rangle - a_1 \langle F_m \rangle, \tag{2.1}$$

where S_Y and S_m are the sample standard deviations of Y and F_m, respectively, and R_m is the correlation coefficient between the ensemble mean forecast and the observations. The regression relationship is frequently subject to an analysis of variance with two components defined as follows:

sum of squares due to regression,

$$\mathrm{SS}_{\mathrm{regression}} = \sum_{j=1}^{M} (F_{m,j}^{'} - \langle Y \rangle)^2, \quad \text{and}$$

sum of squares about the regression (residual),

$$\mathrm{SS}_{\mathrm{residual}} = \sum_{j=1}^{M} (F_{m,j}^{'} - Y_j)^2.$$

If the regression estimates are regarded as calibrated forecasts, the means of these two variance components, the regression and the residual, can be regarded as the sample variance of the regression forecasts, $S_{F_m^{'}}^2$, and an associated variance in the forecast errors, S_e^2, respectively, and are related to R_m as shown:

$$S_Y^2 = S_{F_m^{'}}^2 + S_e^2, \tag{2.2a}$$

$$S_{F_m^{'}}^2 = S_Y^2 R_m^2, \quad \text{and} \tag{2.2b}$$

$$S_e^2 = S_Y^2 (1 - R_m^2). \tag{2.2c}$$

Note that these relationships are robust and follow directly from the definitions of the variance, means, and correlation with no requirement for F_m or Y to have Gaussian distributions. Gaussian assumptions are used in significance testing, or in establishing an estimate of the forecast error distribution, but are not required for these relationships to be valid on the dependent data.

Both S_Y^2 and S_e^2 are biased estimates of each variable's true variance, σ_Y^2 and σ_e^2, respectively. For a sample size of M cases,

$$\sigma_Y^2 = \frac{M}{M-1} S_Y^2 \quad \text{and} \quad \sigma_e^2 = \frac{M}{M-2} S_e^2,$$

$$\sigma_{F_m^{'}}^2 = \sigma_Y^2 R_m^2 \quad \text{and} \tag{2.3a}$$

$$\sigma_e^2 = c \sigma_Y^2 (1 - R_m^2), \tag{2.3b}$$

where c represents an additional correction factor to the residual variance to account for uncertainty in R_m. It is frequently the practice in seasonal forecasting to use an estimate of σ_Y^2 from long-term climatology rather than to base the estimate on the sample climatology, thereby partially compensating for the biases in these relationships. The constant c compensates for the remaining bias and is given by

$$c = \frac{M-1}{M-2}.$$

Equation (2.3a) is the explained variance, and (2.3b) gives the unexplained variance of the regression relationship. If the true relationship between F and Y is linear, and the errors are uncorrelated and Gaussian distributed, then the residual distribution will also be Gaussian. This implies that the residuals, ε, can be represented by a Gaussian distribution centered on $F_m^{'}$, $\mathrm{PDF}(\varepsilon) \approx N\{F_m^{'}, \sigma_Y[c(1 - R_m^2)]^{1/2}\}$, following standard terminology where $N(\mu, \sigma)$ represent a normal distribution with mean μ and standard deviation σ. The expected value of σ_e^2 increases with the distance from the sample mean due to uncertainty in the regression coefficients. This effect will be neglected here for simplicity.

An important feature of the regression estimate is that its variance is reduced according to R_m^2. This "skill

damping'' effect can best be seen by expressing the regression equation in terms of a standardized departure from the mean, as shown:

$$\frac{F_m' - \langle Y \rangle}{S_Y} = R_m \frac{F_m - \langle F_m \rangle}{S_m}. \tag{2.4}$$

The standardized anomaly of the regression estimate is damped toward the mean by the factor R_m.

b. Statistical constraints on an ensemble set

Rather than a single forecast, an ensemble prediction consists of a set of related forecasts all paired with a single observation. This constrains the statistics of the ensemble set according to the following series of relationships. The mean squared error of the individual ensemble members, F_i, is related to the ensemble spread and the squared error in the ensemble mean, by (see appendix A)

$$\langle (F_i - Y)^2 \rangle = \langle E^2 \rangle + \langle (F_m - Y)^2 \rangle, \tag{2.5}$$

where $\langle E^2 \rangle$ is the mean ensemble spread,

$$\langle E^2 \rangle = \left\langle \frac{1}{N} \sum_{i=1}^{N} (F_i - F_m)^2 \right\rangle.$$

The sample variance of the individual ensemble forecasts, S_I^2, can be related to the mean spread and variance of the ensemble mean, S_m^2, by a derivation similar to that shown in appendix A except substituting $\langle F \rangle$ for Y:

$$S_I^2 = S_m^2 + \langle E^2 \rangle. \tag{2.6}$$

The correlation coefficients between Y and (a) the individual ensembles, R_I, and (b) the ensemble mean, R_m, are also related as shown (see appendix B):

$$R_m = R_I \frac{S_I}{S_m}. \tag{2.7}$$

Applying (2.6) and rearranging terms, $\langle E^2 \rangle$ is given by

$$\langle E^2 \rangle = S_I^2 \frac{(R_m^2 - R_I^2)}{R_m^2}. \tag{2.8}$$

c. Ensemble regression (EREG)

An ensemble prediction is frequently regarded as a set of possible states resulting from a given initial condition. Of the various solutions, one will be ''best,'' which might be regarded as either the closest to the observation, or best in some multivariate sense, as Roulston and Smith (2003) suggest. If the ensemble members are generated by the same atmospheric model, it is usually assumed that each member has an equal likelihood of being best. A linear regression model can be specifically tailored for use on ensemble prediction considering these specialized assumptions. As with any regression procedure, a tentative model is considered to describe the system, which can be rejected at a later time if these assumptions are not supported by the data.

Without actually identifying a best member, F_b, we postulate that it is related to the observation according to

$$Y = \alpha_0 + \alpha_1 F_b + \varepsilon_b, \tag{2.9}$$

where ε_b represents the errors only for F_b. It is further assumed that ε_b is distributed in the same manner for each potential realization.

Given our tentative regression model, the regression equation that minimizes ε_b is given by (from standard linear regression theory summarized in section 2a)

$$F_b' = a_0 + a_1 F_b \quad \text{and} \tag{2.10a}$$

$$a_1 = R_b \frac{S_Y}{S_I}, \quad a_0 = \langle Y \rangle - a_1 \langle F_b \rangle. \tag{2.10b}$$

Here, R_b is the unknown expected value of the correlation between the best ensemble member and the observation. Note that given our assumption that members are equally likely to be best, on any given case, j, the expected value of F_b can be calculated from F_i, as follows:

$$\text{expv}(F_b) = \frac{1}{N} \sum_{i=1}^{N} F_i = F_m.$$

Here it is reasonable to assume that the best member is determined from the closest ensemble solution after calibration by (2.10a).

The expected value of the grand mean of F_b is

$$\langle \text{expv}(F_b) \rangle = \frac{1}{M} \sum_{j=1}^{M} \text{expv}(F_{b,j}) = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} F_{i,j} = \langle F \rangle.$$

By similar reasoning, the expected value of S_b can determined as follows:

$$\text{expv}(S_b^2) = \frac{1}{M} \sum_{J=1}^{M} \text{expv}(F_{b,j} - \langle F_b \rangle)^2 = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} (F_{i,j} - \langle F \rangle)^2 = S_I^2.$$

Here, it must be emphasized that these relationships apply only for when our assumption that each ensemble member is equally likely to be best are met.

In subsequent discussions, quantities involving best-member statistics (subscript b, except for F_b and $F_b^{'}$) refer to their expected values, and the notation will be simplified so, for example,

$$S_b = \text{expv}(S_b).$$

Similar to the results in section 2a, the expected values of both the explained variance, S_{Fb}^2 and the residual error variance, S_{eb}^2, are given by

$$S_{Fb}^2 = S_Y^2 R_b^2 \quad \text{and} \tag{2.11a}$$

$$S_{eb}^2 = S_Y^2(1 - R_b^2). \tag{2.11b}$$

As shown in appendix C, the expected values of the regression coefficients a_0 and a_1 are the same as the coefficients of the ensemble mean when regressed onto the observations. Thus, from (2.1),

$$a_1 = R_m \frac{S_Y}{S_m} = R_b \frac{S_Y}{S_I}, \quad a_0 = \langle Y \rangle - a_1 \langle F_m \rangle .$$

Linear transformation of the forecasts does not affect the correlation between the forecast and the observation, so, after substitution from (2.7), R_b is given by

$$R_b = \frac{R_m^2}{R_I}. \tag{2.12}$$

Because the *expected value* of the regression coefficients of the best-member equation are the same as those for the ensemble mean, EREG can be implemented by applying the regression equation based on the ensemble mean to each individual ensemble member. Here, R_b represents the expected value of the correlation between F_b' and the observation, provided that the regression model assumptions are accurate. In standardized anomaly form, the EREG equation is

$$\frac{F_b^{'} - \langle Y \rangle}{S_Y} = R_b \frac{F_b - \langle F \rangle}{S_I},$$

indicating that EREG damps individual ensemble members to a lesser extent than the equation applied to the ensemble mean [see (2.4)].

An estimate of $\langle \text{expv}(Y - F_m^{'})^2 \rangle$ can be obtained both from (2.3b) and also from computing the expected value after substituting the individual member regression estimates:

$$F_i^{'} = a_0 + a_1 F_i + \varepsilon_b.$$

Note that the expected value of the error term and the cross products involving this term is zero, and that ε_b already accounts for the expected errors in the ensemble mean (because it is derived from the residual error of a regression), so after substitution

$$\langle \text{expv}(Y - F_m)^2 \rangle = a_1^2 E^2 + \varepsilon_b^2.$$

The regression estimate of the residual variance about the calibrated ensemble mean is related to the regression-corrected ensemble spread, $a_1^2\langle E^2 \rangle$, as shown:

$$c\sigma_Y^2(1 - R_m^2) = a_1^2 \langle E^2 \rangle + \varepsilon_b^2 \tag{2.13}$$

where ε_b^2 represents the remainder of the variance not accounted for by the calibrated ensemble members. Since ε_b^2 is nonnegative, $c\sigma_Y^2(1 - R_m^2) \geq a_1^2\langle E^2 \rangle$.

If the above inequality is not true, then the ensemble members cannot conform to the EREG assumptions and the regression model must be rejected. This occurs when the calibrated ensemble is overdispersive (members near the ensemble mean have a higher probability of being best than those near the ensemble's outer envelope even after applying the regression equation), and that R_b as calculated from (2.12) exceeds one. An underdispersive model presents no problems, since the regression estimate of ε_b^2 will adjust to account for the model's missing variance. However, there is always a possibility that an underdispersive model can be improved by increasing the ensemble spread, shifting more weight to the dynamic prediction of the errors about the calibrated ensemble mean forecast and less to its statistical estimate. Adjustments to the ensemble spread will be addressed in section 2e.

d. Estimated PDF of the calibrated ensemble

If ε_b, in (2.9) is Gaussian distributed, then its regression estimate, ε_b, is $N(0, \sigma_{eb})$, σ_{eb} can be estimated in a manner similar to (2.3):

$$\sigma_{eb} = \sigma_Y[c(1 - R_b^2)]^{1/2}. \tag{2.14}$$

The distribution of observations around the calibrated best member can then be represented by centering the error distribution around $F_b^{'}$. The forecast PDF representing the entire ensemble of N equally likely members takes the form of a series of ''kernel'' Gaussian distributions, each centered on the regression estimate of an individual member. The final PDF is simply the unit-normalized sum of all error distributions, each representing $1/N$ of the total distribution, as illustrated by the example in Fig. 1.

e. Adjustments to the ensemble spread

In this section we will examine the relationship between the ensemble spread and the EREG residual

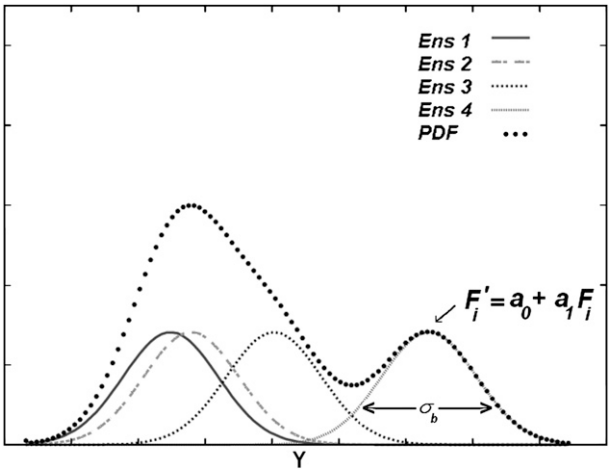

FIG. 1. Schematic illustration of the PDF derived from an ensemble regression of four ensemble members. The PDF is the normalized sum of the Gaussian kernels centered at the regression estimate of each of the four members. Here, F_i' represents the regression forecast based on of the ith ensemble member, F_i. Both a_0 and a_1 are regression coefficients and σ_e is the EREG error estimate for the best member.

error. A linear transformation is proposed to adjust the ensemble spread, if necessary, to better fit the assumptions required for EREG. As discussed earlier, a reduction of the ensemble spread is required to produce a reasonable regression fit when the calibrated ensemble is overdispersive.

Consider a spread adjustment factor, K, applied on all cases prior to regression given by

$$F_i'' = F_m + K(F_i - F_m), \tag{2.15}$$

where F_i'' refers to the transformed ensemble forecasts. The transformation constant, K, alters the correlation between the observation and individual transformed ensemble members, R_I'', and the expected values of R_b and σ_{eb}, and these relationships are derived from (2.6), (2.7), and (2.14):

$$R_I''^2 = R_m^2 \frac{S_m^2}{S_m^2 + K^2\langle E^2\rangle}. \tag{2.16a}$$

Thus,

$$R_b'' = \frac{R_m^2}{R_I''}, \tag{2.16b}$$

$$\sigma_{eb}'' = \sigma_Y[c(1 - R_b''^2)]^{1/2}. \tag{2.16c}$$

The maximum value of K ($K_{\max}$) that is consistent with regression assumptions can be computed by setting $\sigma_{eb}'' = 0$ (implying that $R_b'' = 1$), and can be calculated as shown in (2.17):

$$K_{\mathrm{MAX}} = \left(\frac{\frac{1}{R_m^2} - 1}{\frac{R_m^2}{R_I^2} - 1}\right)^{1/2}. \tag{2.17}$$

When $K_{\max} < 1$, the EREG estimates based on the original model forecasts will be overdispersive and the spread needs to be reduced to assure that the forecast variance is less than the observed variance.

Note that $K_{\max}$ does not account for the sampling variability expected with a limited number of ensemble members. If an ensemble forecast is presumed to be a sample of N solutions drawn randomly from a large population of potential solutions, then the maximum value of K based on sampling considerations, K_N, is related to the ensemble size, N, as shown in appendix D:

$$K_N = \left(\frac{N-1}{N}\right)^{1/2} K_{\max}. \tag{2.18}$$

Equation (2.18) expresses the maximum K value supported by an N-member ensemble randomly chosen from an infinite population of solutions and over the dependent data sample. If $K_N < 1$, the EREG PDF estimate is likely to be overdispersive since the ensemble spread of the original forecasts ($K = 1$) is greater than K_N. In this case the ensemble spread needs to be reduced by applying (2.15) with $K = K_N$ and computing kernel distributions from (2.16a)–(2.16c).

The transformation given in (2.15) can be used to alter the ensemble for specific purposes provided that K stays within the range $0 < K < K_{\max}$. For example, K might be set to $K_{\max}$ to translate the ensemble forecasts into a series of calibrated point (deterministic) forecasts, equivalent to ''inflation'' in MOS equations (Klein et al. 1959). Inflation produces a bias-corrected set of point forecasts whose variance is the same as that of the observations over the dependent data sample. At the other extreme, setting $K = 0$, implies that $R_I'' = R_m$, and ensemble regression becomes standard linear regression on the ensemble mean. This transformation would be appropriate if the data suggested that the information from individual ensemble members worsens the forecast based on the ensemble mean alone. Intermediate values of K can also be tested in conjunction with verification scores that are appropriate for use on probabilistic predictions in an attempt to improve the forecasts. The PDF estimated from an ensemble regression can be generated for a variety of K values, and the value that produces the best result when measured by a given scoring rule and on a given set of forecasts can be selected to provide an estimate of an optimum ensemble spread for that score.

3. Ensemble calibration applied to Niño-3.4 SSTs

a. Forecast description

EREG was tested on sea surface temperature (SST) forecasts for the Niño-3.4 region from NCEP's CFS model (Saha et al. 2006). Niño-3.4 SSTs (mean SSTs between 5°N and 5°S and 170° and 120°W) correlate well with the state of the El Niño–Southern Oscillation (ENSO) (Barnston et al. 1997) and, therefore, are an important indicator for climate anomalies over many parts of the globe.

The CFS is an operational coupled ocean–atmosphere model that is currently run twice daily to produce forecasts for up to 6 months in advance. A CFS ensemble forecast is typically produced to support the Climate Prediction Center's (CPC) operational climate outlooks issued in the middle of each month (Barnston et al. 1994). The CFS ensemble forecast is formed from predictions initialized at various times in the previous month, all valid for the same target periods and thus similar to the strategy used for lagged averaged forecasting (Hoffman and Kalnay 1983; Roads 1988). The use of a lagged average forecast is common in climate forecasts since perturbations in the initial state have little effect on the atmospheric seasonal forecast beyond 1 month (Phelps et al. 2004, and about 2 months for oceanic predictions (Vialard, et al. 2005). Because there is usually close to a 1-month lag between the latest data used for the CPC seasonal forecasts and the start of the first 3-month target season, the effects of different lead times of the ensemble members are expected to be minor in most circumstances, although they may have an impact on early leads.

A retrospective archive of the CFS model is available from three sets of five consecutive daily forecasts initialized near the start, middle, and end of each month between 1981 and 2004. Beginning in August 2004, the CFS model became operational and was run daily, so the ensemble was obtained from the 15 most recent daily runs available early in each month. Lead time is defined as the approximate amount of time, in months, between the data used for the latest CFS model run and the start of the target period. Three-month averages (referred to here as seasonal averages) of SSTs are formed from the monthly means from the CFS.

We have translated the CFS ensemble forecasts for 3-month-mean Niño-3.4 SSTs into a cumulative probability distribution function (CPDF) in a standardized format for ease of data handling. The standard format expresses values of SST that are expected to be equal to or exceed 2%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, and 98% of the time, so that the forecast precision is the same regardless of how many ensembles were used, or how the data distribution was obtained.

The continuous ranked probability score (CRPS) was used to verify the probabilistic forecasts (Matheson and Winkler 1976; Hersbach 2000). The CRPS measures the squared difference between the forecast CPDF and the CPDF of the observation over the complete range of the observations, $-\infty < x < \infty$, where x refers to the range of the forecast values. Note that the CPDF of the observation, y, is simply $F(x) = 0$ for $x < y$ and $F(x) = 1$ for $x \geq y$. A CRPS skill score (CRPSS) was formed by comparison with the scores obtained from climatological probabilities, $\mathrm{CRPS}_{\mathrm{CL}}$:

$$\mathrm{CRPSS} = 1 - \frac{\mathrm{CRPS}}{\mathrm{CRPS}_{\mathrm{CL}}}.$$

The CPDF for the standard forecast format was produced by assuming a linear increase in probabilities between the specified probability values (2%–98%). The distribution tails were supplied by assigning points representing the values where the CPDF reaches 0 and 1. These points were assigned a value that minimizes the CRPS for a linearly increasing CPDF outside of the forecaster-specified interval, assuming that the actual distribution of observations outside the interval is Gaussian.

A CPDF forecast was generated from the ensemble by three different methods. For one method, the ensembles were translated directly into CPDF form assuming a linear increase between the N-ordered ensemble members. The CFS prediction was assumed to be at the median of that member's forecast distribution, with each member representing $1/N$ of the total. The tails of the distribution were obtained by applying the CRPS-minimizing linearly increasing CPDF to the ends of the distribution as defined above (see Fig. 2). The piecewise linear CPDF obtained from the N ensemble members was then interpolated to the standard format and the CRPS was computed from that forecast as described above. We referred to these as the "original" forecasts.

A second method of translation used only the information in the forecast ensemble means, processed by standard linear regression (REG) as described in section 2a with the CPDF obtained from a high-resolution integration of the PDF and then expressed in standard format. In the third method, the EREG method was applied to individual ensemble members as described in sections 2c and 2d.

We processed the historical forecasts for both the REG and EREG using cross validation (Michaelsen 1987), in which each target year was removed from the equation development sample, together with two additional years, chosen randomly. Climatology for a given

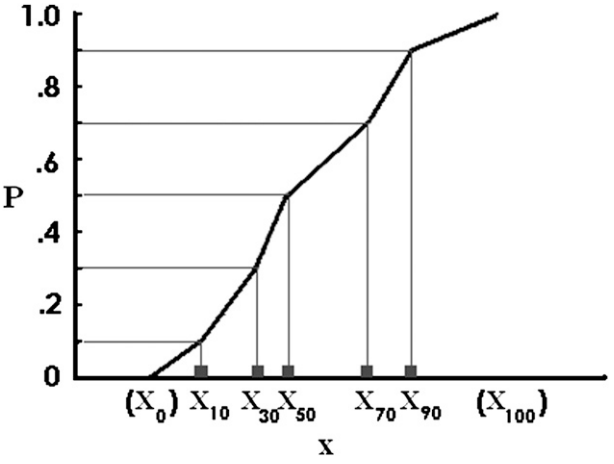

FIG. 2. Schematic diagram illustrating the translation of the original ensemble members (represented by squares) to a cumulative probability distribution function for $N = 5$. The CPDF is produced by a linearly increasing CPDF between ensemble forecasts. The lowest member of a five-member ensemble is assumed to represent the 10th percentile of the CPDF (X_{10}), etc. Both X_0 and X_{100} are set to minimize the CRPS for a linearly increasing CPDF assuming the tails are Gaussian distributed.

cross-validation trial was always computed from the 30 most recent remaining years.

b. Results

Results for all forecasts made between January 1982 and December 2005 are shown in Table 1. The original CFS ensemble members are not competitive with the regression-calibrated forecasts in the CRPS evaluation at any lead time. Differences in skill between the two regression-based postprocessing methods are very small, which is to be expected since they are both based on the same forecast.

The differences in CRPSS between the various methods were tested for significance. The score differences between the models exhibit much less month-to-month variability than the scores themselves do and they have some month-to-month dependence. A rough estimate of the effective sample size was obtained by examination of lag correlations (Thiébaux and Zwiers 1984) and suggests that an effective sample size of about 100 might be appropriate for these data. While accurate assessment of significance would require Monte Carlo resampling tests, these rough tests indicate that the differences between the scores for REG and EREG are not significant at the 5% level at any lead time.

Forecasts were translated into the probabilities that the observation would fall within one of three categories: below, near, or above normal Niño-3.4 SSTs based on the lower, middle, or upper third of the climatological distribution, respectively. This is a common format for seasonal forecasts and is measured here by a three-category ranked probability skill score (RPSS; Epstein 1969b; Murphy 1970) (see the columns under RPSS-3 in Table 1). Results hint that the EREG is slightly favored over REG on lead 1 and beyond when measured by RPSS-3, although again, differences are not significant at the 5% level.

Even though these results show the EREG and REG to be nearly identical for most lead times, EREG makes more direct use of the ensembles and, thus, better represents the information from the CFS. There is some suggestion that EREG improves the three-category probabilities.

The reliability diagrams for the 0- and 5-month lead forecasts are shown in Fig. 3. Because forecasts were generated from a variable width interval with fixed probability thresholds, the sample size is the same for all probability bins in Fig. 3. This is in contrast with most reliability diagrams in the literature, which show the reliability for specific events, and therefore, some bins have more data than others. Reliability diagrams applied to forecasts specified for fixed probability thresholds effectively measure the same information contained in ranked histograms (Anderson 1996; Talagrand et al. 1997; Hamill and Colucci 1997) and are subject to many of the same cautionary issues in their interpretation as discussed by Hamill (2000).

TABLE 1. Verification scores for CFS forecasts of seasonal mean SSTs in the Niño-3.4 region for the period 1982–2005. CRPSS and three-category RPS skill scores (RPSS-3) of probabilistic predictions based on the original CFS ensemble (Orig), CFS probabilities based on linear regression of the ensemble mean (REG), and those based on ensemble regression (EREG) are shown along with the mean absolute error (MAE) with respect to the forecast median value.

	CRPSS			RPSS-3			MAE (°C)		
Lead (months)	EREG	REG	Orig	EREG	REG	Orig	EREG	REG	Orig
0	0.559	0.556	0.509	0.607	0.610	0.551	0.333	0.335	0.370
1	0.500	0.497	0.419	0.553	0.550	0.499	0.377	0.378	0.439
2	0.445	0.444	0.348	0.505	0.501	0.445	0.419	0.418	0.503
3	0.397	0.398	0.295	0.443	0.439	0.368	0.456	0.453	0.546
4	0.349	0.350	0.245	0.400	0.397	0.315	0.491	0.491	0.581
5	0.307	0.308	0.175	0.365	0.360	0.295	0.520	0.520	0.614

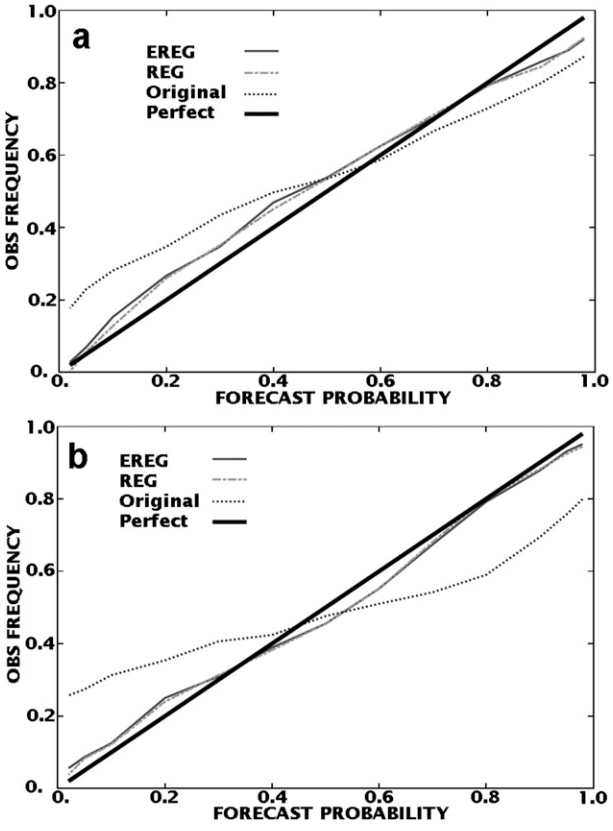

FIG. 3. Reliability diagrams for CFS forecasts for Niño-3.4 SSTs for lead times of (a) 0 and (b) 5 months. Forecasts are produced using three methods: EREG, REG, and direct translation from the original ensembles (original). Data are from cross-validated results for the years 1981–2005 with all initial times combined.

The reliability of the forecasts shows that the poor performance of the original CFS is due primarily to poor calibration leading to an overconfident forecast (a slope of less than one crossing the perfect reliability line near the median). Both regression methods produced reliable probabilities as evidenced in Fig. 3.

The CRPSS for varying K (2.15) for leads of 1 and 5 stratified by initial time are displayed in Figs. 4 and 5, respectively. Here, K expresses the ratio of the ensemble spread between the transformed and original ensembles prior to the regression. Values summarized in Table 1 and Fig. 3 are for $K = 1$, although the yearly average of the data in Fig. 3 is slightly different from the corresponding numbers in Table 1 because this sample includes additional data from February–December 1981. Results displayed are for Niño-3.4 SST forecasts initialized in the months of December, January, or February (D, J, or F); March, April, or May (M, A, or M); June, July, or August (J, J, or A); and September, October, or November (S, O, or N). In general, the CRPSS remains nearly constant from $K = 0$ to about $K = 0.8$, and then falls steadily until $K = K_{\max}$ [see (2.16)]. The value of $K_{\max}$ varied for each initial month and was usually between 1.5 and 2 for these data. Where K in Figs. 4 and 5 exceeded $K_{\max}$, scores were obtained from a kernel width of near zero, duplicating the results obtained from the K value where the kernel width first approached zero. The CPDF for $K = K_{\max}$ is a step function increasing about $1/N$ each time an ensemble member's forecast value is passed. Note that the scale of the plots in Figs. 4 and 5 varies according to the CRPS score range, and that the scores are not dramatically lower than the optimum even for large K. Because the forecast when $K = K_{\max}$ is essentially a series of calibrated point forecasts, a comparison of these values with the original forecasts (orig in Table 1) shows the benefits of using calibration as opposed to the benefit kernel smoothing. The calibrated ensembles are considerably better than the original forecasts, with the yearly average CRPSS for 1- and 5-month leads of 0.535 and 0.274, respectively, compared to corresponding values for the original (uncalibrated) ensemble in this sample of 0.410 and 0.184.

4. Discussion and conclusions

EREG is a statistical model designed for use in ensemble forecast problems. It has been shown that for such a system, the expected linear least squared solution and associated error estimates are relatively simple functions of σ_Y, σ_{F_m}, R_m, and R_I. The PDF of the ensemble forecast is estimated from the normalized sum of the Gaussian errors around each ensemble member in a manner similar to Gaussian kernel smoothing (Silverman 1986), except with kernels centered on the EREG-calibrated ensemble forecasts and kernel widths based on the regression error estimates. A linear transformation of the original model's ensemble spread can be employed in conjunction with a suitable probabilistic verification score to improve the predictions.

EREG closely resembles the ensemble dressing approach to ensemble calibration but has several important advantages. First, the bias correction and kernel estimation procedures are integrated properly for a least squares fit to the data. Second, the EREG model puts the problem into a regression framework, which allows the application of statistical theory developed for regression to be applied to ensemble forecasts (analysis of variance, error estimation, weighted regression, etc.). The kernel dressing approaches of Wang and Bishop (2005), Fortin (2006), and Roulston and Smith (2003) apply kernels to bias-corrected, but not skill-damped, ensembles. This can be expected to significantly degrade the accuracy of the calibrated ensemble, especially in low-skill situations. This

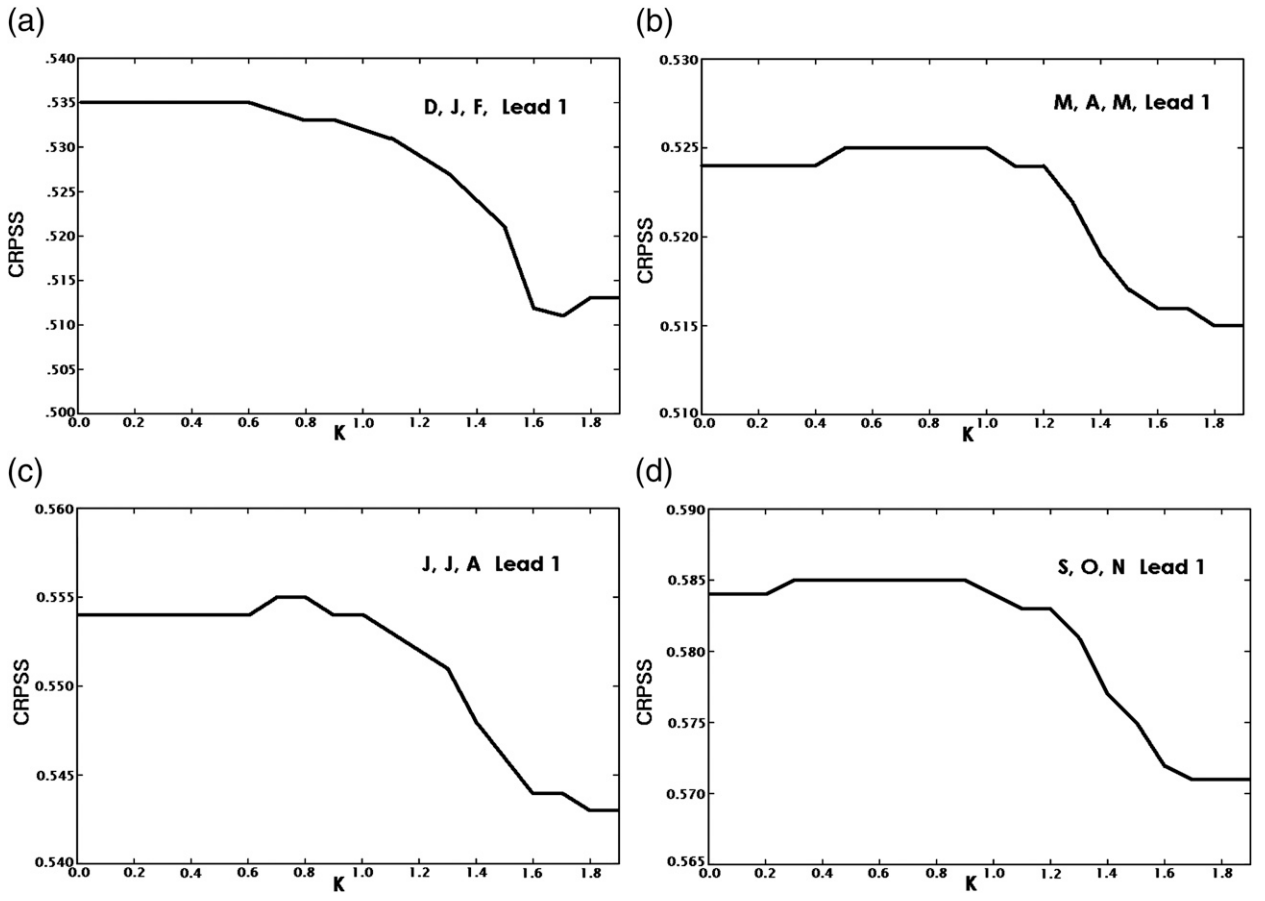

FIG. 4. Cross-validated CRPSSs for Niño-3.4 SST forecasts from the CFS for the period 1981–2005 for 1-month lead forecasts initialized in the winter months (DJF), spring (MAM), summer (JJA), and fall (SON) for varying ensemble spread values. Here, K is the fraction of the original model spread retained in the transformed forecasts, prior to regression calibration.

may explain the ensemble dressing method's relatively poor calibration in comparison with other ensemble calibration methods (Wilks 2006; Wilks and Hamill 2007).

Fortin (2006) makes an argument for nonequal kernel widths, and this has some support from regression theory when uncertainty in the regression line is considered (see Draper and Smith 1981, section 1.4). The theory indicates that kernel widths should increase with increasing distance from the sample mean and not necessarily in relation to the ensemble mean as Fortin's model would indicate.

The Bayesian model averaging (BMA) approach as outlined by (Raftery et al. 2005; Wilson et al. 2007) uses a kernel density fit to the weighted ensemble members, with weights determined by Bayesian processing. The theory presented here suggests that weights from BMA can be used together with EREG (using weighted regression rather than assuming equal weights as presented here) to derive an appropriate final calibration and kernel density fit to the data.

The results presented here give theoretical support for the approach outlined by Glahn et al. (2009). They applied multiple linear equations based on model output from ensemble means to individual members of the ensemble with good results on short-range weather forecasts. Appendix C indicates that the approach described here applies to the multiple-predictor case and therefore is applicable to multiple linear regression. The appropriate kernel distribution widths are not as easy to compute for multiple linear regression because the individual member correlation for R_b in (2.12) cannot easily be estimated from the data without actually generating forecasts from the individual ensemble members in a second pass through the data and computing R_I from those regression estimates (R_m can be estimated from the multiple correlation coefficient). The method of estimating the kernel width used by Glahn et al. (2009) is an alternative to the method presented here.

It is likely that the regression theory presented here can be expanded to include the treatment of ensemble members of varying skill such as would be found in multimodel ensembles. This would be expected to introduce many complications that are beyond the scope

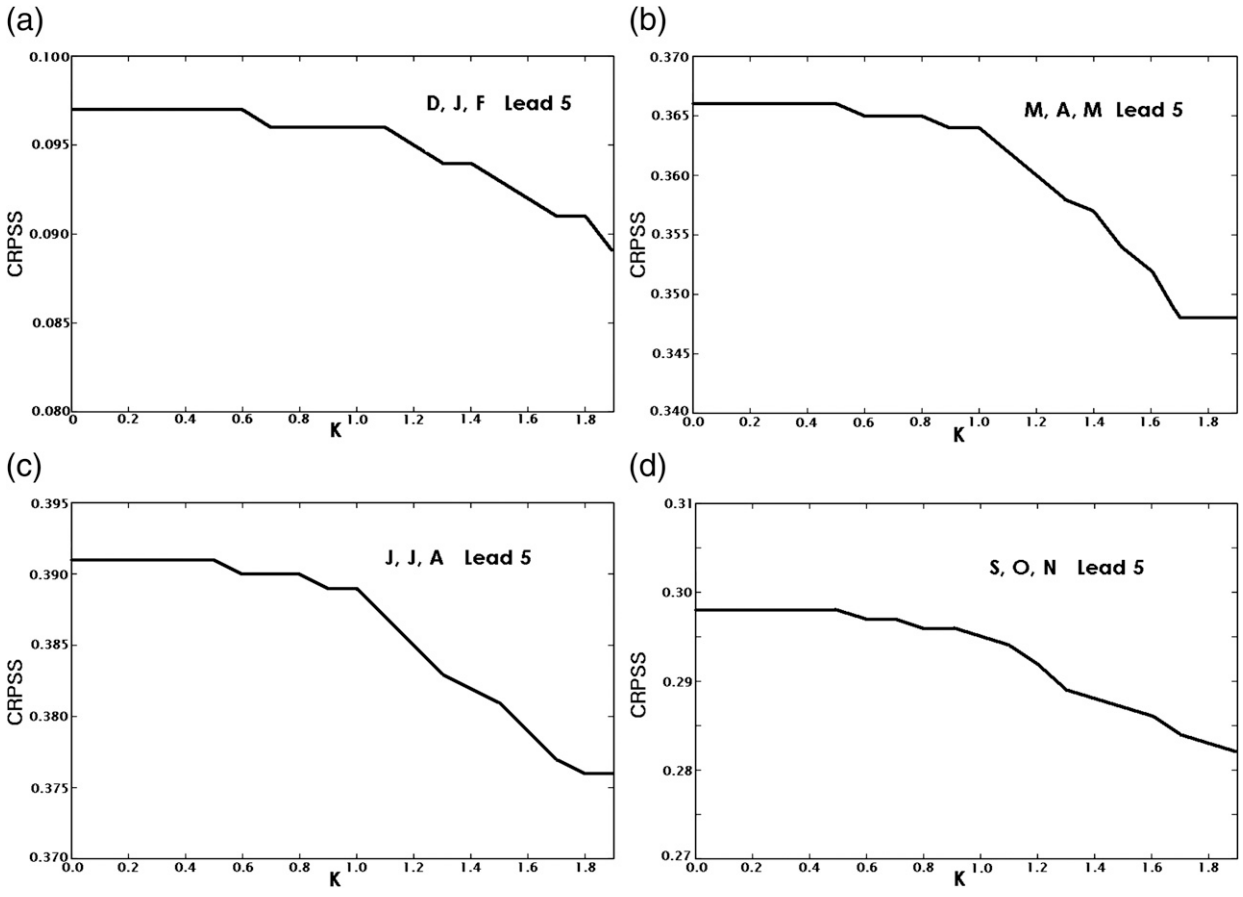

FIG. 5. Same as in Fig. 4 but for a 5-month lead time.

of this work (such as whether to vary the kernel width for less skillful models). If bias-corrected forecasts from ensemble members generated by other models are regarded as additional solutions whose errors are expected to be no different that the others in the event they are the best among the ensemble members, then the derivation of a skill-weighted multimodel ensemble regression is straightforward. This approach is the basis of the operational consolidation forecast recently developed at CPC (O'Lenic et al. 2008).

An examination of seasonal Niño-3.4 SST forecasts from the CFS suggests that the skill levels of the EREG ($K = 1$) and REG models are nearly the same for all lead times. While the score differences between the two methods are small for these data, the EREG procedure has the advantage of utilizing the uncertainty estimate from the dynamic model, rather than pooled statistics from the entire sample.

The CRPSS for CFS Niño-3.4 SST forecasts is not sensitive to spread transformation, and scores obtained from a PDF generated from a calibrated ensemble count (K near its maximum value) are not much worse than those from optimized spread. There is some evidence that the CFS model spread for this element is slightly higher than optimum, and that the CRPSS can be improved slightly by reducing the spread to about 0.8 of its original value prior to regression calibration. Further reductions in spread have little effect on scores.

Acknowledgments. The authors are grateful to Zoltan Toth for supportive discussions regarding ensemble prediction procedures and to the various reviewers for their helpful suggestions. Results for spread optimization were obtained with the help of Georgia Tech student Julie Simon working under the NOAA student scholarship program.

APPENDIX A

Derivation of (2.5)

Starting with the relationship

$$(F_i - Y) = (F_i - F_m) + (F_m - Y),$$

$$\left\langle \sum_{i=1}^{N} (F_i - Y)^2 \right\rangle = \left\langle \sum_{i=1}^{N} (F_i - F_m)^2 \right\rangle + 2\left\langle \sum_{i=1}^{N} (F_i - F_m)(F_m - Y) \right\rangle + \left\langle \sum_{i=1}^{N} (F_m - Y)^2 \right\rangle,$$

$$\left\langle \sum_{i=1}^{N} (F_i - Y)^2 \right\rangle = \left\langle \sum_{i=1}^{N} (F_i - F_m)^2 \right\rangle + 2\left\langle (F_m - Y) \sum_{i=1}^{N} (F_i - F_m) \right\rangle + \left\langle \sum_{i=1}^{N} (F_m - Y)^2 \right\rangle.$$

Noting that for each case, $\Sigma_i(F_i - F_m) = 0$,

$$\left\langle \sum_{i=1}^{N} (F_i - Y)^2 \right\rangle = \left\langle \sum_{i=1}^{N} (F_i - F_m)^2 \right\rangle + \left\langle \sum_{i=1}^{N} (F_m - Y)^2 \right\rangle.$$

After dividing by N to express the relationship in terms of the means of individual members rather than the sums, and noting that the mean ensemble spread, $\langle E^2 \rangle$ is

$$\langle E^2 \rangle = \left\langle \frac{1}{N} \sum_{i=1}^{N} (F_i - F_m)^2 \right\rangle,$$

the relationship becomes

$$\langle (F_i - Y)^2 \rangle = \langle E^2 \rangle + \langle (F_m - Y)^2 \rangle.$$

APPENDIX B

Derivation of (2.7)

Starting with the definitions of R_m and R_I,

$$R_m = \frac{\langle (F_{mj} - \langle F_m \rangle)(Y_j - \langle Y \rangle) \rangle}{S_m S_Y} \quad \text{and} \tag{B.1}$$

$$R_I = \frac{\langle (F_{i,j} - \langle F \rangle)(Y_j - \langle Y \rangle) \rangle}{S_I S_Y}. \tag{B.2}$$

Note that the covariance (numerators in the above relationships) between the ensemble mean and the observations is identical to that of the individual ensemble members and the observations as demonstrated below:

$$\langle (F_{i,j} - \langle F \rangle)(Y_j - \langle Y \rangle) \rangle = \frac{1}{MN} \sum_{j=1}^{M} \sum_{i=1}^{N} (F_{i,j} - \langle F \rangle)(Y_j - \langle Y \rangle)$$

$$= \frac{1}{MN} \sum_{j=1}^{M} \left\{ (Y_j - \langle Y \rangle) \sum_{i=1}^{N} [(F_{i,j} - F_{mj}) + (F_{mj} - \langle F \rangle)] \right\},$$

$\sum_{i=1}^{N} (F_{i,j} - F_{mj}) = 0$; therefore,

$$\langle (F_{i,j} - \langle F \rangle)(Y_j - \langle Y \rangle) \rangle = \frac{1}{MN} \sum_{j=1}^{M} \left\{ (Y_j - \langle Y \rangle) \sum_{i=1}^{N} (F_{mj} - \langle F \rangle) \right\}$$

$$= \frac{1}{MN} \sum_{j=1}^{M} \{ (Y_j - \langle Y \rangle) N (F_{mj} - \langle F \rangle) \}$$

$$= \frac{1}{M} \sum_{j=1}^{M} (F_{mj} - \langle F \rangle)(Y_j - \langle Y \rangle)$$

$$= \langle (F_{mj} - \langle F \rangle)(Y_j - \langle Y \rangle) \rangle.$$

Since $\langle F_m \rangle - \langle F \rangle$, the numerators in (B.1) and (B.2) are equal for an ensemble forecast, and their correlations are related according to (2.7).

APPENDIX C

Expected Values of Best-Member Regression Coefficients

This proof follows similar reasoning to that of testing for bias in regression coefficients as outlined by Draper and Smith (1981, section 2.12). Because of its importance, we will discuss this in terms of a generalized regression.

Let $\mathbf{Y}$ be the vector of the observations (predictands) and $\mathbf{F}$ be a matrix of the predictors from our ensemble, illustrated here for the one-predictor case:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ : \\ Y_M \end{bmatrix} \quad \text{and} \quad \mathbf{F} = \begin{bmatrix} 1 & F_1 \\ 1 & F_2 \\ : & : \\ 1 & F_M \end{bmatrix}.$$

The true statistical model is postulated to be

$$\mathbf{Y} = \boldsymbol{\alpha}\mathbf{F}_b + \boldsymbol{\varepsilon}_b,$$

where $\boldsymbol{\alpha}$ is the vector of the regression coefficients, $\boldsymbol{\varepsilon}_b$ is a vector of the errors, and $\mathbf{F}_b$ is the vector of predictors obtained from the best-member forecasts:

$$\boldsymbol{\alpha} = [\alpha_0\, \alpha_1], \quad \boldsymbol{\varepsilon}_b = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ : \\ \varepsilon_M \end{bmatrix}.$$

The least squares estimate of α is the vector, $\mathbf{a}$:

$$\mathbf{a} = [a_0 a_1] \quad \text{and}$$

$$\mathbf{a} = (\mathbf{F}_b^{\mathrm{T}}\mathbf{F}_b)^{-1}\mathbf{F}_b^{\mathrm{T}}\mathbf{Y},$$

where $\mathbf{F}_b^{\mathrm{T}}$ is the transpose of the matrix $\mathbf{F}_b$.

If the EREG assumptions are correct, then the expected value of $\mathbf{Y}$ for each case, given N equally likely ensemble members, is computed by applying the regression estimate for $\mathbf{Y}$, $\mathbf{Y}' = \mathbf{a}\mathbf{F}_b$, to each member:

$$\mathrm{expv}(\mathbf{Y}) = \mathbf{a}\mathbf{F}_m,$$

where $\mathbf{F}_m$ is the matrix of the ensemble mean predictors (forecasts).

Now, let us examine the regression equation based on the ensemble mean:

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{F}_m + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ is the true model's coefficients, $\mathbf{F}_m$ is the matrix of ensemble mean predictors (the model forecasts in the single-predictor case), and $\boldsymbol{\varepsilon}$ is the vector of errors. Note that some caution is in order when dealing with certain types of complex predictor variables designed to simulate nonlinear responses (e.g., ''dummy'' variables). We restrict our discussions here to predictors obtained directly from the individual members.

This model has a least squared solution for $\boldsymbol{\beta}$ of

$$\mathbf{b} = (\mathbf{F}_m^{\mathrm{T}}\mathbf{F}_m)^{-1}\mathbf{F}_m^{\mathrm{T}}\mathbf{Y}.$$

Following the procedure for examining the bias in regression equations, the expected values of $\mathbf{b}$ and $\mathbf{Y}$ are related as follows:

$$\mathrm{expv}(\mathbf{b}) = (\mathbf{F}_m^{\mathrm{T}}\mathbf{F}_m)^{-1}\mathbf{F}_m^{\mathrm{T}}\mathrm{expv}(\mathbf{Y}).$$

Now, we ask whether the expected value of the vector constants $\mathbf{a}$ and $\mathbf{b}$ are the same. Substituting expv($\mathbf{Y}$) from the best-member equation gives

$$\mathrm{expv}(\mathbf{b}) = (\mathbf{F}_m{}^{\mathrm{T}}\mathbf{F}_m)^{-1}(\mathbf{F}_m{}^{\mathrm{T}}\mathbf{F}_m)\mathbf{a} \quad \text{and}$$

$$\mathrm{expv}(\mathbf{b}) = \mathbf{a}.$$

If the EREG assumptions are correct, then the expected values of the coefficients of the regression equation based on the best member and those of a regression based on the ensemble mean are the same.

APPENDIX D

Maximum Value of K for an N-Member Ensemble Based on Sampling Theory

A maximum value of K based on sampling theory can be related to the ensemble size, N, as follows. An unbiased estimate of the true value of the mean squared ensemble spread, σ_E^2 (population variance), is given by

$$\sigma_E^2 = \frac{N}{N-1}\langle E^2 \rangle.$$

The bias in the estimate, $\langle E^2 \rangle$, arises because of uncertainty in the ensemble mean, so a Gaussian error distribution with standard deviation, σ_u, is assumed to surround each ensemble member to represent this uncertainty. Variance is additive, so

$$\sigma_E^2 = \langle E^2 \rangle + \sigma_u^2 \quad \text{and}$$

$$\sigma_u^2 = \sigma_E^2 \left(1 - \frac{N-1}{N}\right).$$

Since σ_E^2 is the expected value of the true residual variance about the ensemble mean regression estimate, (2.3b) implies

$$\sigma_E^2 = c\sigma_Y^2(1 - R_m^2).$$

We assume the ensembles to be a faithful representation of possible solutions and need to calculate the maximum ensemble spread that is consistent with the skill and a sample size of N, assuming the underlying distributions are Gaussian. From EREG, the residual variance is given by (2.14),

$$\sigma_u^2 = c\sigma_Y^2(1 - R_u^2),$$

where R_u is the expected correlation between the closest ensemble member and the observation given N members:

$$\sigma_u^2 = c\sigma_Y^2(1 - R_u^2) = c\sigma_Y^2(1 - R_m^2)\left(1 - \frac{N-1}{N}\right).$$

We seek a transformation constant, K_N, which when applied to (2.15), will produce the maximum spread sustainable for N members and for normally distributed errors. Noting the relationship R_u and the correlation between the individual members of a *transformed* forecast, R_I'',

$$R_u = \frac{R_m^2}{R_I''}.$$

From (2.6) and (2.7), a relationship for K_N can then be formulated and expressed in terms of statistical parameters from the original ensemble as shown in (2.18):

$$K_N = \left[\frac{\dfrac{N-1}{N}\left(\dfrac{1}{R_m^2} - 1\right)}{\dfrac{R_m^2}{R_I^2} - 1}\right]^{1/2} = \left(\frac{N-1}{N}\right)^{1/2} K_{\max}.$$

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Barnett, T. P., 1995: Monte Carlo climate forecasting. *J. Climate,* **8,** 1005–1022.

Barnston, A. G., and Coauthors, 1994: Long lead seasonal forecasts—Where do we stand? *Bull. Amer. Meteor. Soc.,* **75,** 2097–2114.

——, M. Chelliah, and S. B. Goldberg, 1997: Documentation of a highly ENSO-related SST region in the equatorial Pacific. *Atmos.–Ocean,* **35,** 367–383.

——, Y. He, and D. A. Unger, 2000: A forecast product that maximizes utility for state-of-the-art seasonal climate prediction. *Bull. Amer. Meteor. Soc.,* **81,** 1271–1279.

Benzi, R., and A. Speranza, 1989: Statistical properties of low frequency variability in the Northern Hemisphere. *J. Climate,* **2,** 367–379.

Draper, N., and H. Smith, 1981: *Applied Regression Analysis*. John Wiley and Sons, 709 pp.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting,* **13,** 1132–1147.

Epstein, E., 1969a: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

——, 1969b: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Fortin, V., A. C. Favre, and M. Said, 2006: Probabilistic forecasting from ensemble prediction systems: Improving on the best member method by using a different weight and dressing kernel for each member. *Quart. J. Roy. Meteor. Soc.,* **132,** 1349–1369.

Glahn, B., and Coauthors, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.,* **137,** 246–268.

Glahn, H. R., and D. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasts. *J. Appl. Meteor.,* **11,** 1203–1211.

Gneiting, T., A. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118.

Hamill, T. M., 2000: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, and J. S. Whitaker, 2006: Quantitative precipitation forecasts based on reforecast analogs: Theory and applications. *Mon. Wea. Rev.,* **134,** 3209–3229.

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecasting using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte-Carlo forecasting. *Tellus,* **35A,** 100–118.

Klein, W. H., F. Lewis, and I. Enger, 1959: Objective prediction of 5-day mean temperature during winter. *J. Meteor.,* **16,** 672–682.

Kumar, A., and M. P. Hoerling, 1995: Prospects and limitations of seasonal atmospheric GCM predictions. *Bull. Amer. Meteor. Soc.,* **76,** 335–345.

Leith, C. E., and Coauthors, 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Livezey, R. E., 1990: Variability of skill of long range forecasts and implications for their use and value. *Bull. Amer. Meteor. Soc.,* **71,** 300–309.

——, and M. M. Timofeyeva, 2008: The first decade of long-lead U.S. seasonal forecasts: Insights from a skill analysis. *Bull. Amer. Meteor. Soc.,* **89,** 843–855.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.,* **22,** 1087–1096.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.,* **26,** 1589–1600.

Murphy, A. H., 1970: The ranked probability score and the probability score: A comparison. *Mon. Wea. Rev.,* **98,** 917–924.

O'Lenic, E. A., 2008: Developments in operational long-range climate prediction at CPC. *Wea. Forecasting,* **23,** 496–515.

Phelps, M. W., A. Kumar, and J. J. O'Brien, 2004: Potential predictability in the NCEP CPC seasonal forecast system. *J. Climate,* **17,** 3775–3785.

Raftery, A. E., and Coauthors, 2005: Using Bayesian model averaging to calibrate forecast model ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Roads, J. O., 1988: Lagged averaged predictions in a predictability experiment. *J. Atmos. Sci.,* **45,** 147–162.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamic and statistical ensembles. *Tellus,* **55A,** 16–30.

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

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate,* **19,** 3483–3517.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 175 pp.

Sivillo, J. K., J. E. Ahlquist, and Z. Toth, 1997: An ensemble forecasting primer. *Wea. Forecasting,* **12,** 809–818.

Stern, W., and K. Miyakoda, 1995: Feasibility of seasonal forecasts inferred from multiple GCM simulations. *J. Climate,* **8,** 1071–1085.

Talagrand, O., R. Vauland, and B. Strauss, 1997: Evaluation of probabilistic predictions systems. *Proc. Workshop on Predictions,* Reading, United Kingdom, ECMWF, 1–25. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG29AX, United Kingdom.]

Thiébaux, H. J., and F. W. Zwiers, 1984: The interpretation and estimation of effective sample size. *J. Climate Appl. Meteor.,* **23,** 800–811.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

van den Dool, H. M., 1994: Searching for analogues, how long must we wait? *Tellus,* **46A,** 314–324.

Vialard, J., and Coauthors, 2005: An ensemble generation technique for seasonal forecasting with an ocean–atmosphere coupled model. *Mon. Wea. Rev.,* **133,** 441–453.

Wang, X., and C. H. Bishop, 2005: Improvements of ensemble reliability using a new dressing kernel. *Quart. J. Roy. Meteor. Soc.,* **131,** 965–986.

Wilks, D. S., 2006: Comparison of ensemble MOS methods in the Lorenz 96 setting. *Meteor. Appl.,* **13,** 243–256.

——, and T. M. Hamill, 2007: Comparison of MOS-ensemble methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390.

Wilson, L. J., and Coauthors, 2007: Calibrated surface temperature forecasts from the Canadian Ensemble Prediction System using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 1365–1385.