Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts

THOMAS M. HAMILL

University of Colorado and NOAA–CIRES Climate Diagnostics Center, Boulder, Colorado

JEFFREY S. WHITAKER

NOAA–CIRES Climate Diagnostics Center, Boulder, Colorado

XUE WEI

University of Colorado and NOAA–CIRES Climate Diagnostics Center, Boulder, Colorado

(Manuscript received 4 June 2003, in final form 28 November 2003)

ABSTRACT

The value of the model output statistics (MOS) approach to improving 6–10-day and week 2 probabilistic forecasts of surface temperature and precipitation is demonstrated. Retrospective 2-week ensemble ''reforecasts'' were computed using a version of the NCEP medium-range forecast model with physics operational during 1998. An NCEP–NCAR reanalysis initial condition and bred modes were used to initialize the 15-member ensemble. Probabilistic forecasts of precipitation and temperature were generated by a logistic regression technique with the ensemble mean (precipitation) or ensemble mean anomaly (temperature) as the only predictor. Forecasts were computed and evaluated during 23 winter seasons from 1979 to 2001.

Evaluated over the 23 winters, these MOS-based probabilistic forecasts were skillful and highly reliable. When compared against operational NCEP forecasts for a subset of 100 days from the 2001–2002 winters, the MOS-based forecasts were comparatively much more skillful and reliable. For example, the MOS-based week 2 forecasts were more skillful than operational 6–10-day forecasts. Most of the benefit of the MOS approach could be achieved with 10 years of training data, and since sequential sample days provided correlated training data, the costs of reforecasts could also be reduced by skipping days between forecast samples.

MOS approaches will still require a large dataset of retrospective forecasts in order to achieve their full benefit. This forecast model must remain unchanged until reforecasts have been computed for the next model version, a penalty that will slow down the implementation of model updates. Given the substantial improvements noted here, it is argued that reforecast-based MOS techniques should become an integral part of the medium-range forecast process despite this cost. Techniques for computing reforecasts while minimizing the impact to operational weather prediction facilities and model development are discussed.

1. Introduction

Improving weather forecasts is a primary goal of the National Oceanic and Atmospheric Administration (NOAA) and other weather services. One commonly emphasized way to improve weather predictions has been to improve the accuracy of the numerical forecast models. Much effort has been expended to improve the estimate of the initial condition (e.g., Daley 1991; Parrish and Derber 1992; Courtier et al. 1994; Houtekamer and Mitchell 2001), to conduct forecasts with higher-resolution numerical models (e.g., Weisman et al. 1997; Kalnay et al. 1998; Buizza et al. 2003), and to incorporate more complex physical parameterizations of processes that occur below the grid scale. Within the last decade, ensemble forecast techniques (e.g., Toth and Kalnay 1993, 1997; Molteni et al. 1996; Houtekamer et al. 1996) have also been embraced as a tool for making probabilistic forecasts and for filtering the predictable from the unpredictable scales (via ensemble averaging).

There are forecast situations that are so intrinsically difficult that skill has not improved much despite the investment in large new computers and despite the millions of person hours invested in model development over the last 40 years. Medium-range weather forecasting is one such endeavor. The skill of these forecasts is marginal because of the inevitable rapid growth of errors through chaos (e.g., Lorenz 1969) and because of the steadier growth of model errors. In order to make a skillful medium-range forecast, forecasters must thus

Corresponding author address: Dr. Thomas M. Hamill, NOAA–CIRES CDC, R/CDC 1, 325 Broadway, Boulder, CO 80305-3328.
E-mail: tom.hamill@noaa.gov

be able to adjust for model systematic errors and be able to distinguish between features that are predictable and those that are unpredictable. As will be shown later, unprocessed numerical guidance is often not particularly useful; for example, probability forecasts that are derived from the National Centers for Environmental Prediction (NCEP) ensemble forecasts' relative frequency have no skill and are highly unreliable.

The format of forecasts issued by the NCEP Climate Prediction Center (CPC) implicitly reflects a judgment of what can be predicted skillfully and what cannot. Day-to-day details of synoptic-scale features are considered largely unpredictable, while shifts in the probability density function of averages over several days may be predictable. Consequently, CPC produces probability forecasts of time averages of the deviations from climatology. Specifically, CPC makes 6–10-day and week 2 (8–14 day) forecasts of daily average surface (2 m) temperature and precipitation tercile probabilities. These are forecasts of the probability that the temperature and precipitation averaged over these periods will be below the 33d or above the 67th percentile of the distribution of climatological observed temperatures and precipitation. Forecasters at CPC synthesize information from the NCEP ensemble prediction system as well as models from other weather services and other statistical tools. As will be shown, the skill of operational week 2 forecasts is currently quite low.

Another possible way of improving weather forecasts is to adjust the forecast model output based on a database of retrospective forecasts from the same model. The adjustment of dynamically based forecasts with statistical models has a rich history. Model output statistics (MOS) techniques (Glahn and Lowry 1972; Woodcock 1984; Glahn 1985; Tapp et al. 1986; Carter et al. 1989; Vislocky and Fritsch 1995, 1997) have been used widely since the 1970s. However, in recent years, the National Weather Service (NWS) has de-emphasized the use of MOS techniques based on fixed models; such an approach requires a large sample of forecasts from the same model to achieve their maximal benefit. This implies that a large number of retrospective forecasts must be run prior to implementation of a new model version and that the current forecast model be ''frozen'' until retrospective forecasts are computed for any planned new model version; changing the model numerics may change the forecasts' error characteristics, invalidating the regression equations developed with the prior model version. Consequently, decision makers at many weather prediction facilities have judged that forecast improvements will come much more rapidly if the model development is not slowed by the constraints of computing these retrospective forecasts.

Statistical algorithms like MOS improve on raw numerical forecasts by implicitly removing model bias and filtering the predictable from the unpredictable. Given the difficulty of doing this without statistical models, and given the marginal skill of current week 2 weather forecasts, we reconsider the value of statistical weather forecasting for this application. Specifically, we examine here whether a reduced-resolution ensemble prediction system calibrated from a set of prior numerical forecasts can produce forecasts that are more skillful than the products generated by human forecasters based on a variety of state-of-the-art, higher-resolution models. A T62, 28-level version of NCEP's Medium-Range Forecast (MRF) modeling system based on 1998 model physics was used to run ensemble ''reforecasts'' over the period 1979–2001. Statistically adjusting current T62 forecasts based on these prior forecasts is shown to produce substantial improvements in forecast skill, greatly exceeding the skill of the operational forecasts. We document the skill of these forecasts and examine how many retrospective forecasts are necessary to approach optimum skill. Given the improvements produced through the use of statistical techniques, we propose that reforecasting and the application of MOS-like statistical techniques should become an integral part of the medium-range numerical weather prediction process.

Below, section 2 will outline the forecast modeling system and provide details on the general statistical approaches used. Section 3 presents results, and section 4 concludes with a discussion of the implications of this research.

2. Experiment design

a. Forecast model, initial conditions, and verification data

A T62 resolution (roughly 200-km grid spacing) version of NCEP's MRF model (Kanamitsu 1989; Kanamitsu et al. 1991; Caplan et al. 1997) was used to generate an ensemble of 15-day forecasts over a 23-yr period from 1979 to 2001. Further details regarding the model formulation can be found in Newman et al. (2003).

A 15-member ensemble was produced every day of the 23 years with 0000 UTC initial conditions. The ensemble initial conditions consisted of a control initialized with the NCEP–National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996) and a set of seven bred pairs of initial conditions (Toth and Kalnay 1993, 1997) recentered each day on the reanalysis initial condition.

Forecasts were evaluated in two ways. First, the proposed MOS-based forecasts, described below, were evaluated over winter seasons [December–January–February (DJF)] in a 23-yr period from 1979 to 2001. A set of 484 stations in the conterminous United States (CONUS), Alaska, and Hawaii were used for this comparison. The large majority of these stations are in the conterminous United States (all dots in Fig. 1). These 484 stations were chosen as the subset of available cooperative network (co-op) stations with at least an 80%

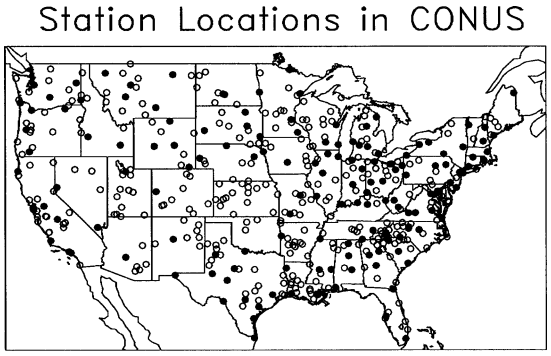


FIG. 1. Locations at which statistical forecast algorithms were evaluated in the conterminous United States. Filled circles indicate the subset of 153 stations where comparison of reforecast and CPC are performed for 100 days during the 2001 and 2002 winters. The union of filled and unfilled circles are the stations where the reforecast MOS algorithm is evaluated from 1979 to 2001.

complete record (Eischeid et al. 2000) from 1979 to 2001. Second, the MOS-based forecasts were compared against CPC operational forecasts for a set of 100 days during the winters of 2001 and 2002 (JFD 2001 and JF 2002). This comparison was performed at the subset of 153 stations where CPC forecasts were available (darkened dots in Fig. 1). The observed climatology used in these experiments was determined from 1971–2000 data, consistent with CPC practice.

b. *Logistic regression model and forecast/evaluation process*

Following the format of operational 6–10-day and week 2 forecasts produced at CPC, we produced forecasts of the probability distribution of precipitation and surface temperature at the stations. Probabilities were set for three categories, the lower, middle, and upper tercile of the distribution of observed anomalies from the mean climatological state. The method for determining the upper-and lower-tercile anomaly boundaries ($T_{2/3}$ and $T_{1/3}$, respectively) is discussed below.

A logistic regression technique (e.g., Wilks 1995; Applequist et al. 2002) was used for this experiment; the spatially interpolated ensemble mean forecast (precipitation) or forecast anomaly (surface temperature) was the only predictor. Separate regression analyses were performed for each observation location. By regressing on the ensemble mean rather than a single forecast, we exploited the ability of ensemble averaging to filter out the smaller, unpredictable scales and retain the larger, predictable ones.

The logistic regression model sets the probability that the observed anomaly V will exceed $T_{2/3}$ or $T_{1/3}$ according to the equation (here, for the upper tercile)

$$P(V > T_{2/3}) = 1 - \frac{1}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x)}, \qquad (1)$$

where x is the ensemble mean forecast or forecast anomaly, $\hat{\beta}_0$ and $\hat{\beta}_1$ are fitted regression coefficients.

We tested other possible predictors to use in the logistic regression analysis. A regression analysis using only the control forecast rather than the ensemble mean provided less skill, as discussed later. A regression analysis in the space of the leading canonical correlates (Wilks 1995) was less skillful than one assuming independent coefficients at each grid point. Similarly, some obvious candidates such as ensemble spread did not improve forecast accuracy. Figure 2 illustrates why spread was not a useful predictor; correlations of ensemble mean error and spread were uniformly low, never exceeding 0.2. It was not clear whether a stronger spread–skill relationship ought to exist at these scales and lead times. It is known that accurate estimates of spread require larger ensembles than accurate estimates of the mean (Compo et al. 2001) and that the current breeding method for generating perturbations is sub-

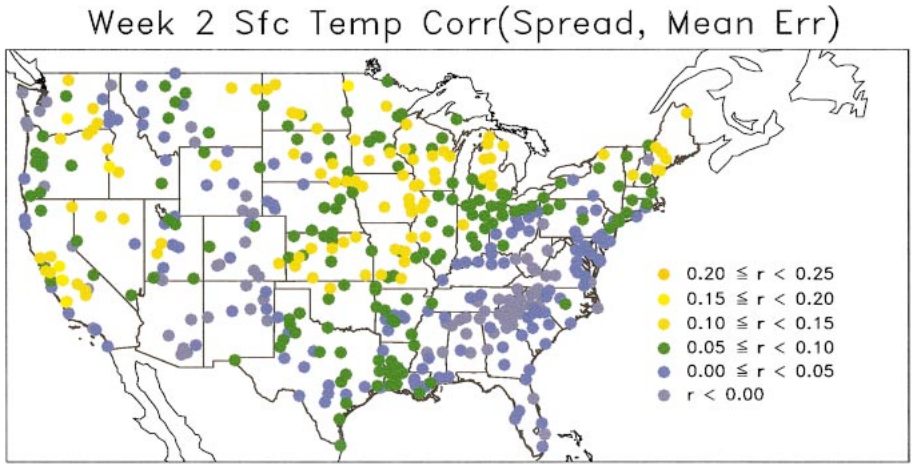


FIG. 2. Rank correlation between spread ensemble mean error using DJF samples from 1979 to 2003.

optimal (Hamill et al. 2000; Wang and Bishop 2003), so perhaps a stronger spread–skill is possible.

The process for producing and evaluating MOS forecasts is described here for week 2 forecasts of upper-tercile probabilities of surface temperature. Lower-tercile probabilities and 6–10-day probabilities were handled in an identical manner. Precipitation was handled somewhat differently and is described later. A separate regression analysis was performed for each day and each station. The regression parameters were determined using a dataset of ensemble mean forecast and observed week 2 anomalies from climatology. From these we compute the associated binary verification data: [$P(V > T_{2/3} = 1$] if the observation anomaly was above the upper tercile, and [$P(V > T_{2/3} = 0$] if the observation anomaly was less than or equal to the the upper tercile. Regression coefficients were determined through a cross-validation approach (Wilks 1995) to ensure the independence of the training and evaluation data. For example, given 23 years of available forecasts, when making forecasts for a particular year, the remaining 22 yr were used as training data. The same 22 yr were used to define the forecast climatology. The observation climatology was fixed using 1971–2000 data.

The generation and evaluation of tercile probability forecasts followed a three-step process. The process is described for a week 2 forecast; an identical process was used for the 6–10-day forecasts. The three steps were

1) *Train:* (a) Calculate a daily running mean climatology of the week 2 forecast and week 2 observed values individually for each station. The observed climatology used observations from 1971 to 2000; the forecast climatology used forecasts from 1979 to 2001. The year for which the forecast is being made is excluded from the forecast climatology. For a given year and day of the year, the forecast climatology was the week 2 value averaged over all sample years and the 31 days (15 before, 15 after) centered on the date of interest. The process was repeated for each year and day of the year. (b) Determine the forecast and observed anomaly by subtracting the respective climatologies. (Repeat this for each year, day, and station.) (c) Generate a training dataset of 22 × 31 samples of week 2 ensemble mean forecast anomalies and week 2 observed anomalies using a 31-day window centered on the day of interest. (Repeat for each year, day, and station.) (d) Set the observed upper-tercile anomaly $T_{2/3}$ as the 67th percentile of the sorted observed anomaly data. (Repeat for each year, day, and station.) (e) Create the 22 × 31 binary verification data samples. Each sample verification is categorized as being above the upper tercile [$P(V > T_{2/3} = 1$] or below or equal to it [$P(V > T_{2/3} = 0$]. (Repeat for each year, day, and station.) (f) Determine $\hat{\beta}_0$ and $\hat{\beta}_1$ through logistic regression using the ensemble mean anomaly as the only predictor. (Repeat for each year, day, and station.)
2) *Forecast:* Produce tercile probability forecasts for each year, day, and station in DJF using Eq. (1).

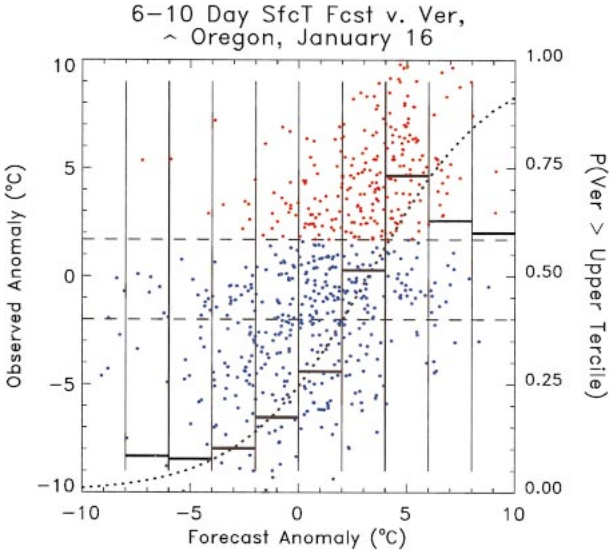


FIG. 3. Illustration of logistic regression method. Ensemble mean 6–10-day forecast anomaly and corresponding 6–10-day observed anomaly are plotted for 16 Jan at Medford, OR. Upper and lower terciles are denoted by dashed lines. Red dots are samples with observed anomalies above the upper tercile; blue dots below. Vertical lines denote bin thresholds for setting tercile probabilities based on the relative frequencies of observed values above the upper tercile. Thick horizontal lines denote the probabilities associated with each bin (refer to probabilities labelled on the right-hand side). Dotted curve denotes the upper-tercile probabilities determined by logistic regression.

Figure 3 illustrates the process for determining the regression model for surface temperatures, here for 6–10-day forecasts at Medford, Oregon, on 16 January. A scatterplot of the ensemble mean 6–10-day forecast anomaly was plotted against the corresponding week 2 observed anomaly using the 22 years × 31 days of samples. From the observed data, the upper and lower terciles were calculated (horizontal dashed lines). Sample points where $P(V > T_{2/3}) = 1$ are denoted with red dots, and points where $P(V > T_{2/3}) = 0$ with blue dots. If one were to set the upper-tercile probabilities just using the relative frequencies of observed values in a bin around a forecast value (the bin limits denoted by the vertical lines), then the average bin probabilities would be denoted by the horizontal solid lines. For example, counting all the forecasts with an anomaly between −6° and −4°C and tallying how often the observed anomaly exceeds the upper tercile, the probability was approximately 9%. When all the samples were supplied to the logistic regression, probabilities were determined as a smooth function of the forecast anomaly according to the dotted curve.

3) *Evaluate:* After forecasts have been produced for each day in DJF for each of the 23 yr using this cross-validation process, evaluate the forecast accuracy using the ranked probability skill score (RPSS) and reliability diagrams (Wilks 1995).

To determine the RPSS, let $\mathbf{y}_i = [y_{1,i}, y_{2,i}, y_{3,i}]^T$ represent the probabilities assigned to each of the three categories for the *ith* forecast of *n* samples. Similarly, denote an observation anomaly binary probability vector for the *ith* forecast, $\mathbf{o}_i = [o_{1,i}, o_{2,i}, o_{3,i}]^T$; for example, if the observed anomaly was lower than $T_{1/3}$, $\mathbf{o}_i = [1, 0, 0]^T$. Define cumulative forecast and observation functions:

$$\mathbf{Y}_i = (Y_{1,i}, Y_{2,i}, Y_{3,i})^T = [y_{1,i}, (y_{1,i} + y_{2,i}), (y_{1,i} + y_{2,i} + y_{3,i})]^T \quad (2)$$

and

$$\mathbf{O}_i = (O_{1,i}, O_{2,i}, O_{3,i})^T = [o_{1,i}, (o_{1,i} + o_{2,i}), (o_{1,i} + o_{2,i} + o_{3,i})]^T. \quad (3)$$

The ranked probability score of the forecast is then defined as

$$\mathrm{RPS_f} = \sum_{i=1}^{n} \sum_{j=1}^{3} (Y_{j,i} - O_{j,i})^2. \quad (4)$$

The RPSS can then be calculated as

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS_f}}{\mathrm{RPS_{clim}}}, \quad (5)$$

where $\mathrm{RPS_{clim}}$ is calculated assuming the climatological forecast probabilities for the three tercile anomalies are [1/3, 1/3, 1/3].

Precipitation forecasts used a slightly modified regression method. Ensemble mean precipitation forecasts and observed values were used without removing the climatological mean. Also, because precipitation forecast and observation data tend to be nonnormally distributed, the precipitation forecasts and observations were power transformed before applying the logistic regression. Specifically, if *x* denotes the ensemble mean forecast, we generated a transformed forecast $\tilde{x}$ according to $\tilde{x} = x^{0.25}$, and $\tilde{x}$ was used as the predictor. The process is illustrated in Fig. 4. The power transformation of observations is convenient for illustration purposes but does not change forecast skill.

Some stations were so dry that precipitation terciles were impossible to define. For example, in late January in Phoenix, Arizona, approximately $p = 60\%$ of the week 2 observed samples had no rainfall. Hence, the 33d percentile of the distribution was zero, but so were the 50th and 60th percentiles. Special rules were needed for such cases. We decided to use the following rule: in the case of more than 33% zero samples in the observed climatology, the lower-tercile threshold was defined as the smallest nonzero precipitation value. Hence, for the lower tercile, step 1(e) above produces a binary set of observed probabilities analogous to a probability of precipitation (POP); that is, a 1 was assigned to nonzero precipitation events and a zero was assigned to zero precipitation events. Similarly, when evaluating the skill score relative to climatology at such points, the

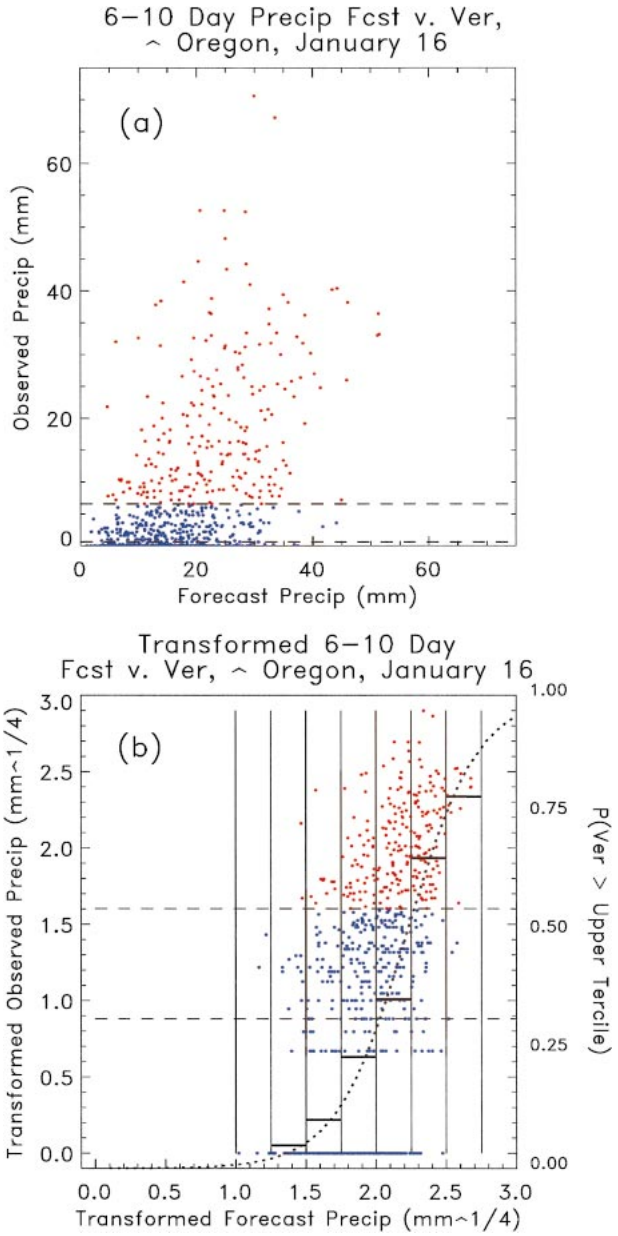


FIG. 4. Illustration of logistic regression process for 6–10-day precipitation. Data is for Medford, OR, on 16 Jan. (a) Scatterplot of ensemble mean forecast precipitation amount vs observed amount (determined from an average of all observations inside the 2.5° × 2.5° grid box). Horizontal dashed lines denote lower and upper terciles. (b) As in (a), but after power transformation. Additionally, vertical black lines denote bin thresholds as in Fig. 3, and thick horizontal solid black lines denote estimated probabilities determined by relative frequency. Dotted curve denotes the upper-tercile probabilities determined by logistic regression.

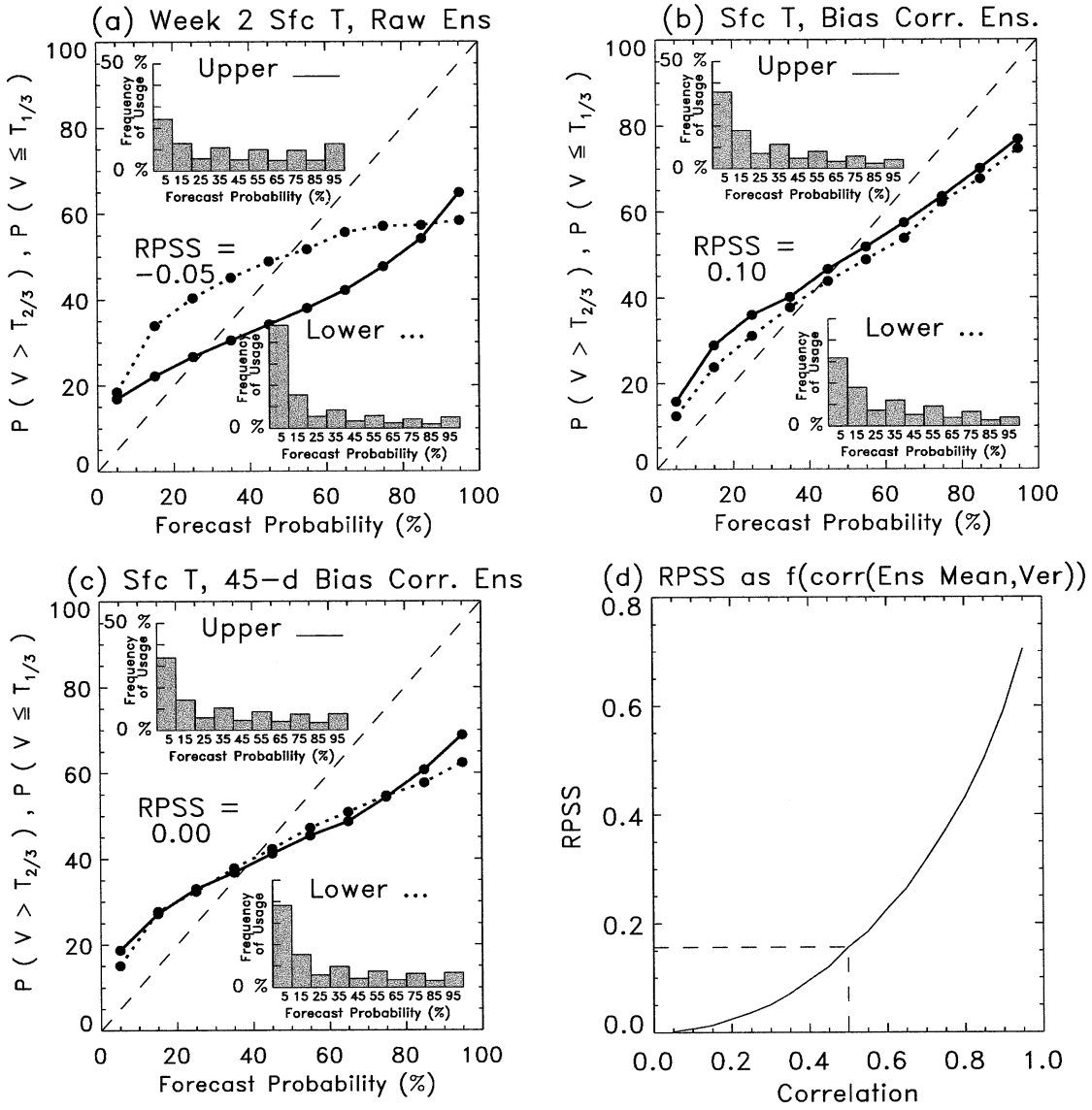


FIG. 5. Reliability diagrams week 2 tercile probability forecasts for surface temperature. DJF forecast data was used from the period 1979–2001, evaluated over the Northern Hemisphere north of 20°N. Dashed line denotes lower-tercile probability forecasts, solid line denotes upper-tercile probability reliability. Inset histograms indicate frequency with which extreme tercile probabilities were issued. (a) Probabilities estimated from raw ensemble relative frequency; (b) as in (a), but with model bias removed before computation of ensemble frequency; (c) as in (b), but where bias correction is based on the past 45 days; and (d) expected RPSS as a function of the correlation between the ensemble mean forecast and verification (see text for details).

probabilities assigned to the three categories were no longer [1/3, 1/3, 1/3], but [p, 2/3 − p, 1/3]. During DJF, there were no stations where there were more than 2/3 of the samples with zeros.

3. Results

Before discussing the results using the MOS algorithm, we note that probability forecasts derived from the raw ensemble have essentially no skill. For example, Fig. 5a shows the reliability diagram for surface temperature forecasts over 23 years as derived from ensemble relative frequency, verified over the Northern Hemisphere north of 20°N latitude. The reliability of tercile forecasts for the upper and lower terciles were different from each other (due to model bias), and the RPSS was near zero (for this reason, operational CPC forecasts are not based primarily on the raw ensemble probabilities; instead they rely on guidance from a number of ensemble forecast systems and statistical methods. Removing the model bias (by computing forecast anomalies relative to the 23-year model climatology rather than the observed climatology) improved the forecasts, resulting in an RPSS of 0.10 (Fig. 5b). However, bias correction

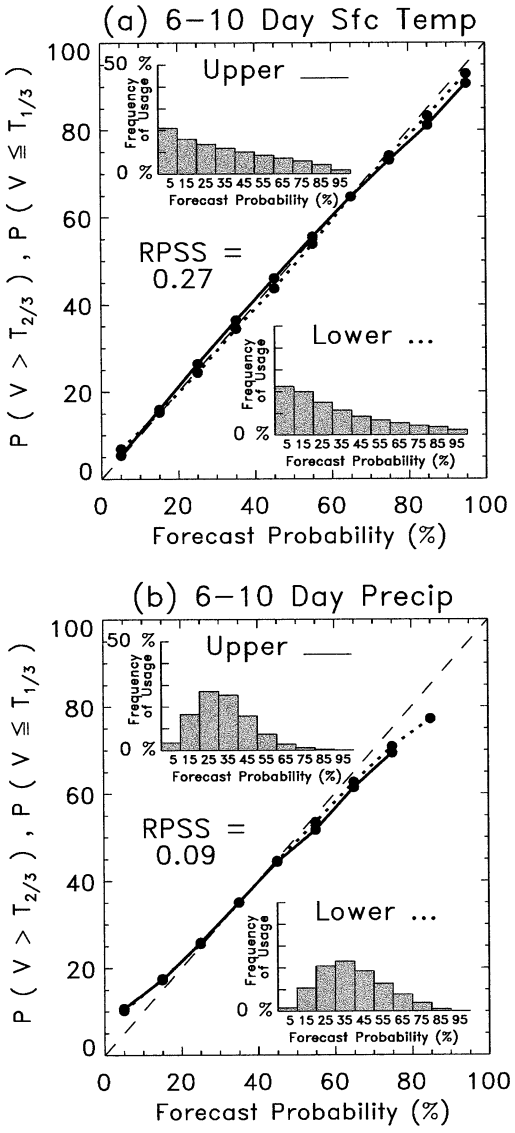


FIG. 6. As in Fig. 5b, but for CDC 6–10-day reforecast based MOS tercile probability forecasts of (a) surface temperature and (b) precipitation, evaluated at 484 stations in the United States and Guam.

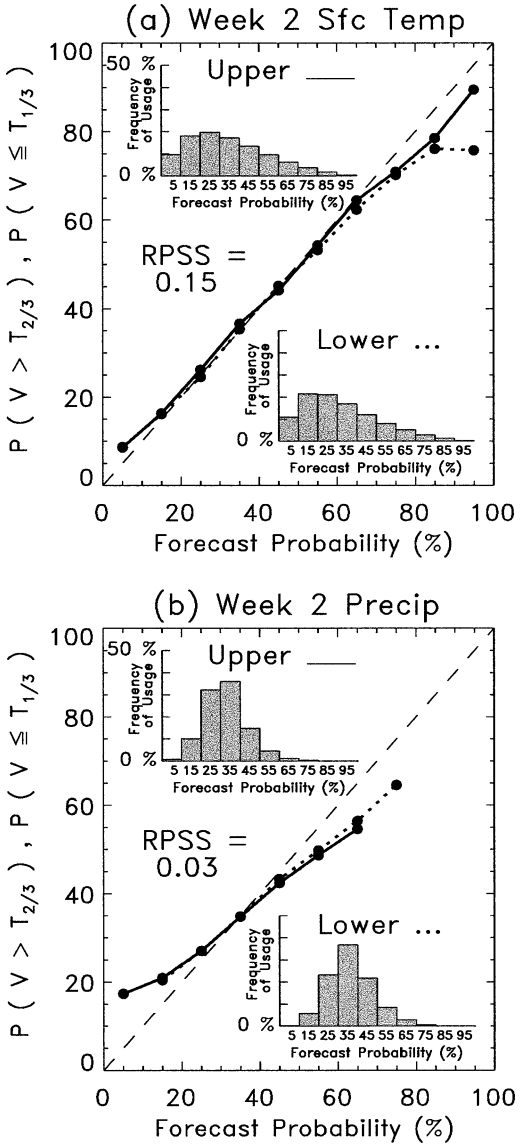


FIG. 7. As in Fig. 6, but for week 2 MOS forecasts.

alone did not greatly improve the calibration (reliability) of the forecasts. If the bias correction were determined, say, from only the last 45 days of forecast, the forecasts were poorly calibrated and had no skill (Fig. 5c).

a. Skill using full reforecast dataset

How much should we expect the RPSS to improve by using the reforecast MOS methodology? If we assume that the forecast and observed surface temperatures have Gaussian statistics, and that the spread of the ensemble does not vary much from day to day (an assumption supported by the low spread–error correlations shown in Fig. 2), it is possible to relate the temporal correlation between the ensemble mean forecast and the verifying analysis to the expected RPSS of tercile probability forecasts. This is done by creating correlated time series (representing the ensemble mean forecast and the corresponding verification) by drawing two random samples from a bivariate normal distribution having a specified correlation. Since the variance of the forecast and analysis distributions are fixed (and known), the tercile probabilities can be calculated given the ensemble mean forecast value by integrating the cumulative distribution function for a Gaussian. Figure 5d shows the expected RPSS for tercile probability forecasts as a function of correlation calculated in this manner. After

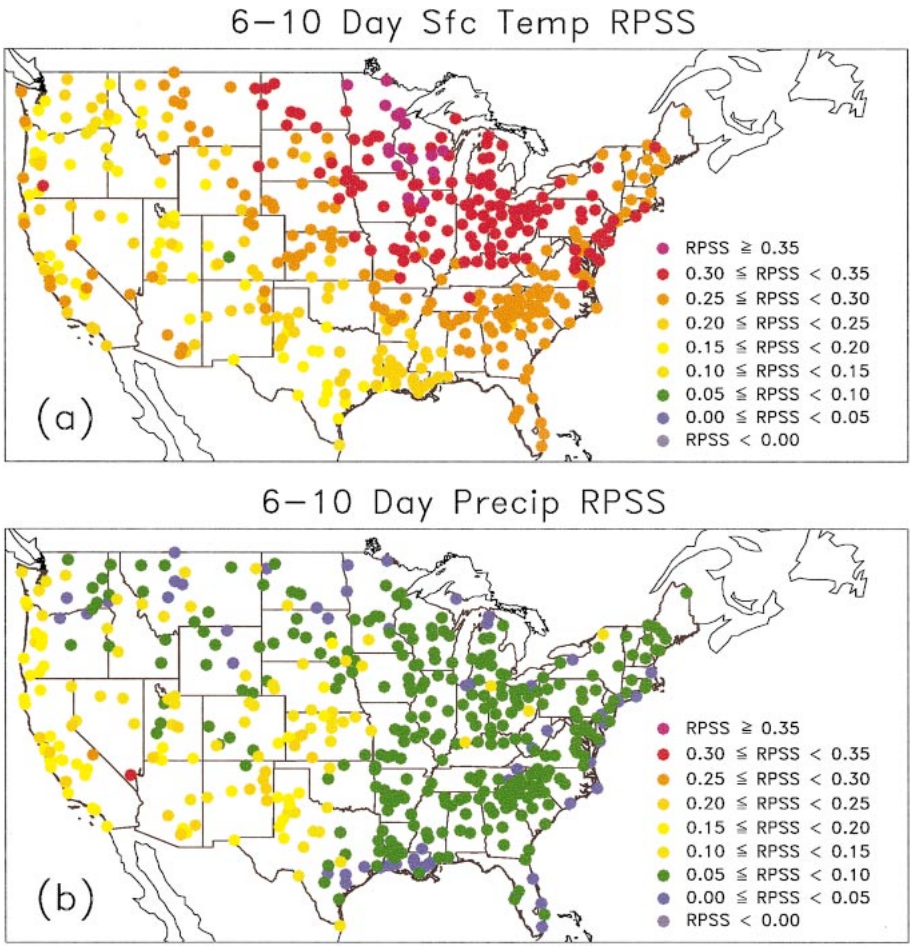


FIG. 8. The 6–10-day MOS forecast RPSS as a function of station location: (a) surface temperature and (b) precipitation.

removing the forecast bias, the average temporal correlation between the week 2 ensemble mean and analyzed surface temperature for all Northern Hemisphere grid points in winter is 0.504. From Fig. 5d, we would then expect an RPSS after MOS calibration of ~0.16.

How much skill, then, do the MOS-based reforecasts have? Figure 6 presents the skill and reliability of 6–10-day probabilistic forecasts using the reforecast MOS methodology, and Fig. 7 presents the same for the week 2 forecasts. These forecasts are significantly more skillful than those produced by raw ensemble counts or bias-corrected ensembles. For example, the MOS-based week 2 surface temperature forecasts achieve an RPSS of 0.15 and are highly calibrated, suggesting that they have extracted nearly all of the potential skill that can be extracted from this forecast dataset, as indicated by Fig. 5d. Further improvements in RPSS are not likely without increasing the correlation between the ensemble mean forecast and the analysis.

Extreme probabilities are issued less frequently than they were when based on the raw ensemble (Fig. 5a), but when they are issued, they are highly reliable. Figures 6 and 7 also illustrate that there was more skill in temperature than precipitation and substantially more skill for 6–10-day forecasts than for week 2 forecasts. Forecasts using the ensemble mean as a predictor in the logistic regression were somewhat more skillful than those using the control run (6–10-day RPSSs were 0.24 and 0.06 for temperature and precipitation using the control, and 0.11 and 0.02 for week 2 forecasts, respectively). The forecasts were sharper (i.e., extreme probabilities issued more frequently) for 6–10-day forecasts than for week 2 forecasts; this was to be expected, for the longer-lead forecasts should more closely resemble the climatological distribution of [1/3, 1/3, 1/3] always being issued. Surface temperature forecasts were sharper than precipitation forecasts. Generally, many prior studies have found that precipitation is one of the most difficult elements to predict. These results reinforce this general conclusion.

The geographic variations in forecast skill are illustrated in Figs. 8 and 9. Surface temperature forecasts were most skillful in the eastern United States and the

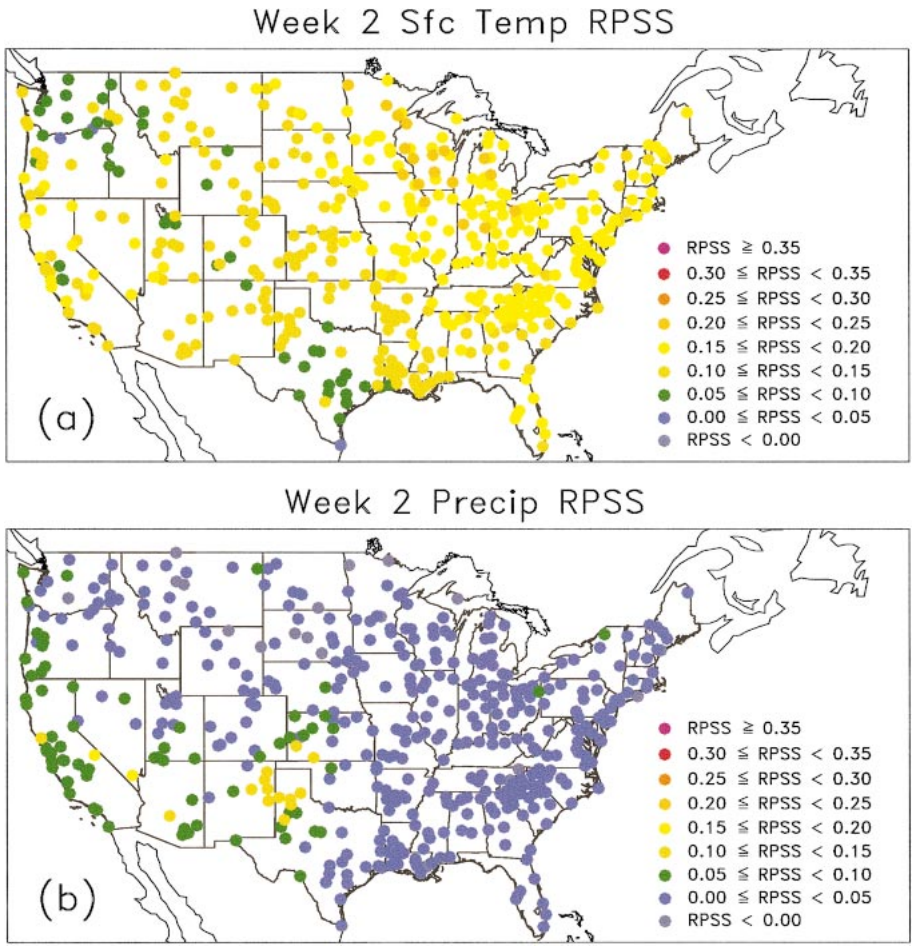


FIG. 9. As in Fig. 8, but for week 2 forecasts.

Ohio Valley. Precipitation forecasts were most skillful along the West Coast.

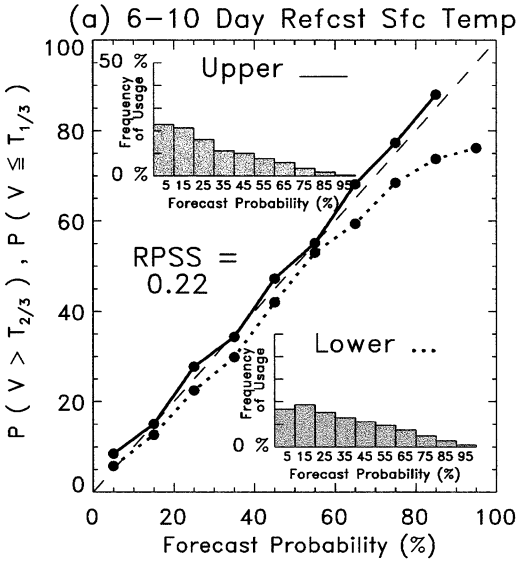
b. Comparison against operational NCEP forecasts

The skill of temperature and precipitation forecasts in Figs. 6 and 7 may seem unimpressive at first glance. However, most prior studies have shown marginal or nonexistent skill for these longer-lead forecasts. For example, Eckel and Walters (1998) showed that even after recalibration with a short training dataset, daily precipitation forecasts based on the MRF had negative skill beyond day 6. The key question is whether CDC's reforecast based MOS forecasts improved upon the presumed state of the art, operational NCEP/CPC forecasts. We compared the two using a set of 100 days of DJF data from the 2001–2002 period at the subset of 153 stations shown in Fig. 1 (darkened dots). Independent data prior to the year 2001 were used to train the CDC reforecast MOS algorithm.

Figures 10 and 11 show reliability diagrams and RPSSs for the CDC reforecast and operational CPC 6–10-day forecasts, respectively. Figures 12 and 13 provide the week 2 forecast diagrams. With a smaller sample size, the CDC reforecasts were less reliable than they were when validated over all 23 yr, as was to be expected from a smaller sample size (Wilks 1995, Fig. 7.9f). However, the reforecasts were significantly sharper and more reliable than the operational CPC forecasts and hence much more skillful. In fact, the CDC reforecasts were more skillful at week 2 than the CPC forecasts were at 6–10 days. Equivalently, this indicates that over these two winters, *the application of the MOS approach increased the effective forecast lead time by several days.*

c. Skill with smaller training samples

The less computationally demanding it is to compute these reforecasts, the more likely the operational centers are to adopt these techniques. We thus examined how much less forecast skill will result when less than the full 22 years of training data are used. We found that the logistic regression scheme occasionally was unable to generate forecasts with only 1 year of data; more was needed for computational stability. Figure 14 plots the



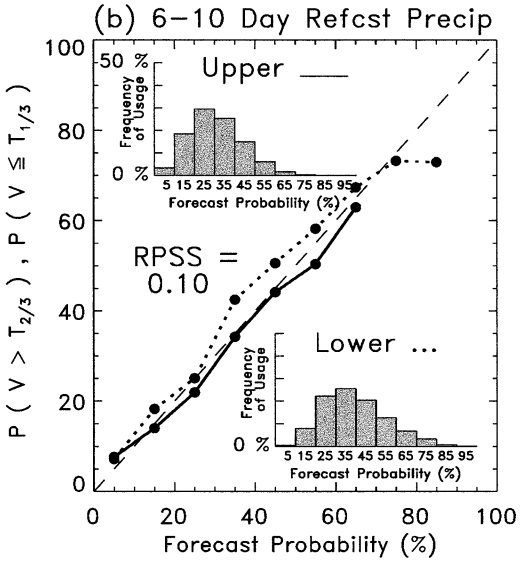
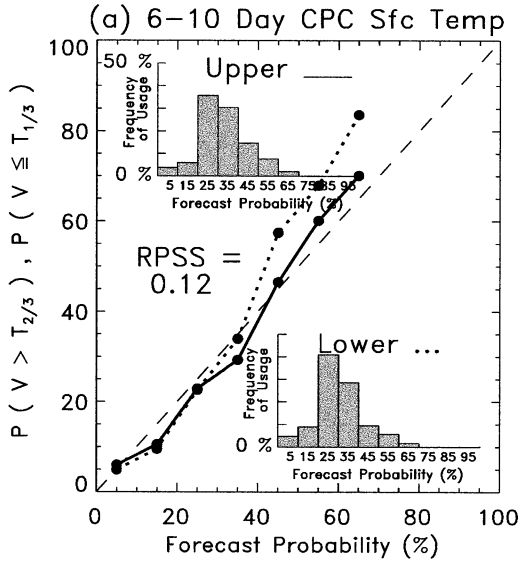


FIG. 10. As in Fig. 6, but for CDC 6–10-day MOS forecasts validated during the winters of 2001–2002.



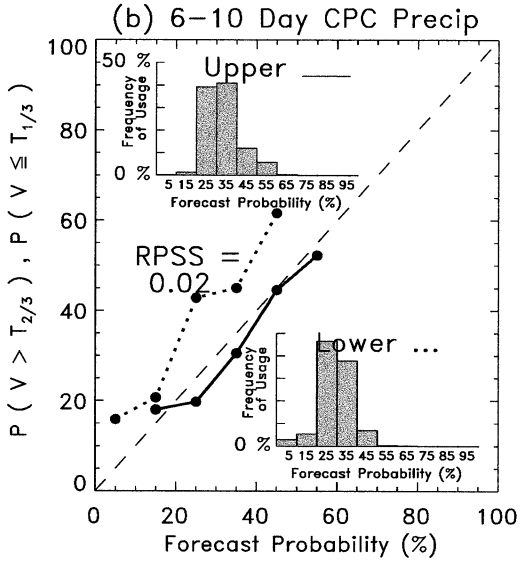


FIG. 11. As in Fig. 10, but for NCEP/CPC operational 6–10-day forecasts.

RPSS of the forecasts for training samples of various sizes. Most of the usable skill has been obtained once ~10 years of training data is available. Most likely, another decade of reforecasts are not worth the extra computational burden.

Another possibility is that if, say, computational considerations prohibit more than 4 yr of reforecasts, these reforecasts could span anywhere from 4 to 20 yr of meteorological conditions by skipping 1 to 5 days between cases (though the control run and breeding cycle would need to be run every day). Figure 15 illustrates that indeed, if there is a fixed upper limit to the number of reforecasts that could be computed, MOS based on a sample of reforecasts that were composed of a training dataset with more days between samples was more skillful than those based on a set with fewer days between samples. For surface temperature, given 5 days between samples, 4 years worth of training data produced probabilistic forecasts that were almost as skillful as those obtained with the full 22-yr training data. The improvement by skipping days is a result of a sample that spans a wider range of meteorological scenarios. Forecasts separated by 1 day had strongly correlated errors; the 1-day lagged autocorrelation of week 2 precipitation error was 0.75, with a temperature error autocorrelation of 0.85.

4. Discussion and conclusions

Improving forecasts through the use of MOS techniques applied to frozen models has been de-emphasized

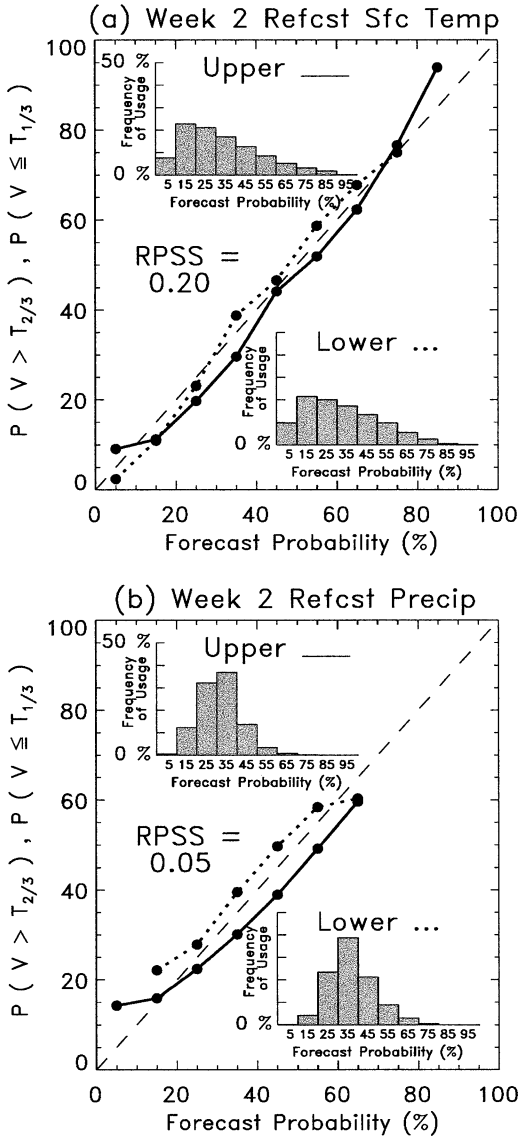


FIG. 12. As in Fig. 10, but for CDC week 2 MOS forecasts.

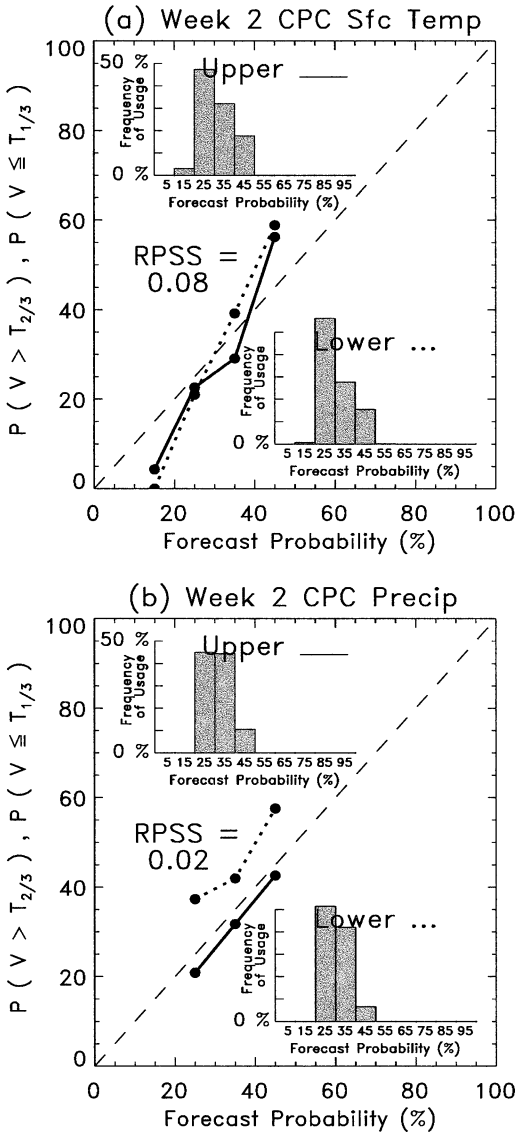


FIG. 13. As in Fig. 10, but for NCEP/CPC operational week 2 forecasts.

during the past decade. These techniques require that a large sample of forecasts and associated verification data be available for the regression analysis. This additional computational burden could potentially slow down the implementation of model changes. In this article, we have shown that dramatic improvements in medium-to extended-range probabilistic forecasts are possible using MOS techniques and a frozen forecast model. Using a T62 version of the MRF model and 22 years of training data, it was possible to make probabilistic week 2 forecasts that were more skillful than the current 6–10-day operational forecasts during the 2001–2002 winters. This improvement occurred despite the fact that operational forecasts utilize larger ensembles with higher-resolution, more highly developed models, and multiple models—but without knowledge of their error characteristics.

Based on these results, NCEP has agreed to take over the daily production of the reforecasts and will be integrating this process into their daily operations. (Until they do, prototype reforecast products will be available online at http://www.cdc.noaa.gov/~jsw/refcst.)

Down the road, this or a similar technique may be desirable to apply to a newer version of the forecast model, perhaps at higher resolution. In subsequent tests, can the computational expense of these reforecasts be reduced? Yes. Most of the skill improvement was retained with only 10 years of data. Also, if the training process were constrained to, say, computing only 4 years of reforecasts, it was preferable to compute reforecasts over a 20-yr span, skipping 5 days between samples. In

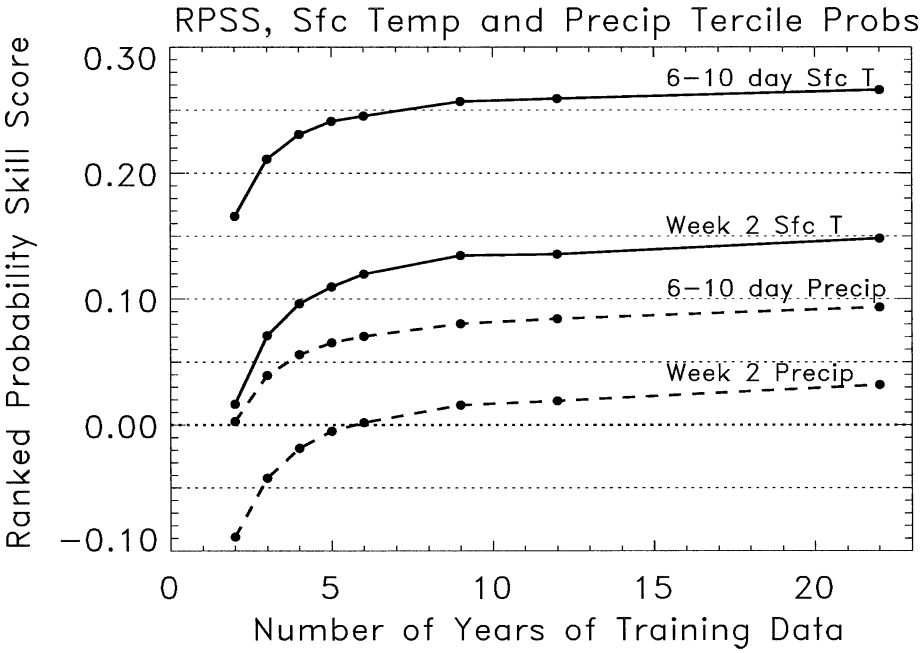


FIG. 14. RPSS as a function of the number of years of training data used.

this way, a broader diversity of weather scenarios was sampled. Computational expense may also be reduced by using a smaller ensemble. We have found that the skill of 6–10-day MOS forecasts of surface temperature using only the control run was comparable to those obtained using the 15-member ensemble mean (although the differences for precipitation and week 2 forecasts are larger). This is consistent with the notion that the benefit of ensemble averaging is a function of the ratio of the predictable signal (i.e., the ensemble mean anomaly) to the unpredictable noise (i.e., the ensemble spread). Ensemble averaging produces the largest rel-

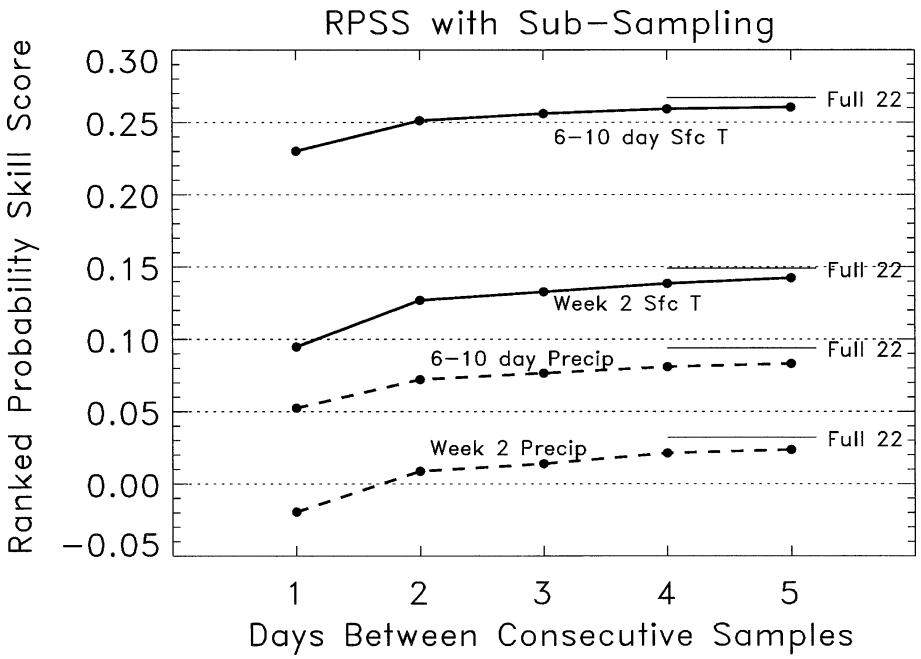


FIG. 15. RPSS when 4 years of training data were used, with 1, 2, 3, 4, and 5 days between successive samples in the training dataset. The lines labeled ''Full 22'' indicate the skill (taken from Figs. 6 and 7) when the full 22 years of cross-validated training data were used.

ative increase in anomaly correlation skill when this ratio is small, and the benefit of adding more ensemble members decreases as this ratio increases (cf. Fig. 15 in Compo et al. 2001). Although we have not investigated the impact of ensemble size here, it is possible that most, if not all, of the skill of the MOS forecasts could be recovered with a significantly smaller ensemble (Du et al. 2000, their Fig. 5).

Are there other ways of minimizing the impact on operations and model development? Reforecasts are easily parallellizable, and the ideal reforecast computing system need not compete for time on the production computer system. As long as ensemble initial conditions have been precomputed, then many days of reforecasts can be computed simultaneously on computers or CPUs separate from the production system. For this experiment, our hardware consisted of 72 CPUs. Each compute node consisted of a 2.2-GHz Pentium processor with a 20-GB hard drive and 1-GB RAM. The individual CPUs were not connected with any special hardware to speed message passing. This equipment, including 2.5 TB of extra storage, cost in total about $90,000 in 2002 U.S. dollars. It took approximately 10 months to compute the full 23 yr of reforecasts at T62L28 resolution on this equipment. While operational centers may prefer to compute such reforecasts with higher-resolution models, by thinning the reforecasts as previously discussed, it is likely that computer clusters on order of several hundreds of thousands of dollars may be suitable for generating these reforecasts. The reforecasting tax thus consists of the upfront cost of hardware and a small staff to maintain the hardware and to run the forecasts and develop the regression algorithms. The more often the operational forecast model is changed, the more expensive the reforecast effort becomes, since the reforecasts must then be completed more rapidly.

If the freedom to continually update the model is deemed extraordinarily important, another possible compromise is for operational facilities to run two model versions. One model version would be continuously updated, as is done currently. A second version would be the dedicated ''reforecast'' run. This model would operate unchanged until a full dataset of reforecasts are available for the next model version.

It is possible that some model changes may further improve the MOS-based forecasts. For example, the MOS approach may be worth retesting with higher-resolution ensemble forecasts (Buizza et al. 2003) and improved ensemble perturbation methods that produce better spread–skill relationships (e.g., Wang and Bishop 2003). If multiple, independent forecast models are available, perhaps MOS approaches using multimodel data may provide additional benefit.

Though this article has focused on the direct benefit of MOS approaches, there are numerous other benefits from computing a large number of reforecasts. Reforecasts may facilitate the model development process, for systematic errors that may not be apparent when model changes are tested on just a few cases may be more obvious with the larger sample afforded by reforecasts. Extreme weather events are of course more numerous in longer training datasets, so forecast characteristics during these important events can be determined. (CDC is making the current reforecast dataset freely available for download at http://www.cdc.noaa.gov/reforecast.) This dataset may be useful for exploring other MOS approaches, for predictability research, and for a host of other applications.

In summary, we have shown that MOS approaches can result in dramatic improvements to 6–10-day and week 2 forecasts. Such approaches require a large dataset of retrospective forecasts and observations. Given the substantial value added, weather forecast services may wish to evaluate how they can incorporate these statistical techniques into their forecast process.

Acknowledgments. This project would have been much more difficult without the assistance of many other scientists. Jon Eischeid (CDC) provided us with the station observation data. Scott Handel (NCEP/CPC) provided us with the operational data, and Ed O'Lenic and Wes Ebisuzaki (NCEP/CPC) assisted us in using NCEP's computers to obtain observations and analyses in near–real time. We gratefully acknowledge stimulating discussions with other CDC personnel, including Klaus Weickmann, Klaus Wolter, Matt Newman, Gil Compo, Shaleen Jain, Gary Bates, Andrea Ray, Robert Webb, and Marty Hoerling. Dan Wilks and Matt Briggs (Cornell) and Bob Vislocky (The Forecast Institute, Inc.) provided informal reviews, and we thank two anonymous reviewers for their substantive feedback. We thank CDC director Randy Dole for his leadership in developing the ''Weather-Climate Connection'' program that funded this research.

REFERENCES

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Wea. Forecasting,* **17,** 783–799.

Buizza, R., D. S. Richardson, and T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparisons with poor-man's ensembles. *Quart. J. Roy. Meteor. Soc.,* **129,** 1269–1288.

Caplan, P., J. Derber, W. Gemmill, S.-Y. Hong, H.-L. Pan, and D. Parrish, 1997: Changes to the 1995 NCEP operational medium-range forecast model analysis-forecast system. *Wea. Forecasting,* **12,** 581–594.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting,* **4,** 401–412.

Compo, G. P., P. D. Sardeshmukh, and C. Penland, 2001: Changes of subseasonal variability associated with El Niño. *J. Climate,* **14,** 3356–3374.

Courtier, P., J.-N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J. Roy. Meteor. Soc.,* **120,** 1367–1387.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

Du, J., S. L. Mullen, and F. Sanders, 2000: Removal of distortion error from an ensemble forecast. *Mon. Wea. Rev.*, **128,** 3347–3351.

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

Eischeid, J. K., P. Pasteris, H. F. Diaz, M. Plantico, and N. Lott, 2000: Creating a serially complete, national daily time series of temperature and precipitation for the United States. *J. Appl. Meteor.*, **39,** 1580–1591.

Glahn, H. R., 1985: Statistical weather forecasting. *Probability, Statistics, and Decision Making in the Atmospheric Sciences*, A. H. Murphy and R. W. Katz, Eds., Westview Press, 289–335.

——, and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11,** 1203–1211.

Hamill, T. M., C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular vector, and perturbed observation ensembles. *Mon. Wea. Rev.*, **128,** 1835–1851.

Houtekamer, P. L., and H. L. Mitchell, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129,** 123–137.

——, L. Lefaivre, and J. Derome, 1996: The RPN ensemble prediction system. *Proc. ECMWF Seminar on Predictability*, Vol. II, Reading, United Kingdom, ECMWF, 121–146. [Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom.]

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77,** 437–471.

——, S. J. Lord, and R. D. McPherson, 1998: Maturity of operational numerical weather prediction: Medium range. *Bull. Amer. Meteor. Soc.*, **79,** 2753–2769.

Kanamitsu, M., 1989: Description of the NMC global data assimilation and forecast system. *Wea. Forecasting*, **4,** 334–342.

——, and Coauthors, 1991: Recent changes implemented into the global forecast system at NMC. *Wea. Forecasting*, **6,** 425–435.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21,** 289–307.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122,** 73–119.

Newman, M., P. D. Sardeshmukh, C. R. Winkler, and J. S. Whitaker, 2003: A study of subseasonal predictability. *Mon. Wea. Rev.*, **131,** 1715–1732.

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's Spectral Statistical Interpolation Analysis System. *Mon. Wea. Rev.*, **120,** 1747–1763.

Tapp, R. G., F. Woodcock, and G. A. Mills, 1986: The application of model output statistics to precipitation prediction in Australia. *Mon. Wea. Rev.*, **114,** 50–61.

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

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125,** 3297–3319.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.*, **76,** 1157–1164.

——, and ——, 1997: Performance of an advanced MOS system in the 1996–97 national collegiate weather forecasting contest. *Bull. Amer. Meteor. Soc.*, **78,** 2851–2857.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60,** 1140–1158.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125,** 527–548.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Woodcock, F., 1984: Australian experimental model output statistics forecasts of daily maximum and minimum temperature. *Mon. Wea. Rev.*, **112,** 2112–2121.