5.5

# ENSEMBLE RE-FORECASTING : IMPROVING MEDIUM-RANGE FORECAST SKILL USING RETROSPECTIVE FORECASTS

Thomas M. Hamill, Jeffrey S. Whitaker, and Xue Wei

*NOAA-CIRES Climate Diagnostics Center, Boulder, Colorado*

## 1. INTRODUCTION

Improving weather forecasts is a primary goal of the U.S. 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, to conduct forecasts with higher-resolution numerical models, and to incorporate more complex physical parameterizations of processes that occur below the grid scale. Within the last decade, ensemble forecast techniques 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 and because of the steadier growth of model errors. In order to make a skillful medium-range forecast, forecasters must thus be able to adjust for model systematic errors and be able to distinguish between features that are predictable and those that are unpredictable.

The format of forecasts issued by the NCEP Climate Prediction Center (CPC) implicitly reflect a judgment of what can be predicted skillfully and what cannot. Daily 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 (2m) temperature and precipitation tercile probabilities. These are forecasts of the probability that the temperature and precipitation averaged over these periods will be below the 33rd 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, or "MOS" techniques (Glahn and Lowry 1972; Woodcock 1984; Glahn, 1985; Carter et al. 1989; Vislocky and Fritsch 1995) have been used widely since the 1970s. However, in recent years, the U.S. 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 making skillful medium-range forecasts without statistical models, we reconsider the value of statistical weather forecasting for this application. Specifically, we will 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 reduced-resolution (T62) version of NCEP's Medium-Range Forecast (MRF) modeling system based on 1998 model physics was used to run a set of ensemble "re-forecasts" over the period 1979-2001. Statistically adjusting current T62 forecasts based on these prior forecasts will be shown to produce substantial improvements in forecast skill, greatly exceeding the skill of the operational forecasts. Given the improvements produced through the use of statistical techniques, we propose that re-forecasting and the application of MOS-like statis-

*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

tical techniques should become an integral part of the medium-range numerical weather prediction process.

## 2. EXPERIMENT DESIGN

### *a. Forecast model, initial conditions, and verification data*

A T62 resolution 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-year period from 1979 to 2001.

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 Centers for Atmospheric Research (NCAR) reanalysis (Kalnay et al. 1996) and a set of 7 bred pairs of initial conditions (Toth and Kalnay 1993, 1997) re-centered each day on the reanalysis initial condition.

Here, we will concentrate on comparing the proposed MOS-based forecasts against CPC operational forecasts for a set of 100 days during the winters of 2001 and 2002. This comparison was performed at the subset of 153 stations where CPC forecasts were available (darkened dots in Fig. 1). Independent data prior to the year 2000 were used to train the CDC reforecast MOS algorithm. We also present some results summarized over a set of 355 stations and the full 23 years of December-January-February (DJF) forecasts from 1979 to 2001. 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

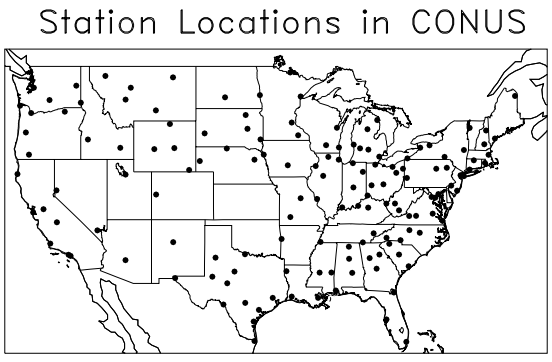


Figure 1. Locations in conterminous U.S. where CDC MOS and CPC forecasts were compared.

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 and $\hat{\beta}_0$ and $\hat{\beta}_1$ are fitted regression coefficients.

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 data set of ensemble mean forecast and observed week 2 anomalies from climatology. From these we compute the associated binary verification data (was the observed anomaly above the upper tercile ($P(V > T_{2/3} = 1)$ or below or equal to it ($P(V > T_{2/3} = 0$?) 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 years were used as training data. The same 22 years were used to define the forecast climatology.

The generation and evaluation of tercile probability forecasts followed a 3-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 was excluded from both. For a given year and day of the year, the 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 data set of 22 $\times$ 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]. (d) Create the 22 $\times$ 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]. (e) 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).

Figure 2 illustrates the process for determining the regression model for surface temperatures, here for 6-10 day forecasts at Medford, Oregon on January 16. A scatterplot of the ensemble mean 6-10 day forecast anomaly was plotted against the corresponding week 2 observed anomaly using the 22 years $\times$ 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 percent. 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 years using this cross-validation process, evaluate the forecast accuracy using the ranked probability skill score (RPSS; Wilks 1995) with climatology as a reference, and reliability diagrams (ibid).

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 non-normally 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 $\widetilde{x}$ according to $\widetilde{x} = x^{0.25}$, and $\widetilde{x}$ was used as the predictor.

## 3. RESULTS

Figures 3 and 4 show reliability diagrams and RPSSs for the CDC re-forecast and operational CPC 6-10 day forecasts, respectively. Figures 5 and 6 provide the week 2 forecast diagrams. The re-forecasts were significantly sharper and more reliable than the operational CPC forecasts and hence much more skillful. In fact, the CDC re-forecasts 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.*

The skill of the CDC MOS forecasts varied with geographic location (Figure 7). For temperature, cold/warm outbreaks in the midwest were the most skillfully predicted, while for precipitation, west-coast forecasts were more skillful, perhaps associated with the PNA pattern.

We also considered the extent to which forecast skill could be retained if smaller data sets were used. Figure 8 plots the skill of forecasts for different numbers of years of training data, and Figure 9 indicates how much skill is present with four years of forecasts when 1, 2, 3, 4, or 5 days are used between sample forecasts. The results suggest that the full 23 years of retrospective data is not needed; by 10 years nearly all of the benefit is obtained. Also, if a limited number of years of

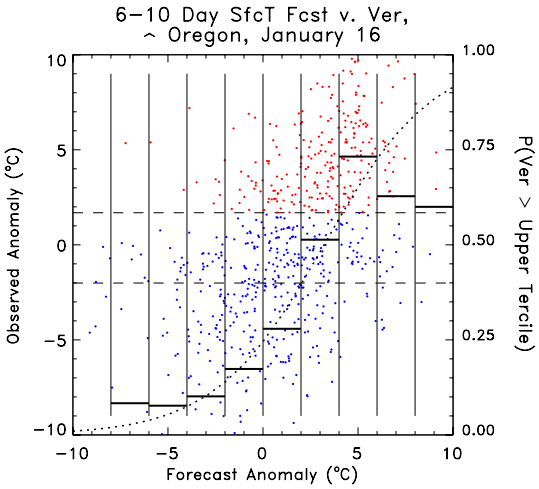


Figure 2. Illustration of logistic regression method. Ensemble mean 6-10 day forecast anomaly and corresponding 6-10 day observed anomaly are plotted for 16 January at Medford, Oregon. Upper and lower terciles are denoted by dashed lines. Red dots are samples with observed anomalies above the upper tercile; 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 labeled on the right side of the plot). Dotted curve denotes the upper tercile probabilities determined by logistic regression.

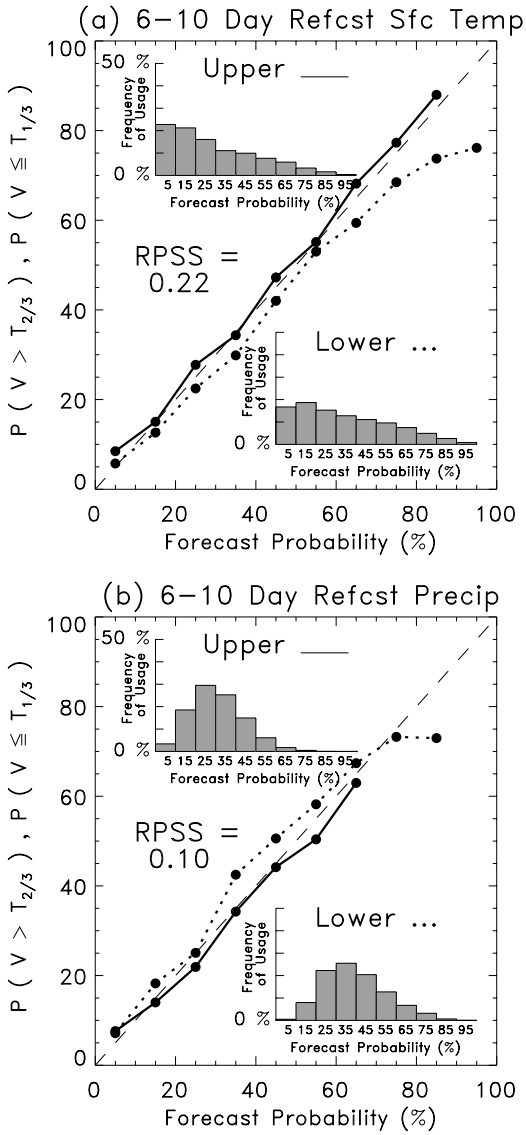


Figure 3. Reliability diagrams CDC's MOS-based 6-10 day tercile probability forecasts for (a) surface temperature and (b) precipitation. Dashed line denotes lower tercile probability forecasts, solid line denotes upper tercile probability reliability. Inset histograms indicate frequency with which extreme tercile probabilities were issued.

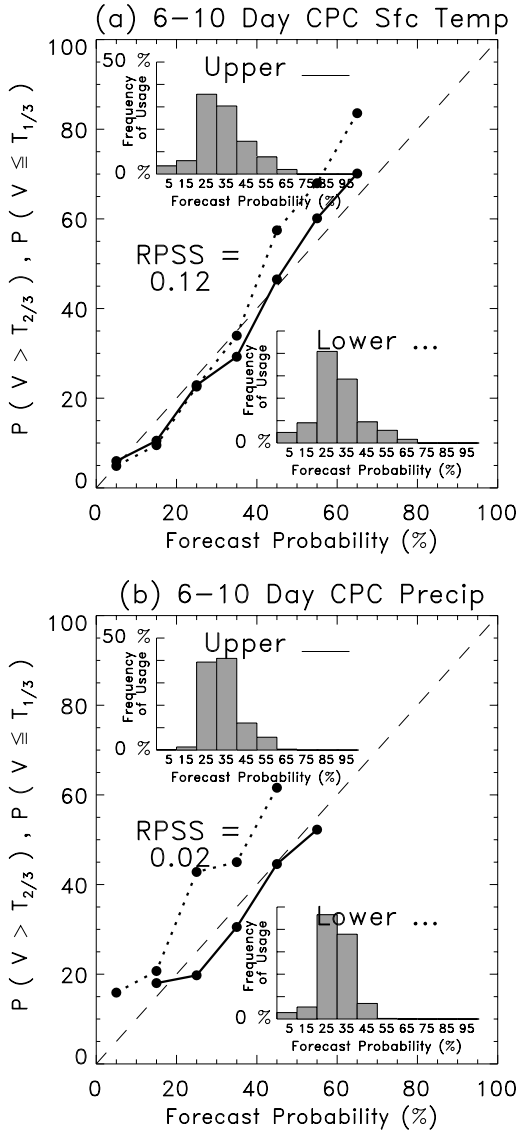


Figure 4. As in Fig. 2, but for operational NCEP CPC 6-10 day re-forecast based MOS tercile probability forecasts.

retrospective forecasts can be computed, it is wiser to use forecasts over a longer span of time with more days between sample forecasts. In this way, a wider diversity of weather scenarios are spanned.

More results are discussed in an upcoming journal article to appear in Monthly Weather Review, probably in the spring of 2004. Until then, a version of the full manuscript can be downloaded from http://www.cdc.noaa.gov/∼hamill.

## 4. DISCUSSION AND CONCLUSIONS

In this article we demonstrated dramatic improvements in medium- to extended-range forecasts are possible using MOS techniques. Using a low-resolution 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 are based on larger ensembles and higher-resolution models - but without knowledge of their biases and error statistics.

Though this article has focused on the direct benefit of MOS approaches, there are numerous other benefits from computing a large number of re-forecasts. Re-

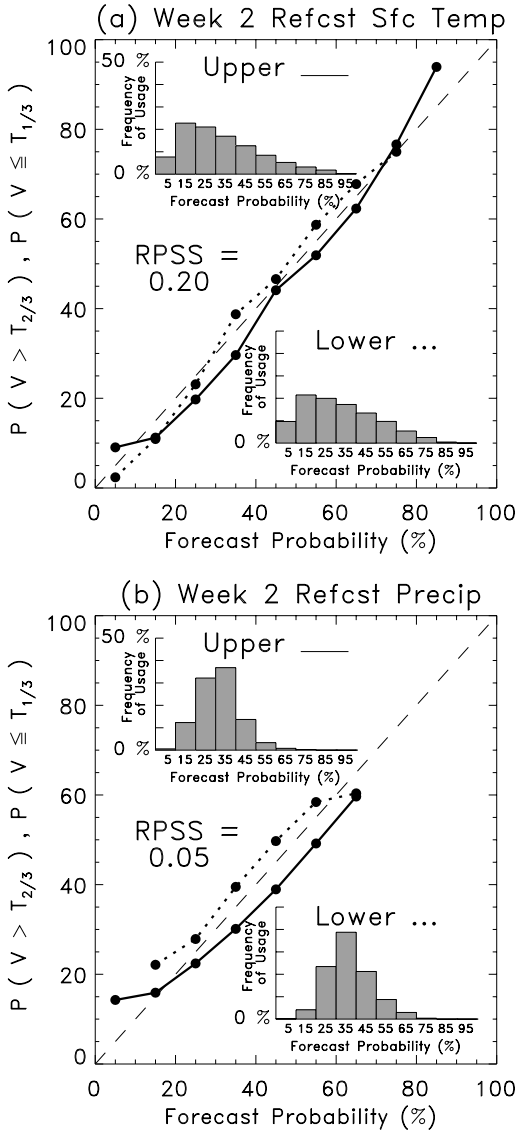


Figure 5. As in Fig. 2, but for week 2 CDC MOS-based forecasts.

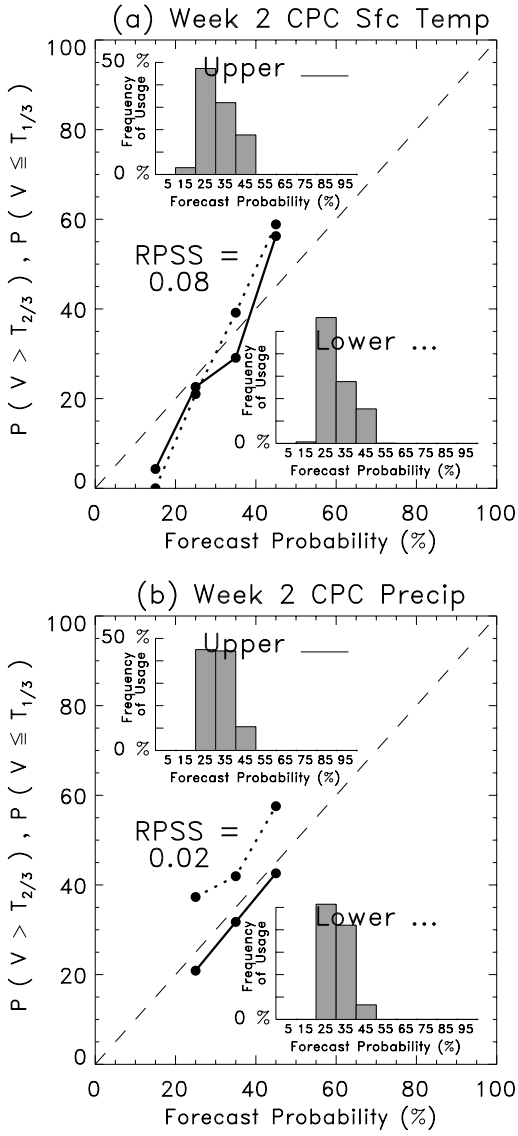


Figure 6. As in Fig. 3, but for week 2 NCEP CPC forecasts.

forecasts 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 re-forecasts. Extreme weather events are of course more numerous in longer training data sets, so forecast characteristics during these important events can be determined. CDC is making the current re-forecast data set freely available for download at http://www.cdc.noaa.gov/reforecast . This data set may be useful for exploring other MOS approaches, for predictability research, and a host of other applications.

In summary, we have showed that MOS approaches can result in dramatic improvements to 6-10 day and week 2 forecasts. Such approaches require a large data set 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. NOAA/CDC is working with NCEP to integrate these techniques into operations.

## 5. 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.

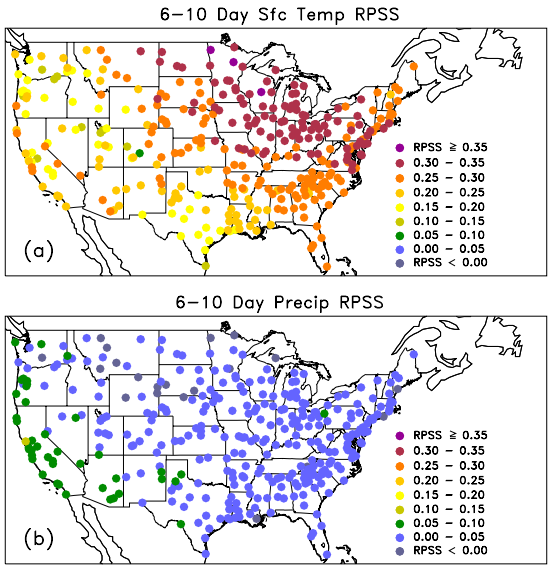


Figure 7. RPSS of CDC MOS 6-10 day forecasts, evaluated in DJF from 1979-2001 at a set of 355 stations, primarily in conterminous U.S.

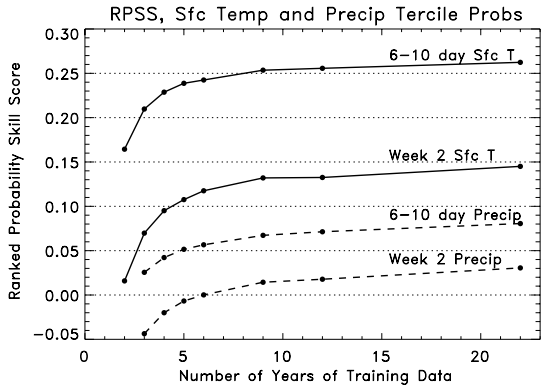


Figure 8. RPSS as a function of the number of years of training data used.

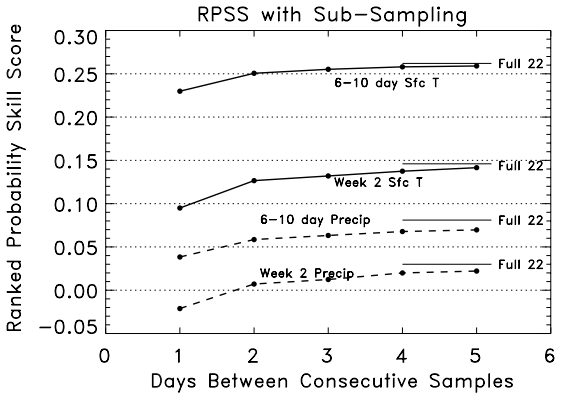


Figure 9. RPSS when 4 years of training data were used, with 1, 2, 3, 4, and 5 days between successive samples in the training data set. The lines labeled "Full 22" indicate the skill when the full 22 years of cross-validated training data were used.

## 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.

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.

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

——, 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.

Kalnay, E., and co-authors, 1996: The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**, 437-472.

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.

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., J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc.*, **76**, 1157-1164.

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.