THE FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

POST-PROCESSING IMPROVEMENTS TO AN ENSEMBLE FORECAST USING AN ARCHIVE OF PAST FORECASTS AND VERIFICATIONS

By

ADAM DOUGLAS ALLGOOD

A Thesis submitted to the
Department of Meteorology
in partial fulfillment of the
requirements for the degree of
Master of Science

Degree Awarded:
Spring Semester, 2007

The members of the Committee approve the Thesis of Adam Douglas Allgood defended on March 29, 2007.

Jon E. Ahlquist
Professor Directing Thesis

T. N. Krishnamurti
Committee Member

Xiaolei Zou
Committee Member

The Office of Graduate Studies has verified and approved the above named committee members.

To my grandfather Abraham Nagel, who gave me good genes. May his memory be a blessing.

ACKNOWLEDGEMENTS

There are numerous people who have provided me with so much support during my Master's Degree work, to whom I am very grateful. My Major Professor, Dr. Jon Ahlquist, has provided me with a tremendous amount of information, not only in my research, but also computing, writing, economics, and too many other subjects to enumerate. I am grateful for his patient assistance and willingness to spend large amounts of time working with me in spite of a full schedule. I am also grateful to my other committee members, Drs. Krishnamurti and Zou, who took the time to listen to my research results and provided me with meaningful feedback. Additionally, Chris Allen and Alec Bogdanoff provided essential assistance with setting up the data archive and research web site.

I am extremely fortunate to have a wife so patient, kind, and supportive of me in all of my work. Blair has helped me keep a good perspective on life in times when the "little" details seemed so daunting. She and my three wonderful children, Caroline, Ellie, and Tzippi fill my life with joy, whether dropping into the office for a visit, calling me on the phone, or seeing me off and welcoming me home.

I would like to express gratitude to my parents and parents-in-law, who have provided me with advice and so much emotional and financial support during my work at Florida State University. Finally, I am grateful to Hashem, who has given me life, blessings, and strength.

This research was supported by NOAA CSTAR grant NA03NWS4680018.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Ensemble forecasts are the primary tool used operationally to assess forecast uncertainty. Studies of ensemble forecasts, however, have shown that forecast verifications too frequently lie outside of the ensemble's range of possibilities, meaning that uncorrected ensemble forecasts suggest more confidence than is justified. To make ensemble forecasts more representative of the actual range of possibilities, we present a technique to post-process ensemble forecasts by replacing member forecasts with verifications of what actually occurred when past forecasts were similar. To maximize the information that can be extracted from an archive of past forecasts and verifications, we allow analogs to come from different locations in space.

We evaluated our procedure to post-process NCEP ensemble precipitation forecasts for the United States for 15-day periods in July 2005 and January 2006. Our analog correction technique significantly improved the ensemble's ability to forecast the probability of precipitation, in particular correcting the NCEP Global Ensemble's "wet" bias at low precipitation amounts. Brier Skill Scores for 6-hour accumulated precipitation during the winter indicated that uncorrected ensemble forecasts were less skillful at predicting the probability of precipitation than forecasting zero precipitation as indicated by negative Brier Skill Scores (roughly -2.5). Post-processed forecasts had Brier Skill Scores as high as 0.34. The tendency of the ensemble to underforecast heavy precipitation events, however, was not well corrected by our post-processing technique. Examinations of analog locations during heavy precipitation events indicated that analogs were taken from regions where precipitation patterns differed from those at the forecast point. This indicates that analogs must be chosen using more information than merely the similarity of ensemble precipitation forecasts to past forecasts.

CHAPTER 1

Introduction

1.1 Forecast Uncertainty and Ensemble Forecasting

Numerical weather prediction (NWP) techniques utilize solutions of deterministic equations to model the state of the atmosphere at future times based on observations of the initial state of the atmosphere. As observational coverage increases and numerical techniques improve, providing a more accurate assessment of the initial conditions and subsequent evolution, NWP has become far more reliable than in decades past. Since observational networks and NWP processes do not capture the smallest scales, however, there will always be a degree of error in the modeled initial conditions of any numerical prediction system. Lorenz (1963) showed using a simplified set of deterministic equations that atmospheric flow may be highly unstable with regard to small amplitude modifications of the initial input. Thus, small errors in the initial conditions of a NWP system may grow in size as the model is integrated until they affect the largest scales resolved in the model. Running a simple 28-variable numerical model with small variations in input, Lorenz (1965) showed that such errors grow rapidly, rendering a forecast unusable after a period as short as a few days, or as long as a month depending on the particular atmospheric state. Lorenz also noted that in a chaotic system, smaller scale errors tend to grow even more rapidly, causing a faster degradation in forecast skill with time. It is essential, therefore, when considering applications of NWP to account for not only the single deterministic forecast output by the model, but also the forecast uncertainty that arises due to the errors in the initial conditions. For example, a citrus farmer, who must take costly preventative action to protect his crops in the event of a freeze, requires not only a single temperature forecast, but also information about the degree of uncertainty in that forecast, in order to assess the risks in his decision properly.

Ensemble forecasts are a collection of two or more forecasts that verify at the same time

(Sivillo and Ahlquist 1997), which, due to differences in the initial conditions or numerical methods, tend to produce divergent solutions. When the differences in initial conditions among ensemble member forecasts are within the bounds of observational errors, hence being equally plausible, the resultant forecasts are also considered equally plausible outcomes. An ideal collection of such forecasts can therefore provide meaningful information about the distribution of forecast uncertainty, and useful products such as a consensus (average over all ensemble members) forecast. The probability of specific events can also be computed using an ensemble forecast. For example, given an ensemble with N member forecasts, where N_E of those forecasts predict the occurrence of an event, such as a wind speed or precipitation accumulation exceeding a certain threshold, the forecast probability of the event occurring can be expressed as:

$$P_E = \frac{N_E}{N} \tag{1.1}$$

The National Centers for Environmental Prediction (NCEP) Environmental Modeling Center in Camp Springs, MD has issued global ensemble forecasts operationally since December 1992. The Global Ensemble Forecasting System (GEFS) based on the Global Forecast System (GFS) model, will be described in more detail in Chapter 3. Evaluations of operational ensembles (Toth and Kalnay 1993, 1997) have shown that the consensus forecast is more skillful than single deterministic forecasts produced at a higher resolution. These ensemble systems, however, currently perform well below their potential capabilities, which are demonstrated in the following section. The purpose of our research is to enhance the performance of operational ensembles.

1.2 Evaluating Operational Ensembles using Rank Histograms

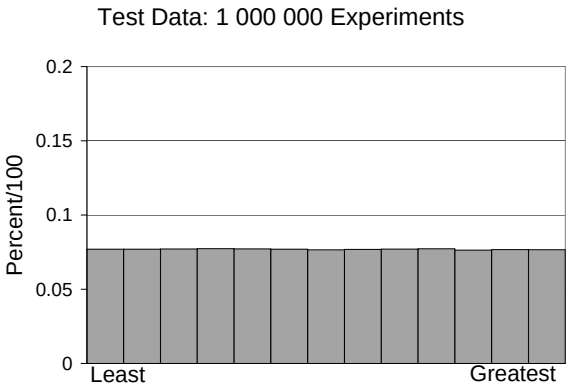
A useful test tool for the evaluation of ensemble forecasts with regard to their representativeness of the true forecast uncertainty is the Rank Histogram diagram (Hamill and Colucci 1997). To compute a Rank Histogram, consider a sample of realizations, x_n, for $n = 1, ..., N$, of a random process with a particular distribution, and another realization y. Sort the values of x_n such that $x_1 \leq x_2 \leq x_3 \leq ... \leq x_N$. The value y will then fall into a bin from 1 to $N + 1$ defined by its rank with respect to the sorted values of x_n. For example, if the value of y is less than x_1, which means it is less than every x_n value, its rank would be

1, or smallest. A fundamental property of the Rank Histogram is that if y is drawn from the same distribution as the sample x_n, then y has an equal probability of being ranked in any bin. Therefore, when numerous x_n samples and y values are drawn from the same distribution, and we compute the fraction of total y values that are ranked in each bin, the Rank Histogram will have a uniform distribution among the bins. If the x_n samples determining the bins are not representative of the distribution from which the y values were drawn, then the Rank Histogram will have a non-uniform distribution among the bins, such as certain bins receiving a greater fraction of the total y values.

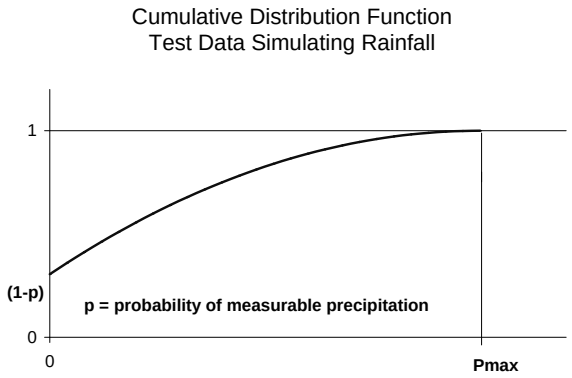
When computing Rank Histograms, assigning the value y to a rank becomes more complex when one or more of the x_n values are equal to y. Following Hamill and Colucci (1997), we break such ties by generating random numbers and assigning one to each x_n value equal to y, and one to y as well. A negligible amount is added to each x_n value with an assigned random number greater than the number assigned to y, and the same negligible amount is subtracted from each x_n value with an assigned random number less than the number assigned to y. This allows y to be assigned to one of the potential ranks randomly and uniformly.

When evaluating ensemble forecasts with Rank Histograms, the x_n sample becomes the collection of forecasts in the ensemble, and the y value represents the observation of what actually occurs when the forecast verifies. If an ensemble forecasting system is truly representative of the forecast uncertainty which arises in the numerical model, which is ideal, then the verification has an equal probability of being ranked anywhere among the sorted ensemble member forecasts. A Rank Histogram computed using many ideal ensemble forecasts and verifications will then have a flat bin distribution. Ensembles that are unrepresentative of the forecast uncertainty, for example an ensemble which too frequently underforecasts extreme events, will yield Rank Histograms with non-uniform bin distributions.

To demonstrate the Rank Histogram resulting from x_n and y values being drawn from the same distribution and also to test our software, we conducted an experiment using a test dataset produced by a random number generator. Samples of independent, identically distributed random numbers were used to represent x_n and y values. These data were adjusted to simulate the distribution of rainfall by setting a specified fraction of the random numbers to zero, and adjusting the remaining values so that the Probability Density Function



(a) Rank Histogram of test data for 1,000,000 experiments.



(b) Cumulative Distribution Function of test data.

Figure 1.1: Results of a test of our Rank Histogram generation software using random numbers taken from the same distribution to represent ensemble forecasts and observations in each experiment.

decayed linearly to zero at a maximum threshold p_{max}. Figure 1.1(a) shows the resultant rank histogram, which has a flat appearance as expected due to the "observations" and "forecasts" being drawn from the same distribution. Employing a Chi-Square Goodness of Fit test to determine the probability that the bins of this Rank Histogram are uniformly distributed:

$$\chi^2 = \sum_{i=1}^{N+1} \frac{(O_i - E_i)}{E_i} \tag{1.2}$$

where O_i represents the observed bin value, and E_i corresponds to the expected value given a hypothesis of a uniform distribution, $E_i = N_{exprmts}/N + 1$ we find that the probability of the null hypothesis—the rank histogram bins do not come from a uniform distribution—is less than 1 percent, and can be rejected.

Rank Histograms computed using experimental versions of NCEP operational ensembles exhibited non-uniform rank bin distributions (Hamill and Colucci 1997, 1998, hereafter HC97 and HC98). Figure 1.2 (HC97 Fig. 3) shows the distribution of verification ranks among 15 ensemble members for 24 hr precipitation accumulations over a 15 day case study. The spikes in the first (driest rank) and last (wettest rank) bins indicate that the ensemble underforecasted the true inherent uncertainty. In too many cases, the observation was either

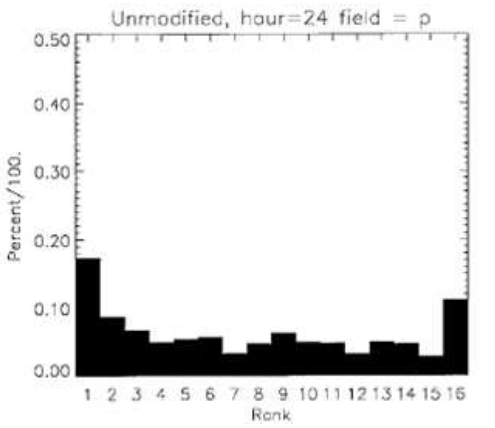


Figure 1.2: Rank Histogram of 24hr accumulated precipitation forecasts taken from an ensemble consisting of forecasts produced by the NCEP ETA and Regional Spectral Model (Hamill and Colucci 1997 Fig. 3). Note the spikes in ranks 1 and 16, which reveal the ensemble's inability to represent the forecast uncertainty.

wetter or drier than every forecast in the ensemble. Although HC98 noted an improvement in forecast skill by using the ensemble's consensus forecast in place of the higher resolution operational forecast, they cautioned that the Rank Histogram evaluations indicated that the currently operational ensemble failed to "forecast the forecast skill."

Given these serious deficiencies in the NCEP operational ensembles, there have been attempts to adjust the ensemble output to improve its distributions. These techniques are called post-processing, since they are applied to the output after the model is run rather than attempting to change the way the model is run itself. Eckel and Walters (1998) employed a statistical bias correction of 24 hr accumulated precipitation forecasts using information obtained from an archive of Rank Histograms generated from a large training dataset. Ensemble behavior in the past was used to adjust the current forecast to be more representative of the observed uncertainty in the training dataset. Forecast lead time and forecast variability were taken into account when choosing how to post-process the forecasts. This technique successfully improved the performance of the ensemble short range forecasts (e.g., 0–3 days lead time), and under certain conditions improved mid range forecasts. The technique only improved forecasts during high precipitation events in the very short range, however. Eckel and Walters' Rank Histogram Bin Correction (HBC) is currently used operationally with the NCEP GEFS to produce 24hr accumulated precipitation forecasts.

We seek in our research to develop a new technique to post-process ensemble forecasts utilizing the same information used to evaluate the ensemble in the Rank Histograms: an archive of past forecasts and observations of what actually occurred. Instead of statistically adjusting the forecast, we replace the ensemble members with relevant past observations themselves, selecting which observations to use with an analog technique. This method of post-processing ensemble forecasts is outlined in detail in the following chapter, and the results of experiments we conducted using this technique are presented in Chapter 4. Ahlquist conceived of the idea in early 2004. Similar ideas were conceived independently by Z. Toth (personal communication), Leonard Smith (personal communication, October 2004) and Hamill, Whitaker, and Mullen (2006). Important distinctions between their work and ours will be discussed in the following chapter. Theoretically, this technique can be applied to any atmospheric variable present in the numerical model, and Ahlquist proposes to expand this technique to additional variables in future research; however, since precipitation is currently among the poorest forecasted fields, it is an important place to begin tackling the challenge of improving the operational ensembles.

Our use of analogs differs significantly in concept from the analog forecasting technique employed and rejected by Lorenz (1969). In that seminal paper, Lorenz further developed the idea of the atmosphere as a chaotic and non-periodic system. In a deterministic framework, if the state of the system were ever to become exactly the same as a previous state (a "perfect" analog), then the system would evolve in exactly the same manner as it did previously following the analog. In effect, a perfect analog in a deterministic system would indicate periodicity, and hence predictability. Lorenz attempted to forecast future states of the atmosphere using the evolution of the atmosphere following "close" analogs in the past, since no perfect analogs existed in his data. These close analogs could be considered the same atmospheric state as the current atmosphere, plus a small error representing the difference between the states. When the evolution of the current atmosphere was compared to the previous evolution of the atmosphere following the analog, the errors grew rapidly, rendering the technique unfeasible. Lorenz concluded that the likelihood of finding analogs good enough for use in prediction of the current evolution of the atmosphere was small. Van Den Dool (1989) found that better analogs could be found through using a patchwork of small areas, and employing independent analogs for each of the smaller areas independently, although the technique was still not very good. In our work, we do not consider or use

the past evolution of the atmosphere following the analog. Rather, we seek examples when the numerical model made similar forecasts in the past to our current forecast, and are only interested in what actually occurred at the time the forecast verified. Therefore, our technique does not encounter the rapidly increasing errors which occur as the atmosphere evolves following the analog state. We are only interested in the information present at the time when the model state was similar to the current forecast.

CHAPTER 2

Method

Our method of post-processing ensemble forecasts replaces the original collection of member forecasts with a sample of verifications for similar forecasts made in the past, using the following procedure. First, we choose an ensemble forecast at a particular point in space and a particular lead time to be post-processed, $F_n(\lambda^*, \varphi^*, t^*)$, for $n = 1, ..., N$, where N is the number of forecasts in the ensemble. Then, we search an archive of past forecasts produced by the same ensemble system, computing the similarity of the chosen forecast to each forecast in the archive. To determine the similarity between two forecasts, consider each ensemble forecast to be a column vector of length N, where each component of the vector consists of a member forecast. Two such vectors can be compared by calculating a metric, a scalar value representing a measure of the difference between two vectors. There are an infinite number of possible metrics. For example, for N-dimensional vectors x and y, consider a fairly general metric:

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{K} \sum_{n=1}^{N} |x_n - y_n|^{\alpha_n} \tag{2.1}$$

where K is a normalizing constant and the exponents α_n could be anything. The two most common choices both set $K = N$. The mean absolute error (MAE) sets $\alpha_n = 1$ for all n, and the mean squared error (MSE) sets $\alpha_n = 2$ for all n. Another possibility might weight different components of the vector differently. Specifically, if the vector elements x_n and y_n are sorted such that $x_1 \leq x_2 \leq ... \leq x_N$ and $y_1 \leq y_2 \leq ... \leq y_N$, greater weight can be given to different parts of the distribution by varying the α_n values. For example, by increasing α_N such that it is greater than all other α_n elements, the greatest weight is given to the largest elements in the distribution of each vector. If α_1 is increased, greater weight will be given to the smallest vector elements when computing the metric. We use a root mean squared

(RMS) metric,

$$D_{RMS}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - y_n)^2} \tag{2.2}$$

to determine the similarity of the sorted elements of ensemble forecast $F_n(\lambda^*, \varphi^*, t^*)$ to similarly sorted forecasts in the archive. The forecasts are sorted as described above. The metric therefore becomes:

$$D(\lambda, \varphi, t) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(F_n(\lambda, \varphi, t) - F_n(\lambda^*, \varphi^*, t^*))^2} \tag{2.3}$$

where $F_n(\lambda, \varphi, t)$ is the n-th sorted forecast for location (λ, φ) at an earlier time t. We allow λ and φ to range through all possible points in space and time contained in our forecast archive, provided the forecast shares the same "season" with the current forecast, defined as lying within ± 45 days of the chosen forecast, with a six month offset for forecasts in the opposite hemisphere. The search is also limited to a region bounded by longitudes of $\varphi = \pm 45^\circ$ of φ^*, to ensure that forecasts verify close to the same time of day. Later, we need to relax this condition so that analogs for North America can come from locations such as Asia. These constraints provide a simplistic accounting of diurnal and seasonal effects while maximizing the number of points to be searched for analogs.

After computing all possible metrics for a chosen $(\lambda^*, \varphi^*, t^*)$, we sort the metric list from least to greatest, ranking the archive forecasts from most similar to least similar. The archive forecast with the smallest metric is chosen as the "best" analog to the current forecast. Next, we exclude any forecasts lying within a certain region of space-time around the best analog to ensure that all of the selected analogs are uncorrelated. The next remaining archive forecast with the smallest metric is chosen as the second best analog, and this procedure is repeated until the desired number of analog forecasts is obtained. Although a large sample of analog forecasts is preferable, the ideal number of analogs to select is limited by the loss of similarity that occurs as metric values increase, which is especially problematic for forecasts of rare events. We now possess a set of points in space-time, $(\lambda_j, \varphi_j, t_j)$ whose ensemble forecasts are similar to the ensemble forecast for $(\lambda^*, \varphi^*, t^*)$. Because of the space-time separation, these forecasts are regarded as independent of each other. Since all of these forecasts were produced by the same numerical model, we assume that the biases inherent in each forecast are similar. This premise is the basis for the effectiveness of our technique.

Having identified points with similar forecasts, we extract the verification values for those points, $V_1, V_2, ..., V_M$, where $V_j = V(\lambda_j, \varphi_j, t_j)$. This collection of verifications then replaces the original forecast as the new ensemble forecast: $F^*_m(\lambda^*, \varphi^*, t^*) = V_m$, $m = 1, ..., M$. In this way, we correct the numerically generated ensemble member forecasts with observations of what actually took place when the original ensemble system produced similar forecasts to the current forecast. If the analogs we chose from the archive are meaningful in the sense that they contain similar biases to the current forecast which we assumed above, then replacing the forecast with the analog verifications will remove those biases from the forecast. The forecast generated by a numerical weather prediction technique that is failing to reflect the true distribution of possible outcomes is corrected by taking a collection of actual outcomes of equivalent forecasts in the past to represent the uncertainty.

This approach to post-processing ensemble forecasts is similar to the procedure developed independently by Hamill et al. (2006), hereafter HWM06, but there are several conceptual and implementational differences. When comparing current ensemble forecasts to past forecasts in an archive, HWM06 defined the forecast vector as a collection of ensemble average values over a group of gridpoints covering a spatial region, as opposed to the collection of individual ensemble forecasts at a single gridpoint. Since topographical features significantly affect the spatial distribution precipitation patterns, analogs can only come from the same collection of points in space. Therefore, in order to have the capability to obtain sufficient analog samples, particularly for rare events, a long forecast record in time is required. HMW06 developed such an archive, a "reforecast" dataset comprised of 25 years of forecasts reproduced using an unchanging ensemble forecasting system. This technique produced significant improvements of precipitation forecasts, which HWM06 attributed primarily to the ability to "downscale" the forecast to the resolution of the observational grid. A disadvantage of this technique is that extreme events are rare in any given location, even in a multi-decadal archive. By allowing analogs to come from locations other than one particular point in space, the length of the required archive can be reduced. Also, by preserving the individual ensemble member forecasts, the ensemble spread is preserved, which is important for predicting the probability of extreme events. A disadvantage to our technique is a loss of ability to downscale the post-processed forecasts, since each gridpoint is considered independently of all other points. A second disadvantage of our current implementation is the potential for topographical features to affect biases present in analog forecasts taken from

different spatial locations. The ability of this technique to improve operational ensemble forecasts and the effects of allowing analog locations to come from multiple regions are discussed in Chapters 4 and 5.

CHAPTER 3

Datasets

3.1 NCEP GEFS Accumulated Precipitation Forecasts

To evaluate our technique to reduce biases in the NCEP operational ensemble, we collected ensemble precipitation forecasts from late November 2004 to the end of March 2006. This section provides details about the NCEP Global Ensemble (GEFS) and the particular adaptations we were required to make to the dataset for our study. NCEP unveiled a major upgrade to the GEFS configuration on 1 May 2006, so the descriptions of the GEFS presented here do not necessarily apply to the products available subsequent to the upgrade.

Our archive of the NCEP GEFS comprised multiple runs of the Global Forecast System (GFS) model's precipitation forecasts, which were then translated from spectral coordinates onto Cartesian grids and made available publicly in GRIB (GRIdded Binary) format. The ensemble consists of a control run of the GFS model and ten perturbation forecasts, five of which consist of positive perturbations to the control initial state, and five consisting of negative perturbations to the initial state. These perturbations are generated through a technique of breeding fast growing errors (Toth and Kalnay 1993), which is briefly described here. A numerical model is iterated to a short time in the future. The same model is run a second time, with a small random perturbation to the initial state that was used in the first run. As demonstrated by Lorenz (1969), the difference between the modeled atmospheric states may grow larger as the model runs are each integrated forward in time. This difference (error) between the iterated model conditions is computed and then rescaled to match the magnitude of the initial perturbations. This entire process is repeated beginning at the new time period, iterated further into the future, except instead of perturbing the initial state of the second model with random errors, the rescaled perturbations determined from the

previous runs are used. After multiple repetitions of this process, the calculated perturbation field begins to reflect particular errors which grow fastest in the model, and hence are the most significant when assessing forecast uncertainty. In this way, the most important perturbations are "bred" from the model after starting with only random perturbations. To generate ensemble member forecasts, the original initial conditions of the model have these bred perturbation values either added to them or subtracted from them. Therefore, utilizing a control run and five different breeding cycles, the NCEP GEFS in the archive have a total of eleven unique member forecasts all produced by the same numerical prediction scheme. Additionally, a higher resolution operational run of the GFS model was interpolated onto the same Cartesian grid and distributed with the ensemble at 0000 UTC, adding a twelfth member forecast once a day.

We archived NCEP "high resolution" GEFS accumulated precipitation forecasts produced four times daily, at 0000, 0600, 1200, and 1800 UTC, with 1° by 1° forecasts with lead times up to 180 hours (7.5 days). Each data record in the archive therefore contained a 65,160 point rectangular grid of accumulated precipitation produced by an ensemble member forecast. The accumulation periods alternated between 6 and 12 hours depending on the forecast lead time. For example, ensemble records of accumulation forecasts for 0–6 hours in the future were followed by records of 0–12 hour accumulations, then 12–18 hours and 12–24 hours. To obtain records entirely of 6 hour precipitation accumulations, we subtracted point by point the record of 6 hour accumulations comprising the first half of the 12 hour record from the 12 hour accumulation values to obtain the second 6 hour half.

$$p_{i,j}(t+6,\ t+12) = p_{i,j}(t,\ t+12) - p_{i,j}(t,\ t+6) \tag{3.1}$$

As a result of this adjustment, negative precipitation values were found at approximately 1 percent of the gridpoints on any given map. The vast majority of these negative values had a magnitude of only 0.1 mm, the smallest nonzero magnitude allowable by the precision of the dataset; however, there were some instances of negative precipitation values of ≈ 10 mm (≈ 0.5 in). Y. Zhu (2006, personal communication) indicated that the source of the problem lay in the interpolation from a Gaussian grid to a 1° by 1° Cartesian mesh, particularly in areas of strong precipitation gradients. Since negative precipitation values are physically impossible, whenever such data were encountered, the values were reset to zero. Although the impact of this dataset problem is expected to be small, it is possible that these interpolation

errors may add an artificial degradation to the forecast skill in these records. The May 2006 upgrade of the NCEP GFS Ensemble distributes records of 6 hour accumulations for all lead times, which has since eliminated this problem.

3.2 Verifications Using CMORPH Precipitation Estimates

CMORPH (Climate Prediction Center morphing method) precipitation estimates were chosen as a verification dataset corresponding to the NCEP Ensemble forecasts due to the extensive spatial data coverage present in the dataset. As availability of verifications is a limiting factor when choosing analog locations, it was advantageous to choose a verification field which had comprehensive spatial coverage. The CMORPH dataset (Joyce et al. 2004) provides precipitation estimates at a high temporal and spatial resolution through the use of satellite-sensed microwave and infrared radiation. Passive radiation sensors are placed on satellite platforms with two different types of orbits. Geostationary satellites orbit the Earth at a radius of 42,200 km with the same angular velocity as the Earth, maintaining a constant position directly over a single location on the equator. Polar orbiting satellites revolve approximately meridionally around the Earth at roughly 1,000 km above the Earth's surface with a greater angular velocity, passing over different swaths of the Earth's surface as the planet rotates. Due to the extremely high resolution that would be required to gather microwave data from a distant geostationary platform, present technology limits microwave sensors to polar orbiting satellites. Since these satellites pass over a given point infrequently, microwave datasets tend to have poor temporal resolution. Microwave-based precipitation estimates, however, are vastly superior to the estimates derived from infrared based cloud top temperature techniques, although the latter are available at a higher temporal resolution. The solution employed by Joyce et al. (2004) in the CMORPH technique is to estimate precipitation rates only from microwave data, but propagate these features between microwave overpasses using motion vectors derived from infrared data. The precipitation features are morphed as they are propagated between microwave overpasses using linear weighted averages of the features as they appear in the previous and subsequent observation times. For example, given a precipitation feature at two observation times, P_0 and $P_{\Delta t}$, the appearance of the feature is morphed pixel by pixel in two evenly spaced time

steps by the following:

$$P(t=0) = P_0 \tag{3.2}$$
$$P(t=0+\frac{1}{3}\Delta t) = \frac{2}{3}P_0 + \frac{1}{3}P_{\Delta t} \tag{3.3}$$
$$P(t=0+\frac{2}{3}\Delta t) = \frac{1}{3}P_0 + \frac{2}{3}P_{\Delta t} \tag{3.4}$$
$$P(t=\Delta t) = P_{\Delta t} \tag{3.5}$$

Comparisons of this technique to other microwave sensed techniques and techniques involving infrared based precipitation estimates by Joyce et al. (2004) indicated a significant improvement in quality when using rain gauge data as a reference. Assessments of the correlation of CMORPH data to ground-based radar precipitation estimates indicated a strong dependence on the availability of frequent microwave overpasses (see Figure 3.1). As the frequency of microwave overpasses increased to 30 minutes or less, the correlation between CMORPH data and radar estimates exceeded 0.7. Once the time between microwave overpasses exceeded 2.5 hours, however, the CMORPH quality was less correlated with radar than techniques using infrared sensed data. Due to this limitation, periodic degradations in the quality of the CMORPH dataset are possible when polar orbiting satellite coverage is sparse. Another disadvantage of the CMORPH technique is the potential for small-scale precipitation features to develop and dissipate between microwave satellite overpasses, which would not be included in the CMORPH archive.

CMORPH data are available as records of 30 minute precipitation rate estimates, 3-hour averaged precipitation rates, and 24-hour averaged precipitation rates on a 0.25° by 0.25° Cartesian grid from 60°S to 60°N. In order to obtain a verification dataset of 6-hour accumulation forecasts, pairs of 3-hour precipitation rate records were combined to produce records of 6-hour accumulations in the following way:

$$P(t,\ t+6)(mm) = 6(hr) \times \frac{d\overline{P}(t,\ t+3)}{dt} \times \frac{d\overline{P}(t+3,\ t+6)}{dt}(\frac{mm}{hr}) \tag{3.6}$$

Unlike the NCEP Ensemble data, where datapoints correspond to precipitation calculated at the gridpoints on the mesh, the CMORPH datapoints represent an average value within each quarter-degree gridbox (see Figure 3.2). When computing a verification of an NCEP forecast point, therefore, we consider the verification to be the average of the four closest CMORPH gridboxes, excluding missing data. This provides a 0.5° by 0.5° precipitation

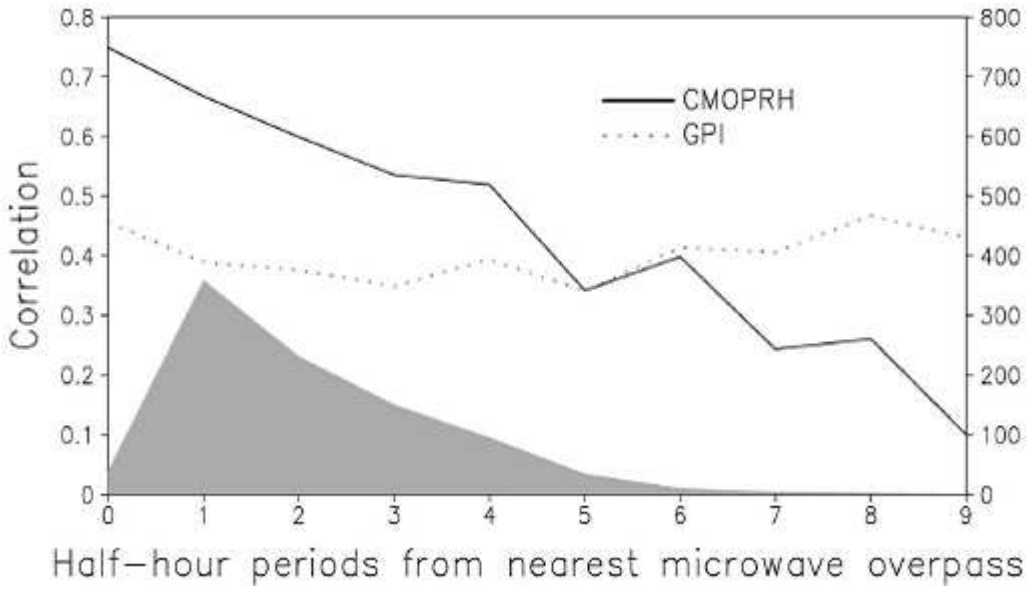


Figure 3.1: Correlation scores between radar estimated precipitation and CMORPH estimates (solid black line). Dashed line represents estimates derived using infrared satellite techniques. The shaded area represents the number of cases used in the calculation for each half-hour period (from Joyce et al. 2004 Figure 13).

accumulation centered on the NCEP gridpoint. CMORPH data, additionally, are computed to a thousandth of a millimeter. Since NCEP precipitation forecasts have a precision of a tenth of a millimeter, the CMORPH data were rounded to match data precisions when scoring.

To evaluate the effectiveness of our analog technique to correct the ensemble distribution, we generated post-processed ensemble forecasts for two 15-day periods: 1–15 July 2005 (a summer case) and 1–15 January 2006 (a winter case), for a network of 72 points spaced evenly at 5 deg by 5 deg intervals over the contiguous United States (see Figure 3.3). At each forecast cycle, therefore, we post-processed 1,080 forecasts for each time period from 0–6 hours to 174–180 hours in the future. Since these forecasts were selected from the middle of the archive, analog forecasts could come from model runs that took place both before and after the forecasts. To prevent any correlation among analog forecasts and the forecast targeted for post-processing, no forecasts issued within 5 days of the target forecast were searched for analogs. The results of these experiments are discussed in detail in the following

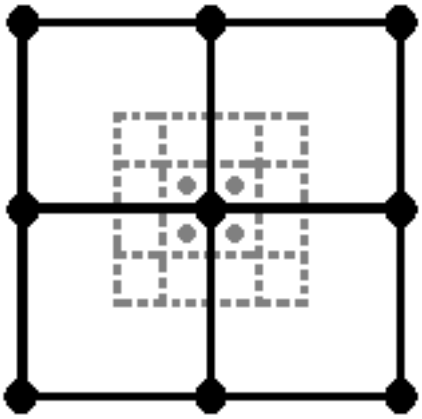


Figure 3.2: Relationship between NCEP 1 deg datapoints (black dots) and CMORPH precipitation estimate gridbox averages (gray shaded areas). CMORPH based observations at a NCEP gridpoint are taken to be the average of the four closest gridboxes.

chapter.

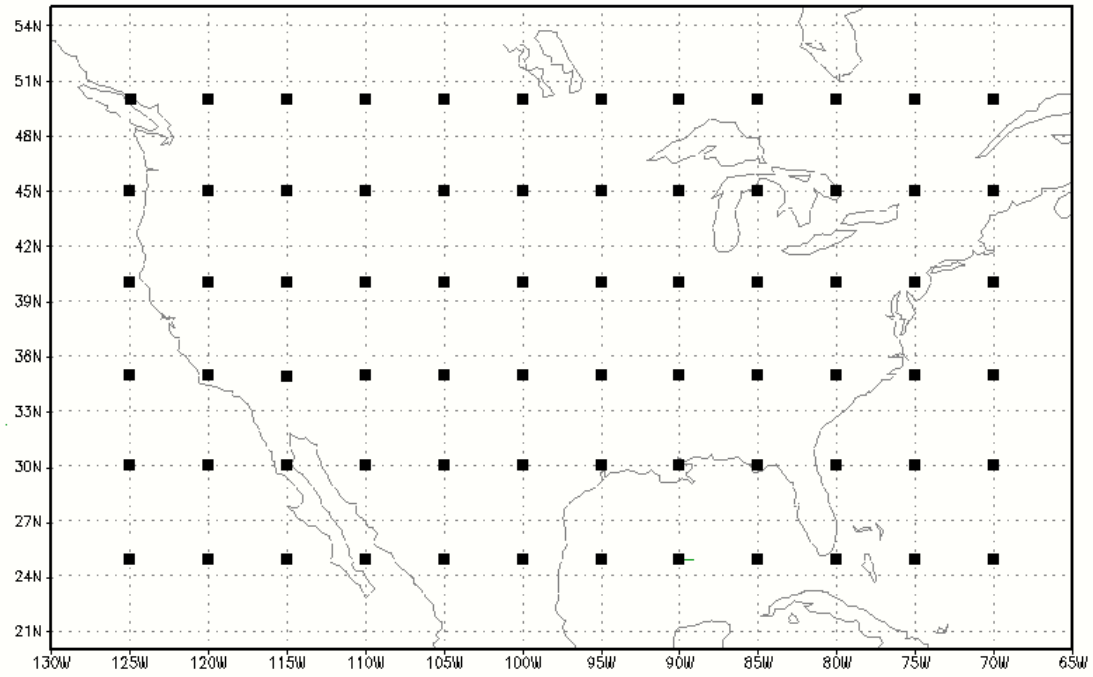


Figure 3.3: Map of spatial points chosen for our experiments.

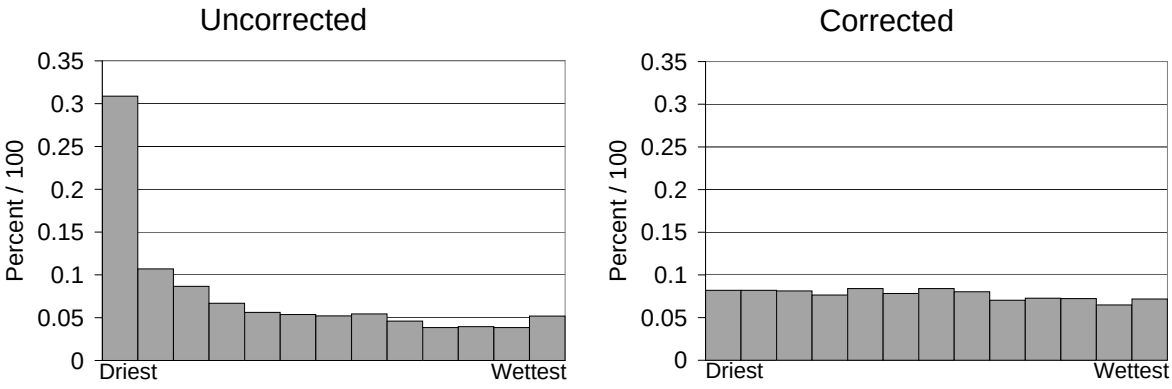
CHAPTER 4

Results

Forecast evaluation involves the comparison of forecast samples to verifications of what actually occurred. As mentioned in the discussion of metrics in Chapter 2, there are a myriad number of ways to compare two vectors of information. Different characteristics of a forecast's departure from the verification can be emphasized depending on the chosen evaluation technique. When forecasting precipitation, for example, one technique may evaluate only a forecast's ability to predict the occurrence of precipitation, while other techniques may evaluate forecast skill for heavy precipitation events. Scoring ensemble forecasts, which provide more information than a single deterministic forecast, is even more complex. This chapter presents evaluations of the experiments described in the previous chapter using several different techniques. The results analyze the ability of both the original NCEP GEFS and the post-processed ensemble to represent forecast uncertainty, generate probabilistic forecasts, and have a skillful consensus forecast.

4.1 Rank Histograms

Rank Histograms, described in Chapter 1, have been a benchmark for evaluating an ensemble's ability to "forecast the forecast skill" (HC98, p. 713). Nonuniform bin distributions indicate that the ensemble is not providing good samples of the possible outcomes given the inherent forecast uncertainty. We computed Rank Histogram diagrams for the NCEP GEFS 6hr precipitation accumulation forecasts and also for the post-processed forecasts in the summer and winter cases. Figures 4.1 and 4.2 demonstrate Rank Histograms computed for these experiments over a 24hr period beginning with forecasts 1 day into the future. When HC97 computed Rank Histogram diagrams for 24hr accumulated precipitation using an ensemble system consisting of forecasts generated by regional models (see Figure 1.2), greater



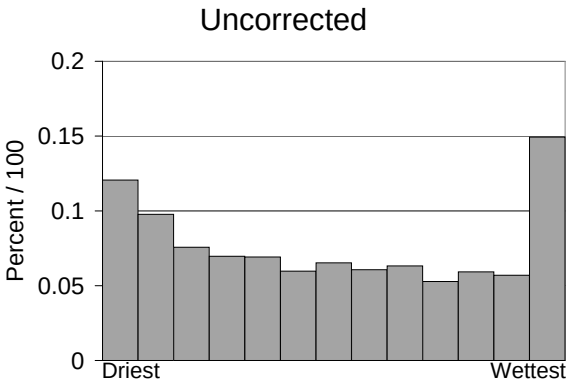
(a) Uncorrected NCEP Ensemble Forecast.

(b) Observed Analog Corrected Ensemble Forecast.

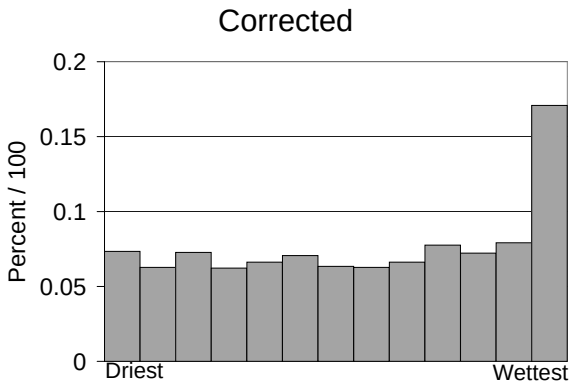
Figure 4.1: Rank Histograms for forecasts of 6hr accumulated precipitation issued at 0000Z on 1–15 January 2006. Figure includes 24–30hr, 30–36hr, 36–42hr, and 42–48hr lead times combined. Verifications were drawn from the CMORPH based archive.

verification counts were observed in the first and last ranked bins, indicating ensemble biases both to underforecast heavy precipitation and overforecast light precipitation. Figures 4.1(a) and 4.2(a) show that the NCEP GEFS 6hr accumulation forecasts exhibit the same tendency to overforecast light precipitation amounts, particularly during the winter season, where a pronounced maximum in verifications ranked drier than every ensemble member exists. The possibility exists of a significant dry bias in the CMORPH dataset, which would skew the verifications towards the driest ranks; however, since the percentage of verifications ranked driest in the summer case is far less pronounced than in winter, it is more likely that the bias exists in the winter ensemble forecasts, and not in a systematic dataset error. The bias towards underforecasting heavy precipitation amounts present in past ensemble studies such as HC97 is strongly pronounced in the NCEP GEFS during the summer but is not apparent during the winter case.

The post-processing technique described in Chapter 2 appears to largely correct the NCEP GEFS precipitation overforecasting (wet) bias in both seasons, producing a nearly uniform Rank Histogram bin distribution for the winter case (Figure 4.1(b)). The pronounced spike in driest verifications evident in Figure 4.1(a) vanishes due to the analog corrections, and verifications appear to be redistributed evenly among the remaining bins.



(a) Uncorrected NCEP Ensemble Forecast.



(b) Observed Analog Corrected Ensemble Forecast.

Figure 4.2: Rank Histograms for forecasts of 6hr accumulated precipitation issued at 0000Z on 1–15 July 2005. Figure includes 24–30hr, 30–36hr, 36–42hr, and 42–48hr lead times combined. Verifications were drawn from the CMORPH based archive.

Although the smaller wet bias during the summer case is likewise corrected, the pronounced precipitation underforecasting (dry) bias for heavy precipitation, revealed by the maximum of verifications in the highest rank in the summer, does not appear to be affected by the post-processing. In fact, the total percentage of wettest verifications increases slightly as a result of the analog correction technique. This increase, however, occurs as each bin receives a greater fraction of the verifications due to the redistribution of values ranked driest. For example, if a post-processed forecast's ensemble members are all zero and the verification is also zero, the verification will be placed randomly into one of the bins, with an equal probability of falling into each bin, including the last.

To demonstrate this, Tables 4.1 and 4.2 provide more detail about the verifications occupying the driest and wettest Rank Histogram bins in the summer. Inches are used rather than millimeters because of the application to operational forecasting in the United States. The first column of each table categorizes all CMORPH verifications on the study grid (Figure 3.3) during 1–15 July 2005 that fell into the first bin of the rank histogram. For example, the first column in Table 4.1 shows that 94.3% of the verifications falling into the first bin of the NCEP GEFS Rank Histogram are accumulations of 0.01 in or less, whereas 100% of the verifications falling into the first bin of the post-processed ensemble

Rank Histogram are accumulations of 0.01 in or less. The second column of each table breaks down the verifications that fall into the last Rank Histogram bin by accumulation amount as well. The column on the right categorizes all of the CMORPH verification data collected during 1–15 July 2005. The verification categories range from values less than the NWS measurable precipitation threshold (0.01 in) to amounts exceeding 1.0 in in a 6 hr period. This way, we can determine whether the distribution of accumulations populating the first and last Rank Histogram bins are unrepresentative of the total distribution of CMORPH precipitation, such as when a disproportionate percentage of verifications falling into the last bin are heavy precipitation accumulations. As expected, the total distribution of CMORPH accumulations in the right column of each table is heavily skewed towards light (or zero) precipitation, with nearly 85% of all verifications being below the threshold of measurable precipitation. Only 0.5% of all verifications fell into the highest category, accumulations of an inch or greater. For both the uncorrected NCEP GEFS and the post-processed forecasts, the contribution from precipitation events under the measurable threshold to the driest ranked bin is even greater than their contribution to the total number of events. For example, every verification ranked driest in the summer compared to the post-processed ensemble came from the smallest precipitation category of 0.01 in or less, whereas accumulations in this category made up only 84.5% of all summer CMORPH verifications. This shows that the majority of ensemble overforecasting occurs when every ensemble members predicts some precipitation, and the observed rainfall is less than 0.01in. For verifications ranked wettest, however, the heaviest precipitation events, defined here as amounts greater than 1 inch in 6 hours, contribute disproportionately, with 91% of the 151 total extreme events observed in the summer being underforecast by the NCEP Ensemble. Although the total count of underforecasted events increases slightly due to the post-processing, the percentage of underforecasted heavy events decreases slightly to 88%. The majority of that increase in underforecasted events instead comes from verifications of 0.01 in or less. This indicates that the primary source of the increase in the wettest bin of the post-processed ensemble's Rank Histogram in Figure 4.2(b) comes either from negligible precipitation events, or cases when every post-processed ensemble member and the verification are zero, and the verification is placed randomly into the last bin.

Figures 4.1 and 4.2 represent the Rank Histogram of 6hr precipitation forecasts summed over an entire day, which filters out any diurnal variations. The distributions of verification

Table 4.1: Breakdown of CMORPH verifications by accumulation amount during 1–15 July 2005 which fell into the first (first column) and last (second column) bins of the NCEP GEFS Rank Histograms. The third column shows the contribution from each accumulation category to all CMORPH verifications during this summer case.

Verification (inches)	Fraction of cases in first bin	Fraction of cases in second bin	Fraction of total
$V < 0.01$	0.943	0.450	0.845
$0.01 \leq V < 0.1$	0.050	0.252	0.096
$0.1 \leq V < 0.25$	0.006	0.117	0.028
$0.25 \leq V < 0.5$	0.002	0.098	0.018
$0.5 \leq V < 1.0$	0.000	0.050	0.007
$V > 1.0$	0.000	0.033	0.005

Table 4.2: Breakdown of CMORPH verifications by accumulation amount during 1–15 July 2005 which fell into the first (first column) and last (second column) bins of the post-processed ensemble Rank Histograms. The third column shows the contribution from each accumulation category to all CMORPH verifications during this summer case.

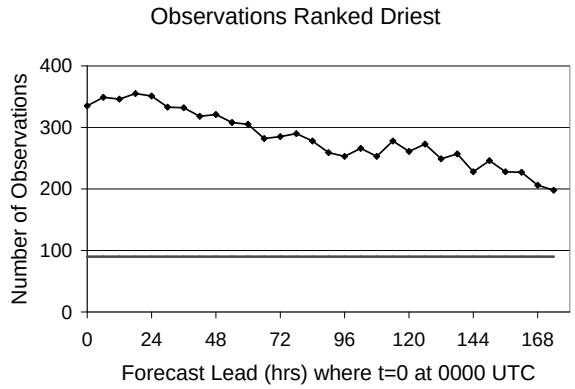
Verification (inches)	Fraction of cases in first bin	Fraction of cases in second bin	Fraction of total
$V < 0.01$	1.000	0.478	0.845
$0.01 \leq V < 0.1$	0.000	0.275	0.096
$0.1 \leq V < 0.25$	0.000	0.104	0.028
$0.25 \leq V < 0.5$	0.000	0.077	0.018
$0.5 \leq V < 1.0$	0.000	0.039	0.007
$V < 1.0$	0.000	0.027	0.005

values in the outer bins, however, exhibit a pronounced diurnal signal in the summer case (see Figure 4.4). More verifications are ranked in the first bin over the spatial domain during 1200–1800 UTC in the summer than at other times of the day. This same time period also corresponds to a daily minimum of verifications ranked in the last bin. These signals are closely related to the diurnal variation in the total number of precipitation events that occur over the contiguous United States. When the total number of measurable precipitation verifications during a six-hour period decreases, the ensemble dry bias with respect to all of the verifications increases, and the wet bias decreases. The effect of diurnal variations in precipitation averaged over the spatial domain will be discussed in more detail in the following

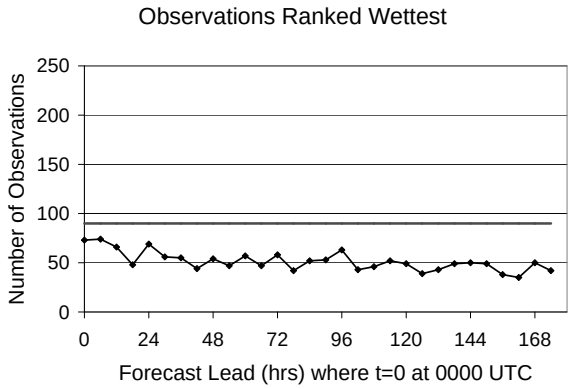
section. Diurnal signals are not apparent in the outer bins of winter case histograms (see Figure 4.3). Comparing the total number of verifications in each bin compared to what is expected given a uniform Rank Histogram, denoted by a solid line in each figure, Figure 4.4(c) shows that there are fewer verifications in the summer case for the post-processed ensemble Rank Histogram driest bin than would be expected. This is due to the disproportionate number of verifications ranked in the last (wettest) bin, which leave fewer verifications to be ranked in the remaining bins, despite the bins having a relatively uniform distribution. Likewise, the disproportionate number of verifications ranked driest in the NCEP GEFS winter case, apparent in Figure 4.3(a) causes the number of verifications ranked wettest shown in 4.3(b) to be lower than expected. The post-processed ensemble outer bins during the winter case (Figures 4.3(c) and 4.3(d)) are both close to what is expected given a uniformly distributed Rank Histogram diagram, which is consistent with the uniform Rank Histogram shown in Figure 4.1(b). Another property of the variation in ensemble biases with forecast time is an apparent decrease in bias as lead time increases, indicated by a decrease in the number of verifications falling into the first and last bins with time. It is unlikely that the ensemble members more accurately sample the forecast uncertainty with greater lead time, given the large biases present in earlier iterations. This apparent decrease in bias may rather be attributed to a greater divergence of ensemble member forecasts as their respective forecast errors grow. With a greater ensemble spread, the verification is more likely to rank inside the range of forecasts, but the forecast itself is less useful since confidence among the member forecasts is quite low.

4.2 Brier Skill Scores

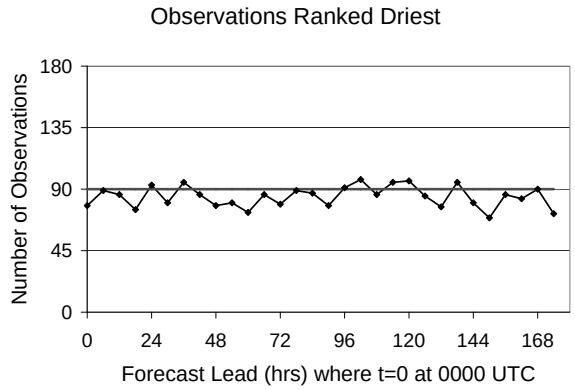
Probabilistic forecasting is one of the most widely used tools that incorporates information about forecast uncertainty. Instead of providing a single forecast, for example, "it will rain 2 or more inches tomorrow," a probabilistic forecast expresses a degree of uncertainty in the forecast: "there is a 60% chance it will rain 2 inches or more tomorrow." Chapter 1 notes that when ensemble members properly represent the forecast uncertainty, they can be used to determine the probability of an event's occurrence by dividing the number of ensemble members forecasting the event into the total number of ensemble members. Evaluations of these probabilistic forecasts can be applied to assess the ensemble's ability to represent the



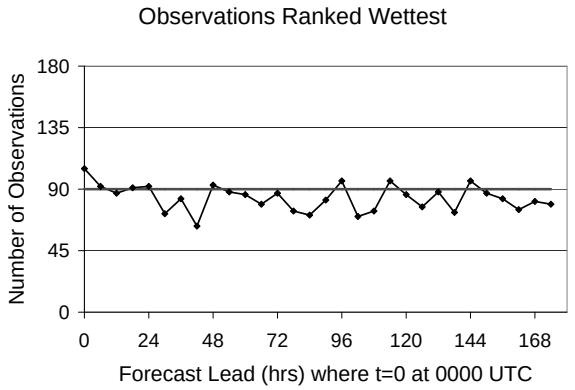
(a) Verifications ranked driest among NCEP GEFS member forecasts.



(b) Verifications ranked wettest among NCEP GEFS member forecasts.



(c) Verifications ranked driest among the post-processed ensemble member forecasts.



(d) Verifications ranked wettest among the post-processed ensemble member forecasts.

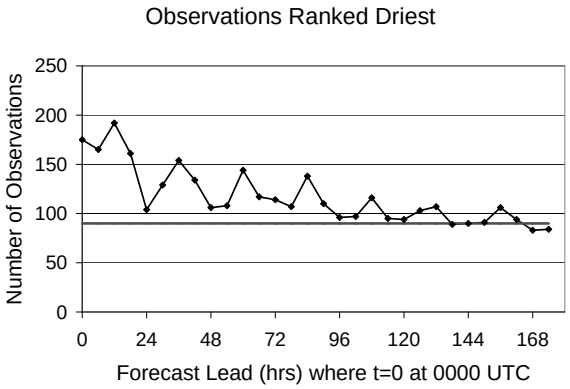
Figure 4.3: Time series depicting the number of verifications which fall into the first and last Rank Histogram bins as a function of forecast lead time. Forecasts were made at 0000 UTC, 1-15 January 2006 (winter case).

distribution of possible outcomes of a forecast.

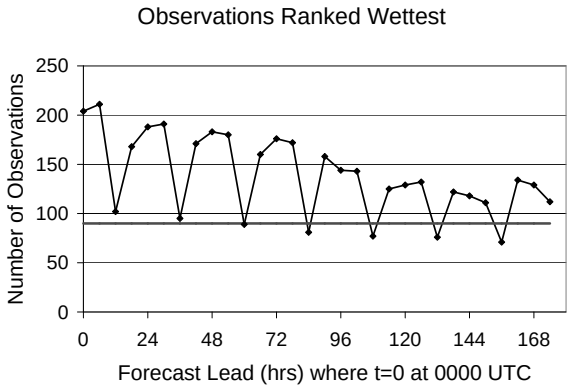
Brier Scores (Brier 1950) and Brier Skill Scores are useful tools for evaluating probabilistic forecasts. Given a collection of forecast probabilities of events and their verifications, the Brier Score is defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - V_i)^2 \tag{4.1}$$

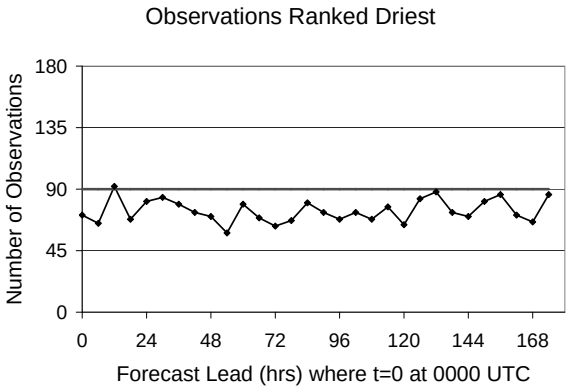
The value p_i represents the forecast probability that event i will occur. V_i represents the verification, and is set to 1 if the event occurred, 0 if the event did not occur. The Brier



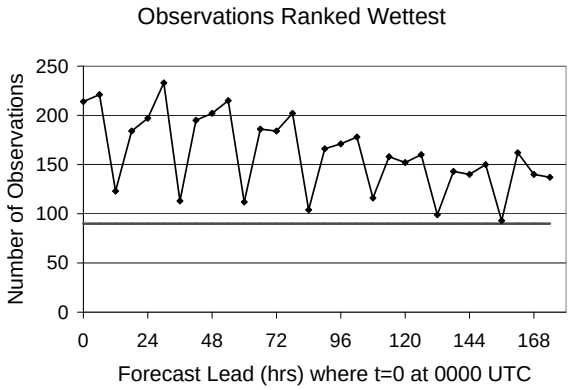
(a) Verifications ranked driest among NCEP GEFS member forecasts.



(b) Verifications ranked wettest among NCEP GEFS member forecasts.



(c) Verifications ranked driest among the post-processed ensemble member forecasts.



(d) Verifications ranked wettest among the post-processed ensemble member forecasts.

Figure 4.4: Time series depicting the number of verifications which fall into the first and last Rank Histogram bins as a function of forecast lead time. Forecasts were made at 0000 UTC, 1-15 July 2005 (summer case).

Score lies between 0 and 1. A Brier Score of 0 represents a perfect forecast sample, e.g., the forecast probability of occurrence was always 1 when the event occurred, and 0 when the event did not occur. A score of 1 thus represents the worst possible probabilistic forecast. Note that a perfect Brier Skill Score can be obtained if the event being forecast is so rare (or the sample size is so small) that no ensemble member forecasts it and it never verifies.

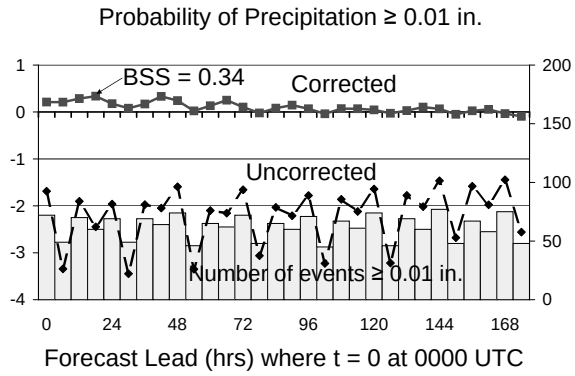
When only considering a Brier Score between 0 and 1, it is frequently difficult to ascertain the skill of a forecasting scheme, since almost any forecasting system will produce scores less

than 1, and the usefulness of a probabilistic forecast is dependent on the events forecasted. For example, suppose a forecast for a rare event such as tornadoes was made daily for 365 days at a single location where the probability of a tornado occurrence was always set to zero. Since the event forecasted is so rare, the Brier Scores are likely to be low, i.e., apparently excellent. For example, if a tornado strikes the location once in the year, the Brier Score of this forecast would be 0.0027, but this low score is only a result of the rarity of the event rather than forecast skill. It is beneficial, therefore, to define a reference forecast with a known value, such as climatology or persistence, and compare the new forecast's Brier Score to this reference. In this way, the Brier Skill Score is defined as:

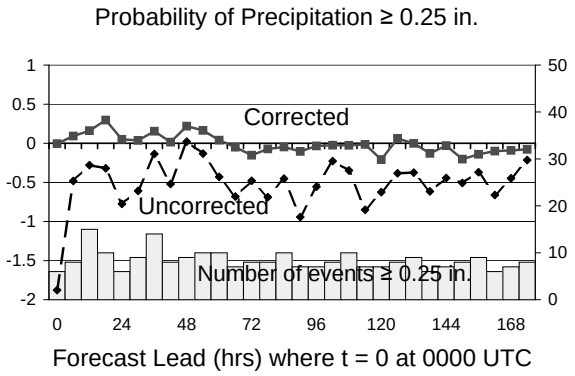
$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{4.2}$$

with a range of $-\infty$ to 1. When the Brier Score of the evaluated forecast is equal to the reference forecast's Brier Score, possessing the same skill, the Brier Skill Score has a value of 0. Therefore, when a Brier Skill Score is greater than 0, the evaluated forecast is considered more skillful than the reference forecast. Negative Brier Skill Scores occur when the reference Brier Score is lower than the evaluated forecast's Brier Score; and such forecasts are considered less skillful than the reference forecast. A drawback to the Brier Skill Score is instability when the overall sample size is small, or the forecasted event is rare, which yields only a small number of forecasts and verifications of the event, or none at all. This may result in BS_{ref} being zero in the denominator, or a small number leading to large negative Brier Skill Scores.

By defining ensemble-based probabilistic forecasts as the fraction of member forecasts which predict an event, N_E/N, Brier Skill Scores were computed for the NCEP GEFS precipitation forecasts and the post-processed ensemble. Precipitation events were defined as 6hr accumulations equaling or exceeding certain thresholds, and two different reference forecasts were chosen for comparison. Figure 4.5 demonstrates the Brier Skill Score as a function of forecast lead time for the winter case, where the reference forecast sets the probability of any precipitation to 0 in all cases (forecasts of zero precipitation). The two events shown in the figure are 6hr accumulated precipitation equaling or exceeding the NWS threshold for measurable precipitation (0.01in), and equaling or exceeding 0.25in. Considering forecasts of the probability of the first event shown in Figure 4.5(a), the post-processed ensemble forecast has a significantly higher skill than the NCEP GEFS. In fact, the



(a) Forecasting the probability of precipitation equaling or exceeding 0.01in in 6hrs.



(b) Forecasting the probability of precipitation equaling or exceeding 0.25in in 6hrs.

Figure 4.5: Brier Skill Scores for the NCEP GEFS (dashed line) and the post-processed ensemble (solid line), with values on the left axis. The reference forecast used is forecasts of zero probability of precipitation. The number of actual occurrences for each event forecasted in a 6hr period is included (columns), with values on the right axis. Forecasts came from 1-15 January 2006.

latter is consistently less skillful than the reference forecast of zero precipitation probabilities throughout the forecast period. The corrected forecast, however, is more skillful than this reference forecast through nearly all of the forecast period, with Brier Skill Scores hovering just above zero after 72hrs. A maximum skill of 0.34 occurs at the 18–24hr accumulation period, corresponding to 1800–0000 UTC, which is during the afternoon and evening over the Contiguous United States. The post-processed forecasts also improve forecast skill for probabilities of heavier events, such as accumulations of 0.25in or greater shown in Figure 4.5(b). This corrected ensemble forecast is also more skillful than the reference forecast during the first three days of the forecast, but becomes less skillful than the reference beyond 72hrs. Forecasts of the probability of heavy precipitation become less skillful than forecasting zero precipitation when there are cases of ensemble members predicting heavy rainfall that subsequently do not verify. Given the large fraction of heavy events which do occur yet are underforecasted by the ensemble, it is interesting to note that ensemble members forecast heavy precipitation in situations where the event does not occur.

Further examination of Figure 4.5 reveals an unusual characteristic of these error scores in comparison to Brier Skill Scores shown in previous work, such as HWM06. There is a

pronounced diurnal periodicity in the forecast skill, particularly in the uncorrected NCEP ensemble forecasts. This signal is most pronounced when scoring probabilistic forecasts of lower thresholds of precipitation. In particular, skill reaches a local minimum for forecasts verifying at 0600–1200 UTC relative to forecasts of zero precipitation. Local maxima in skill occur at 0000–0600 UTC in the NCEP GEFS forecasts, and 1800–0000 UTC for the corrected ensemble. The signal in the uncorrected NCEP ensemble is closely related to the total number of measurable precipitation events verifying at each forecast lead time. Over the contiguous United States, the highest frequency of measurable rainfall events occurs at 0000–0600 UTC, and the lowest frequency of events occurs at 0600–1200 UTC. Since the reference forecast of zero precipitation becomes less appropriate when there are more precipitation events, the resulting fluctuations in the reference Brier Score contribute to the diurnal signal. The daily variation in skill is far less pronounced in the post-processed ensemble forecasts, becoming less evident in forecasts of higher precipitation thresholds. The same signals are apparent in the 11-member ensemble forecasts run at different times of day (0600, 1200, 1800 UTC) as well.

Figures 4.7(a) and 4.7(b) represent the Brier Skill Scores also using zero precipitation as a reference but for 1–15 July 2005. As in the winter, the post-processing technique produces significant improvements in skill during the summer. The NCEP ensemble skill continues to oscillate widely on a diurnal cycle, but the summer pattern is offset. As with the winter case, the oscillations are closely related to the precipitation frequency during various 6hr periods of the day. During the summer season, certain regions of the contiguous United States receive greater amounts of convective precipitation, such as the Southeast, which has a maximum thunderstorm frequency in the afternoon, and the central United States, which receives more convective precipitation in the overnight hours (Wallace 1975, see Figure 4.8). The corrected ensemble also oscillates, though not as sharply as the NCEP ensemble. The highest daily skill for both ensembles occurs at the times of the greatest number of events (1800–0000 UTC), which is similar to what occurred during the winter case. The periods of highest and lowest precipitation frequency on average over the United States are offset from each other in the winter and summer seasons, corresponding to the offset in diurnal skill oscillations. Brier Skill Scores were also computed for precipitation events equaling or exceeding 0.5in in the summer. Both the overall skill and diurnal signals are reduced at this higher precipitation threshold, and the total number of events decreases to fewer than 25 at

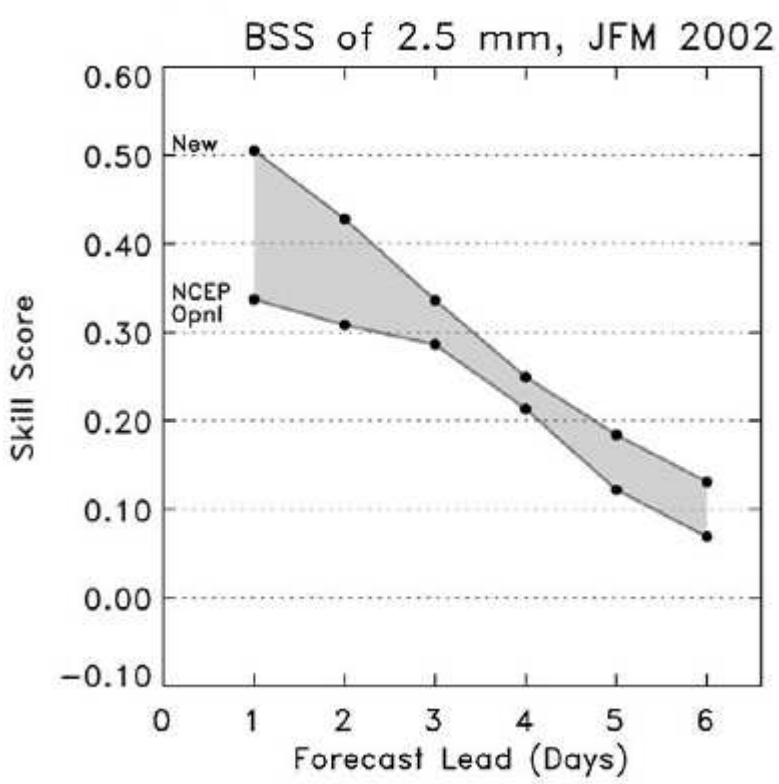
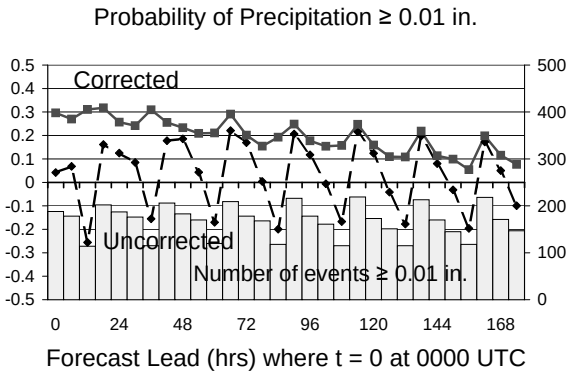


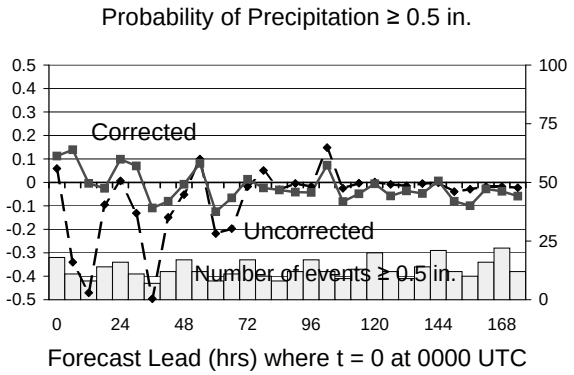
Figure 4.6: Example of Brier Skill Scores computed for NCEP GEFS 24hr accumulated precipitation forecasts and the post-processing technique employed by Hamill et al. (2006). The reference forecast is season-averaged climatology. Note the sharp decrease in forecast skill with lead time, which is not apparent in the NCEP GEFS 6hr accumulation forecasts. (From Hamill et al. 2006 Figure 5a).

each time period. After three days lead time, the Brier Skill Scores for both ensembles are essentially zero.

Brier Skill Scores were also computed using a reference forecast of diurnal persistence. A diurnal persistence forecast considers previous states of the atmosphere at the same period day in order to construct the prediction. For example, a forecast for today at 0000–0600 UTC is made by taking what occurred previously at 0000–0600 UTC as the forecast. This forecast is used to predict what will happen at 0000–0600 UTC throughout the entire forecast period. The probabilistic reference forecasts using persistence were set to 0 if the event did not occur and 1 if the event did occur. Since diurnal effects are accounted for by this reference forecast, it was expected that some of the diurnal signal in the Brier Skill Score values for the



(a) Forecasting the probability of precipitation equaling or exceeding 0.01in in 6hrs.



(b) Forecasting the probability of precipitation equaling or exceeding 0.5in in 6hrs.

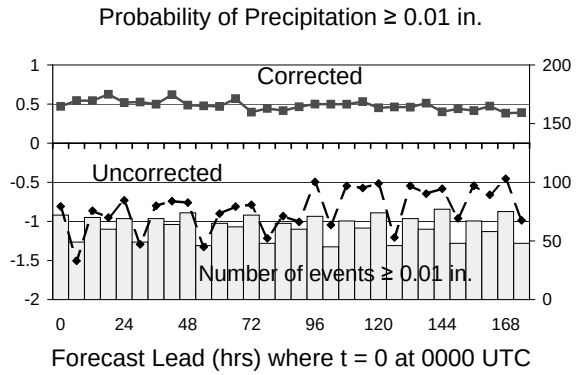
Figure 4.7: Brier Skill Scores for the NCEP GEFS (dashed line) and the post-processed ensemble (solid line), with values on the left axis. The reference forecast used is forecasts of zero probability of precipitation. The number of actual occurrences for each event forecasted in a 6hr period is included (columns), with values on the right axis. Forecasts came from 1-15 July 2005.

ensemble forecasts would be reduced. 6hr accumulation forecasts may be heavily influenced by small-scale features, however, which may not repeat. Since small-scale precipitation events may vary largely from day to day, resulting in large forecast errors when using persistence forecasts, this may explain why the Brier Skill Scores using persistence as a reference tended to be higher than those using zero precipitation. The Brier Skill Scores with reference of diurnal persistence are shown in Figure 4.9. For forecasts of measurable precipitation (Figure 4.9(a)), the post-processed ensemble forecasts perform remarkably better than the uncorrected NCEP forecasts. Whereas the NCEP GEFS is less skillful than the persistence forecast, the corrected ensemble's Brier Skill Scores peak above 0.5. While the diurnal oscillations in skill are more suppressed in the corrected ensemble, they are still largely present in the NCEP ensemble forecasts, with the same pattern found when using a reference forecast of zero precipitation (Figure 4.5(a)). Turning to Figure 4.9(b), which displays the skill of heavy precipitation forecasts during 1-15 January 2006, the diurnal signal is only apparent after a period of three days. These signals are also highly correlated with the pattern of heavy precipitation events. Note that the highest frequency of heavy precipitation events does not coincide with the highest frequency of measurable precipitation events. This

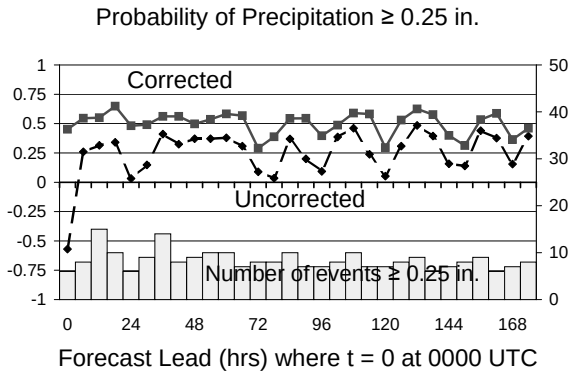
Figure 4.8: Thunderstorm frequency and diurnal period over the contiguous United States (Wallace 1975 Figure 2). The orientation of the barbs indicates the time of day on a 24hr scale of maximum thunderstorm frequency at a given location. Barbs pointing "from the north" for example, indicate a midnight maximum. Barbs pointing "from the south" indicate a noon maximum.

is consistent with the different maxima of stratiform (usually light) and convective (usually heavy) precipitation events determined by Wallace (1975). Although the NCEP GEFS was more skillful than persistence when forecasting heavy precipitation, it was still outperformed by the post-processed forecasts, whose Brier Skill Scores were near 0.5.

During the summer case (Figure 4.10), both the uncorrected and corrected ensembles are very skillful in comparison to the persistence reference. The post-processed ensemble in general outperforms the uncorrected NCEP forecasts. For the higher precipitation threshold (0.5in or greater, Figure 4.10(b)), the skills of both ensembles are very close beyond three days in the forecast. Although the corrected ensemble's forecasts are noisier with time, a diurnal cycle is still evident, particularly after three days, with the daily minimums in skill corresponding to the time of fewest events. The daily maxima in skill, however, particularly in the NCEP GEFS, occur in the 0000–0600 UTC time period, which immediately follows the period with the highest number of events.

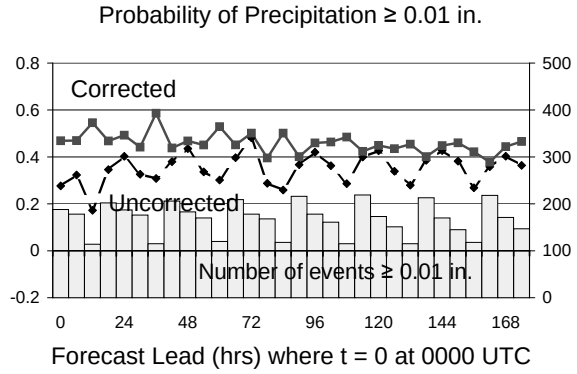


(a) Forecasting the probability of precipitation equaling or exceeding 0.01in in 6hrs.

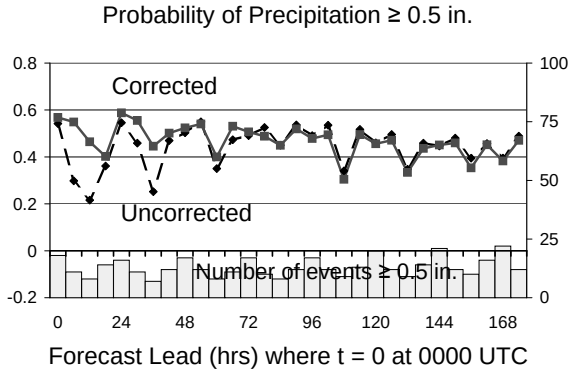


(b) Forecasting the probability of precipitation equaling or exceeding 0.25in in 6hrs.

Figure 4.9: Brier Skill Scores for the uncorrected NCEP Ensemble (dashed line) and the corrected ensemble (solid line), with values on the left axis. The reference forecast used is forecasts of diurnal 6hr persistence. The number of actual occurrences for each event forecasted in a 6hr period is included (columns), with values on the right axis. Forecasts came from 1-15 January 2006.



(a) Forecasting the probability of precipitation equaling or exceeding 0.01in in 6hrs.



(b) Forecasting the probability of precipitation equaling or exceeding 0.5in in 6hrs.

Figure 4.10: Brier Skill Scores for the uncorrected NCEP Ensemble (dashed line) and the corrected ensemble (solid line), with values on the left axis. The reference forecast used is forecasts of diurnal 6hr persistence. The number of actual occurrences for each event forecasted in a 6hr period is included (columns), with values on the right axis. Forecasts came from 1-15 July 2005.

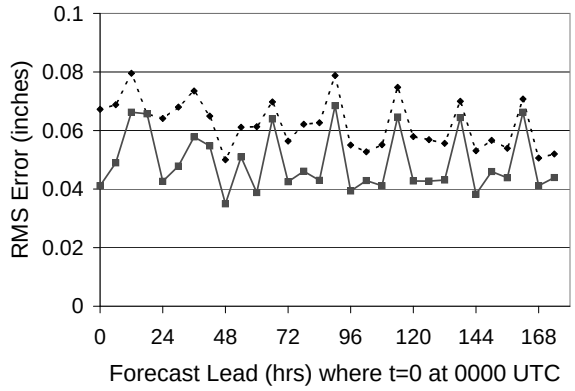
An unusual characteristic present in almost all of the Brier Skill Score figures is a lack of degradation in forecast skill as lead time increases, particularly for the NCEP GEFS forecasts. There are some cases, especially apparent in Figure 4.9(a), where the uncorrected NCEP GEFS skill averaging out the diurnal signal even improves after 72 hours. When the statistical properties of a deterministic system are stationary, it is not expected that the skill of a numerical forecast will improve with time, since the conditions used to initialize the forecast become older. An example of a situation when the statistical properties of the atmosphere are not stationary is a location on the Gulf Coast in the southeastern United States during the winter and the summer. During the winter season, the standard deviation of daily-averaged temperatures at a location on the Gulf Coast is larger than the standard deviation of summer daily temperature averages. Temperatures in the southeast during the winter vary widely when both continental polar and maritime tropical air masses dominate at different times. During the summer, however, the average temperature on the Gulf Coast varies very little from day to day. Therefore, it is possible to make a summer temperature forecast for a location on the Gulf Coast years in advance that is more skillful than a 7-day winter temperature forecast. Since the case studies presented here involve 15-day periods within the middle of a summer and winter season, however, it is unlikely that the statistical properties of the atmosphere vary much. Since the increase in Brier Skill Score is small, and the NCEP GEFS forecast skill varies widely in proportion to the number of precipitation events occurring, it is possible that some of the increase is due to a corresponding small increase in the number of precipitation events at many time periods after 72 hours in comparison to earlier time periods, seen in the columns in Figure 4.9(a).

Brier Skill Score evaluations of NCEP GEFS 24hr accumulated precipitation forecasts with a reference of climatology performed by HWM06 (see Figure 4.6) indicate a degradation of skill with time. The post-processed forecast skill presented here does decrease somewhat with time in comparison to a reference forecast of zero precipitation, but not to the degree shown in HWM06. The general steadiness of the forecast skill with time may indicate that the uncorrected NCEP GEFS 6hr forecasts have no skill with reference to climatology, even in the shortest range. Since the post-processed forecasts outperform both the NCEP GEFS precipitation forecasts and the reference forecasts, even when it becomes steady with time, it is useful since it applies a correction of the climatology of the NCEP GEFS forecast to the climatology of the verification dataset.

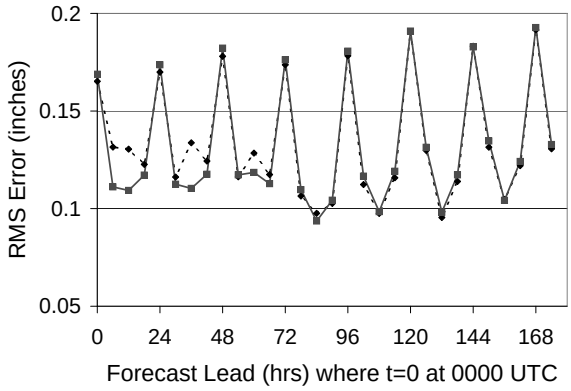
4.3 RMS Error Scores

Since the ensemble consensus (the average of all of the ensemble members) is a popular product and has been shown to perform better than a single numerical forecast produced at a higher resolution, we evaluated the consensus forecasts of the uncorrected NCEP and post-processed ensemble using the RMS error. The results are displayed in Figure 4.11. The post-processing technique we employed reduces the RMS error scores in the winter case (Figure 4.11(a)) by nearly one-third for some time periods, and a reduction in RMS error was observed through the duration of the forecast. There was very little reduction in RMS error after post-processing the ensemble in the summer case (Figure 4.11(b)), especially after three days lead time. In both cases, a strong diurnal variation in RMS errors was present. For the summer case, the worst error scores coincided with the 6hr periods that had the highest frequency of heavy ($\geq$ 0.5in) precipitation events. This is expected, since error magnitudes tend to be larger during heavy precipitation events, and these errors are given more weight in the RMS calculation. The signal in the winter case is more complex, with maximum errors occurring during 1200–1800 UTC on the first two days, coinciding with the maxima of heavy ($\geq$ 0.25in) precipitation events, but switching to 1800–0000 UTC for the rest of the forecast period. During the first two days of the forecast period (e.g., note the columns in Figure 4.9(b)), the 1200–1800 UTC periods are the only 6hr accumulation times with more than 10 events.

To summarize the results of this research, the technique of replacing ensemble forecasts with a collection of observed analog forecast verifications, where we allow analogs to come from any location with similar forecasts resulted in a substantial improvement in the ensemble forecast's ability to assess the true forecast uncertainty more accurately. The most significant improvements occurred in the short range (0–3 days into the future), though there was some improvement throughout the 7.5 day forecast period. The majority of the improvement came as a result of a correction in the NCEP Ensemble's wet bias, where the ensemble members frequently forecasted more precipitation than actually occurred. There was little improvement in correcting the tendency of the NCEP Ensemble to underforecast heavy precipitation, however. These results are significant, however, when considering that only a 1.5 year archive was used as an analog pool. These results are not surprising because there are vastly more low precipitation examples in an archive of any length than high precipitation



(a) Forecasts valid 0000 UTC, 1-15 January 2006.



(b) Forecasts valid 0000 UTC, 1-15 July 2006.

Figure 4.11: Root Mean Squared Error scores for uncorrected NCEP Ensemble 6hr accumulated precipitation forecasts (dashed line) and corrected ensemble forecasts (solid line).

examples.

Use of this technique has the potential to generate significant corrections to the ensemble distributions without large expenses in computer time and memory to develop a large archive. The implications of these results, such as relating the diurnal skill oscillations to dataset issues, and the effectiveness of the post-processing technique in general, as well as ideas for future work to reduce the problem of underforecasted extremes, will be discussed in the following chapter.

CHAPTER 5

Conclusions

The reasoning behind our technique to post-process ensemble forecast output using observed verifications of similar past forecasts is based on the premise that the analog forecasts are so closely related to the forecast we are correcting that they can be considered identical with a negligible error. We assume that if one forecast is identical to another, then the underlying conditions which caused the ensemble to generate that forecast are also identical, and its verification is an equally plausible outcome for the other forecast. In this way, a new ensemble of equally plausible outcomes is generated, but instead of outcomes produced by deterministic equations, we have outcomes drawn from the observational dataset itself. The corrected ensemble thus becomes adjusted to fit the observed forecast uncertainty: the various "true" outcomes rather than bias-containing modeled outcomes. Theoretically this technique has the potential to improve operational ensemble forecasting significantly. There are two limiting factors, however: a paucity of extreme event observations and the degree of difference in underlying conditions between analog forecasts. Similar ensemble forecasts, for example, may exhibit different biases in alternate spatial locations. If these locations are chosen as analogs, the observation of what occurred at one point will not be an equally plausible outcome at another point.

Instead of searching for analogs only in the same location where the forecast is made, we allowed analogs to come from a large range of locations, without regard to topography. By doing this, we significantly reduced the length of the archive necessary to search for analogs. Hamill et al. (2006), for example, used a 25-year archive to search for analogs, while we used a 1.5 year archive, with both cases resulting in significant forecast improvement. A disadvantage of allowing analogs to come from multiple locations is the loss of ability to downscale the corrected forecast. Downscaling occurs when the observational dataset has a

higher resolution than the original ensemble forecast and analogs are found over a collection of spatial points rather than a single point. When replacing a spatial block of ensemble forecasts with verifications of observed analogs, the new ensemble forecast has the resolution of the verification dataset. Hamill et al. (2006) attributed much of the success in their technique to this downscaling. The computational advantages of having a shorter archive and meaningful bias corrections at the coarser resolution of the original ensemble are critical for operational meteorology, however, which makes this research meaningful. The dilemma thus becomes a question of balance between decreasing the archive length by allowing multiple locations to be searched for analogs and the selection of locations where the ensemble biases are similar.

The post-processing technique presented here effected a substantial improvement in the ensemble-based probability of precipitation forecasts. The NCEP GEFS 6hr precipitation forecasts overpredicted rainfall to the point that forecasting no precipitation at all was more skillful. Replacing these forecasts with analog verifications eliminated this bias, as shown by the removal of the nonuniform spike in the corrected ensemble's driest bin (see Figure 4.1). The correction also caused an improvement in the performance of the ensemble consensus forecast in the winter, where the wet bias was most pronounced. Since precipitation distribution is heavily weighted towards zero or light precipitation, our correction works well for the majority of cases. The dry bias in the ensemble during the summer for heavier rainfall events, however, was largely uncorrected by our post-processing. When precipitation events of an inch or more occurred in a 6hr period, the observation exceeded every ensemble member forecast of that event 88 percent of the time, which is only a 3 percent reduction from the performance of the uncorrected NCEP GEFS. The failure to correct this bias in our experiments can be explained by the two limiting factors described above. First, since heavy precipitation events are far more rare due to the distribution of rainfall, few good analogs exist unless a substantially longer archive is used, which can be difficult to obtain as the forecasting model evolves on a regular basis. Even when allowing analogs to come from multiple locations, a 1.5-year archive was insufficient to produce a sample of good analogs for forecasts of the heaviest precipitation events. A second explanation of the failure relates to the appropriateness of analogs. Figure 5.1 depicts a map of the locations selected for analogs whose verifications replaced an ensemble forecast of a high precipitation event. The forecast was made for a coastal location in the Florida panhandle during the landfall of Hurricane Dennis in July 2005, indicated by a black dot. The locations of the "best" analogs

found during the archive search are indicated by large crosshairs. The original forecast was modeling a tropical cyclone, whereas many of the analogs whose verifications replaced the original forecast were taken from the Amazon rain forest. Although heavy precipitation is associated both with tropical cyclones and convective thunderstorms over the inland rain forest, the different processes which produce such rainfall events were ignored during the search for analogs. Since different biases in the model likely exist for the diverse processes which produce heavy rainfall, choosing inland Amazon locations to serve as analogs for a tropical cyclone event is unlikely to adjust the forecast appropriately. For our technique to significantly improve ensemble forecasts of extreme precipitation, we would need to lengthen our archive beyond 1.5 years and be more judicious in choosing analog locations. A simple way to achieve this is to define precipitation regimes based on physical processes which cause rainfall and categorize forecasts of heavy precipitation, limiting searches to events which fall into the same precipitation category. Another method would be to limit analog searches to regions with similar geography; for example, searching the entire U.S. Gulf Coast for analogs of forecasts made at a point on the Florida panhandle.

Other than the improvement in overall skill, one of the most persistent and striking characteristics of the ensemble 6hr precipitation forecasts is the strong diurnal variability in skill present in the uncorrected NCEP forecasts, and a suppression of that variability in the corrected ensemble. Much of this signal is strongly related to diurnal cycles in the types of precipitation events that occur on average over the contiguous United States. Brier Skill scores improved with reference to forecasting zero precipitation when the number of precipitation events increased. This variation due to the choice of zero precipitation as a reference forecast is to be expected; however, it does not explain why there was a significant reduction in the diurnal variability of the corrected ensemble forecasts. Hamill et al. (2006) applied their corrections to 24hr precipitation, so any diurnal signals would have been smoothed by the longer accumulation period and these signals were not present in their work. Since our technique scales the distribution of the ensemble members based on the observational dataset, if a diurnal oscillation in the quality of the observations were present, it would not be apparent in the corrected forecast. It would show up in the uncorrected NCEP forecasts, however, since variations in observational quality would be reflected by changes in the forecast bias. As shown in Figure 3.1, the quality of CMORPH precipitation estimates decreases sharply as the time between microwave satellite overpasses increases.

Figure 5.1: A map of locations where analogs were chosen to produce a post-processed ensemble forecast of a tropical cyclone high precipitation event in the Florida panhandle (black dot). Analog locations are marked with crosshairs.

Part of the signal in the uncorrected NCEP forecasts may then possibly be attributed to variations in the quality of CMORPH data.

Noting that the signal in skill is highly correlated with the type of precipitation occurring, another reason for the reduction in signal in the post-processed forecasts with respect to the Brier Score reference forecast may be a result of the type of precipitation events which are well corrected by our technique. We would expect the NCEP ensemble's wet bias to be most apparent when there are fewer precipitation events. Referring to Figure 4.5, we find that the greatest drop in the NCEP ensemble's probabilistic forecast skill indeed occurs during the periods with the fewest measurable precipitation events. As noted above, the corrected ensemble made its greatest improvements to the wet bias in the NCEP GEFS, so we would expect the greatest improvement in forecast skill to occur during these same periods. Since the uncorrected ensemble's drop in skill during these periods accounts for most of the signal's amplitude, the elimination of the errors causing the drop will erase most of the signal.

Due to the demonstrated degree of improvement in the ensemble's ability to assess the forecast uncertainty and the potential for further improvements, additional study and development of this post-processing technique is important. The downscaling method of Hamill et al. 2006 adds great skill when valid local analogs exist. When local analogs of quality do not exist, we propose use of analogs from other climatologically similar locations. This will be the subject of our future research. Implementation would require a determination of the degree of analog insufficiency that would necessitate searching other locations. Additional research which develops methods to select analog locations more carefully can also greatly improve our analog technique's ability to correct the ensemble distribution in cases of extreme precipitation, which is of primary importance to operational meteorology centers. Ahlquist (2006, personal communication) proposes a second phase of research designed to reduce the computational resources needed to utilize this post-processing, preparing the technique for operational use. Instead of repeated searches of the archive to produce new ensemble forecasts, a single detailed search will be conducted, from which functions would be calculated to adjust forecast distributions to the distributions observed in the analog verifications.

APPENDIX A

Real-Time Ensemble Forecast Products

As part of this research, large volumes of precipitation forecasts were downloaded four times daily for the purpose of building an archive. Due to the availability of these forecasts in real time, we established a Web site displaying a suite of five ensemble precipitation products. This Web site updates automatically four times daily when data are available, and takes advantage of the global extent of the dataset by displaying graphics for regions worldwide. This appendix provides a brief description of the products that are available on the Web site, and a discussion of using GRIB (GRIdded Binary) compressed ensemble data in the GrADS (GRidded Analysis And Display System) graphics utility. The Web site can presently be viewed by accessing the following URL:
http://ahlquist.met.fsu.edu/research/ensproducts.html

The first set of products generated on the Web page is contoured plots of the ensemble consensus (average) for 6hr accumulated precipitation. At each gridpoint, the average of all non-missing ensemble member forecasts is computed, producing a new grid of data. Figure A.1 demonstrates an example of this product. The contours plotted are precipitation accumulation values.

The next set of ensemble products produced is plots of the ensemble median forecast. If the collection of N forecasts at a point are sorted such that $F_1 \leq F_2 \leq ... \leq F_N$, the ensemble median represents the midpoint forecast in the collection. For an odd number of forecasts in the ensemble, the midpoint forecast is simply $F_{int(N/2)+1}$, where $int(N/2)$ is the result of $N/2$ truncated to an integer value. When there are an even number of ensemble members, the median forecast is taken to be the average of the two forecasts which surround the midpoint of the collection, or:

$$\frac{F_{int(N/2)} + F_{int(N/2)+1}}{2} \tag{A.1}$$

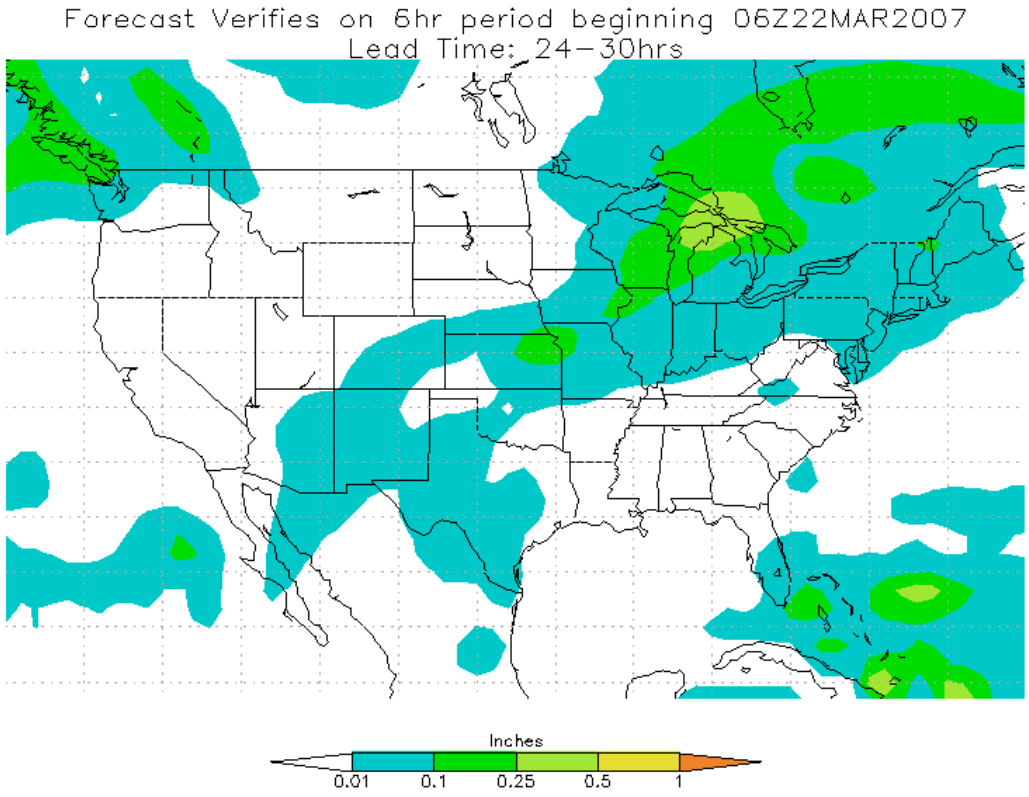


Figure A.1: Example plot of an ensemble consensus 6 hr accumulated precipitation forecast available on the research Web page.

The average of a collection is affected by the extreme values, whereas the median represents a value at which equal numbers of elements in the collection are greater than or less than the value. For example, if most ensemble members at a point forecast no precipitation, and one member forecasts heavy precipitation, the ensemble average at that point will be greater than the median. Figure A.2 demonstrates an example of this product. As with the ensemble average products, the plotted contours represent accumulation amounts.

A third set of products presented on the Web page is probabilistic precipitation forecasts. Five precipitation thresholds are defined: 0.01in, 0.1in, 0.25in, 0.5in, and 1.0in. The probability of precipitation equaling or exceeding each threshold is computed point by point by dividing the number of ensemble members forecasting accumulations greater than or equal to the threshold by the total number of non-missing ensemble forecasts. Figure A.3 shows an example of this product. Note that contours in this product represent probabilities ranging

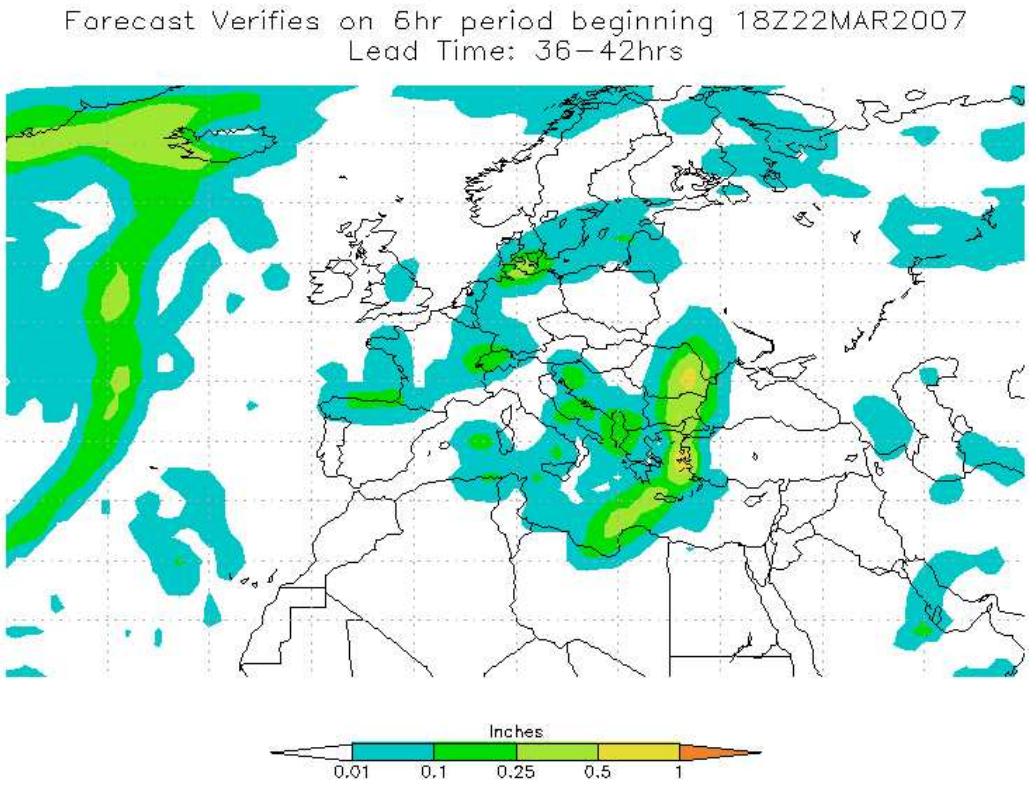


Figure A.2: Example plot of an ensemble median accumulated precipitation forecast available on the research Web page.

from less than 0.1 to greater than 0.9.

The next set of ensemble precipitation products were the first to be created and displayed on this Web site. Ensemble spaghetti diagrams provide users with information about the spread (divergence) of ensemble members as a forecast is integrated forward. As with the probabilistic products, 6-hour accumulation thresholds are defined, and a single contour of that value is plotted for each member of the ensemble on the same map. When the ensemble members are in close agreement, representing a high-confidence forecast, the contours plotted will be tightly clustered. When the ensemble members diverge considerably, great variability in the shape and location of the contours from forecast to forecast is evident. Figure A.4 shows an example of spaghetti 6mm (approximately 0.25in) accumulation contours plotted with a south polar view.

The final group of products produced on the Web page is also plots of the ensemble

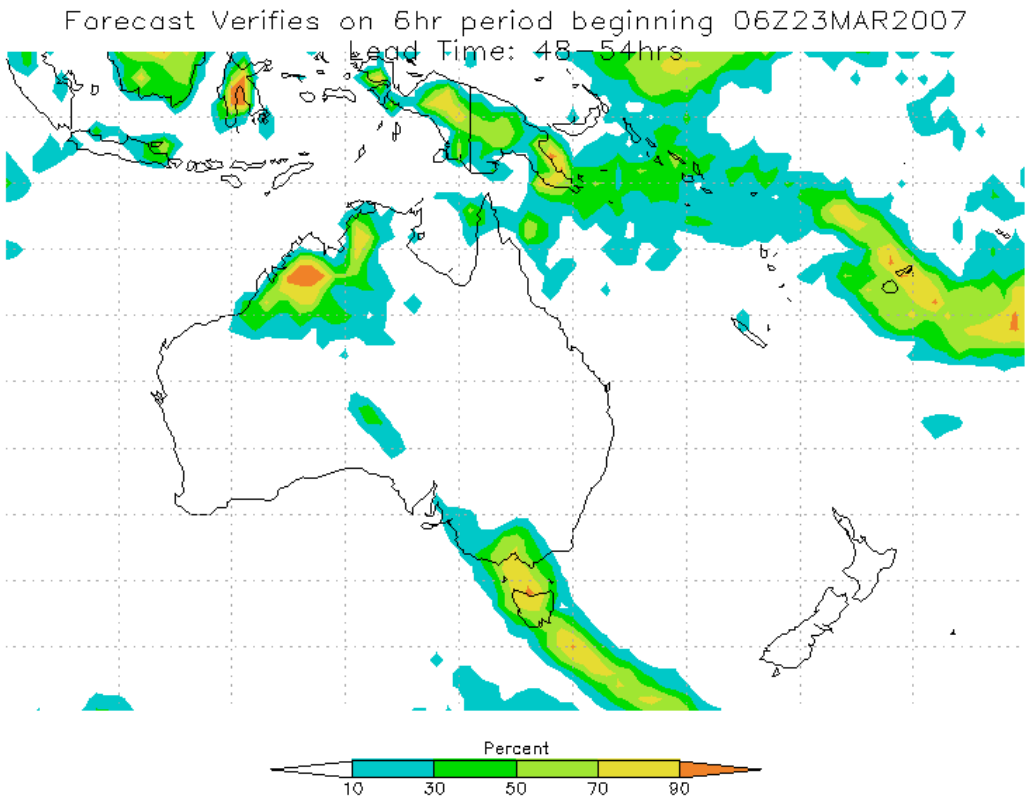


Figure A.3: Example plot of a probabilistic forecast generated by the ensemble members. This forcast represents probabilities of precipitation equaling or exceeding 0.25in in 6 hours.

average precipitation, but accumulations are summed throughout the forecast period. One plot, for example, will show ensemble average accumulation for the first 6 hours in the forecast period. The next slide shows the accumulation through the first 12 hours, and the next slide shows the first 18 hours. This information is useful for river stage forecasting, flood preparations, and the agricultural industry. An example of this product is shown in Figure A.5.

Each product on the Web page is displayed as an animated slide show for forecasts ranging from 0hrs to 60hrs in the future. The updating of graphics is handled by a Perl script, which checks for data files newer than the information on the Web page which may have been downloaded. If a newer forecast file is detected, the Perl script writes a series of instructions which produce GrADS scripts to create the images. GrADS (see: http://www.iges.org/grads) is then run automatically in batch mode, and the plots produced

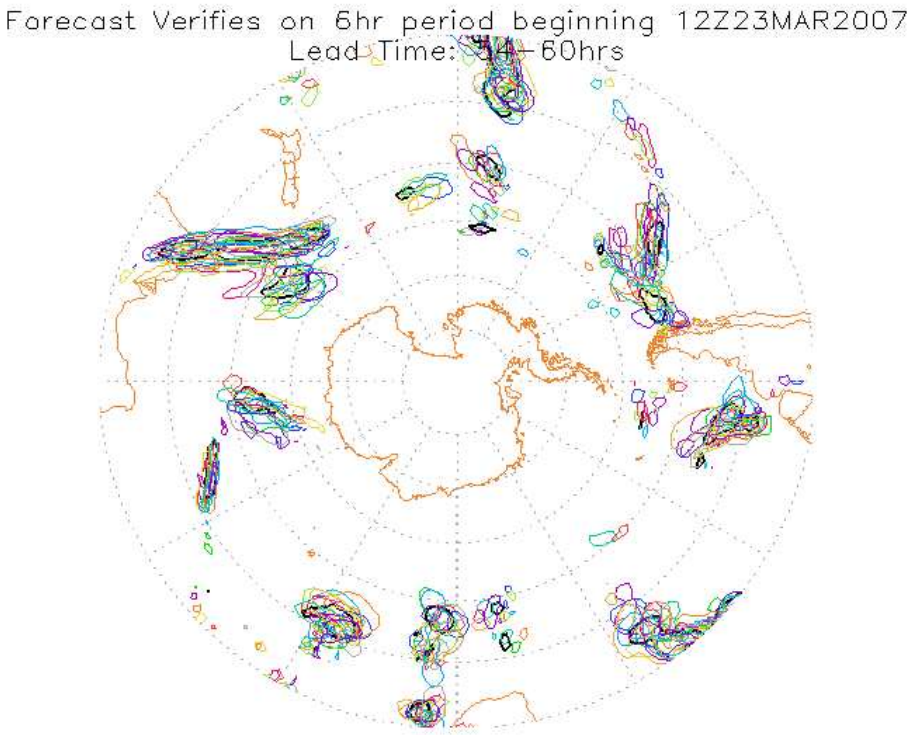


Figure A.4: Example plot of an ensemble spaghetti diagram for 6hr accumulated precipitation.

are output in GIF format. One challenge encountered during this process involves the display of ensemble grids in GRIB format. Since GRIB is becoming the standard format for meteorological datasets, and GrADS is among the most widely used utilities for displaying gridded data, this challenge will be briefly discussed here. GrADS requires a "control" file, which is an ASCII file containing information about the dimensions of the grids corresponding to the data file. Five dimensions are possible, including latitude, longitude, vertical pressure levels (or height), time, and variable. GrADS is capable of working directly with GRIB data, given an additional index file which provides translations for GRIB data onto the dimensions specified by the control file. The necessary index files can be created using freely available software from NCEP, such as "grib2ctl." When working with ensemble data in GRIB format, however, a proper index file cannot be easily developed, since GrADS expects only one grid of each variable from the GRIB file, and ensemble datasets have multiple

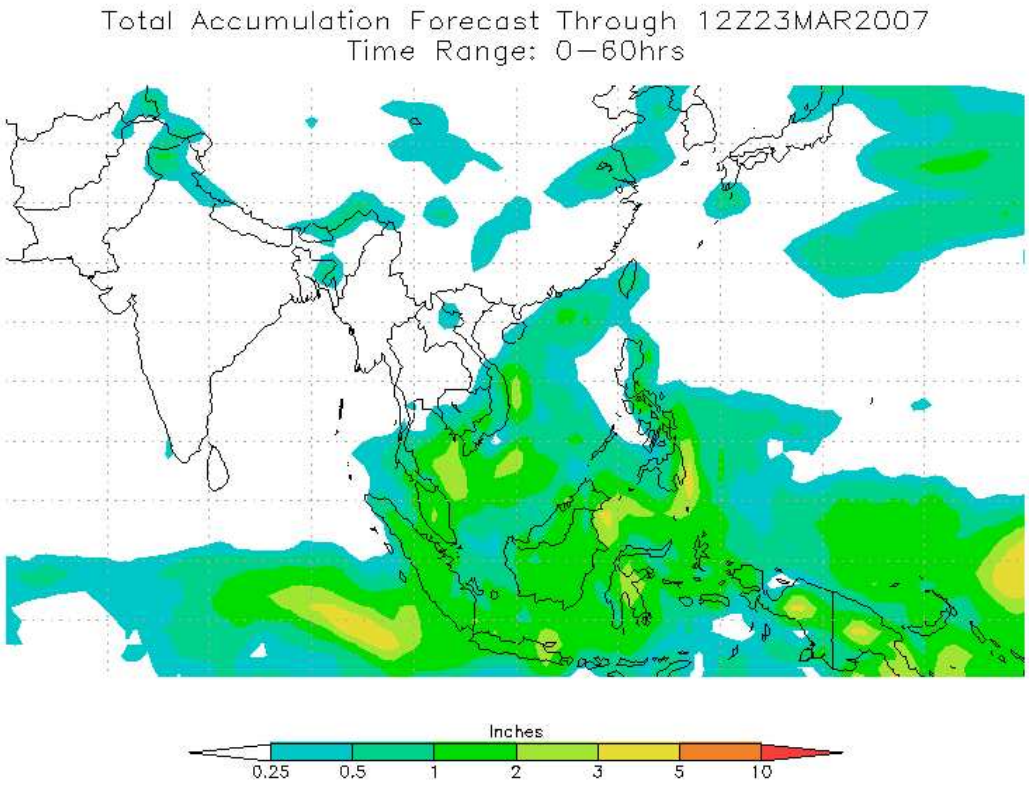


Figure A.5: Example plot of an ensemble consensus total accumulation precipitation forecast available on the research Web page. Accumulations are summed from the beginning of the forecast period.

grids of the same variable. When given an ensemble GRIB file, GrADS will only be able to plot the first ensemble member's data, ignoring the remaining ensemble members. The solution employed for producing ensemble plots on the Web site was to convert the GRIB file to an unformatted binary data file, which approximately triples the file size. Each data point in an unformatted binary file is stored in a 32-bit word, and there is no information regarding variable or dimension. This way, the GrADS control file can provide the necessary dimensions, including specifying multiple ensemble members as different variables. After the graphics are produced for the Web site, the large binary file is discarded.

REFERENCES

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78**, 1–3.

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

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

Hamill, T. M. and S. J. Colucci, 1998: Evaluation of eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126**, 711–724.

Hamill, T. M., S. L. Mullen, C. Snyder, Z. Toth, and D. P. Baumhefner, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.,* **81**, 2653–2664.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129**, 550–560.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87**, 33–46.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie, 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeor.,* **5**, 487–503.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20**, 130–141.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus,* **17**, 321–333.

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.,* **26**, 636–646.

Murphy, A. H., 1973: Hedging and skill scores for probability forecasts. *J. Appl. Meteor.,* **12**, 215–223.

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

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., 1989: A new look at weather forecasting through analogues. *Mon. Wea. Rev.,* **117**, 2230–2247.

Wallace, J. M., 1975: Diurnal variations in precipitation and thunderstorm frequency over the conterminous united states. *Mon. Wea. Rev.,* **103**, 406–419.

Zawadzki, I. I., 1973: Statistical properties of precipitation patterns. *J. Appl. Meteor.,* **12**, 459–472.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.,* **83**, 73–83.

BIOGRAPHICAL SKETCH

Adam D. Allgood

Adam D. Allgood was born on 6 February 1982, in Palm Beach Gardens, Florida. In April 2004, he completed a Bachelor of Science degree *cum laude* in Meteorology at the Florida State University, with minors in Mathematics, Physics and Music. He is currently completing a Master of Science degree in Meteorology at the Florida State University under the advisement of Professor Jon E. Ahlquist.

Adam's research interests include ensemble forecasting, post-processing numerical model output, and the development of meteorological software and graphics products. He has experience with software development in several computer programming languages, including Fortran, Perl, C++, and Java.

Adam lives in Tallahassee, FL, with his wife and three children.