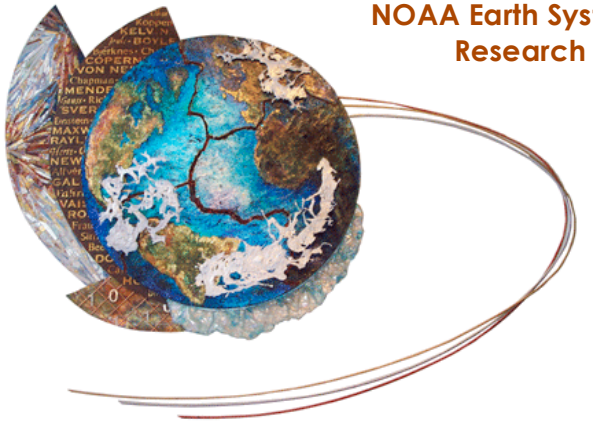

NOAA Earth System
Research Lab

WHITE PAPER

PRODUCING HIGH-SKILL PROBABILISTIC FORECASTS USING REFORECASTS: IMPLEMENTING THE NATIONAL RESEARCH COUNCIL VISION

Thomas M. Hamill, Jeffrey S. Whitaker, and Randall M. Dole

NOAA Earth System Research Lab
Physical Sciences Division
Boulder, Colorado
e-mail: tom.hamill@noaa.gov, (303) 497-3060

DRAFT
15 August 2006

TABLE OF CONTENTS

Executive Summary

The National Research Council recently prepared a report strongly recommending that NOAA redouble their efforts to produce and disseminate high-quality probabilistic forecasts. They stated, "*Uncertainty is a fundamental characteristic of weather, seasonal climate, and hydrological prediction, and no forecast is complete without a description of its uncertainty*."

Currently, most probabilistic forecast information is based on ensemble forecasts, a suite of model simulations integrated forward in time from different initial conditions, and sometimes utilizing different models. Unfortunately, the probabilistic forecasts directly estimated from these ensembles are neither as skillful nor as reliable as they could be. Hence, probabilistic information provided to forecasters should be *calibrated*, corrected using the errors that have been determined from prior forecasts and observations.

For many of the forecast problems that NOAA cares most about such as precipitation forecasts, an effective calibration is much more difficult without a long time series of past forecasts from the same model that is being run operationally. Accordingly, the NRC recommended that "*NOAA should include reforecast data sets to facilitate post-processing.*" These reforecast data sets are old forecasts run from the same model that is used operationally. Recent work at NOAA/OAR/ESRL has demonstrated that calibration based on reforecasts can improve probabilistic forecast skill by an amount equivalent to a decade of numerical weather prediction development.

We propose that many of the new probabilistic forecast products proposed by the NRC should be produced using ensemble forecasts accompanied by reforecasts for calibration. However, NOAA is not prepared to regularly produce and utilize reforecast-based products. Currently NCEP concentrates on disseminating the highest quality numerical forecast, and the suite of forecast models is updated as rapidly as possible with improved numerical methods and higher-resolution models. Hence we recommend cooperation across NOAA organizations to produce these new probabilistic products. NCEP, as NOAA's center of expertise for numerical weather prediction, will continue produce the real-time numerical ensemble forecasts. Perhaps in collaboration with OAR and/or JCSDA, NCEP will also regularly produce reanalysis and reforecast data sets. OAR, with its track record in developing reforecast-based products, will develop the advanced calibration techniques necessary to calibrate the current model forecast, collaborating as necessary with MDL. And finally, MDL will statistically adjust the forecasts and disseminate the corrected statistical forecasts through the National Digital Gridded Database (NDGD).

The result will be state-of the art, calibrated, skillful probabilistic forecasts. This project will require (1) a modest supplement of funding to support personnel to develop and disseminate the probabilistic forecasts, (2) extra computer resources for the regular, dedicated production and archival of reanalyses and reforecasts, and (3) a willingness for NOAA personnel and managers to work collaboratively across organizations, using the talents unique to each – a "one NOAA" approach.

1. Introduction: National Research Council Recommendations

In July, 2006 the National Research Council delivered to NOAA a report entitled, "*Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*" (National Research Council, 2006). The report noted that

> "*Uncertainty is a fundamental characteristic of weather, seasonal climate, and hydrological prediction, and no forecast is complete without a description of its uncertainty*."[1]

The report noted that forecast users have become conditioned to using deterministic forecast products produced by the NWS and others, such as the temperature forecast from the National Digital Forecast Database, shown in Fig. 1 below. Such products are deceiving, for they imply that the temperature can be determined very specifically. Of course, weather forecasts are in error by an amount that will vary from day to day, so a deterministic forecast is incomplete. Many users could benefit if the NWS provided probabilistic forecasts. For example, a one percent chance of a temperature above a 100-degree threshold may be an acceptable risk to a company pouring a high-temperature sensitive concrete, but a 10 percent chance means an expensive, ruined job one time in ten, which may bankrupt the company over time. A deterministic map does not provide them with the information they need to weigh the relative costs of idling their workers until cooler weather vs. pouring concrete and having it ruined by the high temperatures.

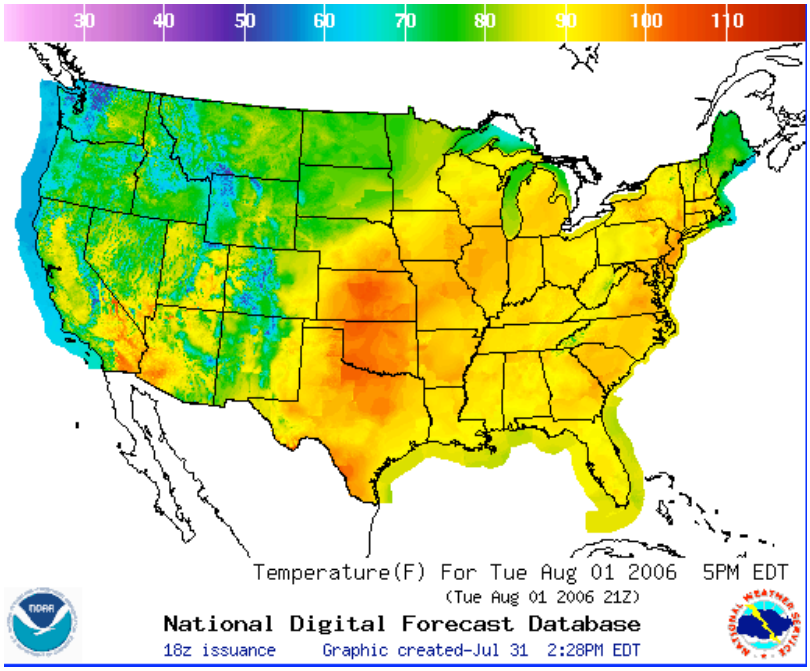

Figure 1: Sample of a 1-day temperature forecast from the NWS's National Digital Forecast Database.

[1] From the opening paragraph of the summary.

Accordingly, the NRC recommended that the weather "Enterprise" (i.e., the collective efforts of public and private forecasters) needed to move toward conveying forecast information probabilistically. The panel noted the difficulty of the task and asked NOAA to take the lead:

> "*Hydrometeorological services in the United States are an Enterprise effort. Therefore, effective incorporation of uncertainty information will require a fundamental and coordinated shift by all sectors of the Enterprise. Furthermore, it will take time and perseverance to successfully make this shift. As the Nation's public weather service, NWS has the responsibility to take a leading role in the transition to widespread, effective incorporation of uncertainty information into hydrometeorological prediction*."[2]

The NRC report produced many other recommendations, but the overall tenor of the report was that NOAA must lead the US effort to produce useful probabilistic forecasts.

This white paper is designed to outline a program to address some of the crucial recommendations in the NRC report. How can NOAA produce and disseminate skillful probabilistic predictions? Below, section 2 will review the current state of the art of probabilistic prediction in the NWS. This process is based on ensemble forecasts, but these forecasts have deficiencies, and probabilistic information is not yet effectively conveyed to weather forecast offices (WFOs) and the public. Section 2 will also review some of the underlying causes for the deficiencies in probabilistic forecasts from ensembles. Section 3 then provides background on recent research in NOAA demonstrating how the major deficiencies in probabilistic forecasts can be addressed through the statistical correction using a database of reforecasts (i.e., hindcasts) and observations. Section 4 sketches out a rough vision of how the major NRC recommendations can be addressed. We envision a forecast process whereby the current ensemble forecast is statistically adjusted using reforecasts. This program will require that NOAA personnel and computer resources be reallocated, and this program envisions a closer cooperation between NOAA/NCEP, NOAA/MDL, and NOAA/OAR, a "one NOAA" approach that must be embraced by scientists and program managers to be successful.

In the subsequent discussion, we will often refer to results from a pilot reforecast data set using a reduced-resolution, 1998 version of the NCEP Global Forecast System (Hamill et al. 2006).

[2] From finding 1 of the summary.

2. Probabilistic forecasting in NOAA, circa 2006: ensemble forecasting and its deficiencies

a. Ensemble forecasting basics

NOAA has, of course, already begun using *ensemble forecasting* as a tool for providing probabilistic information. A collection of different forecasts is generated from different initial conditions and/or different models. The rapid growth of errors due to chaos and model uncertainty will result in a spread of possible forecast states. Ideally, ensemble forecasts may provide an early warning of possible severe weather. A single integration of the forecast model may not indicate that a storm will develop. With an ensemble, perhaps a very small change in the initial condition will be enough to trigger the development of a storm. An ensemble forecast thus may provide value-added information, providing the forecaster with a heads-up of a possible unusual event that was not available when only one forecast was produced. A particularly successful example of this is provided for the December 1999 "Lothar" storm in Europe using the ECMWF ensemble (Fig. 2, from Palmer 2006).

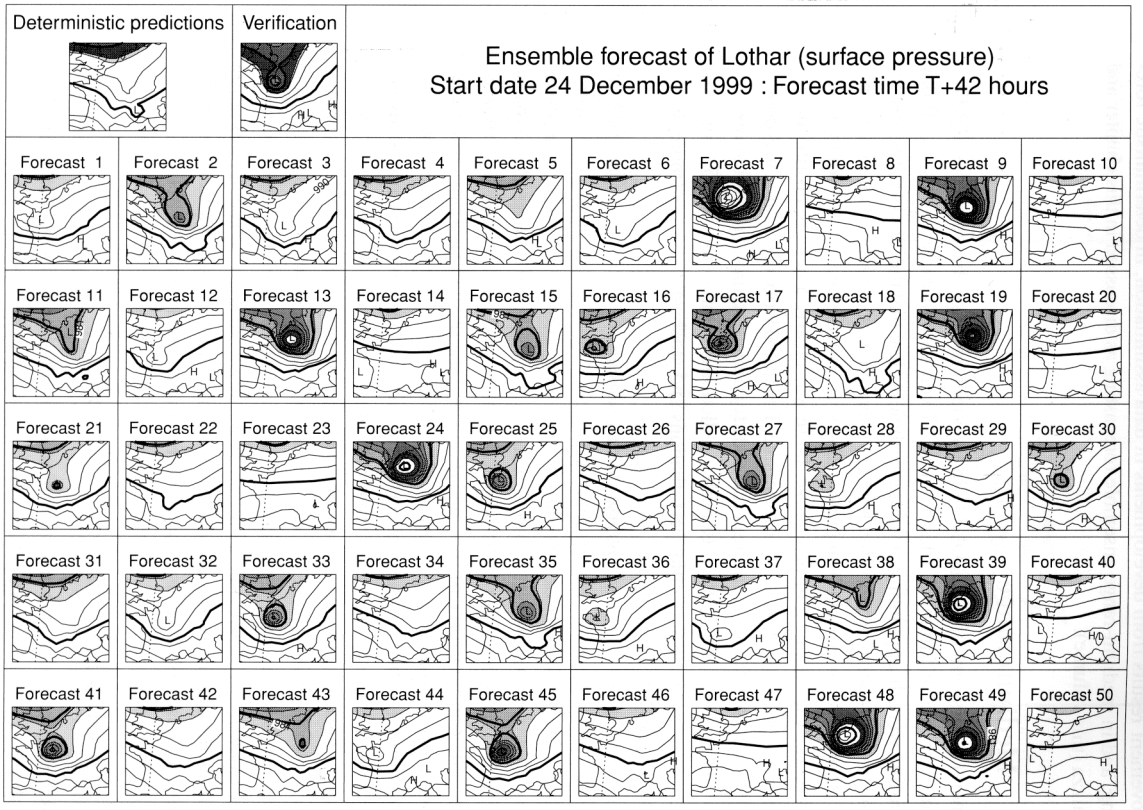

Figure 2: An example of deterministic and ensemble forecasts for a deadly storm in Europe, December 1999 (the map is centered on France, with southern England in the upper left). The deterministic forecast of sea-level pressure (contours, with low values shaded) in the upper left produced no storm over Europe, when in fact a very strong storm developed over northern France. When a 50-member ensemble was run, many of the members produced intense storms, warning forecasters of the possibility of a severe storm.

More quantitatively, we design the ensemble forecast system with the intent that the probability of an event, severe or otherwise, can be determined directly from the ensemble. If we wished to estimate the probability of more than 1 cm of rainfall tomorrow in Washington, D.C., ideally we would count up the number of members forecasting more than 1 cm (say, 3 members) and then divide this by the total number of members (say, 15 members), producing a 20 percent chance of 1 cm or greater rainfall. This forecast should be as *sharp* as possible while still remaining *reliable*. A sharp forecast is one that will frequently deviate from the long-term climatological probabilities, issuing mostly definitive yes (probability=1.0) or no (probability = 0.0) forecasts rather than waffling. A reliable forecast issues probabilities that match the long-term event frequency, so that, for example, over all the times when a 10 percent chance of rain was issued, it actually rained 10 percent of the time. The more sharp a forecast is while remaining reliable, the more skill it possesses.

Unfortunately, while ensemble forecasts are computed at NCEP, there are two major problems:

(1) The regular dissemination of a wide suite of probabilistic products is not yet part of the NOAA concept of operations. Ideally, such products would be distributed through the National Digital Guidance Database (NDGD), and worded National Weather Service zone forecasts would convey forecast information probabilistically.

(2) Even if the ensemble-based probabilities were disseminated through the NDGD, the current forecasts are neither particularly skillful nor reliable.

b. The root causes of deficiencies in ensemble forecasts

Ensemble forecast deficiencies have at least four basic causes. The first cause can be traced to deficiencies in the forecast model(s) that are used to generate the member forecasts. An ensemble forecast system inherits the deficiencies of the model(s) used to integrate each member. For example, if the model consistently produces a forecast that is too cool or too dry, one can expect that an ensemble forecast also will be biased toward high probability of cool and dry events. As the forecast model is improved, the ensemble forecasts should become more skillful.

A second, more subtle problem is that ensemble forecast systems typically produce forecasts with too little spread (spread is defined as the standard deviation of the ensemble members about their mean), and the spread of the ensemble is not indicative of the expected error in the forecast, as it should be (Whitaker and Loughe 1998). Even after a sophisticated correction of bias, too often the observed weather lies outside of the span of the ensemble, leading to unreliability of the forecasts (Fig. 3). This spread deficiency could be due to many problems with the ensemble forecast system. The initial perturbations may not realistically sample the distribution of plausible initial states, as they should. The growth of differences among forecasts may be constrained by the use of lower-resolution models; the larger the grid spacing, the less errors can grow at the

smallest scales and then interact with the larger scales. And the forecast model may be coded to assume that certain sub-gridscale processes such as convection and turbulent mixing operate deterministically, when in fact they operate more stochastically, thereby limiting the growth of spread (Palmer 2001).

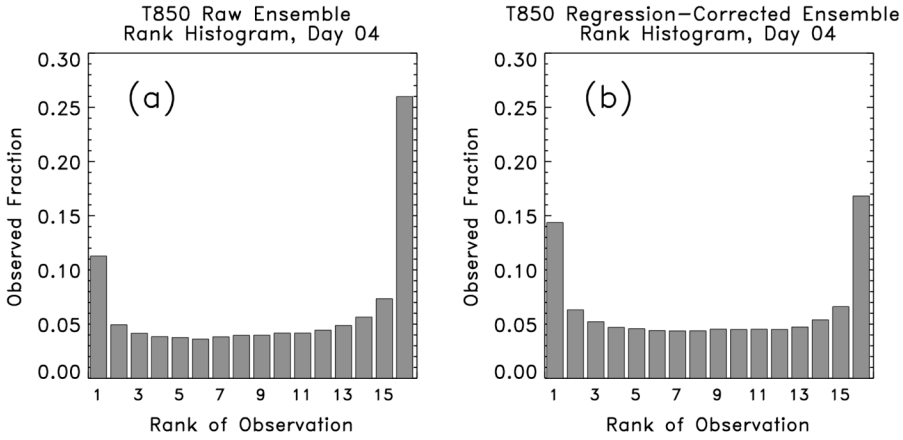

Figure 3: Rank histograms of Day-4 850 hPa temperature ensemble forecasts (a) before, and (b) after a regression-based bias correction to the ensemble mean. The rank histogram tallies the rank of the observed data relative to a sorted ensemble over many locations and cases. The high population in the uppermost bin of panel (a) indicates that the observation was often warmer than all the ensemble members, indicating a cold bias in the forecasts (a flat rank histogram is desirable). Panel (b) shows that a bias correction evens out the percentage of too-cold and too-warm cases, but still too often the forecasts lie outside the range of the ensemble, indicating a general lack of spread. The forecast model is a circa 1998 version of the NCEP Global Forecast System (Hamill et al. 2006), but the general properties shown here are fairly common among ensemble forecast systems. For more information on the interpretation of rank histograms, see Hamill (2001).

A third source of error in probabilistic forecasts is caused by the limited size of the ensemble. Ideally, the ensembles are randomly sampling from the distribution of possible weather states, and all things being equal, the probabilities will be estimated much more accurately with many members than with few members.[3] A final problem is that the ensemble forecast system is producing a gridded, area-average representation of the weather rather, and in many situations the user requires a forecast for a specific location.

c. Approaches to deal with deficiencies of ensemble forecast systems.

There are three general approaches one might take to address these problems, and these approaches are not mutually exclusive. They are: (1) develop more accurate models. (2) Rely upon our operational forecasters to adjust the forecasts before dissemination to the public, and (3) develop an objective system for correcting the errors before they are disseminated.

[3] In actuality, the tradeoff is more complex, for given a fixed amount of computer resources, running a larger ensemble will require computing that ensemble at a coarser resolution, with more concomitant error in each member.

If we rely upon the first option, disseminating information directly from the ensemble and working to improve that ensemble, the history of numerical weather prediction development suggests that we will have to wait a very long time for dramatic improvements. The biases in the many of the forecast parameters we care about most, such as surface temperature, wind speed, and precipitation, are all dependent on accurately modeling some of the most minute aspects of the weather prediction system that we still know very little about, such as the way the size distribution of cloud droplets or the way that wind will interact with a variety of vegetation types. Hence, while model deficiencies are reduced with each passing year, the improvements are incremental, and they presuppose a continued investment in new observational platforms, basic science, larger computers, and a dedicated staff of talented scientists to improve the computer models. Extrapolating from past experience, it is unrealistic to expect that model deficiencies will be eliminated in 10 or even 50 years; the remaining scientific challenges are too great.

Can we do something practical to increase skill while we work toward a bias-free model? There are many independent efforts going on worldwide to produce weather forecasts, and one approach that has been suggested is to collaborate, to combine the forecasts from multiple centers. The hope is that each model will have a distinctly different model bias, and hence the model uncertainty is accounted for and forecast biases are reduced through averaging. With this in mind, the US and Canada have entered into a partnership called "NAEFS," the North American Ensemble Forecast System[4]. It is expected that the United Kingdom Met Office and the US Navy will add their forecasts in the next few years. This approach is exciting, but it is still somewhat untested.[5] Will the forecast models truly have independent biases, or since numerical techniques are shared internationally, will the forecasts still tend to resemble each other? Multi-model systems will continue to be developed, but *it is risky for NOAA to assume that a multi-model system alone will provide the high-quality probabilistic forecasts that are desired*.

The second possible approach is to rely on our human forecasters to correct for errors in any probabilistic information that is to be disseminated publicly. There are several disadvantages of this approach. First, typically the forecaster is very busy, and modifying a probabilistic forecast is a substantially more complicated task than modifying a deterministic forecast. Second, it's not clear that many forecasters will have the necessary information and experience to correct errors in the probabilistic forecasts. In the NDFD we already have problems with the subjective modification of gridded forecasts (note, for example, the discontinuities of South Dakota temperature forecasts in Fig. 1, produced by different forecast offices making different subjective forecasts). Modifying probabilistic forecasts presupposes that the forecaster will be monitoring not only forecast bias, but also other problems like spread deficiencies. This is an unrealistic assumption.

---

[4] Presentation from the most recent NAEFS workshop can be viewed at http://www.cmc.ec.gc.ca/~cmcdev/naefs/, (username "naefs", password "cmc").

[5] See also THORPEX research on a related project called TIGGE, http://www.wmo.ch/thorpex/pdf/tigge_summary.pdf

The third approach is to produce an objective, computer-based *calibration* of the probabilistic forecasts, adjusting the probabilities based on the discrepancies between time series of past forecasts and observations. This calibration would improve the probabilistic forecast skill and reliability. Ideally, only a short record of past forecasts would be needed, for current and past forecasts should come from the same model, and that model may be updated several times a year. However, practical experience has shown that for some of the tougher forecast problems such as long-lead forecasts or forecasts of precipitation, a short set of forecasts is simply not adequate to achieve an effective calibration (we shall return to demonstrate this in section 3). Hence, many years or decades of prior forecasts would be useful, forecasts from the exact same model that is run operationally. We call these "reforecasts," noting the similarity with the concept of reanalysis (Kalnay et al. 1996). With a large reforecast database, there is a greater likelihood of finding past forecast events similar to today's forecast event, even if that event is relatively rare, a necessary pre-requisite to effective calibration. A series of articles have shown that reforecasts are highly beneficial for improving the skill of medium range probabilistic forecasts (Hamill et al. 2004, Whitaker et al. 2006) and short-range probabilistic precipitation forecasts (Hamill et al. 2006, Hamill and Whitaker 2006). The NRC was convinced of the importance of reforecast data sets and issued some specific recommendations to NOAA advocating their computation. For example, they state that

> "*an easily accessible observation and forecast archive is a crucial part of all post-processing or verification of forecasts*"

and they recommend that

> "*NOAA should include reforecast data sets to facilitate post-processing.*"[6]

The disadvantage of the reforecast-based approach is that it is computationally expensive to produce, and once the data set is produced, there is less impetus to change the forecast model given the expense of recomputing the reforecasts. Also, to fully get the benefit of reforecasts, companion reanalysis data sets must also be produced. These tasks could slow down the numerical weather prediction development process. Nonetheless, the initial tests with reforecasts have demonstrated that dramatic improvement in the skill of forecasts are possible, skill improvements equivalent to ~10 years of model development and computer upgrades (Hamill et al. 2004, 2006). Also, there may be ways to minimize the impact on the NWP development process, which we shall return to in section 3d.

The next section outlines the rationale for reforecasting.

[6] Recommendation 3.4 of the NRC report, from section 3.1.4.

3. **A rationale for institutionalizing reforecasts in NOAA**

a. Problems with short training data sets and compositing over different locations

A simple example, shown below in Fig. 4, demonstrates why short training data sets are often inadequate and a large reforecast data set is needed. Suppose we have a month and a half of prior forecasts and observations for a location in the desert of the western US. Today's ensemble forecasts a significant rain event, while no similar rain event was forecast during the training period. How then can past forecasts provide any pertinent information for correcting possible errors in the current forecast?

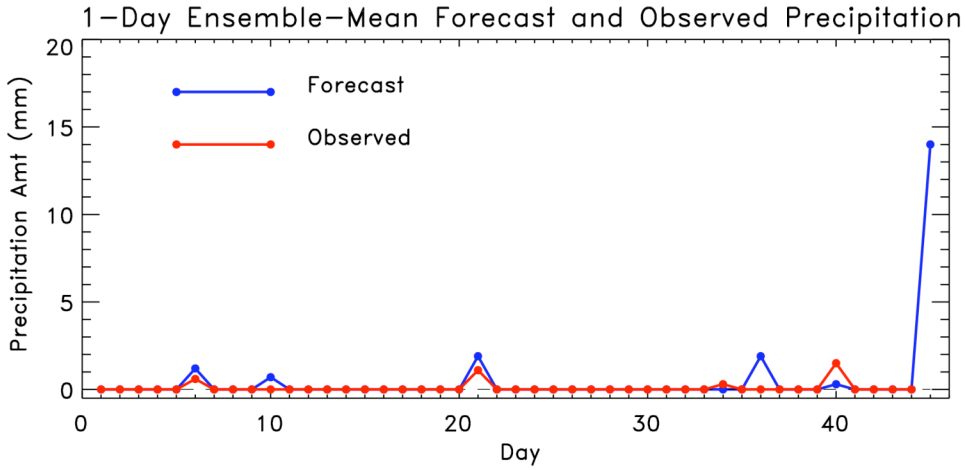

Figure 4. A time series of forecasts and observations for a dry location. The forecast-observation differences for past month and a half are probably useless for determining a bias correction for the latest forecast, which is far wetter than any of the prior forecasts.

Many proposed calibration methods attempt to enlarge the sample size by compositing forecast/observed samples over different locations. However, such methods have not been demonstrated to be highly successful in correcting bias and improving forecast skill. One reason is that nearby locations, especially those in mountainous regions, may have very different errors. Two locations on opposite sides of a mountain range may have different errors due to the forecast model's simplified representation of wind flow around and over that range. An example of this is provided in Fig. 5. A technique from NCEP uses the difference between cumulative distribution functions (CDFs) of the forecast and observed to make bias corrections (for a critical review of this technique, see Hamill and Whitaker 2006). Panel (a) shows the CDFs for a location along the California coast, just north of San Francisco. The CDF indicates the collective probability that the precipitation will be less than or equal to the specified amount. The 90th percentile of the observed CDF is approximately 4 mm, while the 90th percentile of the forecast CDF is approximately 7 mm. This suggests that at this threshold the forecast may have a moist bias that should be corrected to 4 mm. The NCEP technique re-maps members' forecast amount based on the differences in CDFs, and the bias correction is shown in panel (c). The CDF for Sacramento are shown in panel (b), and its bias correction is also shown in panel (c). Note that two locations that are separated by a relatively small distance have very different implied bias corrections, illustrating the

potential downside of a technique that attempts to make a generalized bias correction using data across many locations.

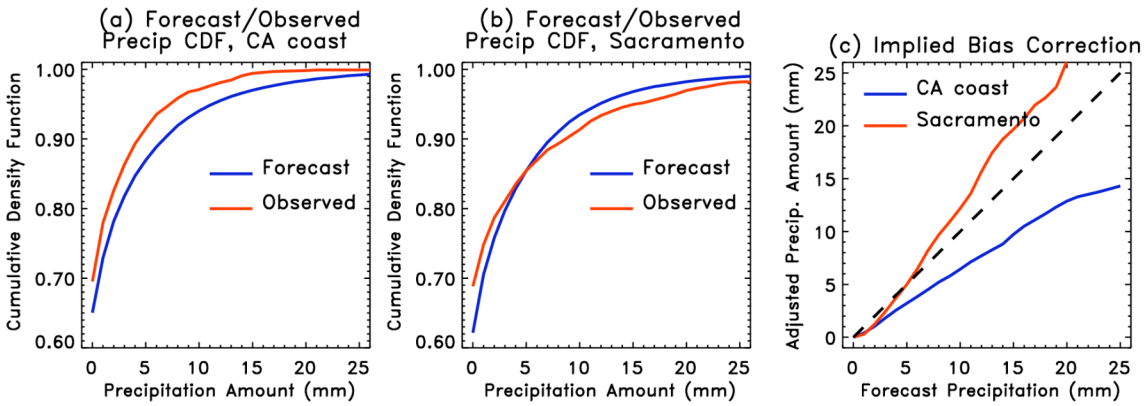

Figure 5: Illustration of a bias correction technique used at NCEP and designed to work with small training data sets aggregating forecasts/observations at different locations. Panels (a) and (b) provide the cumulative density function (CDF) of 1-day forecasts of precipitation for 1 January (CDFs determined from reforecast data and observations in Dec-Jan). Panel (a) is for a location on the CA coast, just north of San Francisco, and panel (b) is for Sacramento, CA. Panel (c) provides the implied function for a bias correction from the forecast amount to a presumed observed amount. Note the very different corrections implied at two nearby locations.

Sample size could be increased through the use of more than the past month's forecast and observations. However, if some of the older forecasts were computed with a different model version, that model may have different forecast-error characteristics than is used for the current forecast. That is, the current forecast would then be partly correcting for the bias in a previous model version, limiting its accuracy.

Synthesizing these general results, we conclude that to extent that model errors are consistent from one day to the next or consistent across many diverse locations, they can be corrected with relatively short training data sets. If they vary quite a bit from one day to the next and one location to the next, they will require a larger sample of past forecasts to find enough similar forecast events in the past to attempt a calibration.

b. When larger training data sets are helpful: Example 1, week-2 forecasts.

We can see the particularly beneficial effect of reforecasts when calibrating week-2 forecasts. Here, we found that model errors were so large that corrections a bias correction of the raw ensemble using a short training data set was simply not able to produce a forecast with appreciable positive skill (Fig. 6). A bias correction with a longer (22-year cross validated) training data set improved the skill, but the forecasts remained somewhat unreliable. Only when a full calibration of bias and spread deficiencies was applied using the reforecasts did they become both appreciably skillful and reliable.

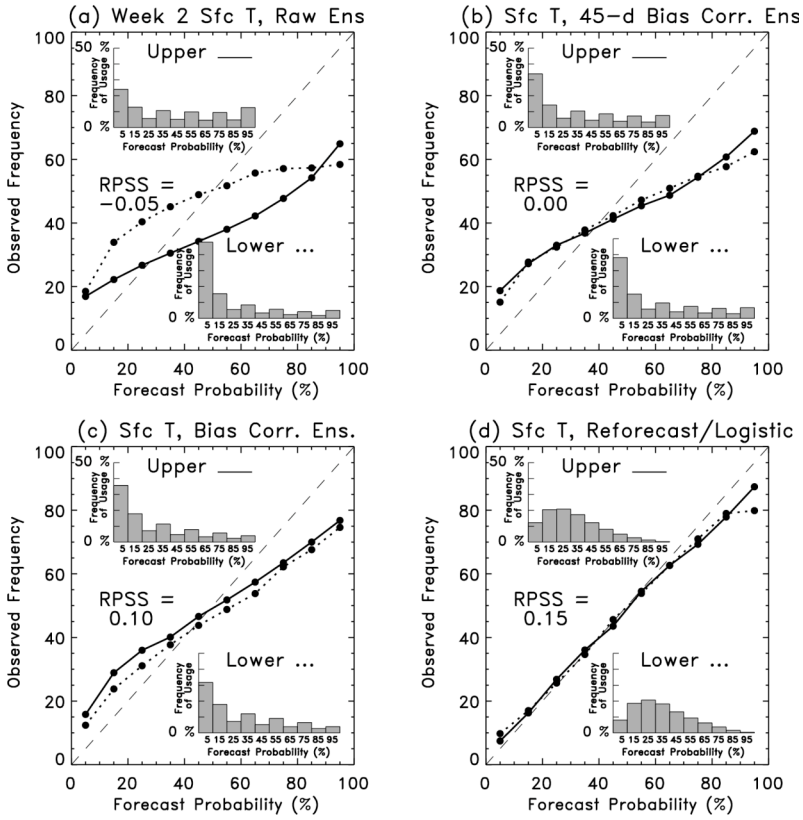

Figure 6. Reliability diagrams for week-2 forecasts. In each panel the lines plot the observed event frequency as a function of the forecast probability. Lines along the 45-diagonal are preferred (perfect reliability). The dotted line is a forecast for below-normal temperature probabilities (the "lower" tercile), and the solid line is for above-normal temperature probabilities (the "upper" tercile). The inset histograms provide information on how often a given forecast probability is issued, a measure of forecast sharpness. The overall skill is summarized by the reported ranked probability skill score (RPSS), where 1.0 is the skill of a perfect probabilistic forecast and 0.0 is the skill of a climatological forecast. (a) Reliability diagram of probability forecasts derived from the raw ensemble relative frequency. (b) Diagram of probability forecasts derived from the ensemble relative frequency after a bias correction based on the previous 45 days of observations and forecasts. (c) Diagram of probability forecasts derived from the ensemble relative frequency after a bias correction based on the previous 22 years of observations and forecasts. (d) Diagram of probability forecasts based on a logistic regression using 22 years of training data from the reforecast data set, which produces a forecast with both a bias and spread correction.

c. When larger training data sets are helpful: Example 2, short-range PQPFs.

Long training data sets are also extremely valuable for the calibration of short-term probabilistic quantitative precipitation forecasts (PQPFs). Here, the imperfections in the probabilistic forecast are likely to be very dependent on the specific synoptic situation that day, which may be quite different from the synoptic situation the day before.

NOAA/OAR/Earth System Research Lab's (ESRL's) recent work has demonstrated that it is possible to statistically downscale a coarse-resolution forecast using the reforecasts, producing skillful, very high-resolution predictions with detail that matches the local climatology. These calibrated forecasts are improved both in skill and

reliability. Figures 7 and 8 demonstrate how much improvement is possible. Figure 7 shows the probabilistic forecast skill from raw ensemble forecasts, while Fig. 8 shows the skill after correction using an analog-based technique. The raw forecasts have very little forecast skill relative to the skill of a climatological forecast, and often the forecasts are less skillful than climatology, especially in the warm season. The corrected forecasts are very skillful (Fig. 8) and very reliable (not shown, but see Hamill and Whitaker 2006).

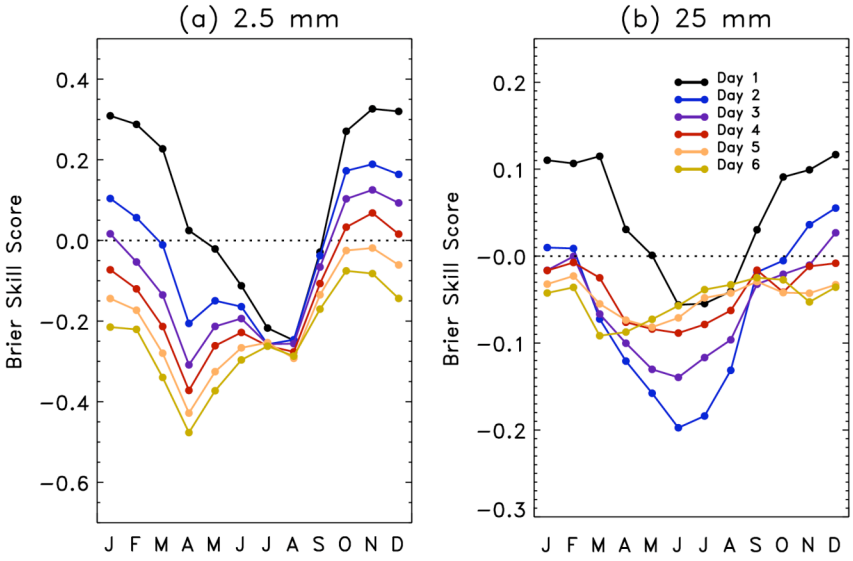

Figure 7: Brier Skill Score (Wilks 2006) of precipitation forecasts generated directly from ensemble forecast output. (a) Skill of 2.5 mm forecasts. (b) Skill of 25 mm forecasts. Note the dashed line denoting zero skill. Forecasts are verified with 32-km North American Regional Reanalysis precipitation analyses (Mesinger 2006).

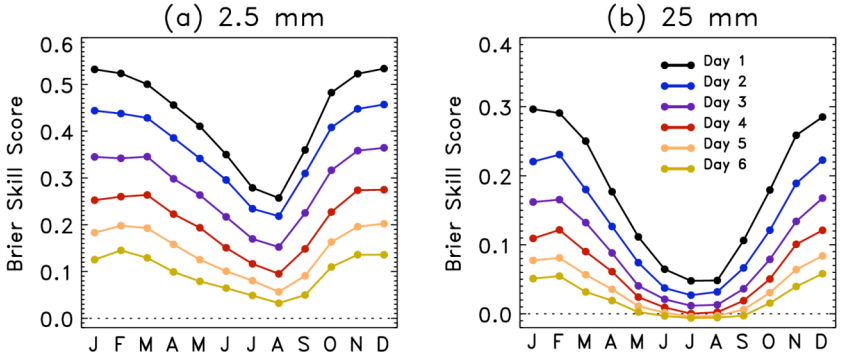

Figure 8: As in Fig. 7, but for calibrated precipitation forecasts using a smoothed rank analog technique described in Hamill and Whitaker (2006).

Recently, ESRL started producing an experimental probabilistic quantitative precipitation forecast product based on the reforecasts, and downscaled to ~5 km grid spacing using a PRISM climatology (Daly 2002). These forecasts are available in real time at www.cdc.noaa.gov/reforecast/narr, and they are being used and evaluated at the NCEP Hydrometeorological Prediction Center. An example of these forecasts and observed data is shown in Figs. 9-10. Notice that while the forecast model is relatively

low resolution, it is possible to downscale the actual forecasts to provide very high-resolution forecasts; the key tool was a high-resolution data set of analyzed precipitation.

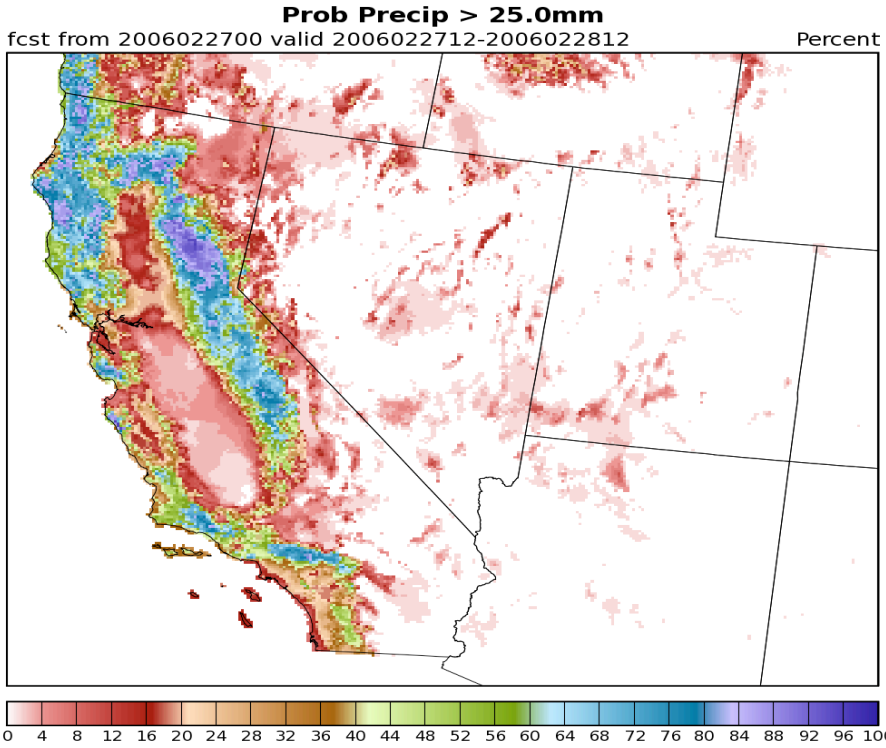

Figure 9: 12-36 h forecast of the probability of greater than 25 mm precipitation for the period 1200 UTC 27 February 2006 to 1200 UTC 28 February 2006, using a reforecast-based analog technique described in Hamill et al. (2005).

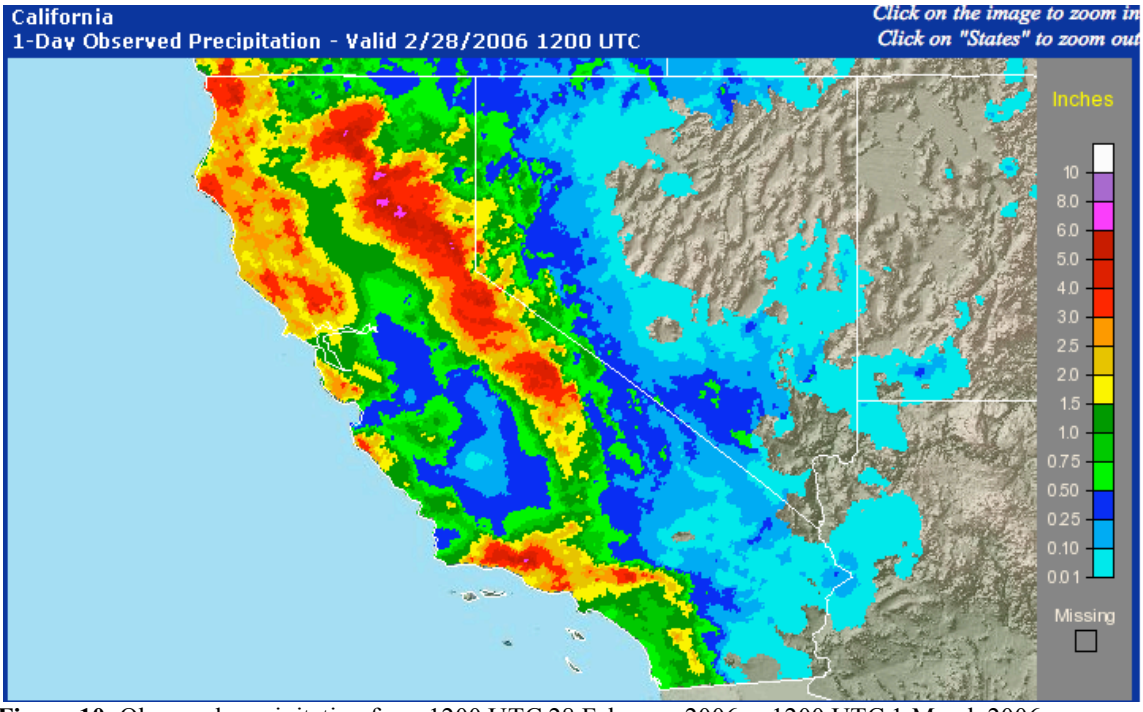

Figure 10: Observed precipitation from 1200 UTC 28 February 2006. – 1200 UTC 1 March 2006.

d. Questions about reforecasts.

The idea of reforecasting being central to the production of NWP products is a relatively new idea. Below, we attempt to answer some of the possible questions the skeptical reader may have about reforecasts.

(1) Will reforecast data sets still be as useful as our models are improved?

Certainly, calibration of a more accurate model will then provide less of a beneficial effect, for there will be less error to correct. But how much less benefit? Few other organizations have produced extensive reforecasts, so it is difficult to evaluate this. However, ECMWF recently produced a small reforecast data set with a 2004 version of their forecast model, with more than double the resolution of the 1998 NOAA GFS reforecast data set discussed above. Given ECMWF's substantial lead in forecast skill, this model represents about 10 or more years of improvement in numerical weather prediction technology. ECMWF's forecasts spanned only 12 years of dates in January-February-March. Only 5-member ensembles were computed, and only one forecast ensemble was produced every two weeks, for a total of 84 forecast/observed samples at each location. GFS reforecast data was sub-sampled to the same ensemble size and limited set of dates. Figure 11 below shows the reliability and skill of forecasts from the raw ensemble, and Fig. 12 provides the same after calibration. The improvement from calibration of ECMWF's week-2 forecasts (11.2 percent) was nearly as large as it was with the GFS model (14 percent). For more details, please see Whitaker et al. (2006).

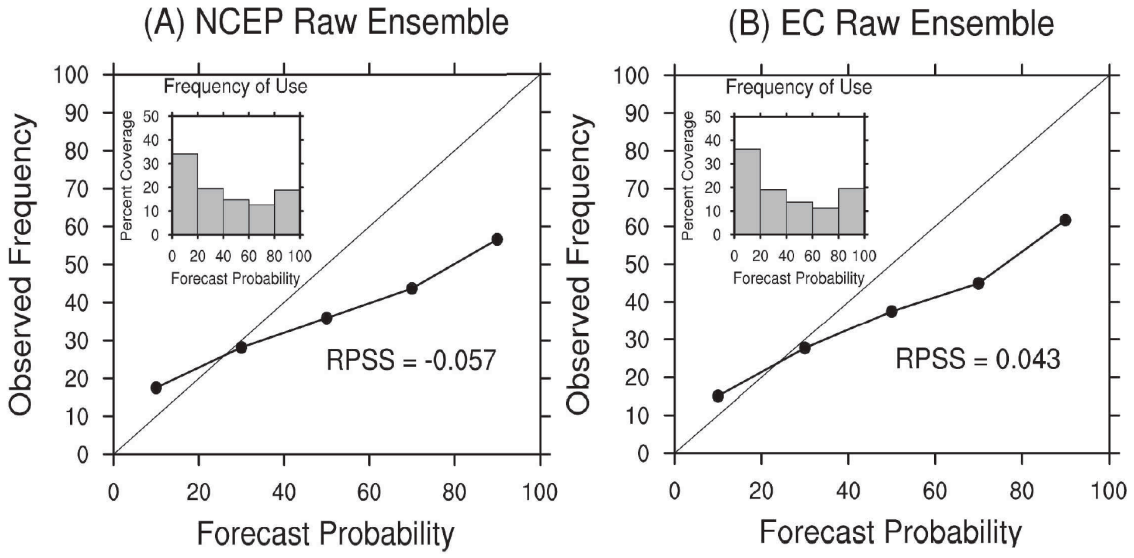

Figure 11: Reliability diagram for probabilistic forecasts of week forecasts of above and below normal temperatures (here, data from upper and lower tercile forecasts are combined). Data taken from 12 years of January(a) NCEP raw 5-member ensemble, and (b) ECMWF 5-member raw ensemble.

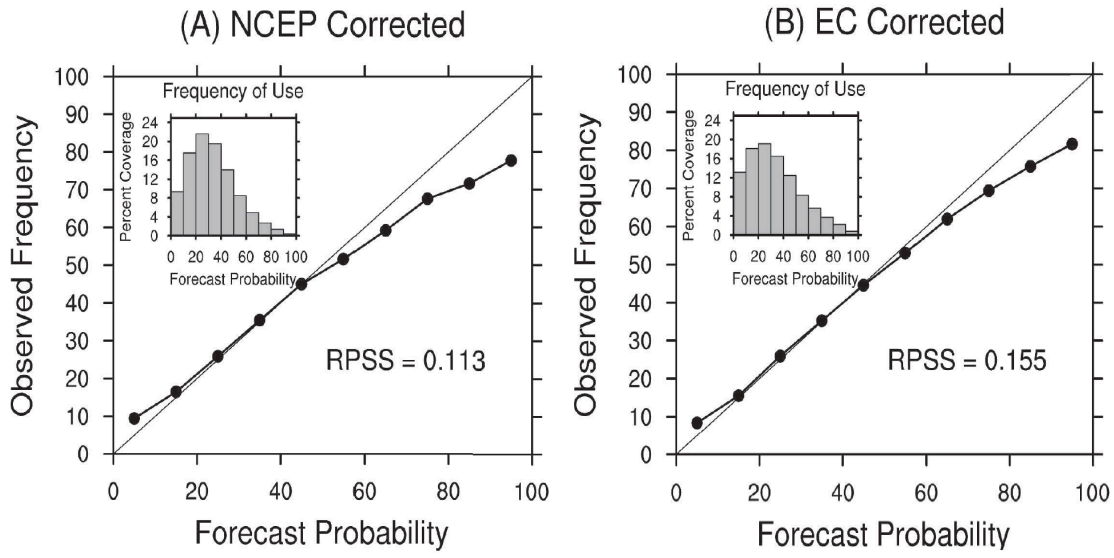

Figure 12. As in Figure 11, but after forecasts were calibrated with a logistic regression technique trained on reforecast data.

(2) Can a shorter reforecast data set be crafted to effectively substitute for a longer one?

Figures 11-12 may lead the reader to question whether long reforecast data sets are really necessary, for a somewhat effective calibration of the forecasts was there shown to be possible with a much smaller reforecast data set, with only 84 forecast/observed samples at each location. Unfortunately, had the 84 samples been drawn from the previous 84 days of forecasts, the calibration would not have been nearly as effective. Week-2 forecast errors tend to be highly correlated from one day to the next, so 84 previous days of forecasts may provide an effective sample size of only 20 to 30 forecasts spanning a much smaller range of weather scenarios.

If a limited reforecast data set is all that can be computed, for week-2 calibration the reforecasts could safely skip days between forecast and thereby span more years. The usefulness of such an approach is demonstrated in Fig. 13, taken from Hamill et al. (2004). After producing probabilistic forecasts using 22 years of daily reforecasts (a 23-year reforecast data set with one year subtracted for cross validation), we compared the accuracy that could be obtained if only four years of reforecast data were available. Those four years could be the prior four years with a reforecast computed every day, or they could span 8 years with a reforecast computed every second day, 12 years every third day, and so on. Figure 13 shows that four years of reforecast data produce almost as skillful a calibration as the full 22 years if 5 days are skipped between sample forecasts, for then the samples will be nearly independent of each other and will have sampled the range of possible weather scenarios over a full two decades.

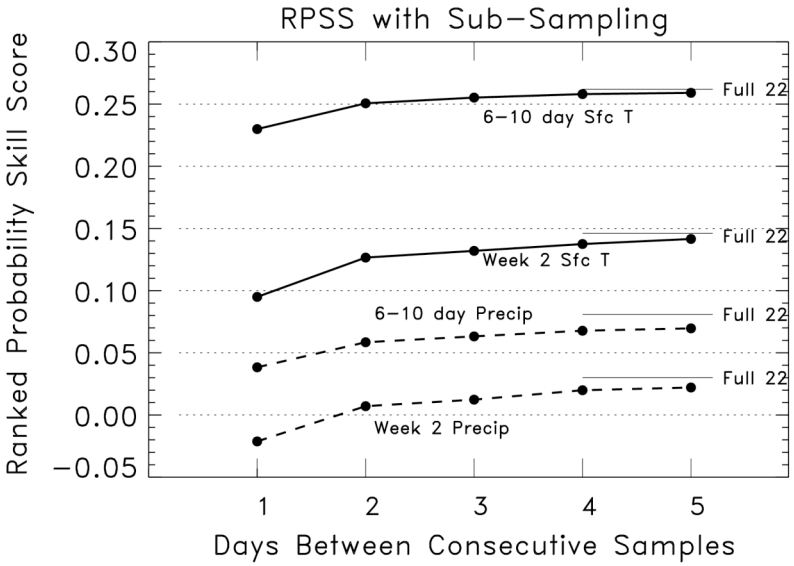

Figure 13: Ranked probability skill score (RPSS) when four years of training data were used, with 1, 2, 3, 4, and 5 days between successive samples in the training data set. The lines labeled "full 22" indicate the skill when the full 22 years of cross-validated training data were used.

Compared with week-2 averages, which will not vary much from one day to the next, precipitation forecasts vary greatly. What is important for calibration is simply having a large number of past forecast events that are similar to the current forecast event. If the current forecast event under consideration is garden-variety weather, then a short training data set is more adequate, as shown in Fig. 14a, taken from Hamill et al. (2006). However, for rarer events, there is a much more substantial benefit of a long training data set. In Fig. 14(b), one can see that for the probability of greater than 25 mm rainfall, a 2-day probabilistic forecast based on a 24-year training data set is as accurate as a 1-day forecast from a 3-year training data set.

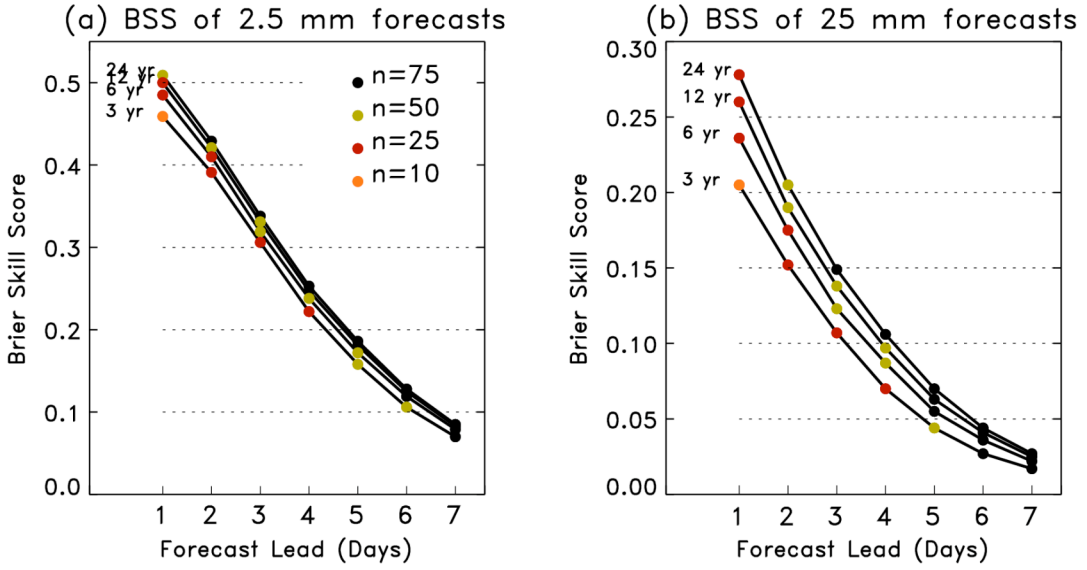

Figure 14. Brier skill scores of the analog reforecast technique for various lengths of the training dataset. Probabilistic forecasts were calculated for analog ensembles of size 10, 25, 50, and 75; the skill of the ensemble size that was most skillful is the only one plotted. The color of the dot denotes the size of the most skillful analog ensemble.

In general, the more rare the event, the more useful a long reforecast data set will be for extreme-event calibration. Since much of the value that people get from weather forecasts comes in situations where the weather is unusually adverse, large reforecast data sets will thus help the most with the calibration of problems that are of the greatest concern.

(3) How many reforecast members should be computed?

Our experience with reforecasts suggests that the most of the benefit of the reforecast data set can be obtained from a relatively small ensemble. The calibration techniques that we have developed use the ensemble-mean state as a predictor and do not rely on the values of individual members. Much of the benefit of improving the ensemble mean is obtained with relatively small ensembles, say 5-10 members; 25 or 50-member reforecast ensembles are not necessary, nor is the operational production of a 50-member ensemble forecast. However, the reforecast data set that we work with uses the breeding method to generate initial perturbations (Toth and Kalnay 1997), and the reforecast model resolution is quite coarse, T62. It is possible that future ensemble reforecast data sets constructed with more advanced perturbation techniques and models and will be able to extract more informational content out of large ensembles.

(4) Are companion reanalyses necessary each time we generate a new reforecast?

Producing new reanalyses along with the reforecasts will be highly beneficial. In Fig. 15, we show the anomaly correlations (AC; 1.0 = perfect) of forecasts from a 2004 version of the NCEP GFS system. The improvement in initial conditions through the use of a better data assimilation/forecast system have made current 120-h forecasts as good as or better than old 108-h forecasts, a 12-h increase in lead time. This effect is even larger, approximately 24 h, in the Southern Hemisphere. From this we can conclude that if we generate a new reforecast data set from initial conditions from an old reanalysis, neither the raw forecasts nor calibrated ones from the reforecasts will be as skillful as they could be.

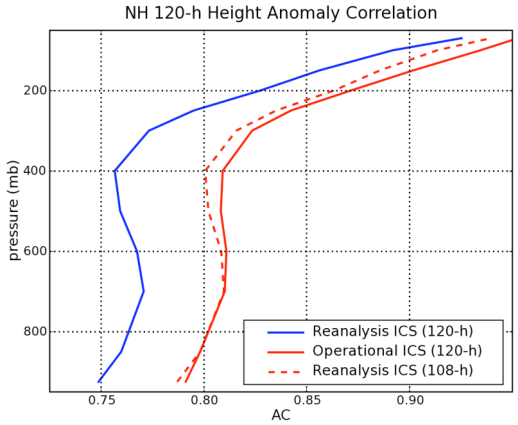

Figure 15: Anomaly correlations (ACs) of Northern-Hemisphere geopotential height forecasts. Blue line: AC of Northern Hemisphere 120-h forecasts started from the NCEP-NCAR reanalysis, which uses a 1998 version of the forecast model. Solid red line: AC of 120-h forecasts from the current NCEP data assimilation system. Dashed red line: the AC of 108-h forecasts from the reanalysis.

(5) Will reforecasting slow the process of implementing model improvements?

There are ways of implementing this where it does not appreciably slow the model development process. One possible way would be to maintain parallel operational numerical models. The first model would be high resolution and updated frequently, but not accompanied by reforecasts; we denote this model as "HR" for high-resolution. The second set of numerical forecasts would be computed from a lower-resolution, infrequently updated model that is accompanied by reforecasts, the "RF" model. Real-time forecasts from the RF model would be insignificant in cost relative to the cost of executing the HR model. Perhaps every 3 or 4 years, a new reanalysis/reforecast data set would be generated using computers others than those comprising the NCEP production system. Once the new reforecasts were computed, then the RF model would be updated. Under this scenario, the automated, calibrated probabilistic forecast products in the NDFD would be generated from the RF data, but forecasters would continue to have access to the HR data in order to get "synoptic intuition" about the weather. The disadvantage of this approach is that the reforecast data set would change relatively infrequently, thereby only sporadically leveraging improvements in the forecast model. Two different forecast model versions would also have to be maintained. For more discussion of this approach, see the conclusion section of Hamill et al. (2006).

Another possible method, proposed by Renate Hagedorn at ECMWF and Zoltan Toth at NCEP, would be to dedicate computer resources to compute reforecasts directly for the model anticipated to be operational one month hence. For example, on reforecasts for the date 1 February would be computed on the operational system on 1 January. A month later, when the date was 1 February, the operational forecast on this date could be calibrated using reforecasts for 1 January – 1 March. The advantage of this method is that reforecasts would be available for the current operational model, and this model could be frequently updated.

(6) How would the reforecasts be computed and stored?

If NCEP's production system were used, the computation of reforecasts could utilize computer resources that NCEP has already planned for other priorities such as increases in model resolution and the introduction of more computationally intensive physical parameterizations. Hence, the use of non-production systems may be preferable. Fortunately, the computation of reforecasts is easily parallelizable, and it may be able to be performed on a cluster of personal computers at a relatively modest cost (order hundreds of thousands of dollars). Alternatively, NOAA could allocate resources on existing systems such as the Climate Test Bed or ESRL's high-performance supercomputer system. Another modest allocation of resources (order several hundred K) would be necessary to archive the reforecasts.

(7) What forecast models should be used? To what lead times should the forecasts extend?

Following the current concept of operations at NCEP, we suggest reforecasts be performed with two models, a limited-area model for short-range forecasts and a global model for forecasts out to two weeks (or perhaps longer). Since there is a recently produced reanalysis for the Eta model (Mesinger et al. 2006), a reforecast using this model would be relatively straightforward to produce.

Our initial reforecast data set was computed to 2 weeks lead. In principle, the techniques used here may be helpful in correcting week-3 and week-4 forecasts, though these forecasts have very little skill relative to the climatology, and simpler statistical techniques may provide just as much skill (Newman et al. 2003).

4. **A road map for NOAA cooperation in producing calibrated probabilistic forecasts**

Achieving the NRC vision of expressing all forecast information probabilistically is an ambitious goal. Per a NRC recommendation, we also suggest that an upper-level manager in NOAA should coordinate these activities, perhaps through the Operations and Service Improvement Process (OSIP) or the New or Enhanced Products and Services Process[7]. This manager would organize the program and delegate tasks according to the unique abilities of NOAA's constituent organizations. This manager would also coordinate the budget for personnel as well as computational and storage facilities.

What are some of the unique abilities of various NOAA organizations? First, MDL is responsible for the production of statistically corrected forecasts, known as "Model Output Statistics," or MOS. MDL is also currently developing gridded MOS products for the NDFD. With this expertise, MDL is the logical choice for the facility to actually produce, disseminate, and verify future reforecast-based probabilistic forecast products.

NCEP currently produces the real-time forecasts, and they would continue in this role, producing real-time forecasts, and perhaps computing the reforecasts as well, depending on the concept of operations.

The regular production of reanalyses for the satellite era is necessary as a companion to the production of reforecasts. Whether this should be done at NCEP, JCSDA, or elsewhere, NOAA should institute a program for repeating the reanalysis task periodically in service of the reforecasts, and of course many other users. Such a recommendation is consistent with that produced by CLIVAR. [8]

OAR/ESRL demonstrated the first reforecast products and techniques, techniques that have provided quantum jumps in forecast skill. ESRL in collaboration with MDL has the expertise to develop new reforecast-based techniques to make probabilistic

[7] See NWS Instruction 10-102, www.weather.gov/directives/sym/pd01001002curr.pdf

[8] See http://www.usclivar.org/Pubs/ReanalysisWorkshop_Rep.pdf

forecasts other parameters, such as wind speed and precipitation type. ESRL also has developed the expertise to provide guidance on the structure of reforecast data sets, such as how many members are needed and what model parameters should be stored.

We foresee other organizations possibly leveraging the reforecast data sets, using it for their particular applications. For example, perhaps the Storm Prediction Center (SPC) of NCEP will use reforecasts to develop products for the advanced warning of severe weather, and NCEP/HPC for specific probabilistic precipitation products. OHD and the River Forecast Centers may use them for river forecasting applications. NOAA's Great Lakes Environmental Research Lab (GLERL) may use reforecasts to develop a tool to forecast changes in lake levels. The military or EPA may use reforecasts for chemical dispersion modeling. The reforecast data and the real-time model data, per the NRC recommendation, would also be publicly available for academics or private companies to use to develop tailored products to specific users.

Our hope and expectation is that these reforecast-based products will be accurate and internally consistent (e.g., no probability of snow forecast if there is little probability of below-freezing temperatures). Forecasters in WFOs will thus not need to manually modify NDFD products, as they do now with the deterministic temperature forecasts. This may save a significant amount of labor in the WFOs, and hence cost to the NWS. However, the WFOs should monitor the accuracy of forecasts, and we expect that all participants will foster open communication in service of the goal of better forecasts.

5. Conclusions

The National Research Council has forcefully recommended that NOAA provide probabilistic weather forecast products to the public. Unfortunately, probabilistic forecasts produced directly by existing ensembles or using short training data sets are unlikely to produce forecasts of sufficient quality to meet user needs. Reforecast-based statistical corrections of ensemble forecasts have been demonstrated to dramatically improve the skill and reliability of probabilistic forecasts generated from ensembles. Hence, we recommend that NOAA should consider a reforecast-based approach. This approach should leverage the talents and capabilities extant at NCEP, MDL, and OAR.

6. References

Daly, C. W. P. Gibson, G. H. Taylor, G. L. Johnson, and P. Pasteris, 2002: A knowledge-based approach to the statistical mapping of climate. *Clim. Res.*, **22**, 99-113.

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 X. Wei, 2004: Ensemble re-forecasting: improving medium range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2005: Reforecasts, an important new data set for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33-46.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.*, in press. Available at www.cdc.noaa.gov/people/tom.hamill/reforecast_analog_v2.pdf.

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

Mesinger, F., and coauthors, 2005: North American regional reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343-360.

National Research Council (Committee on Estimating and Communicating Uncertainty in Weather and Climate Forecasts), 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts.* ISBN 0-309-66327-X. Available from www.nap.edu/catalog/11699.html.

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

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: a proposal for non-local stochastic dynamic parameterisation in weather and climate prediction models. *Quart. J. Royal Meteor. Soc.*, **114**, 691-713.

Palmer, T. N., 2006: Predictability: from theory to practice. Chapter 1 of *Predictability of Weather and Climate*. Cambridge Press, T. N. Palmer and R. Hagedorn, eds.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292-3302.

Whitaker, J. S., X. Wei., and F. Vitart, 2005: Improving week-two forecasts with multi-model re-forecast ensembles. *Mon. Wea. Rev.*, **134**, 2279-2284.