White paper on

A Recommended Reforecast Configuration for the NCEP Global Ensemble Forecast System

Thomas M. Hamill[1], Trevor Alcott,[2] Mark Antolik,[6] James Brown,[3] Mike Charles,[4]
Dan C. Collins,[4] Mark Fresch,[5] Kathryn Gilbert,[6] Hong Guan,[8] Hank Herr,[5]
Wallace Hogsett,[7] David Novak,[7] Melissa Ou,[4] David Rudack,[6] Phillip Schafer, [6]
Michael Scheuerer,[1] Geoff Wagner,[9] John Wagner,[6] Tom Workoff,[7]
Bruce Veenhuis,[6] and Yuejian Zhu[8]

[1] *NOAA Earth System Research Lab, Physical Sciences Division*
[2] *NWS Western Region Headquarters*
[3] *Hydrosolved, Inc.*
[4] *NCEP Climate Prediction Center*
[5] *NWS Office of Hydrologic Development*
[6] *NWS/OST Meteorological Development Lab*
[7] *NCEP Weather Prediction Center*
[8] *NCEP Environmental Modeling Center*
[9] *AceInfo, Inc.*

8 April 2014

Executive Summary

The skill of many weather elements including temperature, winds, and precipitation can be improved dramatically through statistical post-processing. With post-processing, the current forecast is adjusted using statistical relationships estimated from relationships between past forecasts and observations. It has previously been shown that an extensive database of past forecasts consistent with the operational system, i.e., "reforecasts," can be helpful in making effective statistical corrections, especially for longer-lead forecasts and for forecasts of more uncommon events such as heavy precipitation. Many organizations use experimental products generated from the current-generation reforecasts (produced by NOAA/OAR) or expect to use them in the future. These organizations include the CPC, WPC, MDL, and OHD and its customers. The high-profile, Sandy Supplemental-funded "Blender" project will deliver higher-quality guidance if reforecasts are available.

Though a reforecast is available for the current NCEP Global Ensemble Forecast System (GEFS; 2012 version), when this system changes in the near future, the existing reforecasts will be statistically inconsistent with the real-time guidance. Given the value of reforecasts and the desire to make the computation of reforecasts a regular part of operations, we recommend that until a reforecast has been created for an updated version of the NCEP GEFS, NCEP should continue to run an 11-member ensemble for the 00 UTC cycle of the current GEFS, so as to maintain a forecast system that has statistical consistency with the past reforecasts. Further, we recommend that NCEP should aggressively move to institute the regular production of reforecasts and reanalyses (necessary for reforecast initialization, and many other weather and climate applications).

NOAA staff has recently conducted a range of sample-size sensitivity tests in order to determine a configuration that uses the least possible computational resources while producing acceptable post-processed guidance. Based on these experiments, we recommend a reforecast with 5 members, spanning 20 years, skipping 5 days between reforecasts, and performed for the 00 and 12 UTC cycle only. The real-time ensemble computes 84 members/day. This suggested configuration will add 40 members/day, i.e., ~50% computational expense increase, not including the expense of generating the reanalyses. NCEP/EMC is encouraged to identify additional computing resources to support the generation of these reforecasts, or alternatively may choose to slow the pace of planned resolution enhancements in order to fit in the extra reforecast computations, or adjust its GEFS in other ways.

Reforecasts will require initialization with a reanalysis that is statistically consistent with the operational analysis system. This will be a major labor and computational cost, but there are other NOAA requirements for a new reanalysis. Additionally, a high-quality, high-resolution retrospective surface-based weather analysis is needed to facilitate post-processing, such as by an enhanced version of the Real-Time Mesoscale Analysis (RTMA). NCEP is encouraged to plan for these.

1. **Introduction**.

This white paper provides recommendations for how NOAA can institute reforecasts for its operational global ensemble system, with recommendations for other prediction systems as well. This introductory section will discuss the benefits of reforecasting, the issues and challenges in generating the reforecasts, and the anticipated operational users of reforecast products. Section 2 will provide a brief description of the sample-size sensitivity tests that were conducted, with more detail provided in references and online appendices. Section 3 provides a recommended plan of action for instituting regular reforecasts at NCEP/EMC. It will outline the anticipated computational and personnel resources necessary to achieve this and will discuss the pros and cons of some options for making the necessary computational resources available. Section 4 will provide a short conclusion.

a. *The benefit of reforecasting*[1]

The weather and climate prediction community have made continued, significant improvement in the quality of numerical forecast guidance. This has come as a result of increased resolution, improved physical parameterizations, improved chemistry and aerosol physics, improved estimates of the initial state estimate due to better data assimilation techniques, and improved couplings between the atmosphere with the land surface, cryosphere, and ocean, and more. Nonetheless, judging from the pace of past improvements, medium-range forecast systematic errors will not become negligibly small within the next decade or two. For intermediate-resolution simulations such as those from current-generation global ensemble systems, users of forecast guidance may notice biased surface temperature forecasts, or precipitation forecasts with insufficient detail in mountainous terrain, and perhaps too much drizzle or too little heavy rain. They may notice over- or underestimated cloud cover, or that near-surface winds are characteristically much stronger than forecast. They may notice that hurricanes are too large in size but less intense than observed.

In such circumstances, reforecasts are especially helpful for statistically adjusting weather and climate forecasts to observed data, ameliorating the errors and improving objective guidance (Hamill et al. 2006, Hagedorn 2008, Hamill et al. 2013). Reforecasts, also commonly called hindcasts, are retrospective forecasts for many dates in the past, ideally conducted using the same forecast model and same assimilation system used operationally. Reforecasts have been shown to be particularly useful for the calibration of relatively uncommon phenomena such as heavy precipitation (Hamill et al. 2008) and longer-lead weather-climate phenomena (Hamill et al. 2004), where there is small forecast signal and comparatively large noise due to chaos and model error. In both cases, the large sample size afforded by reforecasts is useful for finding a suitably large number of

[1] Some of the text for this section was borrowed verbatim from Hamill et al. (2013), with permission from the lead author (the government retains copyright of this article).

past similar forecast scenarios. With associated observational data, one then can estimate a conditional distribution of the possible observed states given today's numerical guidance, assuming past forecasts have similar errors to current forecasts. Even when no observed data is available for calibration, reforecasts can be useful for determining the climatology of a model. A 20 ms^{-1} surface wind would be exceptionally strong in most locations on earth, but if the forecast model severely over-forecasts wind speeds, such an event may be of less concern. A reforecast can thus be used for estimating the forecast climatology, placing the current forecast in context (LaLaurette 2003ab).

An example of the benefits of reforecasts for precipitation forecasting is shown in Fig. 1. Here, using second-generation GEFS reforecasts and 1/8-degree Climatology Calibrated Precipitation Analysis (CCPA; Luo et al. 2014) precipitation analyses over the contiguous US, we show the skill of probabilistic forecasts for exceeding the 95th percentile of the analyzed climatological distribution, which varies from location to location and month to month. Data are over the CONUS, 2002-2013. The reforecast-calibrated probabilities were generated using a cross-validated rank-analog approach (Hamill and Whitaker 2006, Hamill et al. 2013). As shown, they are much more skillful than those from the 11-member raw ensemble. By using the 11-member reforecast rather than the 21-member operational ensemble to get a sample over the full 2002-2013 period, there was some degradation of skill, but this amounts to approximately 5 percent improvement in the best-case scenario where no model error is present (Richardson 2001, Fig. 1).

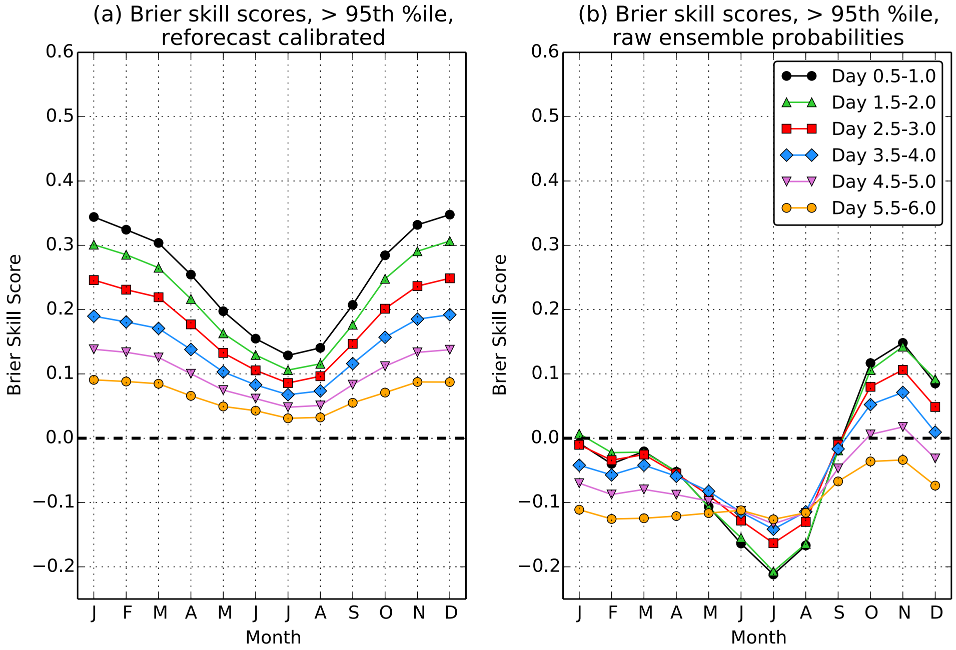

Figure 1. Brier skill scores for exceeding the 95th percentile of the climatological distribution, using 2002-2013 data over the CONUS. (a) reforecast-calibrated probabilities, and (b) raw ensemble probabilities.

Figure 2 below shows that the post-processed forecasts are also quite reliable, though in this circumstance they somewhat under-forecast high probabilities. In contrast, the raw ensemble probabilities are unreliable and tend to drastically over-forecast the high probabilities, though the raw forecasts are much sharper, i.e., they make far more high-probability forecasts. Raw ensemble forecast reliability and skill are slightly degraded (again, Richardson 2001) from the use of 11 members here instead of the real-time 21 members, but this is a comparatively small effect.

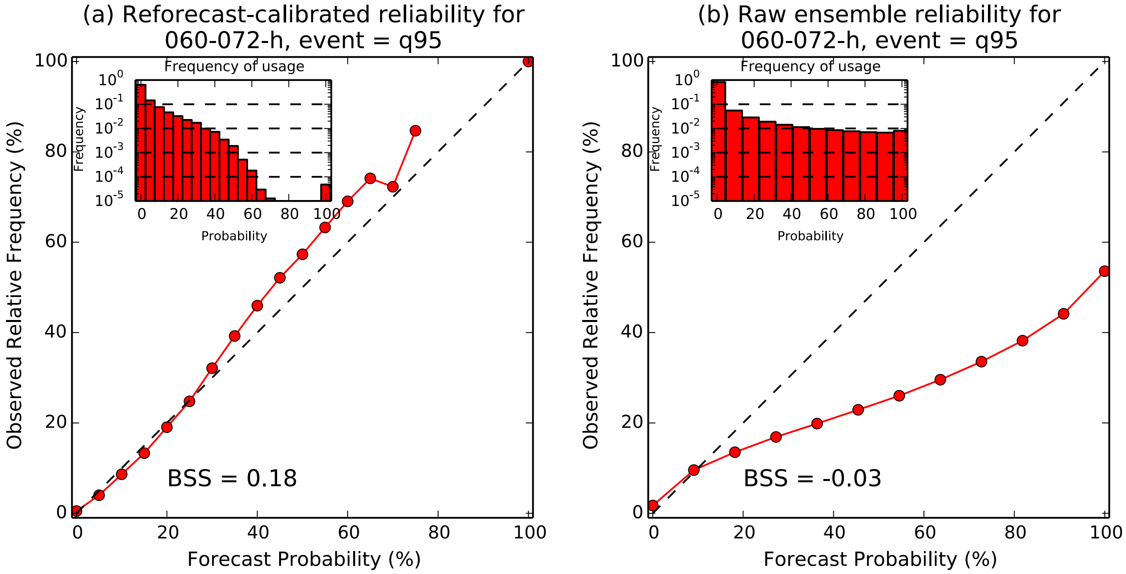

Figure 2: Reliability diagrams for the event of precipitation > 95th percentile of the climatological distribution, composited over the CONUS, and all dates from 2002-2013. (a) reforecast calibrated, and (b) raw ensemble.

Having determined that the use of reforecasts is highly beneficial, the NCEP Climate Prediction Center (CPC) uses the GEFS reforecasts to statistically adjust its 6-10 day and 8-14 day weather forecasts. Figure 3 shows why they use the reforecast product heavily; the skill of reforecast-based products is significantly larger than the skill of probabilistic forecasts from the raw ensemble, as well as bias-corrected forecasts from the NAEFS (North American Ensemble Forecast System). The verification was for the period 2011-2013 using stations over the CONUS.

As a final example, we note the potential of reforecasts for creating valuable new forecast products. For example, currently the NWS does not produce any objective probabilistic tornado forecast guidance at long lead times. Francisco Alvarez, a grad student at St. Louis University, has demonstrated in his Ph.D. research that reliable, skillful long-lead tornado forecasts can be generated using the GEFS reforecasts. Forecast information such as vertical wind shear and CAPE (convective available potential energy) are used in an analog procedure to produce the forecasts. An example of the experimental, real-time tornado forecast probabilities are shown in Fig. 4; these experimental forecasts are available in quasi-realtime from http://tinyurl.com/reforecast-tornado . They will be evaluated by staff at the NCEP Storm Prediction Center during the spring of 2014.

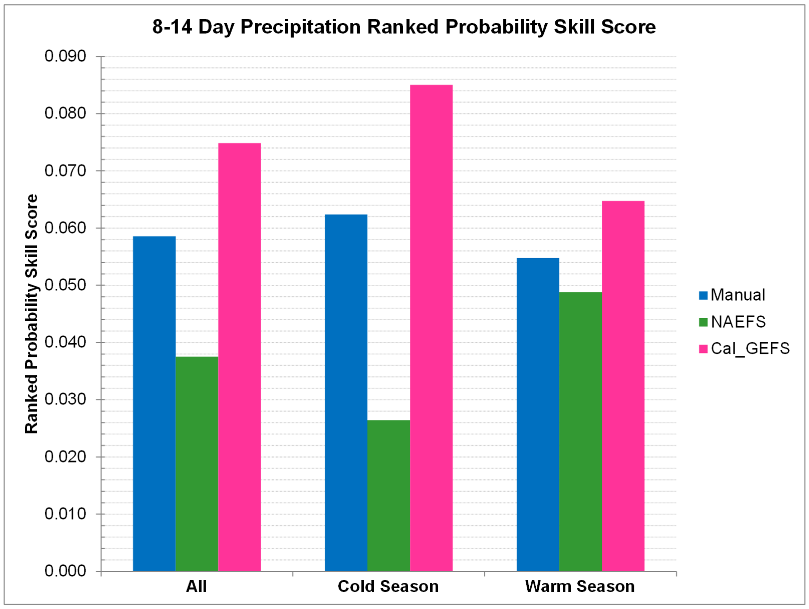

Figure 3: Skill of NCEP/CPC's forecasts of above/near/below-normal precipitation for the 8-14 day period from various methods, including manual forecasts, from NAEFS, and from the reforecast-calibrated GEFS.

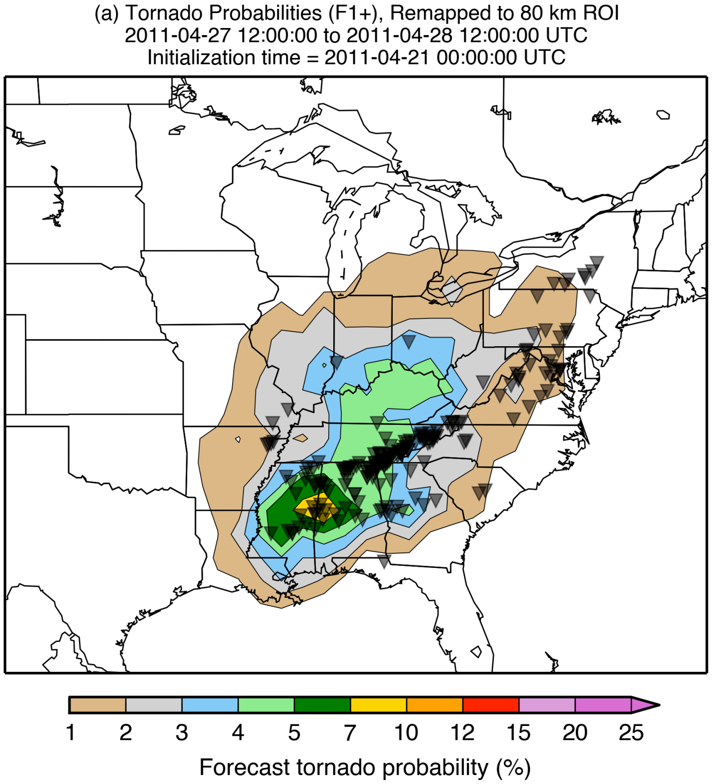
(a) Tornado Probabilities (F1+), Remapped to 80 km ROI
2011-04-27 12:00:00 to 2011-04-28 12:00:00 UTC
Initialization time = 2011-04-21 00:00:00 UTC
1
2
3
4
5
7
10
12
15
20
25
Forecast tornado probability (%)

Figure 4: Objective tornado probabilities for the period 12 UTC 27 April 2011 to 12 UTC 28 April 2011 for forecasts initialized 6 days earlier, based on a reforecast analog procedure. Observed tornado locations are shown with the grey inverted triangles.

b. Issues in the computation of reforecasts.

There are reasons why the NWS has not yet institutionalized the regular production of reforecasts. Simply, they are a significant added computational and resource expense and require extra planning for changes in the forecast system. The more retrospective forecasts that are generated, the greater the computational burden that is added to that of generating the real-time forecasts.

To ensure statistical consistency, ideally a reanalysis should be in place for reforecast initialization. It should use the same data assimilation system and model that is used operationally, and the configuration of the forecast model should remain fixed as well. Generating this multi-decadal reanalysis is another computational and labor-intensive proposition, and using a frozen model version is a practice NCEP prefers to avoid. However, the consequences of not fully fixing the assimilation and forecast system are real and are illustrated in Fig. 5. In February 2011 the initial conditions for the control forecast of the GEFS reforecast were changed from the CFSR to the operational GSI, which used a somewhat different version of the forecast model for its background. Those seemingly subtle changes appear to have affected the short-term forecast bias in many regions, for example in the region over TX and LA illustrated in Fig. 5. The warm-season bias was colder by ~ 1C for 2011-2013 compared with 2008-2010. This degraded the subsequent calibrated forecasts (not shown). Hence, if an organization is to commit to the computational expense of generating a reforecast, it should ideally also commit to using a fully frozen version of the forecast and assimilation system during the period that reforecast is maintained. Otherwise, the full value of the reforecast cannot be exploited.

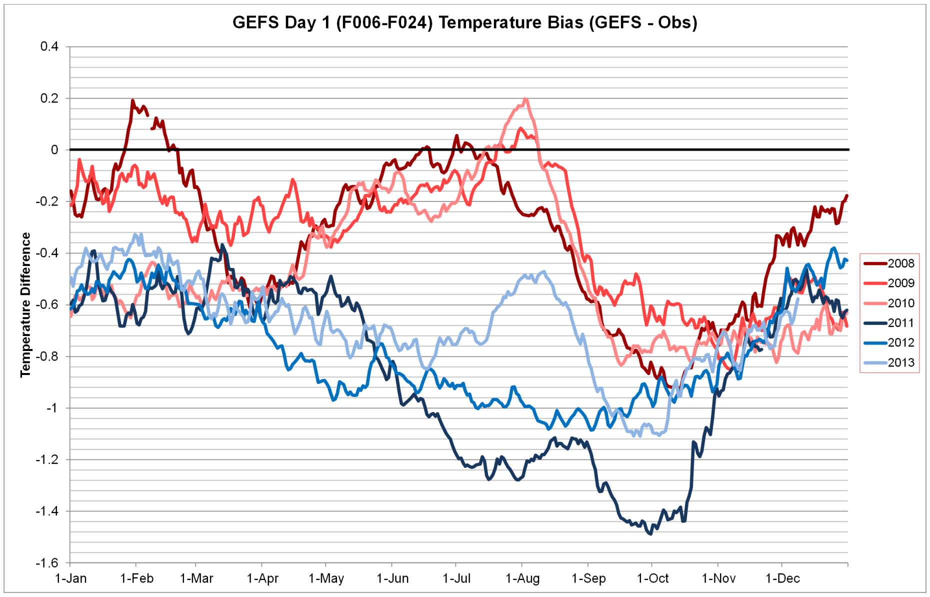

Figure 5: Temporally smoothed (45-day centered average) plots of day +1 temperature forecast bias for the geographic region 103° W to 90° W, 30° N to 37° N.

Another challenge of the reforecast procedure is the expense of archiving the reforecasts. For the current-generation GEFS forecasts, 99 variables were saved at 1-degree resolution and 27 commonly used variables at the higher, ~0.5-degree resolution. Data was saved every 3 h to +72 h lead and every 6 h thereafter. This archive, spanning 29 years and with every-day reforecasts, exceeds 150 TB. The archival costs can be expected to decrease if a smaller, shorter-period ensemble is used, but to increase if the model resolution is increased and customers require the data saved at the native resolution. Given the expense of regenerating the data, a robust backup strategy is needed, ideally including a full tape backup of the entire model states.

Finally, to exploit the reforecasts for statistical post-processing in regions where there are not in-situ observations, one needs a coincident time series of high-quality analysis data (e.g., for surface temperature, winds, and precipitation). This thus means that a high-quality, retrospective surface-based mesoscale analysis should also become a priority for NOAA to generate.

Clearly, generating a reforecast presents significant resource and personnel challenges. Nonetheless, as demonstrated above, we suggest that the benefits of reforecasting are impressive enough for NCEP to work toward their regular computation. For the next generation of the GEFS, and likely for several generations thereafter, the systematic errors are likely to be large enough that there will continue to be significant value from the computation and application of reforecasts. For example, Hamill et al. (2013) discuss that with two generations of reforecasts, using 1998 and 2012 versions of the model, there was less improvement from post-processing the 2012 ensembles than the 1998 version. However, the improvements were still very large, as shown in Fig. 1. Given the difficulty in eliminating model bias, it is reasonable to expect for the next 10 years at least that reforecasts will be needed.

Is a reforecast also necessary for NCEP's other systems such as the deterministic GFS? The answer to this depends on the anticipated applications for the GFS. If the GFS is intended to be used, for example, in blended, post-processed products that provide guidance to forecasters, the GFS data will be strongly de-weighted in such a blend if a sufficient amount of training data is not available; the post-processed guidance from it will be of much lower quality than from a modeling system that does have an accompanying reforecast.

c. *Some users of the GEFS reforecasts.*

As the current GEFS reforecasts were generated by NOAA/OAR, NWS forecasters are expected to treat the reforecast guidance as experimental. Generally, they would like to institutionalize the generation of reforecasts so that the derived products will continue to be available as the system changes. The current and anticipated near-term future customers of reforecast-based products will include:

(i) NCEP Climate Prediction Center (CPC): CPC uses the GEFS reforecasts to support the production of their 6-10 and 8-14 day forecasts and their 3-14 day U.S. hazards forecasts. They plan to use them in future new products as well. As demonstrated above, the reforecast-based products often provide skill that is greater than that from all other current 6-10 or 8-14 day tools, such as the multi-model North American Ensemble Forecast System (NAEFS). Additionally, CPC is currently developing a weighted multi-model outlook product, which will rely heavily on the GEFS reforecast dataset. This multi-model outlook product is intended to eventually serve as the primary guidance for CPC's 6-10 and 8-14 day forecasts. The reforecasts are also required for CPC's experimental week-2 probabilistic extremes tool, used as guidance for the official U.S. Hazards outlook, and are required for a planned week-2 probabilistic hazards product in the future. Expansion of the probabilistic extremes tool to a global domain is planned in the future as guidance for global forecasts, such as the global tropical hazards and benefits outlook.

(ii) NCEP Weather Prediction Center (WPC): WPC uses experimental probabilistic calibrated precipitation forecast guidance produced by the Earth System Research Lab (ESRL) Physical Sciences Division (PSD). They have found that the post-processed probabilistic and deterministic forecast guidance provides relevant bias correction and local downscaling to aid the production of their heavy rainfall forecasts.

(iii) NCEP Storm Prediction Center (SPC): SPC will be evaluating the experimental tornado guidance products produced by ESRL/PSD and F. Alvarez (grad student, St. Louis University). Pending a successful demonstration, SPC may use this to help develop new extended-range tornado prediction products.

(iv) NWS Western Region Headquarters (NWS/WR): NWS/WR uses the reforecast data in their situational awareness tool to establish the GEFS model climatology (http://ssd.wrh.noaa.gov/satable/).

(v) National Digital Forecast Database (NDFD): The NWS has instituted a Sandy Supplemental-funded project to produce automated, downscaled, high-resolution numerical guidance that can be used by forecasters as a first-guess grid in their National Digital Database. The intent will be to blend together and statistically adjust deterministic and ensemble forecast system guidance from a variety of modeling systems, including the GFS and GEFS. Next-generation GEFS reforecasts are expected to make the GEFS the pillar of this system.

(vi) Office of Hydrologic Development (OHD): OHD uses reforecasts both to statistically post-process the meteorological forecast information that is input to their hydrologic prediction systems, and to provide a sufficiently large number of high-impact past cases to validate their forecasts.

(vii) Water Resource Managers: Here is an edited quote provided by James Porter, from the New York City Department of Environmental Protection, that demonstrates that there are important customers outside of NOAA for reforecast-based products, customers that have provided support to NOAA because of its reforecast:

"The New York City Department of Environmental Protection (DEP) manages New York City's water supply, providing more than one billion gallons of high quality water each day to more than 9 million residents and making releases from our reservoirs to support downstream interests. As you know, DEP recently implemented a decision support system called the Operations Support Tool (OST). OST provides guidance to water supply managers by combining ensemble streamflow forecasts from the National Weather Service's Hydrologic Ensemble Forecast Service (HEFS) with near-real-time environmental and reservoir system data to predict reservoir storage and water quality conditions in the future. The system cost over $8 million to develop, including nearly $1 million DEP paid to NOAA to accelerate the development of HEFS forecasts to meet our OST timeframe. OST is a key element in DEP's Filtration Avoidance Determination (FAD), which exempts New York City from building a multi-billion dollar filtration plant. The Operations Support Tool is used daily to guide water supply operations." OST has been developed, calibrated, and tested using the current version of HEFS. The statistical properties of these forecasts are critically important to the performance and reliability of OST. Any changes to HEFS, such as modification of the numerical weather prediction models (e.g. GEFS) that provide input to HEFS, are of utmost concern to us. We support NOAA's efforts to innovate and improve its products, and we welcome any changes that could increase forecast skill and reduce uncertainty. However, given the potential impacts of such changes on OST, we respectfully request the following:

(1) DEP be given notice as far in advance as possible of any planned changes that impact HEFS, including changes to numerical weather prediction (e.g., GEFS) and hydrologic models. Lead times of a year or more could be required for us to assess modified products from NOAA and make changes to OST.

(2) DEP be provided an updated set of HEFS hindcasts that reflect the planned changes. Updated HEFS hindcasts would require GEFS reforecasts if GEFS is changed.

(3) Forecasts from frozen versions of GEFS and HEFS be continued for some agreed-upon period of time to allow DEP to calibrate and validate OST with the new forecast system products.

DEP has a long-standing partnership with NOAA and the National Weather Service, and we look forward to continuing that in the future. Thank you very much for your consideration of our concerns, and please don't hesitate to contact me to discuss this further."

2. **Sample-size sensitivity experiments**.

In this section we summarize the results of sample-size sensitivity tests that have been conducted at ESRL/PSD, NCEP/CPC, and MDL. We start with older work performed by ESRL/PSD scientists. Figure 6 shows the skill of post-processed 6-10 day and 8-14 day probabilistic forecasts of temperature and precipitation and how this varies with the number of years of reforecast training data; this was taken from Hamill et al. (2004) and the experiments with first-generation reforecasts, which are less skillful than the second-generation data (Hamill et al. 2013). For this application, a logistic regression approach was used for calibration. This suggests that if a reforecast data set was available with every-day data, for these applications there was a significant increase in skill from 2 to 5 years of training data, but once 10-12 years were reached, the incremental increase was much smaller.

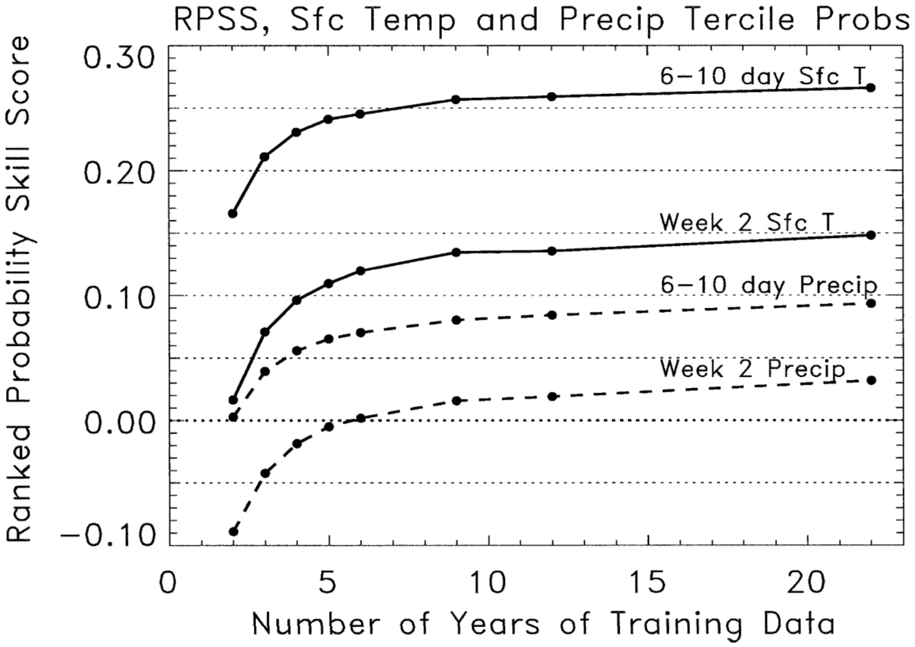

Figure 6: Ranked probability skill score (larger is better) of post-processed surface temperature and precipitation forecasts as a function of the number of years of training data, assuming every-day samples were available.

Figure 7 shows that a smaller amount of training data, 4 years total, can provide effective calibration for these applications if the samples are spread apart by 5 days between consecutive samples. In this way, the reforecasts span a longer period of time with a greater diversity of weather and more independent samples. We shall see below that for other applications such as daily weather (as opposed to the time averages here), spreading out the samples over time does not provide as much positive benefit.

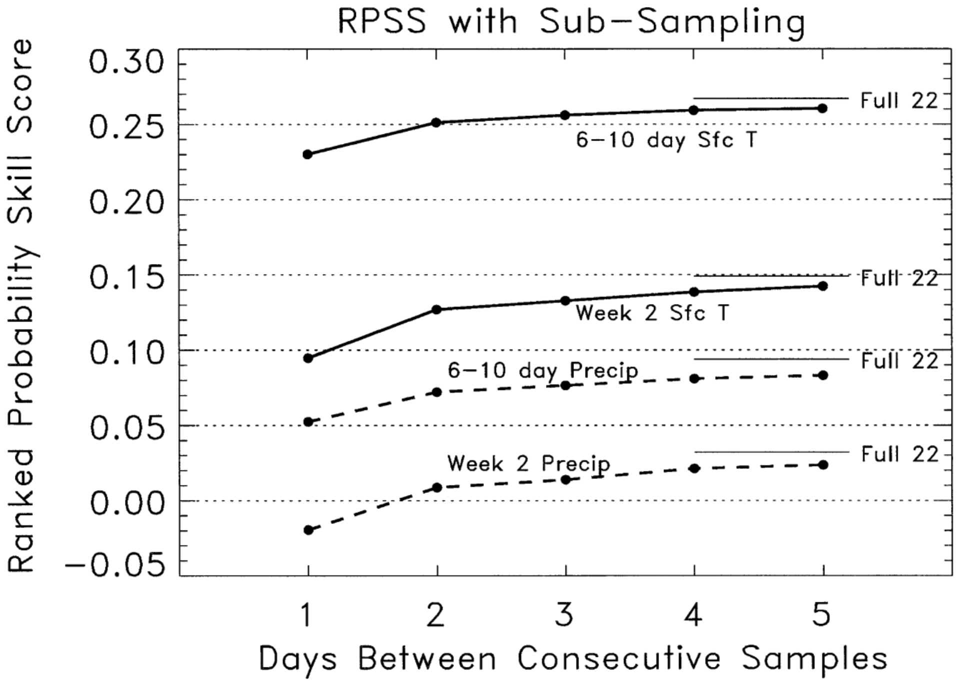

Figure 7: Ranked probability skill score (larger is better) of post-processed surface temperature and precipitation forecasts as a function of the number of days skipped between forecast samples. In all experiments, four total years of training data were used.

Another set of ESRL/PSD's older work involved determining how few members might be used in a reforecast ensemble. The computational expense of the reforecast will of course scale proportionally with the number of members, so if a large ensemble is not absolutely required, computing a smaller-member ensemble reforecast will help constrain the expense. Figure 8 shows results of precipitation skill as a function of ensemble size from an unpublished study from several years ago (http://tinyurl.com/sample-size-reforecast). ECMWF's 2005 reforecasts provided the data, and logistic regression was again used for calibration. When a smaller ensemble was used both for training of the statistical model *and* for the real-time forecasts, there was a more monotonic increase in skill with more ensemble members. When the size of the real-time ensemble was fixed at 15 members, the size of their reforecast, then there was a dramatic increase in post-processed skill when increasing the reforecast size from 1 to 3 to 5 members. But there was much less skill increase in increasing the ensemble size further, from 5 to 7 to 9 members. This suggests that for this application (6-10 day surface temperatures behaved similarly) an acceptably skillful post-processed forecast can be obtained with a 5-member reforecast ensemble; NCEP/EMC need not run a 21-member reforecast as they do for the real-time ensemble. This will save significant computational expense.

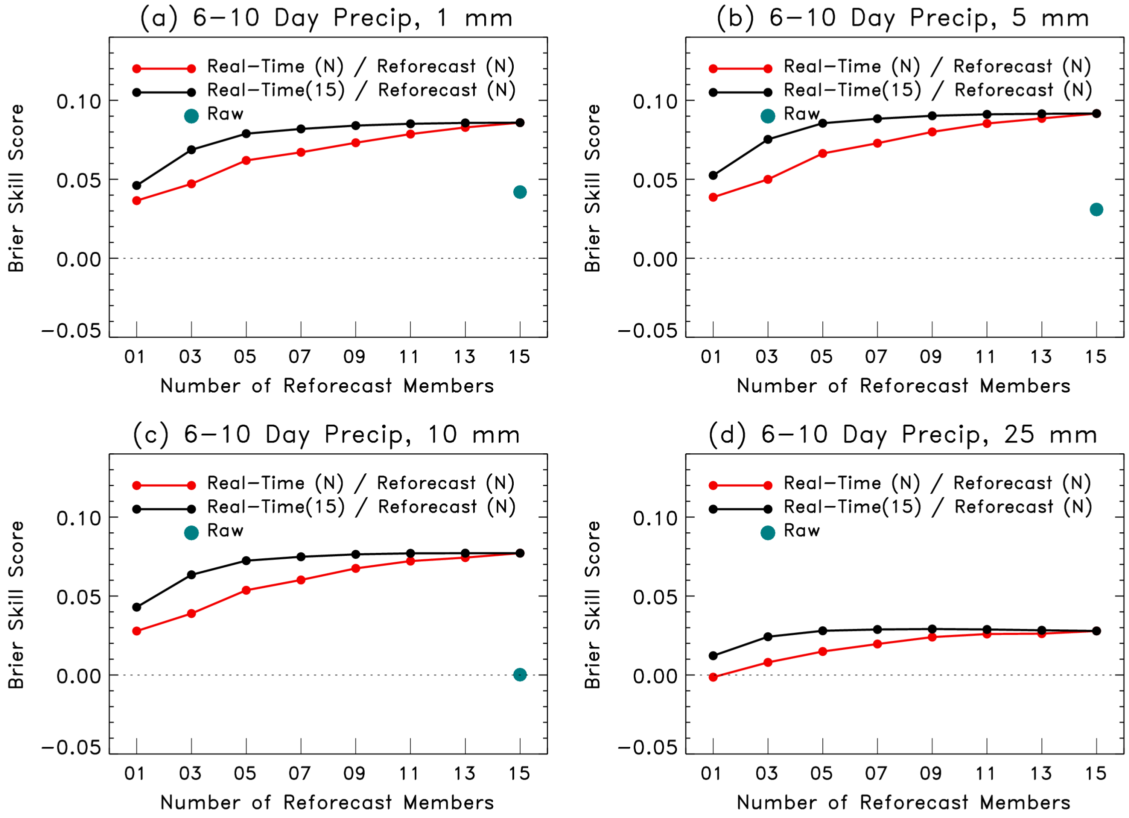

Figure 8: 6-10 day precipitation forecast Brier Skill Score (larger is better) as a function of the number of reforecast members and the number of real-time members. The green dot shows the skill of the raw ensemble forecast guidance, here with 15 members used to set the probabilities. The red curves show the skill when n members are used both as the ensemble size for the reforecast ensemble and for the real-time ensemble. The black curve shows the skill when a n-member ensemble is used for the reforecast and a 15-member ensemble is used for the real-time forecast.

NCEP/CPC also conducted a set of sample-size sensitivity experiments for 8-14 day forecasts. Their results were consistent with those presented above. For example, they found no benefit in spanning more than 20 years for surface-temperature forecasts, and some slight benefit for precipitation forecasts (Fig. 9). In this experiment, the GEFS reforecast data set was thinned to once-weekly samples, and 6 ensemble members were used, applying the regression approach of Unger et al. (2009).

MDL scientists found somewhat less sample-size sensitivity for their instantaneous and shorter-range forecasts, relative to ESRL/PSD and CPC's results. Figure 10 shows that there is very little sensitivity for surface temperature forecasts to the number of years of training data when measured in terms of the continuous ranked probability score (CRPS).

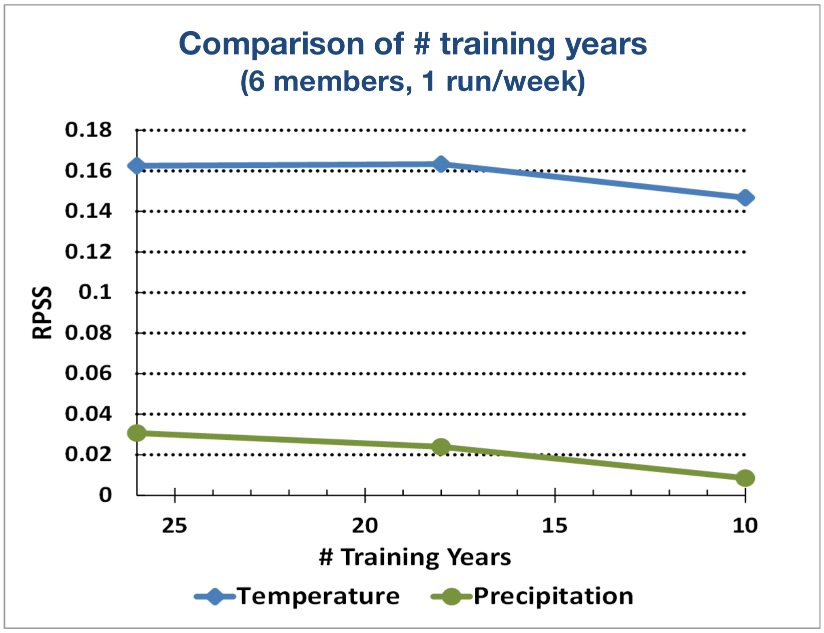

Figure 9: Ranked probability skill score (larger is better) for CPC 8-14 day surface temperature and precipitation skill as a function of the number of years of training data, using GEFS second-generation reforecasts and station observations over the US.

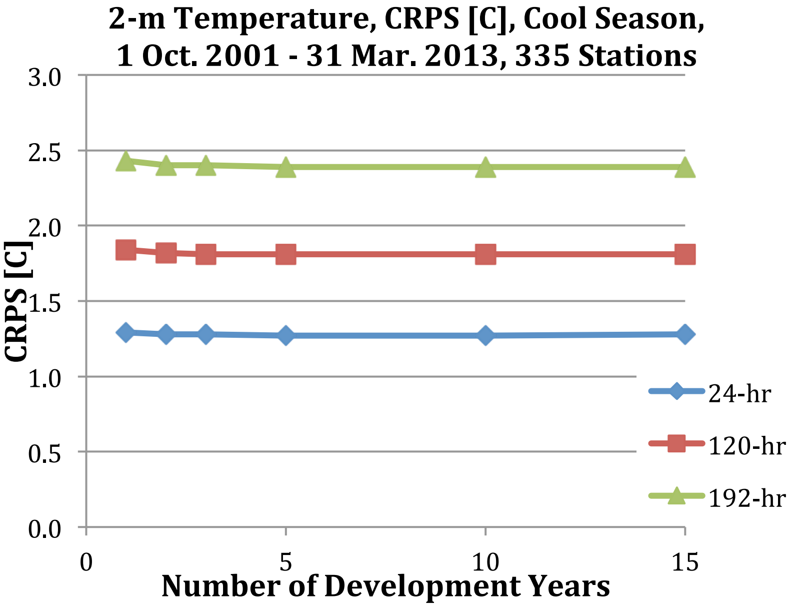

Figure 10. Continuous ranked probability score (CRPS; lower is better) for 2-meter surface temperature forecasts over the US during the cool season.

Figure 11 shows that there is somewhat more sensitivity to the number of years of training data for wind-speed forecasts, here specifically for the higher-speed forecasts, ≥ 10 knots. A linear regression approach was used to calibrate the data, and second-generation GEFS reforecasts and US station data were again used. Forecasts were validated deterministically using mean absolute error. Errors of course increased with longer lead time, but the sensitivity to sample size did as well. With larger, more random errors at longer leads, it required a larger sample to reduce the error. 15 years of data was better than 10 years, but not by a large amount.

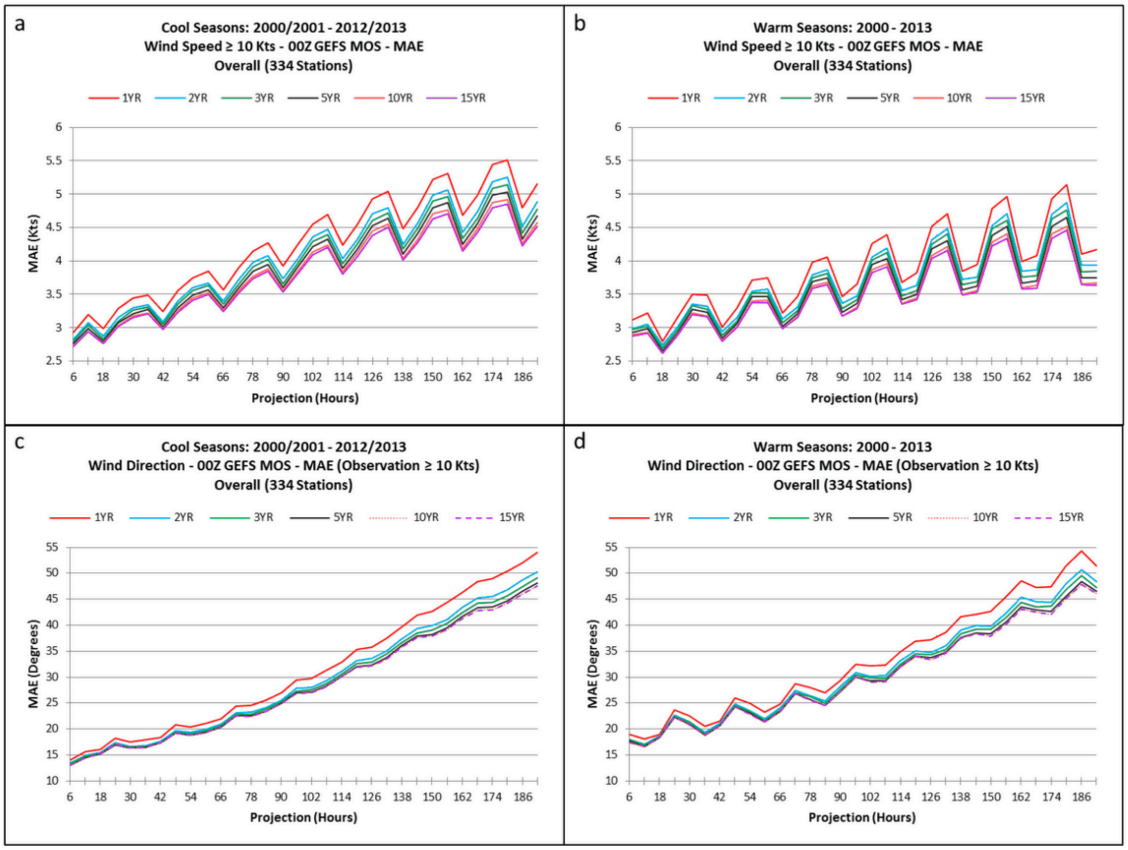

Figure 11. Cool-season (panels a and c) and warm season panels (b and d) mean absolute error (MAE; smaller is better) scores by projection for GEFS MOS wind speed as a function of training sample size. Top panels provide errors of wind speeds when the validation data was ≥ 10 kts, and bottom panels provide errors of wind direction.

Figure 12 considers the sensitivity to training sample size for precipitation type. Training sample sizes of 5 years every day, 15 years every day, and 5 years every third day were considered, as well as the effect of developing the regression equations over several distinct and separate regions of the US to account for possible geographically dependent biases. The results indicate that the use of regional equations provides a substantial skill improvement, though it is possible that when actually viewing the precipitation type forecast maps, there could be geographical discontinuities at the boundaries between regions. There was a smaller sensitivity to training sample size. More sophisticated methods of regionalizing precipitation type, such as training with overlapping tiles to avoid geographical discontinuities, were not tried here for lack of time.

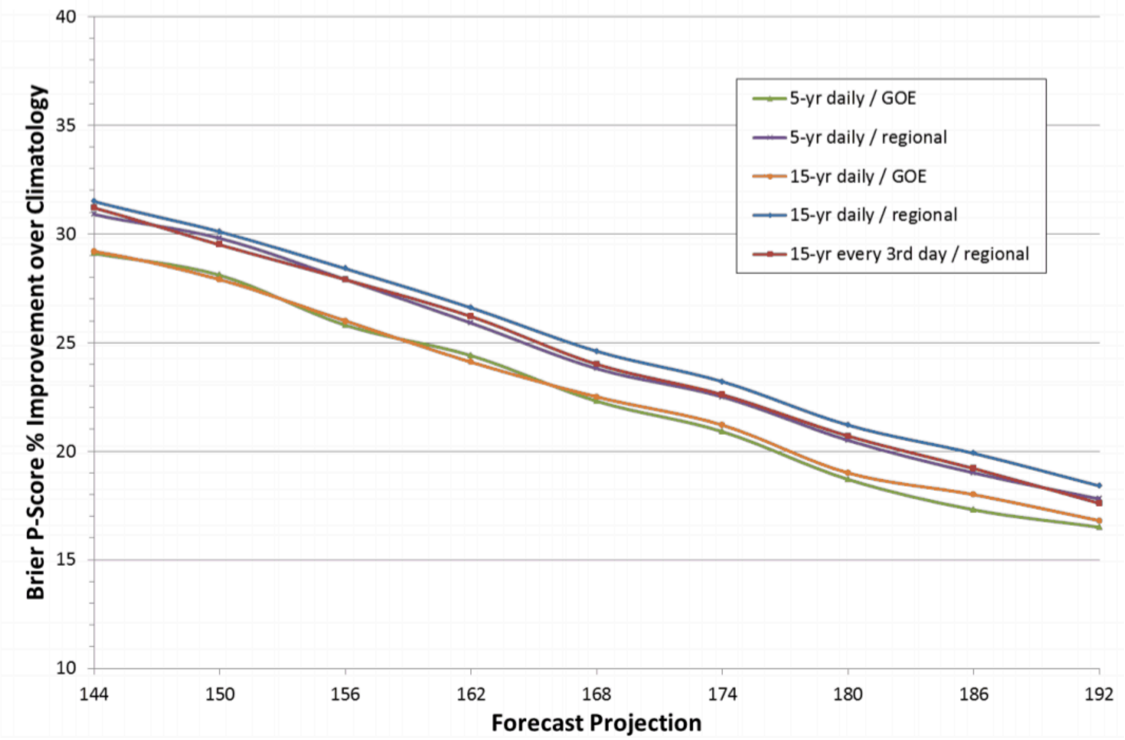

Figure 12: Brier skill score (larger is better) for precipitation type forecasts over the US. “Regional” and “GOE” forecasts were tested, developing the precipitation type either separately for distinct regions of the use or over the entire US, respectively. Training sample sizes included 5 years every day, 15 years every day, and 15 years every third day.

Finally, we consider some recent work at ESRL/PSD to evaluate the sample-size sensitivity for precipitation forecasts over the US. A distribution fitting approach described in Scheuerer (2013) was used. Training data included the 2002-2013 CCPA (Hou et al. 2014) at 1/8-degree grid spacing and coincident second-generation GEFS reforecast data. Four sample sizes, were tested: 1 year of training data, 4 years, and 1 and 4 years with supplemental data added. The supplemental data consisted of forecast-observation pairs from 20 other locations for each grid point. The 20 supplemental locations for each point selected on the basis of similarity of climatology, forecast error, and some minimum separation distance to help ensure some independence of the samples. Figure 13 shows the results, which indicate that 4 years of training data with supplemental locations performed the best, which were slightly better than 1 year with supplemental and 4 years without. One year without supplemental training data was much worse. Figure 14 shows that while the skill of 4 years with and without the supplemental data was not much different, the high-end reliabilities were improved with the extra data. Figure 15 shows the skills of longer-lead precipitation forecasts. Here there was a bit more sensitivity to training sample size, presumably because it takes more samples to determine the predictable signal relative to the chaotic forecast noise, which is much larger at 120-132 h than at 12-24 h.

Another advantage of a larger training sample size is improved meteorological consistency, i.e., post-processed forecasts that have smoother, more natural looking precipitation forecast features. Figure 16 illustrates this for a particular case. The larger training sample size reduces the small-scale noise.

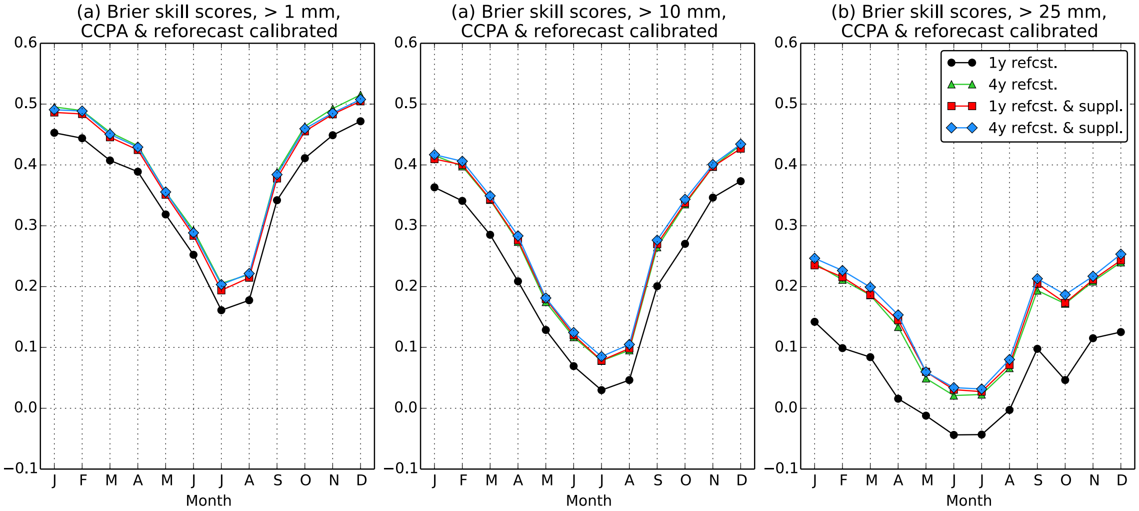

Figure 13: Brier Skill Scores for US precipitation forecasts at a lead time 012–024 h as a function of month of the year. (a) event of > 1mm/12h, (b) > 10mm/12h, and (c) > 25 mm/12h. Different training sample sizes are shown with the different colored curves. Red and blue curves display skill when the training data at each point was supplemented with training data from 20 other points chosen to have similar climatologies and forecast errors.

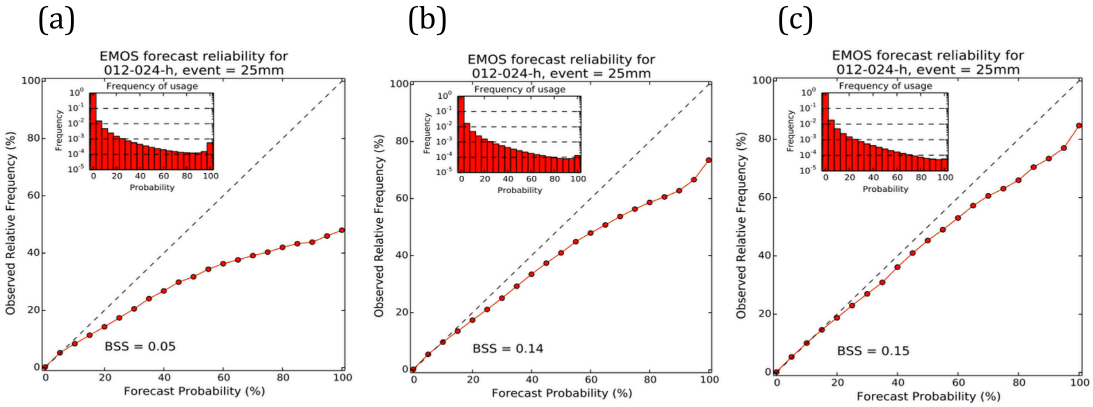

Figure 14: Reliability diagrams for the > 25 mm/12-h forecast event and 12-24 hour forecast lead times. (a) with 1 year of training data, (b) 4 years of training data, and (c) 4 years and 20 supplemental locations for each grid point. Inset histograms provide the frequency of usage for each probability category, in 5-percent intervals.

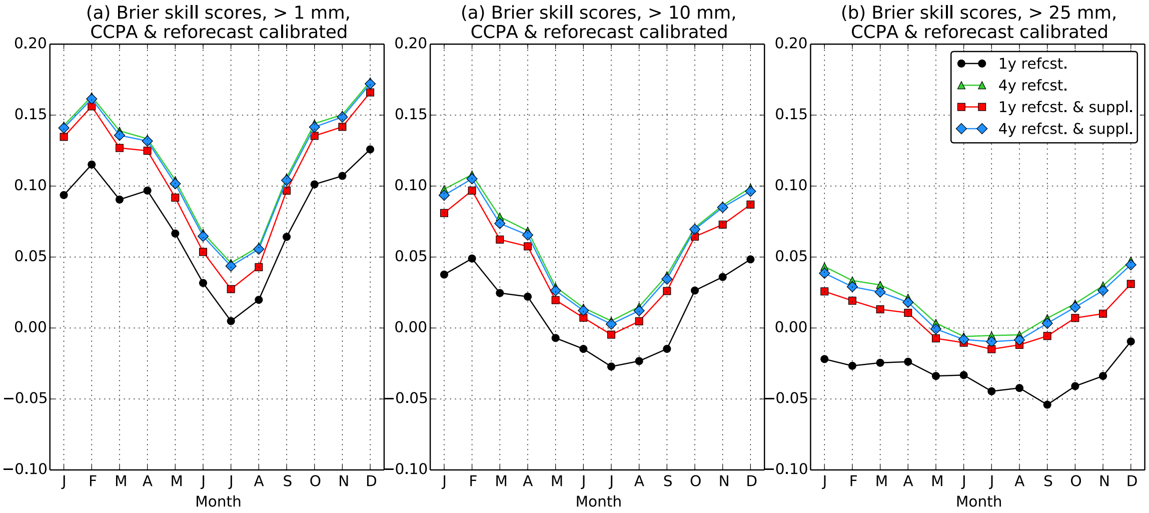

Figure 15: As in Fig. 12, but for 120-132 h forecasts.

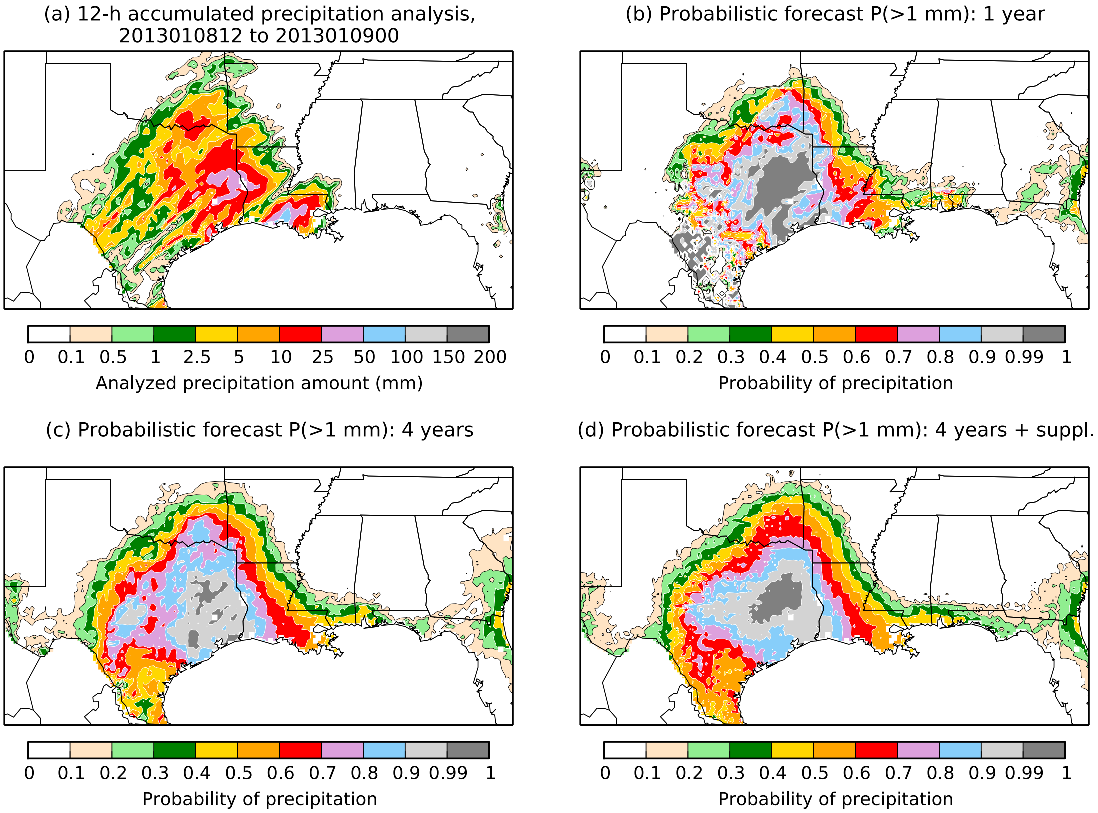
(a) 12-h accumulated precipitation analysis,
2013010812 to 2013010900
0 0.1 0.5 1 2.5 5 10 25 50 100 150 200
Analyzed precipitation amount (mm)
(b) Probabilistic forecast P(>1 mm): 1 year
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.99 1
Probability of precipitation
(c) Probabilistic forecast P(>1 mm): 4 years
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.99 1
Probability of precipitation
(d) Probabilistic forecast P(>1 mm): 4 years + suppl.
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.99 1
Probability of precipitation

Figure 16: (a) 12-h analyzed precipitation amount ending on 00 UTC 09 January 2013, and post-processed forecasts of the probability of precipitation accumulations > 1 mm/12 h obtained with parameter estimates based on training sets corresponding to (b) 1 year, (c), 4 years, and (d) and 4 years plus supplemental locations.

Considering these results collectively, we conclude the following. First, sample-size sensitivity appears to be larger for longer-lead forecasts, and especially for forecasts of time-averaged quantities such as 8-14 day precipitation. There tends to be more sample-size sensitivity for uncommon (and high-impact) forecast parameters such as heavy precipitation than for more common ones such as light precipitation. With an increasing focus on high-impact weather, it may be more important to base a reforecast configuration decision on these extremes than on the more common weather – the extremes, those are the situations where forecasters need accurate guidance the most.

Readers are encouraged to view more detailed reports; hyperlinks to these are provided in the appendix.

3. **Recommendations for instituting reforecasts.**

a. Recommendations.

Based on the results above, the ad-hoc committee that was commissioned at the 2013 NCEP Production Suite review makes the following recommendations concerning reforecasts. Discussion of the recommendations is provided in the next subsection.

(i) Recommendation 1: Until a next-generation reanalysis and reforecast is in place and ready for use, NCEP/EMC should continue the production of an 11-member GEFS ensemble for the 00 UTC cycle in its current (circa 2012) configuration. These real-time forecasts will be approximately consistent with the GEFS reforecast, so existing products can continue to be generated from them. Given the next-generation GEFS will be higher in resolution, this will be a minor computational expense.

(ii) Recommendation 2: NOAA should immediately begin preparations for the production of a next-generation reanalysis to support the reforecast generation process, as well as to facilitate other applications inside and outside of NOAA. The reanalysis configuration should match the operational data assimilation configuration as much as possible. The necessary preparations include determining the computational, storage, and personnel resources needed, as well as organizing the observational data that will be assimilated. The configuration details of the data assimilation system to be used in the reforecast should be decided in consultations between relevant NWS and OAR scientists. We assume that a future reanalysis will be ensemble-based, providing a number of initial analyses suitable for ensemble reforecast initialization.

(iii) Recommendation 3: NOAA should prepare to conduct a reforecast using the anticipated operational configuration of the GEFS. Based on the sample-size experiments and customer needs, we recommend the following configuration for a GEFS reforecast: 20 years, once every 5 days, with 5 members, and twice daily, from the 00 and 12 UTC cycle. This would be an extra 200 members computed every 5 days, compared with the operational 21x4x5 = 420 members computed in those 5 days, i.e., an extra ~50% computational expense. The discussion below will include possible ways to deal with this computational burden.

(iv) Recommendation 4: The GFS should have at least two years of retrospective forecasts computed for it prior to implementation, given that it is also expected to be post-processed by MDL and used for a variety of applications such as in the blender project. Since typically 6 months of prior forecasts are already computed for quality assurance, this request is rather modest. Skipping days between retrospective samples is acceptable.

(v) Recommendation 5: Given the requirements for NDFD guidance of surface weather elements at high (2.5-km) resolution, NCEP should devote the necessary resources to generate a high-quality retrospective analysis of surface weather with its Real-Time Mesoscale Analysis System.

b. Discussion.

We envision readers having several concerns about these recommendations, especially with regards to their computational expense. We try to anticipate these concerns and discuss them here.

First, won't it be a computational burden to continue running (a once-daily, 11-member) 2012-era GEFS, per recommendation 1? Let's consider the relative cost with respect to the next-generation ensemble system. The next-generation system is anticipated to be T574L64, with semi-Lagrangian advection vs. the 2012's T254L42 Eulerian advection. Let us assume that the T574L64 semi-Lagrangian version of the model is roughly 3.5 times more computationally expensive than the T254L42 Eulerian. *Hence, a once-daily, 11-member T254L42 GEFS for purposes of maintaining the reforecast database and serving existing customers will be only ~ 3.7 percent of the daily computational expense of running the 84-member (4 x 21 members) T574L64 ensemble.* There will be additional personnel and storage costs in archiving the data; these have not been estimated here.

The configuration recommended here is not optimal for all customers. Hydrologists and their customers in particular would prefer a reforecast extending over 30 years at least, every day, with at least 10 members. This is desired both for calibration and to validate the characteristics of streamflow forecasts over a large sample of high-impact cases. Pending more computer resources, NOAA might extend its reforecast from the more compact configuration recommended earlier to a more extensive reforecast, so as to meet the needs of the hydrologic community.

The computational expense of generating a reforecast in the suggested configuration (recommendation 3) was previously estimated to be an additional

~50 percent of the real-time GEFS expense. How can this be accommodated? In a scenario with loosely constrained CPU resources, NCEP would simply increase the allocation for the GEFS by 50%. However, there are other possibilities. Though NCEP has developed its models for a finite number of resolutions (T190, T254, T382, T574), intermediate resolutions are possible. Assuming a doubling of resolution increases computational expense by a factor of six for a semi-Lagrangian model, we can estimate, for a given planned forecast resolution, how much that planned resolution would need to decrease in order to accommodate the computation of ~50 percent more members. One can show that this is a factor of 1.17. Suppose the anticipated future resolution of a next-next-generation GEFS was T800. If the resolution were decreased to 800/1.17 = T683, this suggests that the additional reforecast members could be computed with roughly the same amount of CPU time. The improvements from reforecast post-processing are dramatic, much more dramatic than one would anticipate from such a modest change in model resolution. Such a recommendation assumes that terrain, land-use, and other files could readily be generated for T683 as well as T800, and that the model parameterizations can be readily adjusted to this resolution. These are likely to involve some person-months of effort to achieve.

There are yet other ways of potentially saving computational expense. Given the limited demands for off-cycle products, perhaps either the 06Z or 18Z cycles of the GEFS could be eliminated, or they could be run only to a shorter termination time, say 7 days instead of 16.

Another computationally attractive feature of the reforecast computation is that they need not be performed in real time, so they could potentially fill the holes in the "jigsaw puzzle" that defines NCEP/EMC's usage of the supercomputers. Further, while it is easier to perform the reforecasts on the production supercomputer, the reforecast computations could be done on NOAA's research computers, which are less reliable but also less expensive per CPU hour. If one staggers the reforecast computations, say computing the reforecasts for June during the month of March so that the June reforecasts are available far before the month of June for purposes of developing the post-processing methods, then the more-frequent outages of the research computer can be accommodated; there will still be time to make up for such outages.

The major computational expense not yet discussed is the expense of computing the reanalysis necessary for initialization. The existing reanalysis, the Climate Forecast System Reanalysis (CFSR; Saha et al. 2010) is now more than half a decade old, and there are many other needs beyond reforecast initialization for an updated reanalysis data set. The computational and personnel burden is likely to be significant, but this should be considered in the context of all the other users of the data set. Generating a useful reanalysis requires the scientists to deal with many challenging technical problems, such as changing satellites or changing biases for an existing satellite. These can cause the statistical characteristics of the reanalysis to change. Likely the reanalysis generation process will be somewhat iterative, learning about problems by generating a test data set, correcting those problems, and iterating until the reanalysis is judged to be of sufficient quality. It is hoped that because of the diverse set of users, several programs within the NWS and OAR would contribute to funding the generation of the next reanalysis.

Let's suppose that because of computational and personnel expense a reanalysis is regenerated only every ~5 years. Presumably the real-time analysis and forecast model used to generate its background will be updated more frequently, potentially changing the accuracy and bias of the analyses relative to reanalyses. What can be done, short of generating the reanalysis more frequently? This is an area that needs more thought and research. The Canadian Meteorological Centre (CMC) currently faces such a dilemma and are considering their options. They generate a reforecast similar to the recommended configuration here, but they initialize it from ECMWF reanalyses, which have different near-surface biases than in the Canadian system. CMC scientists are exploring whether the reforecasts' near-surface initial conditions should be adjusted so that their climatology is similar to that of the current operational analysis. Adopting some sort of similar procedure may be a practical necessity for NCEP as well should the computation of reanalyses be infrequent.

4. **Conclusions**.

This white paper has demonstrated the value of reforecasts, showed how their computational expense can be minimized, and provided recommendations for how NOAA should proceed with reforecasting. Reforecasting for the GEFS is possible for moderate computational expense. The associated reanalyses needed for the reforecast initialization, which we recommend the NWS proceed to generate, are also required by many other NOAA and external users.

While reforecasts do add computational burdens to NCEP, we hope the reader appreciates the underlying re-conceptualization of the NWP process implicit in these recommendations. *Post-processing needs to be thought of as an integral part of the NWP process*. These data sets are clearly critical to the NWS providing skillful, reliable guidance with its current prediction systems. The same attention to detail that NCEP has given to the development of its dynamical cores, its parameterizations, its assimilation methods, its ensemble prediction systems: these need to be extended to the production of data to support post-processing.

The recommendations herein are a starting point for an anticipated new journey together, with post-processing teams working in concert with the model developers. As with other components of the NWP system, we expect that there will be lessons learned with this suggested reforecast implementation. What we learn this time will make the path smoother in subsequent iterations.

This white paper focused on what should be done, not the specifics (e.g., funding, personnel, computing) of how to do it. The authors are willing to continue to assist with this should such help be desired.

Appendix:

The following are more detailed reports on sample-size sensitivity experiments, from co-authors:

(1) A report on sample-size sensitivity for precipitation type calibration (Philip Shafer, MDL)
[https://docs.google.com/a/noaa.gov/file/d/0BwC_uGfBYRDxS2s3ZXlvX1ZKbUU/edit?pli=1]

(2) A report on sample-size sensitivity for high wind speed and direction calibration (David Rudack, MDL)
https://docs.google.com/a/noaa.gov/file/d/0B6Gj5k7rrFjGcC0wdXFtZzA5eGs/edit?pli=1

(3) A slide deck on sample-size sensitivity for temperature calibration (Bruce Veenhuis, MDL)
https://docs.google.com/a/noaa.gov/file/d/0B3Sh4TjcoR__c29ibVVVVS1PQlE/edit?pli=1

(4) A report on sample-size sensitivity for CPC's 6-10 day and 8-14 day temperature and precipitation forecasts (Melissa Ou, CPC).
https://docs.google.com/a/noaa.gov/document/d/1WsJL_o-cSGHNJ5mU0ylVXwN5HaIxTl28VyInGiVP-0s/edit

(5) A report on sample-size sensitivity for probabilistic quantitative precipitation forecasts (Michael Scheuerer, ESRL/PSD).
https://docs.google.com/a/noaa.gov/file/d/0B9dAMg2-6U-7LUU0LVJueE4zU00/edit?pli=1

(6) A slide deck on NCEP/EMC's experimentation with reforecasts (Hong Guan, Bo Cui, and Yuejian Zhu, NCEP/EMC)
https://docs.google.com/a/noaa.gov/file/d/0B-G2MvTegPBnVmM2d1BCamh0QlU/edit?pli=1

References

Hagedorn, R., 2008: Using the ECMWF reforecast data set to calibrate EPS reforecasts. *ECMWF Newsletter*, **117,** 8-13.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132,** 1434-1447.

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.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.*, **134,** 3209-3229.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136,** 2620-2632.

Hou., D., M. Charles, Y. Luo, Z. Toth, Y. Zhu, R. Krzysztofowicz, Y. Lin, P. Xie, D.-J. Seo, M. Pena, and B. Cui, 2014: Climatology-Calibrated Precipitation Analysis at Fine Scales: Statistical Adjustment of Stage IV towards CPC Gauge Based Analysis. J. Hydrometeor. doi: http://dx.doi.org/10.1175/JHM-D-11-0140.1

Lalaurette, F., 2003a: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. Journ. Royal Meteor. Soc.*, **129,** 3037-3057.

Lalaurette, F., 2003b: Two proposals to enhance the EFI response near the tails of the climate distribution. 8 pp [Available online at www.ecmwf.int/products/forecasts/efi_guide.pdf].

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Royal Meteor. Soc.*, **127**, 2473-2489.

Saha, S., and co-authors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91,** 1015-1057.

Unger, D. A., H. van den Dool, E. O'Lenic, and D. C. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.*, **137**, 2365–2379. doi: http://dx.doi.org/10.1175/2008MWR2605.1