Addressing model uncertainty through statistical post-processing using reforecasts

Thomas M. Hamill

NOAA Earth System Research Laboratory
Boulder, Colorado, US

Abstract

Ideally, ensemble prediction systems for weather and climate will be deployed that produce unbiased, reliable ensemble predictions. It's desirable to achieve this directly by incorporating ensemble methods that sample the analysis uncertainty ensemble and by implementing physically based stochastic parameterizations, i.e., to get the right answer for the right reason. Practically, however, in the near term some bias and unreliability may be inevitable while ensemble prediction systems approach this ideal.

Statistical post-processing can improve ensemble forecast guidance prior to its dissemination. The real-time forecast is adjusted using discrepancies between past forecasts and observations (or analyses). Pending a suitably large sample of past forecasts ("reforecasts") from a stable modeling system, it is possible to dramatically improve forecast guidance, especially if there is significant model bias or lack of spread. The extended abstract below will provide some guidance on the benefits and drawbacks of using reforecasts and statistical post-processing. What niche should reforecasts should occupy in the weather and weather-to-climate prediction process?

1. Introduction

Reforecasts, also called "hindcasts," are data sets of past forecasts, ideally generated using the same modeling and data assimilation system as is used to generate the real-time forecasts. They are commonly used for statistical post-processing to correct modeling system inadequacies and for the generation of other forecast products.

The production of reforecasts is rather common with intraseasonal to seasonal climate forecasts, e.g., http://www.ucar.edu/yotc/iso.html. At long leads (months and beyond), most memory of the atmospheric initial condition is lost. There may be subtle shifts in the distribution of the climate from its long-term mean, but often the model forecast bias is larger than the magnitude of such shifts. A better estimate of the actual climate variability is possible if the projection is corrected for this bias. Unfortunately, it may require many decades of these reforecasts in order to provide a sufficient sample size to estimate the bias in the few remaining low-frequency modes that have some predictability; the reforecast computation is an expensive endeavor on top of an already computationally expensive forecast. Further, if the model changes, the data set should be regenerated; model biases may change from one model version to the next.

The focus for the rest of this paper is on the usefulness of reforecasts at weather and weather-to-climate timescales. Because of the computational expense, the reforecast process has only more recently been used. In the past, statistical corrections of weather forecasts were typically applied to deterministic forecasts and used whatever training data was at hand. Perhaps a year lapsed between significant changes to a numerical weather prediction (NWP) system. That year of forecasts and the associated observations was thus used as input to a Model Output Statistics (Glahn and Lowry 1972, Carter et al. 1989), a multiple linear-regression based correction scheme that would adjust the forecast, and in some instances even provide probabilistic guidance such as the probability of precipitation. If model bias was relatively consistent from one day to the next, then this modest training data set was most likely adequate.

However, there are two general situations where larger training data sets may be particularly helpful. The first is in situations where the predictable signal in a forecast is being overwhelmed by random errors. This may occur in many longer-lead weather-to-climate forecasts and certainly in climate forecasts. More samples are required to isolate the predictable component from the model bias and the chaotic noise. For these longer-lead forecasts where we seek to predict time averages, we also need independent training samples of those time averages. One training sample from 1-7 June and a second from 2-8 June won't be nearly as helpful as a second from 8-14 June. Even then, two such samples may have correlated errors due to both being drawn during a particular climate event, perhaps a La Niña. Hence, to ensure adequate sample size and independence of errors across samples, a multi-year training data set spanning many modes of climate variability is thus a practical necessity.

The second situation where larger training data sets are helpful is for rare events. The improvements in the skill of forecasting rare events with statistical post-processing using lengthy reforecasts is significant, often amounting to an increase of several days lead time (Hamill et al. 2006, Hamill et al. 2008). Why? If today's forecast is for heavy rainfall, say 50 mm in 24 h, it's unlikely that the forecast systematic error from, say, last week's 1-mm forecast will be relevant in diagnosing and correcting today's forecast systematic error. Similar 50-mm forecast events are needed. In areas where the terrain is relatively uniform, it may be possible that the model's systematic errors are similar 500 km to the west or east, and 50-mm forecast events can be found in the recent past at these locations. However, in many situations the model forecast bias may be very dependent on the local geographical features, and surrounding grid points will not be useful as training data. In these situations, it's very helpful to have a long time series with a broad range of past weather forecast events specifically at that grid point, some of which will resemble today's event.

Despite the appeal of reforecasts, their computational expense has kept them from being widely adopted in NWP. Are the benefits of reforecasts and statistical post-processing worth the expense for weather and weather-climate prediction? This extended abstract attempts to provide an honest evaluation of the benefits and drawbacks of addressing model uncertainty through the use of reforecast-based calibration. The abstract will indicate what reforecast data sets are or soon will be available for use, and suggest some of the most interesting and relevant research issues related to the use of reforecasts.

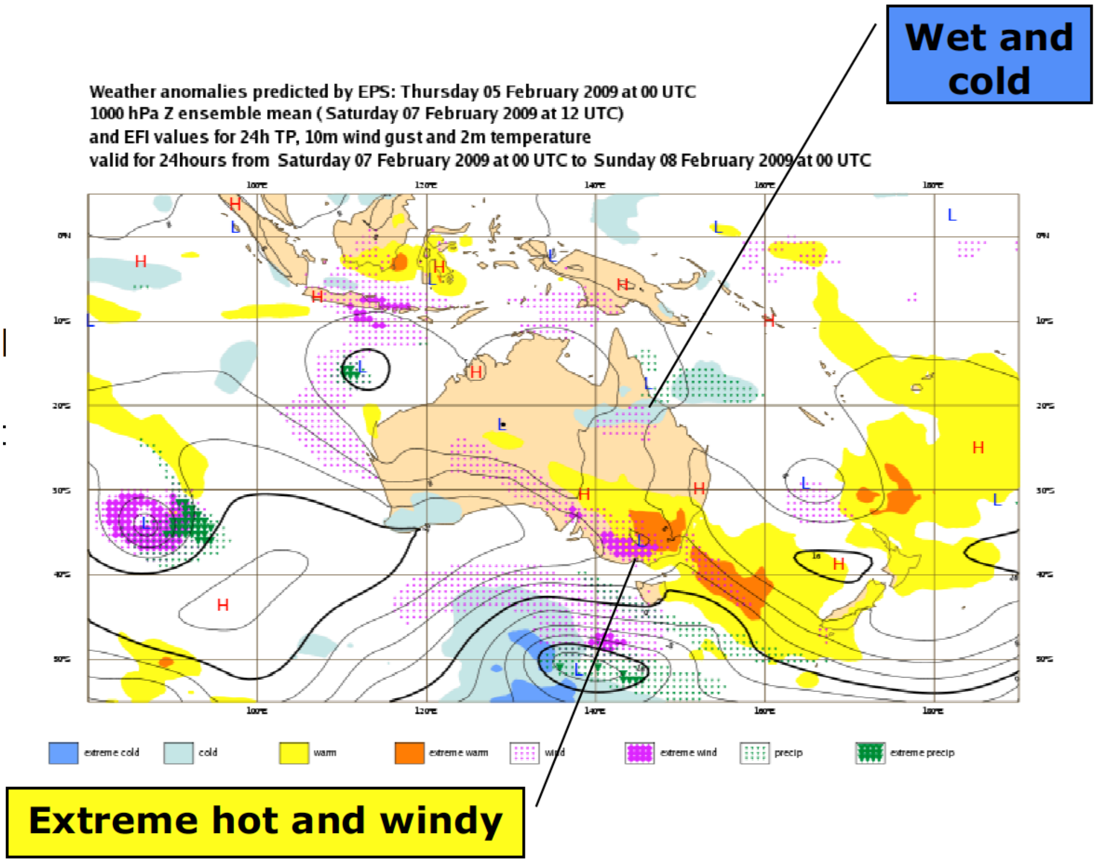
2. Benefits of adopting reforecasts in the numerical weather prediction process

2.1. Quantifying how unusual today's forecast is

Reforecasts can be useful to improving weather guidance even if they are not paired with observations and used in statistical post-processing. The reforecasts provide a relevant model climatology against which today’s ensemble forecasts can be compared. Assuming forecast and observed are correlated, if we can quantitatively assess how unusual today’s forecast is, we have some knowledge of how unusual the observed event may be. One method for doing this is the Extreme Forecast Index, or “EFI” (Lalaurette 2003a, 2003b). The EFI is calculated as

$$EFI = \frac{2}{\pi}\int_0^1 \frac{p - F_f(p)}{\sqrt{p(1-p)}}, \tag{1}$$

where p is the quantile of the cumulative distribution of the model climate estimated from the reforecast, $F_f(p)$ is how the p-quantile of the EPS ranks relative to the model climate (0 the minimum, 1 the maximum). Eq. (1) presents the “Anderson-Darling” version that introduces a weighted statistic that introduces more power in the tails of the distribution. 2/π is the normalization factor to keep $-1 \le EFI \le 1$. Figure 1 provides an illustration an EFI product from ECMWF around the time of the “Black Saturday” bush fires near Melbourne, Australia. The EFI depicts that the region near Melbourne is forecast to be both extremely hot and windy relative to the forecast climatological distributions. A complementary product to the EFI is to “Probability of Return” or “PRET” which uses the reforecasts and a fitted generalized extreme value distribution to estimate the return period for a forecast event (Prates and Buizza 2011).


Weather anomalies predicted by EPS: Thursday 05 February 2009 at 00 UTC
1000 hPa Z ensemble mean (Saturday 07 February 2009 at 12 UTC)
and EFI values for 24h TP, 10m wind gust and 2m temperature
valid for 24hours from Saturday 07 February 2009 at 00 UTC to Sunday 08 February 2009 at 00 UTC
100°E
120°E
140°E
160°E
180°E
0°N
10°S
20°S
30°S
40°S
50°S
H
L
extreme cold
cold
warm
extreme warm
wind
extreme wind
precip
extreme precip

Wet and cold

Extreme hot and windy

***Figure 1**: An example of the EFI forecast product created by ECMWF. Reprinted with permission from a presentation by F. Pappenberger, ECMWF, at the 2009 HEPEX workshop in Toulouse.*

Another example of how reforecasts can be used in the absence of observations is shown in Fig. 2. This shows a medium-range forecast of the frequency of tropical cyclogenesis in various basins. Many meteorological models may, for various reasons, over- or under-estimate the frequency of genesis. One may thus be left wondering whether the average 2.07 genesis events predicted in the Atlantic basin as estimated from the real-time ensemble is higher or lower than the climatological average. The orange bar on this graph shows that climatological forecast average, which is more than a factor of two smaller. Assuming that the ECMWF model has skill in forecasting genesis at these leads, this suggests that the genesis probability is much higher than average. Also note that the plot shows that forecast genesis frequency is lower than average in both the eastern and western Pacific basins.

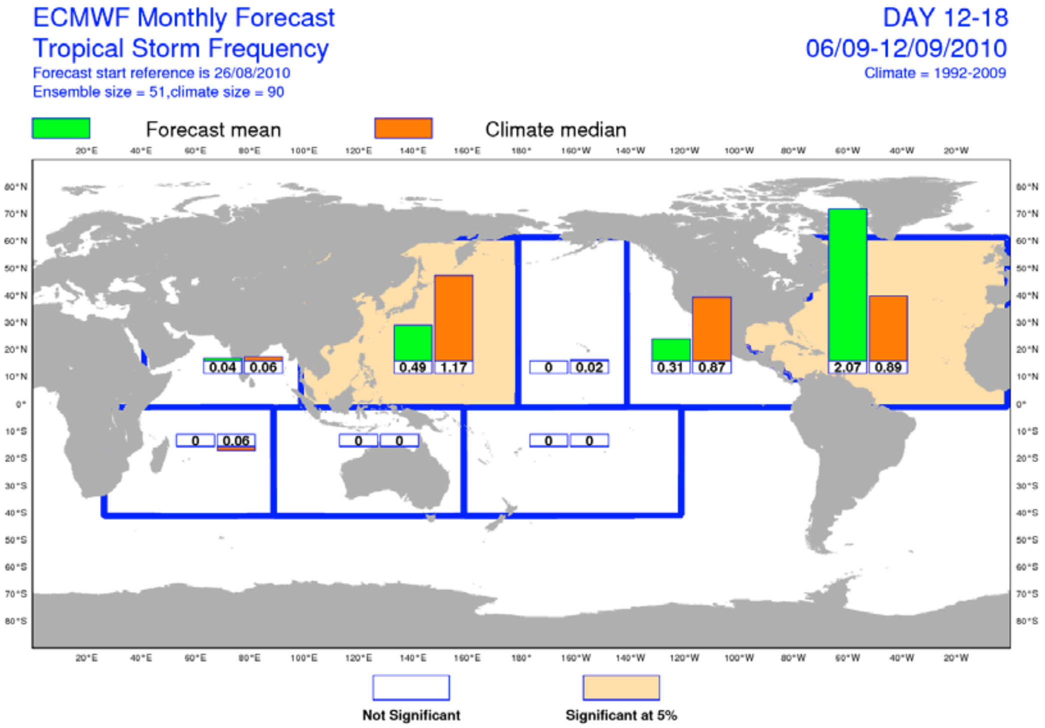


Figure 2: *An illustration of the use of reforecasts to provide context for the real-time forecast of tropical cyclogenesis. In this case, the real-time ensemble-average frequency of 12-18 day forecasts of genesis are shown in the green bars. The climatological forecast average for this time of year as estimated from the reforecasts are shown in the orange bars. Blue lines denote the boundaries of the basins under consideration. Reprinted with permission from D. Richardson, personal communication, ECMWF.*

2.2. Statistical post-processing

The value added by reforecast-based statistical post-processing has now been demonstrated over a variety of applications, including weather-climate forecasts (Hamill et al. 2004), heavy precipitation (Hamill and Whitaker 2006, Hamill et al. 2006, Hamill et al. 2008, Fundel et al. 2010), surface temperatures (Hamill and Whitaker 2007, Hagedorn 2008, Hagedorn et al. 2008), and streamflow (Werner et al. 2005). The author is beginning exploration of the use of reforecasts for wind-energy forecasting, tornado forecasting, and hurricane forecasting.

Below, Fig. 3 shows how the reforecast calibration procedure, via statistical downscaling, can add appropriate small-scale detail to precipitation forecasts. The observed rainfall in panel (a) is heavier along the Coast Range and the high Sierra Nevada mountains than it is in between, in the Sacramento Valley. The ensemble-mean forecast, here from the ~250-km US Global Forecast System model described in Hamill et al. (2006), is shown in panel (b). Given the coarse resolution of the forecast model, it misses all of the terrain-related detail. However, some of this geographic specificity is restored in the statistically post-processed forecast shown in panel (c). A logistic regression approach was used using reforecasts from past years with dates within a month of the date of the forecast. The high probabilities of heavy rainfall (greater than 50 mm 24h^{-1}) mostly correspond with the areas where heavy precipitation occurred. Hamill and Whitaker (2006) provides more detail on the forecast method and verification results, showing that the forecasts dramatically improved the skill and increased the reliability.

There are many other possible methods that could have been applied instead of logistic regression, and the development of new methods is an active area of research. The optimal calibration method may depend on the forecast variable in question and perhaps even the length of the reforecast (Wilks and Hamill 2007) and size of the ensemble. Several recently discussed methods include non-homogeneous Gaussian regression (Gneiting et al. 2005, Hagedorn 2008, Hagedorn et al. 2008), Bayesian Model Averaging, or “BMA” (Raftery et al. 2005, Hamill 2007, Sloughter et al. 2007, Wilson et al. 2007), Bayesian Processor of Forecasts (Krzysztofowicz and Evans 2008), and extended logistic regression (Wilks 1991).

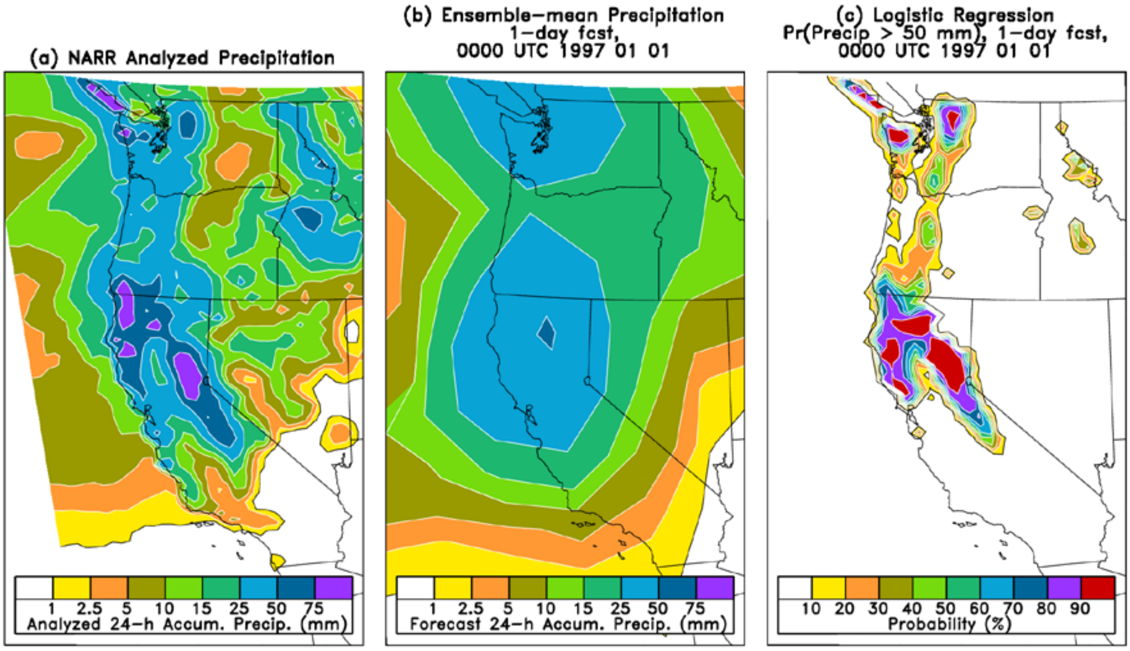

(a) NARR Analyzed Precipitation
1 2.5 5 10 15 25 50 75
Analyzed 24−h Accum. Precip. (mm)
(b) Ensemble−mean Precipitation
1−day fcst,
0000 UTC 1997 01 01
1 2.5 5 10 15 25 50 75
Forecast 24−h Accum. Precip. (mm)
(c) Logistic Regression
Pr(Precip > 50 mm), 1−day fcst,
0000 UTC 1997 01 01
10 20 30 40 50 60 70 80 90
Probability (%)

***Figure 3**: An example of the bias-correction and downscaling possible with reforecasts. (a) 24-h accumulated analyzed precipitation from the North American Regional Reanalysis, valid at 00 UTC 1 January 1997. (b) 15-member ensemble-mean forecast of precipitation amount from the NCEP GFS model. (c) 1-day forecast probability of greater than 50 mm of accumulated precipitation in the prior 24 h, valid at 00 UTC 1 January 1997. Probabilities were estimated with a logistic regression technique trained using GFS reforecasts.*

2.3. Hydrologic applications

The most ardent customers to date for reforecast data have been hydrologists, who use reforecasts to help produce quantitative probabilistic estimates of river streamflow that are as sharp and reliable as possible (Schaake et al. 2007). Reforecasts are valuable in at least two ways to hydrologists. First, hydrologists are often interested in improving forecasts for a particular basin with its own unique geographical features and perhaps unique weather forecast biases. Hence, they seek a long enough time series of reforecasts so that they can estimate whether the forecast input will produced reliable streamflow forecasts for the basin of interest over a range of conditions. Figure 4 shows an example of how reforecasts can illuminate possible biases in weather forecast inputs to hydrologic model. The river discharge for a basin in southern Switzerland is shown when the hydrologic model is forced with the reforecast model guidance and with observed fields. During peak runoff in May and June, there is a substantial over-forecast of runoff when forecasts are used relative to when observations are used. The large, 30-year sample of forecasts permits these biases to be examined over a range of conditions, from dry to wet years.

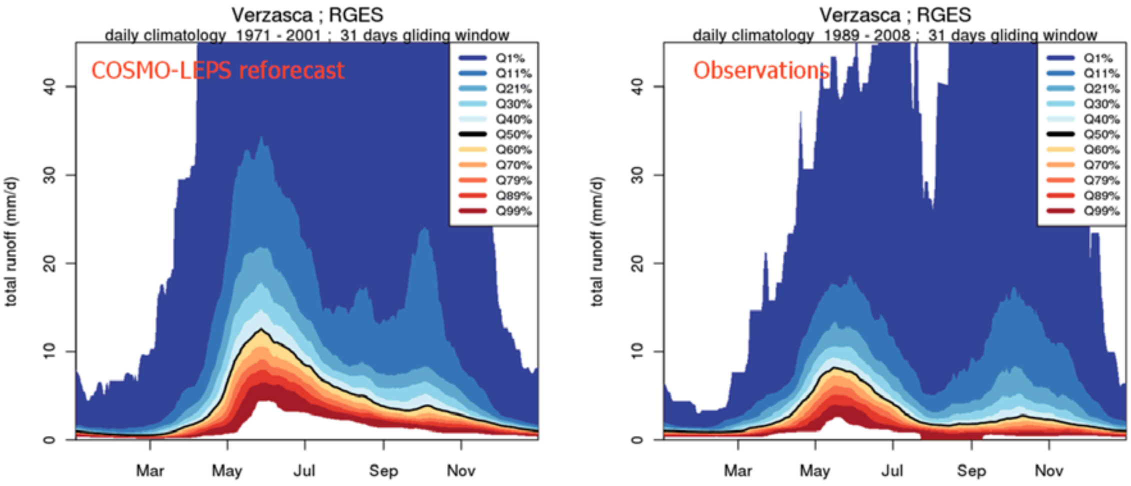


***Figure 4**: Left: Discharge climatology quantiles (30-day running mean) for the Verzasca basins obtained forcing the hydrological model PREVAH with COSMO-LEPS reforecasts (1971-2000). 18- to 90-h lead time forecasts were used. Right: Observed daily discharge climatology (1989-2008). Reprinted with permission from F. Fundel, Swiss Federal Institute WSL.*

A second primary hydrologic use of reforecasts is to provide calibrated precipitation forecast input to hydrologic ensemble prediction systems. Figure 5 illustrates a simple procedure for generating an ensemble of observed conditions that will provide calibrated weather forecast inputs to a hydrologic prediction system. These will be as reliable as when random samples from the observed climatology are used, but they will be much more specific to the weather of the day and hence may provide sharper, improved hydrologic guidance. At the top is the ensemble-mean forecast of interest on a given day. In a limited area, here consisting of an ~7.5-degree box around the area of interest, the root-mean-square fit of past forecast patterns to this day's pattern is evaluated, and the dates with the closest fits are chosen. The second row shows the closest four patterns, though in practice one may

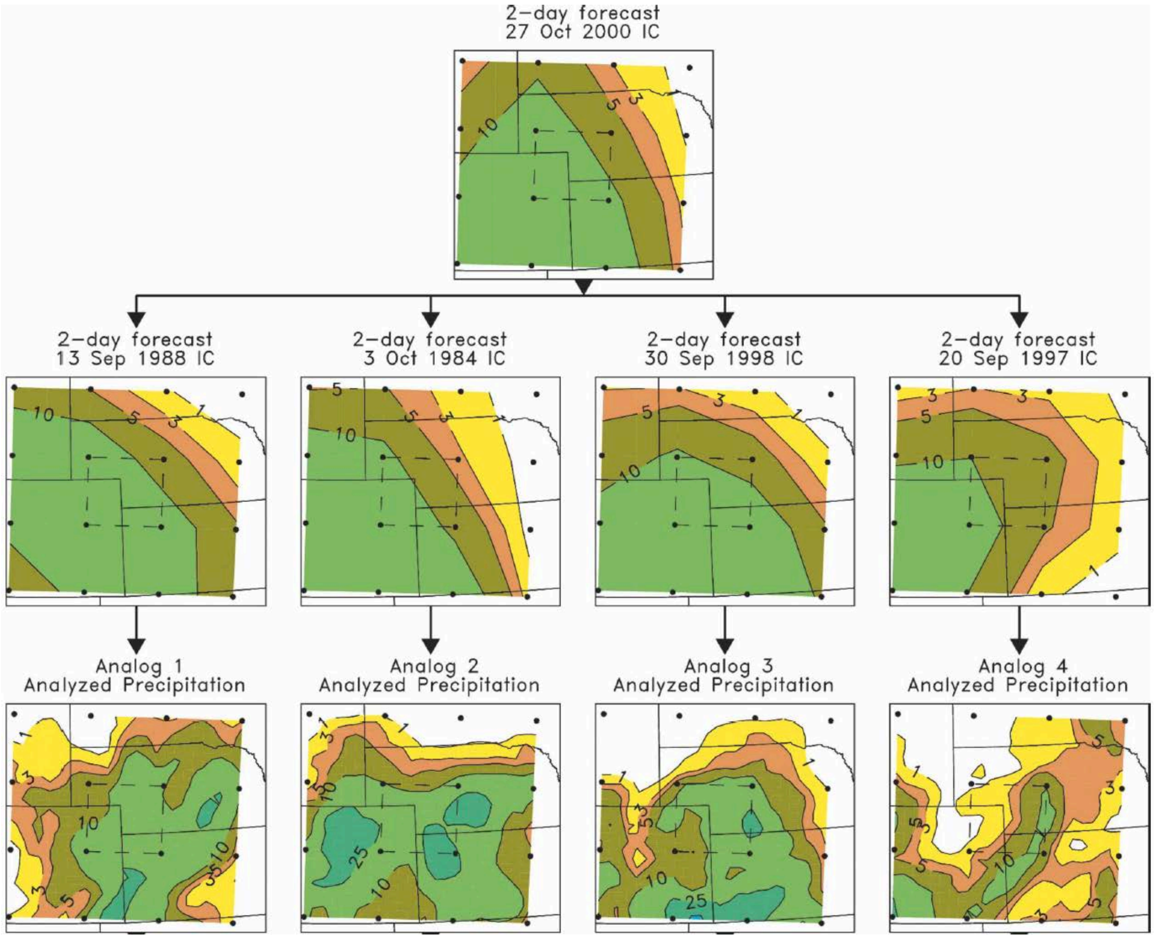


Figure 5: Illustration of the process of generating an ensemble of observed analogs using reforecasts. The top box denotes the ensemble mean forecast on a chosen day where we seek the conditional distribution of the observed given this forecast. The second row shows forecasts that are reasonable pattern matches as determined by comparing reforecasts at similar times of the year. The third row shows the analyzed precipitation on these dates.

choose to use many more patterns. An ensemble of the analyzed conditions is formed on the dates of those forecasts. These are shown in the third row. These could be used as atmospheric precipitation forcings to a hydrologic ensemble prediction system. These data represent a sample of the conditional distribution of the possible observed conditions that may be expected given today’s mean forecast. For more theory, background, and validation, see Hamill and Whitaker (2006).

2.4. Comparison with multi-model guidance

Could simpler methods that incorporate model uncertainty, such as multi-model ensembles, generate forecasts that are as good or better than those from a single model calibrated with reforecasts? Recent results suggest that reforecast techniques are competitive, though the amount of benefit achieved with reforecasts depends on aspects like the variable in question and the training sample size. Figure 6 shows the skill of 2-meter temperature forecasts over Europe from ensemble systems at NCEP, the

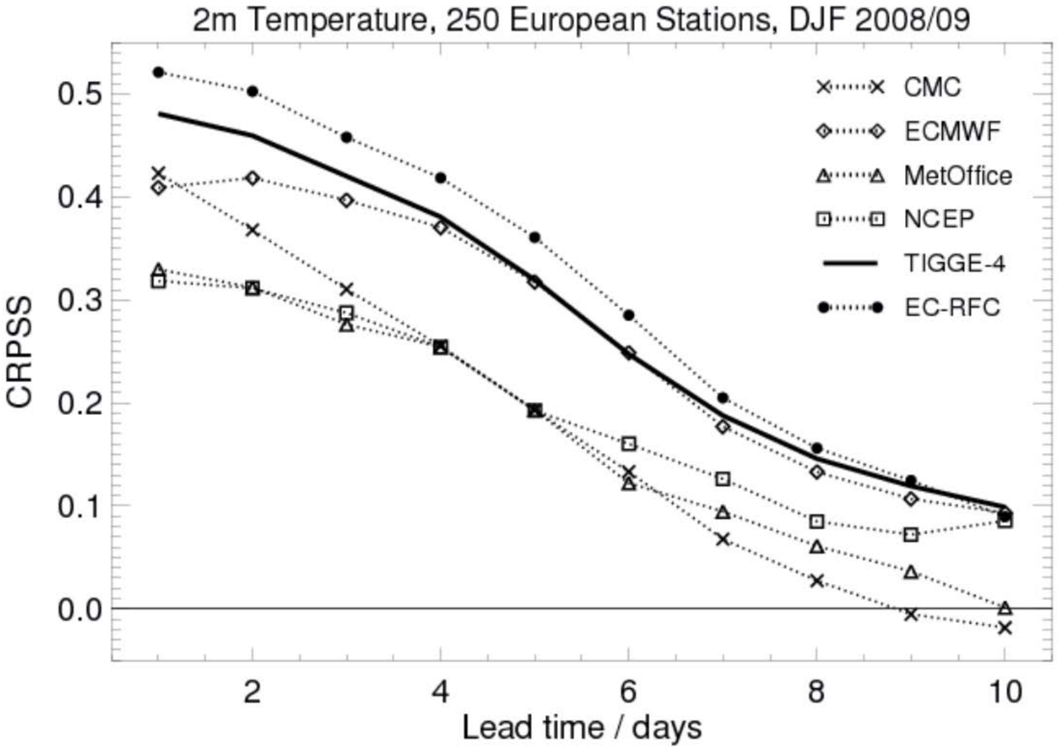


Figure 6: CRPSS 2-meter temperature forecasts at 250 European stations from four different operational centres. Also shown are the multi-model forecasts incorporating a bias correction ("TIGGE-4") and the ECMWF forecasts after statistical post-processing using their reforecasts ("EC-RFC"). Courtesy of R. Hagedorn, ECMWF and DWD.

Canadian Meteorological Centre (CMC), the UK Met Office, and ECMWF. The multi-model forecasts were extracted from the THORPEX Interactive Grand Global Ensemble (TIGGE) archive (Bougeault et al. 2010). Also shown is the skill of 2-meter forecasts from a multi-model combination that includes a bias correction based on the discrepancy at each station between the average forecast and observed during the past 30 days. Finally, the top line, "EC-RFC," shows the skill of ECMWF 2-meter temperature forecasts after their calibration using a blend of the reforecasts calibrated using non-homogeneous Gaussian regression and a simple bias correction using the last 30 days of operational forecast data (ECMWF generates once-weekly, 5-member reforecasts over the past 18 years; see Hagedorn (2008)). The reforecast-calibrated product has the highest continuous ranked probability skill score (CRPSS).

A recent study of precipitation forecasts produced more ambiguous results. Here, 20-member, 24-h accumulated forecasts of precipitation from NCEP, CMC, the UK Met Office, and ECMWF at 1-degree resolution were compared against precipitation analyses over the conterminous US for the period July-October 2010. 80-member multi-model ensemble forecasts were also evaluated, as well as a reforecast-calibrated ECMWF forecast, here trained using 2002-2009 precipitation analyses and the 5-member weekly reforecasts. To boost the sample size, a given grid point was trained with not only the forecast-analyzed data at that grid point, but using the data from 25 other nearby grid points that exhibited similar cumulative distributions of analyzed precipitation. Figure 7 shows the Brier Skill

Scores of these 10-mm forecasts. The skill of both the multi-model and the reforecast-calibrated ensembles were larger than those from the individual ensembles, and this was not a consequence of the larger-ensemble size. 20-member multi-model ensemble forecasts with 5 members from each centre had similar skill. The reforecast-calibrated ensemble was very similar in skill to the multi-model forecasts. Though not shown, applying a logistic-regression calibration to the multi-model ensemble using the last 30 days of forecasts improved the forecasts slightly on days 1-2 but degraded them for days 3-4. When examining reliability diagrams from the ensemble systems (Fig. 8), it's apparent that the reforecast-calibration process improved the reliability, but at the expense of sharpness. The multi-model products were sharper but not quite as reliable.

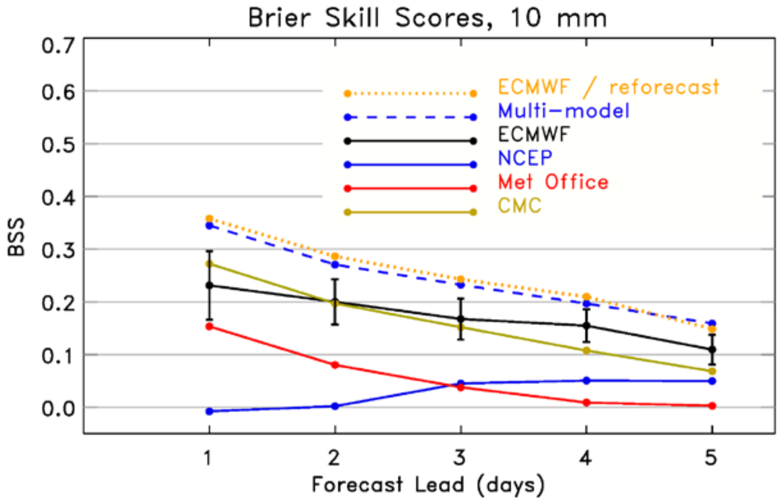


***Figure 7**: Brier skill scores of 24-h accumulated precipitation forecasts for the 10-mm threshold on a 1-degree grid over the CONUS. Confidence intervals represent the 5th and 95th percentiles generated from a paired block bootstrap algorithm between ECMWF and NCEP data, following Hamill (1999).*

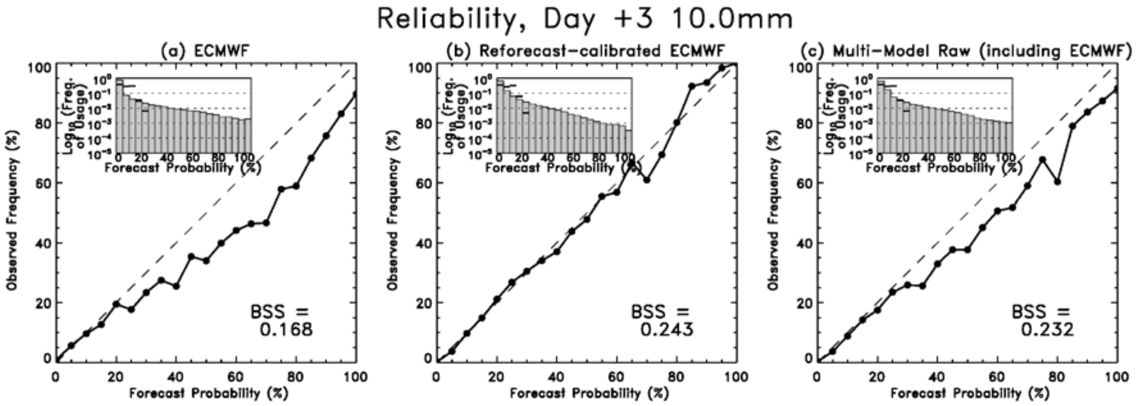


***Figure 8**: Reliability diagrams for 10-mm precipitation forecasts for (a) ECMWF 20-member ensemble; (b) the reforecast-calibrated ECMWF forecasts, and (c) the 80-member multi-model ensemble. The Brier Skill Scores for each are noted, and the bars on the inset histogram indicates how frequently each forecast probability was issued. The solid lines on the histogram indicate the climatological frequency distribution accumulated over all dates and points in the CONUS.*

The relatively similar performance of the multi-model and reforecast-calibrated approaches could be considered evidence for the value and importance of each. From the perspective of any one forecast centre, their forecast skill could be improved significantly were real-time data sharing agreements worked out between the centres. Then again, from the perspective of ECMWF, they have it within their control, simply by running a modest reforecast that added 1/7 computational expense to the ensemble suite production, to gain as much skill as would be gained through multi-model processing, without all of the vulnerabilities that might come from depending on other centre's data to arrive on time. In some sense, ECMWF regards their medium-range reforecasts as computationally free, since they are also used for the monthly forecast calibration process. Should ECMWF choose to compute even more reforecasts, the extra skill gained from post-processing should increase further, e.g. Fig. 7 from Hamill et al. (2006). Ideally, each centre would have its own reforecasts, permitting multi-model reforecast calibration (Whitaker et al. 2006).

3. Drawbacks of adopting reforecasts in the numerical weather prediction process

Above all, the biggest drawback of integrating reforecasting into the operations at NWP centres is the computational expense. While there are accumulating benefits from conducting more reforecasts (more years, more members, more days each year), the computational expense is of course proportional to each, i.e., a 20-year reforecast data set is twice as computationally expensive as a 10-year data set.

Let's consider the computational burden for reforecasts of varying lengths. Suppose, like ECMWF, you generated 50-member perturbed ensembles twice daily, a total of 700 members per week. Suppose now you generated a "small" 20-year, 5-member, once-weekly reforecast. This adds another 100 members per week, increasing your computational expense by 1/7, or ~14%. However, you may have customers such as hydrologists that seek an ensemble with more members, performed for a longer period, and performed daily. Suppose you generated a "moderate" once-daily, 10-member reforecast over a 30-year period. Each week you would have computed an extra 2100 members, increasing your computational expense by 300 percent. If you chose to generate a "large" twice-daily, 50-member reforecast over a 30-year period, this would have required the equivalent of an additional 21,000 members computed each week, a 3000-percent increase. The additional computational burden of 14% for the small reforecast was likely manageable, but 300-percent increase for the moderate reforecast would have needed significant scientific justification, and a 3000-percent increase for the large reforecast is clearly impractical for all centres, whatever the benefit.

What reforecast size provides an appropriate balance between the value gained from improved post-processing versus the increase computational burden? Unfortunately, this answer can be expected to differ for one forecast question to the next; a large reforecast may be helpful for a long-term forecast of heavy precipitation but not especially important for a short-term forecast of 500-hPa geopotential calibration. The answer may differ from one forecast system to the next, too; a poor system may benefit a lot from calibration, a good one much less so. Even the method of statistical post-processing may have some effect; some methods like logistic regression may be able to extract more skill from a given reforecast than perhaps an analog technique. While there have been several examinations of the impact of training sample size, more research is needed. Some examinations of forecast improvement versus training sample size have been presented in (Hamill et al. 2004, Hamill et al. 2006, Hagedorn 2008) though these explorations are incomplete at best.

Reforecasts also become less valuable if the statistics of forecast errors are not stationary, due for example to significant changes in the observational network or data processing. Figure 9 shows that forecast errors in ECMWF systems have changed markedly, both as a result of the improvements in the data assimilation and forecast systems and because of the changes in observing systems. With a multi-decadal reforecast data set, should one find a forecast case very similar to today's forecast weather, the errors from the past case are likely to be larger than the error in today's forecast, so statistical post-processing may result in an over-estimate of the actual error. Also, there may be approximations in the computation of the past forecasts that are not made for the real-time forecasts. Reforecasts are initialized from the "ERA-Interim" reanalysis, which uses a slightly older version of their forecast model and data assimilation system; however, the operational reforecasts use the same model used in the real-time forecasts. Further, they do not generate the perturbed initial conditions from by running the perturbed-observation 4D-Var that they use with the operational forecasts (Palmer et al. 2009). Instead, they apply the current year's perturbation to the past year's initial conditions, making them less relevant to the case-dependent analysis errors in the past forecasts. However, flow-dependent singular vectors are recalculated for the past reforecast dates (Isaksen et al. 2010).

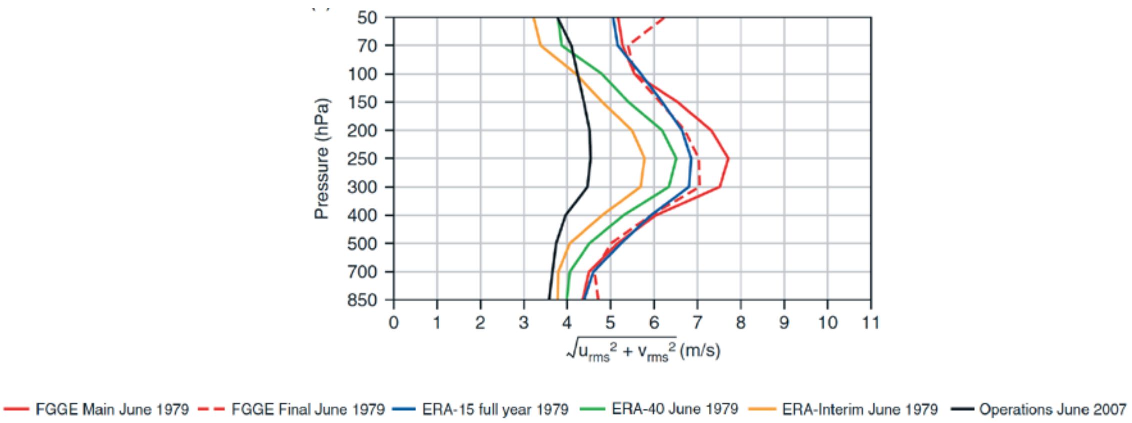


***Figure 9**: Globally averaged RMS errors in upper-air winds from short-range forecasts produced in various ECMWF reanalyses, relative to radiosondes observations. Data are from June 1979. For comparison, background errors in wind estimates from ECMWF operations for June 2007 are also shown. Reprinted with permission from Dee et al. (2011) and the Quarterly Journal of the Royal Meteorological Society (Wiley Press).*

A change in the observed climate may also compound the difficulties of using reforecasts. Anthropogenic global warming is causing more frequent extreme warmth and perhaps even more frequent extreme precipitation, more intense tropical cyclones, and such. Calibration will be more challenging if the past training data set is composed primarily of cooler, tamer weather and the real-time forecast toward forecasts record warmth or extremes of precipitation. Should one be applying, say, regression analysis to perform the statistical post-processing, today's forecast then may be outside the range of the training data, and one is forced into the undesirable situation of needing to “extrapolate the regression.” This may increase the error in the regression analysis, especially should the model biases in the warming climate differ substantially from those in cooler past regimes.

Obtaining long time series of observations or analyses may also be difficult, and without these, much of the benefit of the extended reforecast time series is lost. For example, the US has 4-km gridded analyses of precipitation available that may facilitate post-processing, but this data set only extends back to 2002. Even should one have a longer reforecast data set, it cannot be utilized in such a circumstance. This was the case for the precipitation post-processing example presented in Fig. 7 above.

Finally, not all types of forecast errors will be able to be reduced through statistical post-processing. Figure 10 shows probabilistic precipitation forecasts from an operational and an experimental ensemble prediction system. The leftmost panel shows the forecast from the ~30-km short-range ensemble forecast system (SREF) at NCEP, utilizing convective parameterization. The radar-generated analysis in the rightmost panel shows that the forecast badly misses most of the area of heavy precipitation. The forecast from a 4-km experimental “storm-scale ensemble forecast” (SSEF) system provides a much better forecast, presumably because of the use of explicit convection and other model improvements. The SREF system performed poorly in many other warm-season cases with weak forcing (not shown). If there isn’t some underlying correlation between forecast variables and observed amounts, the statistical post-processing will not be able to adjust the forecast so that it provides skill relative to a climatological forecast.

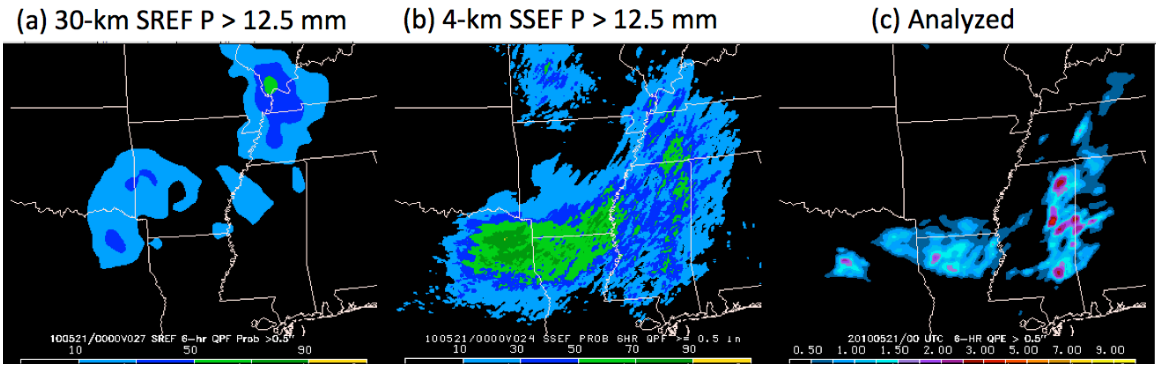

(a) 30-km SREF P > 12.5 mm
(b) 4-km SSEF P > 12.5 mm
(c) Analyzed
100521/0000V027 SREF 6–hr QPF Prob >0.5"
10
50
90
100521/0000V024 SSEF PROB 6HR QPF >= 0.5 in
10
30
50
70
90
20100521/00 UTC 6–HR QPE > 0.5"
0.50
1.00
1.50
2.00
3.00
5.00
7.00
9.00

***Figure 10**: 6-h to 12-h probabilistic forecasts of accumulated precipitation from the (a) operational SREF system, and (b) experimental SSEF system, valid 0000 UTC 21 May 2010. (c) Analyzed precipitation amounts. Domain is in the central southern US. Data provided courtesy of the NOAA NSSL/SPC Hazardous Weather Testbed.*

4. Current and pending reforecast data sets for weather and weather-climate applications

There are current and soon-to-be available reforecast data sets that may be of interest to those interested in statistical calibration of ensembles. These are summarized in Table 1. The ECMWF reforecast is described in Hagedorn (2008). Access for research purposes may be arranged by contacting Data.Services@ecmwf.int, http://www.ecmwf.int/products/data/archive/. The 1998 NCEP GEFS reforecast is described in Hamill et al. (2006) and data may be obtained from http://www.esrl.noaa.gov/psd/forecasts/reforecast/data.html. The 2012 NCEP GEFS reforecast is in production and should be publicly available by late 2011; for more details, contact tom.hamill@noaa.gov. The Atmospheric Technology Services Company (ATSC) reforecast is described in Marzban et al. (2010); data inquiries made to Fanyou Kong, fkong@ou.edu, or Vicki Rose, vicki.rose@atscwx.com. The COSMO-LEPS reforecast is described in Fundel et al. (2010) and can be obtained by contacting andre.walser@meteoswiss.ch.

Producer	# years	# members	Frequency	Realtime or offline?	Resolution	Forecast duration
ECMWF EPS	18 (1992-2010)	5	weekly	Real-time	T639, then T319 after 10 days	30 days
NCEP GEFS (1998 version)	33 (1979-2011)	15	daily	offline	T62	15 days
NCEP GEFS (2012 version)	32 (1980-2011)	10	daily	offline	T254, then T190 after 8 days	16 days
ATSC (WRF model over US)	19 (1987-2005)	10	Every 5th day	offline	15 km	48 h
COSMO-LEPS (Europe)	20 (1989-2008)	1	Every 3rd day	offline	~7 km	90 h

***Table 1**: A list of several current reforecast data sets.*

5. Future directions and conclusions.

While reforecasts have been used extensively for statistical correction of the forecasts, there are several other related applications where reforecasts are unproven but potentially very beneficial. The first is their use in data assimilation. An assumption that is frequently made but often difficult to ensure is that the prior forecast is unbiased. It is possible that some reforecast-based processing of the prior short-term forecast, using simple methods like those previously discussed or more complicated methods such as those discussed in (Li et al. 2009), can reduce bias and improve the data assimilation. Simpler methods may be improved with reforecasts, too, such as a bias correction to the prior based on a time series of analysis increments as discussed in Dee (2005) , section 3.b.v.

There are also many low-frequency modes of atmospheric variability where a short set of past forecasts may be insufficient for understanding the forecast characteristics. For example, the amount of Madden-Julian Oscillation (MJO) activity may vary by nearly an order of magnitude from year to year, and should one wish to understand the forecastability of the MJO, say, specifically when it is in the eastern Indian Ocean, then an extensive reforecast data set may be particularly helpful. The use of reforecasts for diagnosing model biases in such phenomena is relatively unexplored, though a limited set of hindcasts have been proposed for exploring the MJO at intraseasonal time scales (http://www.ucar.edu/yotc/documents/mjo/iso_**hindcast**_exp_plan_3.pdf)

Over the coming decades, we certainly hope that model systematic biases will be reduced enough that extensive processing with reforecasts will no longer be necessary. Experientially, however, systematic errors can be expected to contaminate NWP products for the foreseeable future. These errors may be due to aspects of the forecast models such as cloud microphysics that are either very difficult to obtain sufficient observations for to improve the methods, or which are computationally very expensive to correct. Hence, for the foreseeable future, we expect reforecasts and statistical post-processing to be a valuable part of the end-to-end forecast process, useful and sometimes crucial for delivering reliable forecast guidance.

References

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bulletin of the American Meteorological Society*, **91,** 1059-1072.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical Forecasts Based on the National Meteorological Center's Numerical Weather Prediction System. *Weather and Forecasting*, **4,** 401-412.

Dee, D. P., 2005: Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, **131,** 3323-3343.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, **137,** 553-597.

Fundel, F., A. Walser, M. A. Liniger, C. Frei, and C. Appenzeller, 2010: Calibrated Precipitation Forecasts for a Limited-Area Ensemble Forecast System Using Reforecasts. *Monthly Weather Review*, **138,** 176-189.

Glahn, H. R., and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, **11,** 1203-1211.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133,** 1098-1118.

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

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures. *Monthly Weather Review*, **136,** 2608-2619.

Hamill, T. M., 1999: Hypothesis Tests for Evaluating Numerical Precipitation Forecasts. *Weather and Forecasting,* **14,** 155-167.

——, 2007: Comments on "Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging". *Monthly Weather Review,* **135,** 4226-4230.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation. *Monthly Weather Review,* **136,** 2620-2632.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly Weather Review*, **134,** 3209-3229.

——, 2007: Ensemble Calibration of 500-hPa Geopotential Height and 850-hPa and 2-m Temperatures Using Reforecasts. *Monthly Weather Review*, **135,** 3273-3280.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An Important Dataset for Improving Weather Predictions. *Bulletin of the American Meteorological Society,* **87,** 33-46.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Monthly Weather Review*, **132,** 1434-1447.

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, and L. Raynaud, 2010: Ensemble of data assimilations at ECMWF, 45. pp.

Krzysztofowicz, R., and W. B. Evans, 2008: Probabilistic Forecasts from the National Digital Forecast Database. *Weather and Forecasting,* **23,** 270-289.

Lalaurette, F., 2003a: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quarterly Journal of the Royal Meteorological Society*, **129,** 3037-3057.

——, 2003b: Two proposals to enhance the EFI response near the tails of the climate distribution. 8.

Li, H., E. Kalnay, T. Miyoshi, and C. M. Danforth, 2009: Accounting for Model Errors in Ensemble Data Assimilation. *Monthly Weather Review*, **137,** 3407-3419.

Marzban, C., R. Wang, F. Kong, and S. Leyton, 2010: On the Effect of Correlations on Rank Histograms: Reliability of Temperature and Wind Speed Forecasts from Finescale Ensemble Reforecasts. *Monthly Weather Review*, **139,** 295-310.

Palmer, T. N., and Coauthors, 2009: Stochastic parameterization and model uncertainty. *ECMWF Tech Memo 589*.

Prates, F., and R. Buizza, 2011: PRET, the Probability of RETurn: a new probabilistic product based on generalized extreme-value theory. *Quarterly Journal of the Royal Meteorological Society,* **137,** 521-537.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review,* **133,** 1155-1174.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX: The Hydrological Ensemble Prediction Experiment. *Bulletin of the American Meteorological Society*, **88,** 1541-1547.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, **135,** 3209-3220.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay, 2005: Incorporating Medium-Range Numerical Weather Model Output into the Ensemble Streamflow Prediction System of the National Weather Service. *Journal of Hydrometeorology*, **6,** 101-114.

Whitaker, J. S., X. Wei, and F. Vitart, 2006: Improving Week-2 Forecasts with Multimodel Reforecast Ensembles. *Monthly Weather Review*, **134,** 2279-2284.

Wilks, D. S., 1991: Representing Serial Correlation of Meteorological Events and Forecasts in Dynamic Decision–Analytic Models. *Monthly Weather Review*, **119,** 1640-1662.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Review*, **135,** 2379-2390.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging. *Monthly Weather Review*, **135,** 1364-1385.