Consensus Probabilistic Forecasting

Bill Myers and Barbara Brown
National Center for Atmospheric Research
Boulder Colorado, USA

1. Introduction

To make optimal decisions, end-users of decision support systems require information accurately describing the uncertainty of the underlying weather forecasts. Air temperature, dew point temperature, and wind speed are critical surface weather variables in many economic sectors. The generation of sharp and calibrated probabilistic forecasts and their effective presentation to decision makers are current research challenges. This paper addresses the first of these challenges by describing an operational probabilistic forecast system developed at NCAR/RAL. This system is a probabilistic extension of DICast, an automated consensus forecast system which serves as the operational backbone for several major US weather providers.

Today, forecasts of these continuously-valued weather surface variables are commonly generated and presented as deterministic (scalar) values. For example, the maximum temperature 2 days from now will be (exactly) 25ºC. A more complex forecast representation is required describe the uncertainty in the forecast. Rather than forecasting a scalar, a probabilistic forecast ideally takes the form of a probability density function (pdf). Ensemble methods provide a natural approach for creating these types of forecasts. However, to create meaningful forecasts using ensemble methods generally requires production of a large number of realizations of a model forecast, which can be expensive in time and other resources. Moreover, calibration of the ensemble forecasts is often a concern.

The DICast probabilistic forecast system considers multiple numerical weather model inputs and uses a multi-model or “poor man’s” ensemble approach. Each model is interpreted statistically to generate individual pdfs for the variables of interest (e.g., temperature). The system then combines the resulting forecast distributions using weights based on the past forecast performance of each of the models’ pdf forecasts. The resultant consensus forecast is again a pdf. This weighting procedure allows generation of multimodal forecast distributions.

The main conceptual difference between this probabilistic forecast system and the “traditional” scalar DICast system is that the probabilistic system produces and combines pdf’s rather than scalars. It is well understood that the combination of scalar forecasts produces statistically superior forecasts. The goal of this paper is to demonstrate that the same can be true for probabilistic forecasts.

2. Data

This study used Eta and GFS numerical weather prediction (NWP) model data generated at NCEP. For both of these models, only the 12Z forecast cycle data was considered. With computational and communication latency, the forecasts generated would have been available at roughly 18Z (1300 EST) on any particular day. Raw model output from June 25, 2002 through May 23, 2003 was used in this study. Probabilistic forecasts were produced, based on the model data, from October 3, 2002 through May 23, 2003. These forecasts were produced for 18 cities spread across the United States. The observing station used for verification of forecasts at each of these cities was the METAR at one of the city's main airports. These cities are listed in Table 1. For each city on each day, probabilistic 2m air temperature forecasts were produced out to 60 hours at 3 hour intervals. At each lead time (e.g. the 9 hour forecast, valid at 21Z), there were a total of 4194 forecasts, that is, 233 forecast days at each of the 18 sites.

Table 1

City	**METAR**	**City**	**METAR**
Atlanta	KATL	Minneapolis	KMSP
Boston	KBOS	New York	KLGA
Chicago	KORD	Oklahoma City	KOKC
Cincinnati	KCOV	Philadelphia	KPHL
Dallas	KDFW	Phoenix	KPHX
Goodland	KGLD	Portland	KPDX
Houston	KIAH	Sacramento	KSAC
Kansas City	KMCI	Sioux Falls	KFSD
Los Angeles	KLAX	Washington, DC	KDCA

3. Forecast Generation Method

DICast is a two step forecasting system. First, statistical techniques are used to generate forecasts based on the output from individual NWP models. These statistical methods are a type of updateable MOS in which regression equations are formulated based only on recent model data and observations. Once all these Dynamic MOS (DMOS) forecasts from individual models have been generated, DICast's second step is applied. This integration step uses a fuzzy logic approach to combine the statistically generated forecasts. This combination attempts to produce an optimal consensus forecast.

Analyses of regression-based forecasts of air temperature, dew point temperature, and wind speed have indicated that the errors are approximately normally distributed. Thus, normal distributions were used as a template for the pdf forecasts. The DMOS regression equation's predicted value was used as the normal distribution's mean. The distribution's variance was derived from the equation's predictive variance. This variance depends on the variance of the fitted equation as well as the values of the current day's model output. The variance is naturally smaller near the mean of the data used to develop the regression equation. It increases when applied to data further from the mean of the equation development data set. This aspect of the pdf forecast generation is intuitively pleasing.

However, it is not clear that this estimate of the variance from a single regression equation is the best choice. One would expect that the best variance to use would lie between zero (a deterministic forecast) and the climatological variance. This study evaluates three normally distributed pdfs with the same mean in order to verify the validity of the choice of the predictive interval variance (σ_p) over the alternatives σ_0 and σ_c. That is, these two alternative forecasts are used as standards of comparison. The predictive interval variance typically lies between the two extremes during the 60-hour forecast range of this study. This provides us with an idea of which variance provides the "best" normally-distributed pdf forecast from a single NWP model.

The second step of the DICast forecast process combines the pdfs generated from the individual models into a final, integrated forecast. This combination is performed using a simple weighted sum of the individual pdfs (e.g., Figure 1). The weights used in this study were generated based on the relative skills of the means of the pdfs. Clearly this approach does not take into account the pdfs' spread. Weights should probably be calculated based on the performance of the entire pdf rather than just its mean. The weight generation process is currently being upgraded to generate optimal weights based on minimization of the forecasts' Continuous Rank Probability Score (CRPS). The integrated pdf forecasts generated using this suboptimal weighting scheme are compared to the individual models' pdfs. This part of the study focuses on determining whether the combination of probabilistic forecasts provides a superior forecast.

4. **Verification Methods**

Verification of pdfs is inherently difficult. No probabilistic forecast can be "correct" unless it is deterministic, that is, all of the pdf's mass lies on a single point and the observation exactly matches that value. Instead, pdf verification must be based on a family or collection of probabilistic forecasts all made by the same method or technique. Only by evaluating the whole set of a family of forecasts can the characteristics of that forecast technique be determined.

Further complexities arise in the comparison of forecast generation techniques. Probabilistic forecasts have many verification facets and determination of which forecast technique is superior is generally not completely clear. The forecasts generated by one technique may be better than another according to one verification measure but not another. Thus, the decision as to which technique is superior usually comes down to end user criteria, and this study looks at a variety of verification measures to provide an overall assessment of the quality of the forecasts.

To more fairly evaluate forecasts from sites with differing climatologies, the pdfs were transformed into a "climatologically-normalized" space. That is, each pdf was transformed so that so that its native units were climatological standard deviations relative to the climatological mean. The climatological means and standard deviations applied to a pdf were site, time of day, and seasonally specific. This leads to a pdf with an x-axis in units of σ_c. This approach turns out to be exactly what was recommended in a recently submitted paper by Tom Hamill and a co-author.

The DICast pdf forecasts are evaluated using standard metrics such as the CRPS and its decomposition elements. The normalization process described above also allows an evaluation of performance for a variety of "events" ranging from extreme to near normal (seasonal). Rank histograms provide other insights into the forecast quality. Reliability diagrams and ROC plots were also examined but discussions of these are not included in this paper.

5. **Results**

Results are presented for air temperature forecasts. Results for dew point temperature, maximum temperature, and minimum temperature were similar.

5.1 **Single Model pdfs**

The goal of this part of the study is to examine the performance of the individual probabilistic forecasts from one NWP model.

Four normally distributed pdfs are compared in this part of the study. Three of these are generated from the regression equation. These three use the mean from the regression equation and have different variances. The fourth pdf is a climatological distribution. The distributions will be referred to as follows:

1. Predictive Interval Forecast $\sim N(\mu_p, \sigma_p)$
2. Deterministic Forecast $\sim N(\mu_p, \sigma_0)$
3. Regression Mean, Climatological Variance Forecast (RMCV) $\sim N(\mu_p, \sigma_c)$
4. Climatological Forecast $\sim N(\mu_c, \sigma_c)$

The CRPS scores for the Eta model forecasts are shown in Figure 2. The Predictive Interval forecast is clearly superior to the other forecasts at all lead times. In fact, its CRPS is 15-20% lower at every lead time. The y-axis of this score has units of σ_c. This indicates an average forecast improvement that can be translated into degrees at any particular site and lead time.

The CRPS can be thought of as an integral of the Brier Score across all possible thresholds. It is possible to understand where this improvement in CRPS is coming from by examining the Brier Scores at all thresholds (Figure 3). It can be seen that the Predictive Interval forecasts are much better for typical events (i.e. near the climatological mean). However, the improvement in CRPS is not only for common events at the expense of rare events. The Predictive Interval method does at least as well as or better than the other forecast techniques for unusual events, including those more than $2\sigma_c$. from the climatological mean.

The Brier Scores can be decomposed into the resolution, reliability, and uncertainty terms. These decomposition elements allow further examination of the characteristics of the forecasts. Figure 4 and 5 show the positively oriented resolution verification measure. These results again indicate that the Predictive Interval forecasts have better performance at all lead times. The improvement is about 8-10% better than the nearest competitor, the RMCV forecasts. Thus, the predictive interval method is better able to separate the observations away from the climatological mean. Again, the largest improvement is seen for near climatological events but resolution is not sacrificed for extreme events.

The reliability decomposition shows similar results (Figures 6-7). This measure is negatively oriented and shows similar results for the three forecast pdfs centered on the regression mean. As expected, the pure climatological forecast is highly reliable. Interestingly, the Deterministic forecast has the best (lowest) reliability of the three alternative forecasts. This result is largely due to the aggregation of forecasts at 0% and 100% for the Deterministic forecast. The Predictive Interval forecasts at a particular threshold are more evenly spread throughout the [0-1] interval. This typically causes larger errors within each forecast bin.

While the reliability of a forecast is correctable using statistical techniques, lack of resolution cannot be corrected. The Deterministic forecast may have better CRPS reliability scores but its resolution is quite poor. The Predictive Interval's large advantage in resolution must be considered as a major factor when ultimately choosing a best single model forecast technique.

Results for the GFS model forecasts are not shown but were quite similar. Also, although only the 12 hour forecast Brier Score and decompositions are shown, similar results could be seen at all other lead times.

5.2 Pdf combination

The second part of the study examines the performance of the combined forecasts, to evaluate whether the combination of "best" single model forecasts leads to an improved forecast. First we note that the integrated (combined) forecast leads to a more level rank histogram. The GFS and Eta forecasts are biased and underdispersive. This bias may be due to a seasonal lag inherent in the DMOS regression equation generation. The U-shape in these two rank histograms indicates that the distributions generated are not broad enough. That is, the observation falls in the tails of the distribution somewhat too often. While not perfect, the integrated forecast corrects these problems. It is less biased and better dispersed, as can be seen by its relative flatness. See Figure 8.

The CRPS values for these forecasts at each lead time can be seen in Figure 9. The integrated forecast's CRPS is better (lower) than either of the two predictive interval forecasts at every lead time. Though comparable with the Eta model at analysis time, the integrated forecast's CRPS soon attains and maintains an 8%-12% improvement over the best of the others across the rest of the forecast period. As can be seen in Figure 10, the Brier Scores are again best for near climatological events for all 3 forecasts. The integrated forecast gains its edge in this type of event. For rare events, the Brier Scores are similar.

The integrated forecasts' CRPS reliability, as seen in figure 13, is better than either the Eta or GFS predictive interval reliability at every lead time. Beyond 6 hours into the forecast, the relative improvement is typically more than 10%. Meanwhile, the integrated forecasts' resolution, shown in figure 11, is better than the others at every lead time except the analysis time where it is effectively the same as the Eta. The improvement is for the most part between 3%-5% at all lead times. This effectively shows that the integration step can be used to simultaneously improve both reliability and resolution of the forecasts at all lead times.

Breaking down the reliability (resolution) by variance, Figures 12 and 14, it can be seen that most of the gain in the overall CRPS reliability (resolution) is due to performance for the near-climatology events. For the extreme events, the integrated forecast is "best" at most but not all thresholds. For these rare events, the differences in the reliability (resolution) term are quite small.

6. Conclusions

Evaluation of the pdfs generated by these models indicates that they provide relatively reliable and skillful forecasts when compared to the deterministic forecasts and simple probabilistic forecasts based on climatology. In addition, the rank histograms indicate that the forecast spread is generally approximately correct. Integration of the pdfs provided by the two forecasting systems results in notable improvements, particularly with respect to the resolution and reliability of the forecasts. This improvement is associated with the ability of the integration process to apply larger weights to the forecast model that is providing the best performance.

Results of applying this method for temperature forecasts based on only two NWP forecasts are encouraging; use of additional models or model realizations would be expected to show additional capabilities. Evaluations of forecasts for non-temperature variables such as wind speed are in progress; initial results indicate that application of this approach is also beneficial for these forecasts.

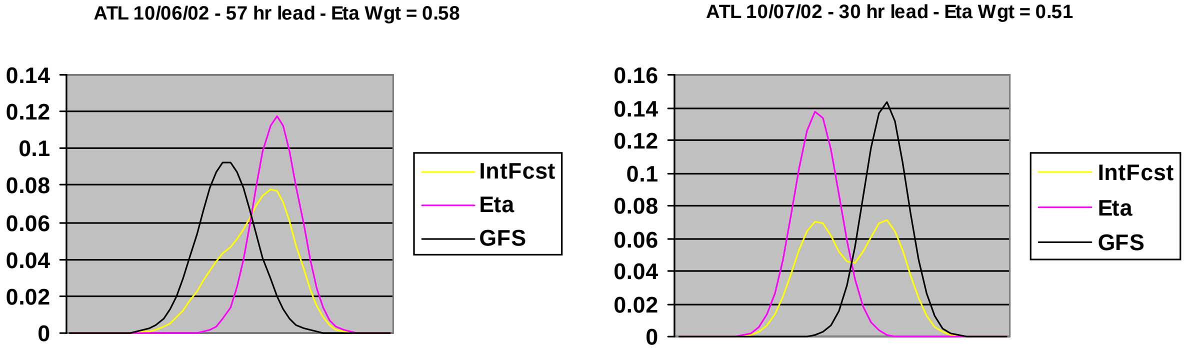


Figure 1: Two examples of pdf combination. Usually, the resultant forecast was unimodal as in the left hand case. However, multimodal forecasts (right) are possible.

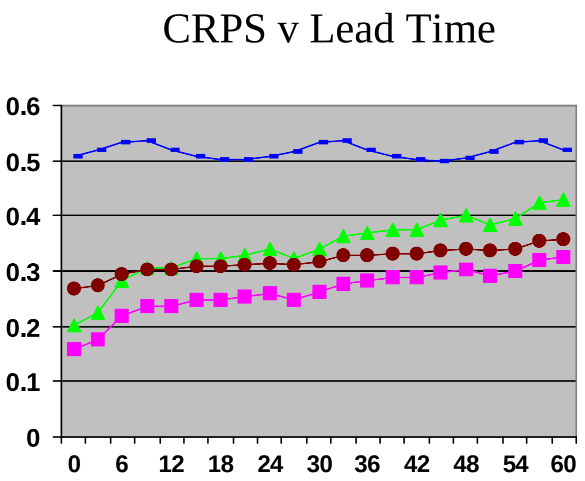

CRPS v Lead Time

Figure 2: Continuous Rank Probability Scores aggregated over all sites. The Predictive Interval forecast produces the lowest CRPS at all lead times.

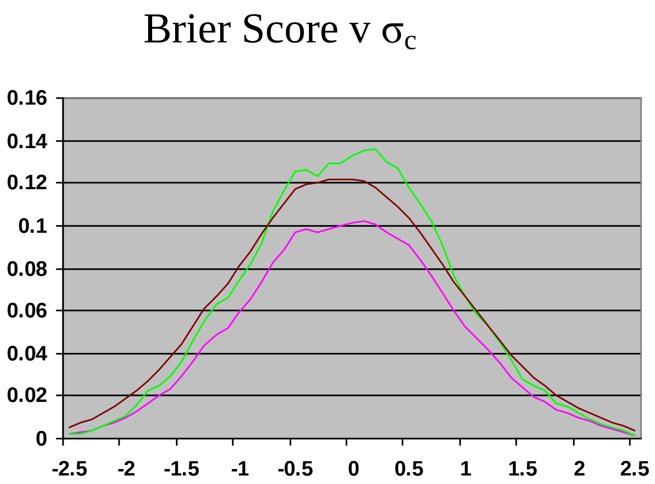

Brier Score v σ_c

Figure 3: The Brier Score from the 12 hour forecast shown in Figure 2.Note that the Predictive Interval Forecasts provide superior forecasts for near climatological predictions and forecasts that are at least as good as the components for more extreme events.

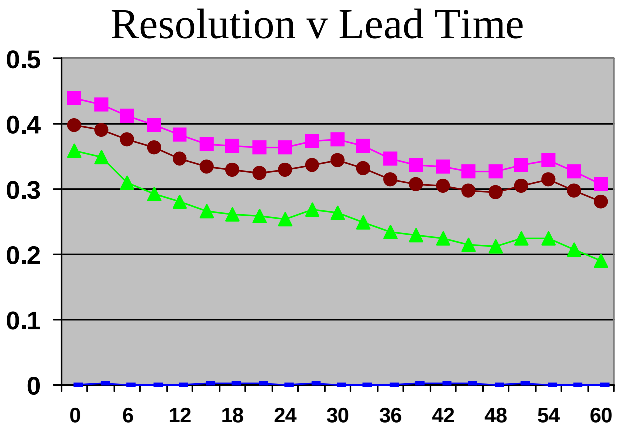

Resolution v Lead Time

Figure 4: Resolution at all lead times. The Predictive Interval has the highest (best) resolution at all lead times.

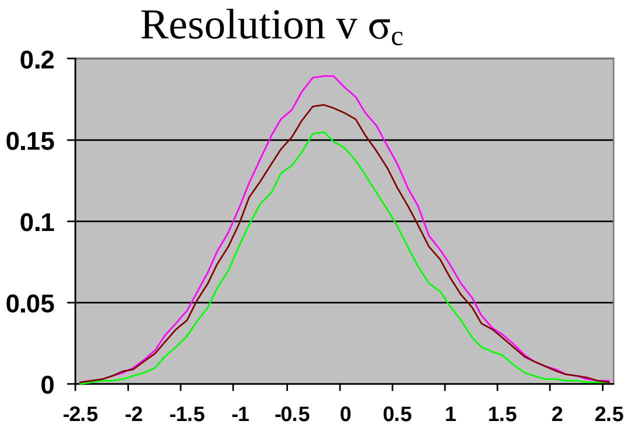
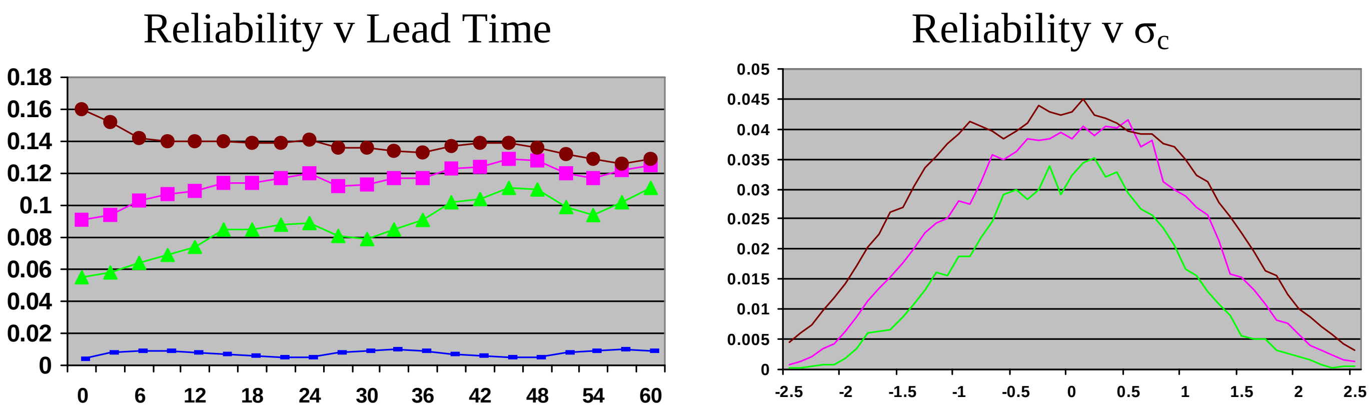

Resolution v σ_c

Figure 5: Resolution for the 12 hour forecast at all thresholds. The Predictive Interval has the highest (best) resolution for near climatological events. It is similar for rare events.

■ Pred Interval
▲ Deterministic
● RMCV
▪ Climo



Figures 6-7: The Brier score reliability component plots show that the climatological forecast has very good (low) reliability as would be expected. The Deterministic forecast has the best reliability of the other 3 forecasts.

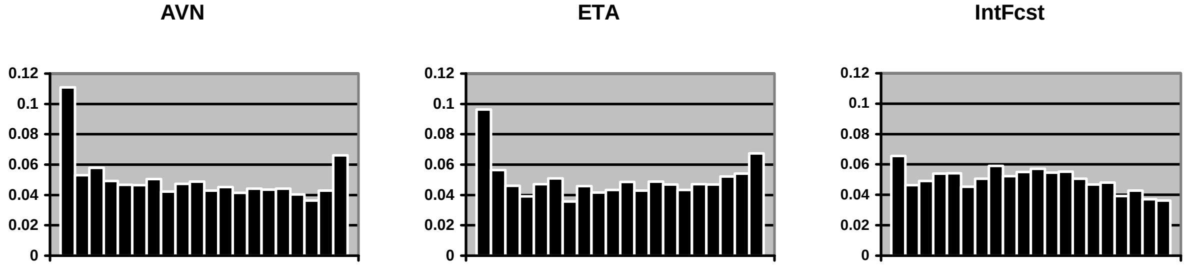


Figure 8: Rank Histograms for the AVN (GFS), Eta, and Integrated Forecasts.

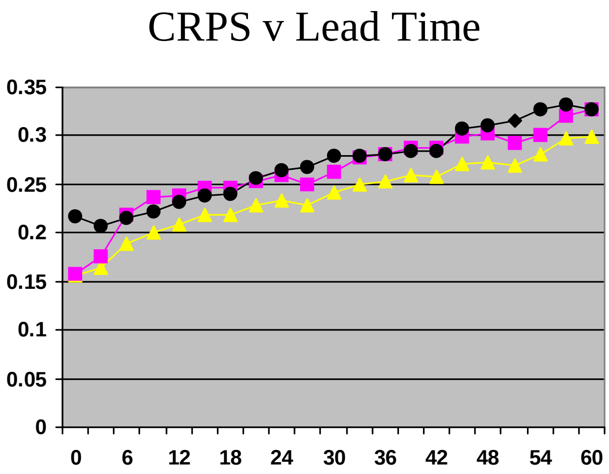


Figure 9: CRPS at all lead times. The integrated forecast is better than either of its two constituents throughout the forecast period.

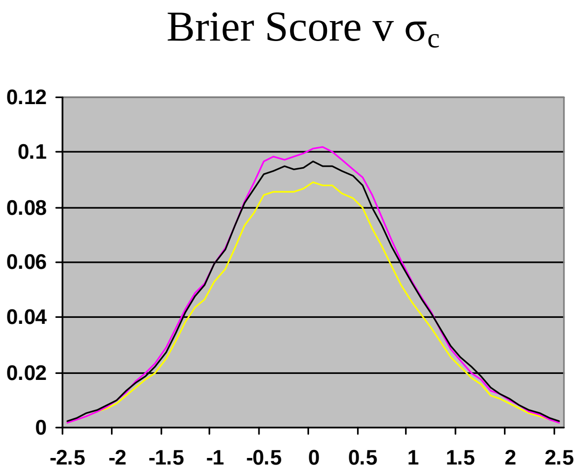
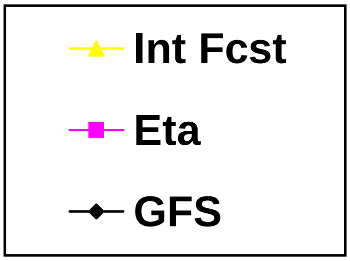


Figure 10: The Brier Scores making up the CRPS at a 12 hour lead time. The integrated forecast is clearly better for common events. For extreme events, its performance is similar to the others.



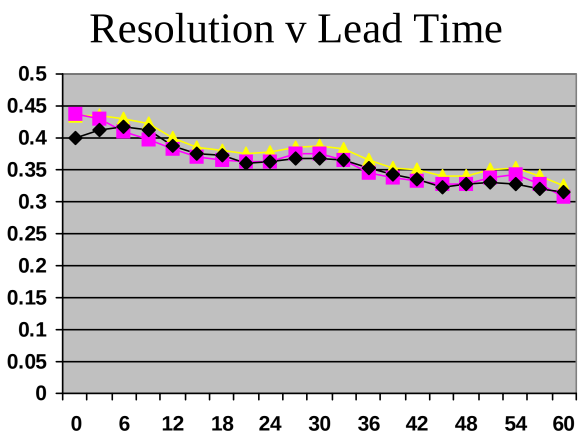


Figure 11: CRPS resolution. After hour six, the integrated forecast is slightly (3-5%) better than the other 2 forecasts.

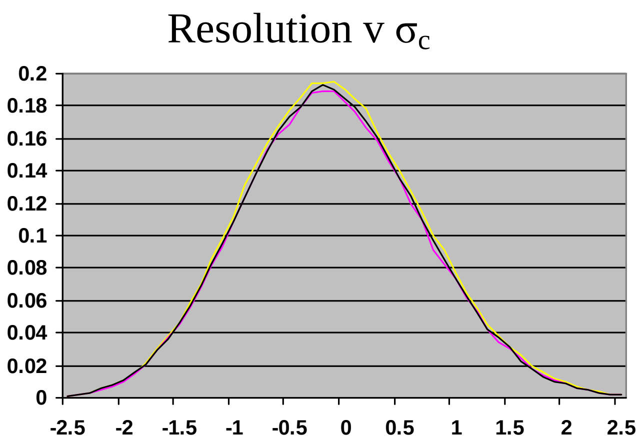


Figure 12: CRPS resolution at hour 12. The integrated forecast is slightly better except in the tails where the resolutions are similar.

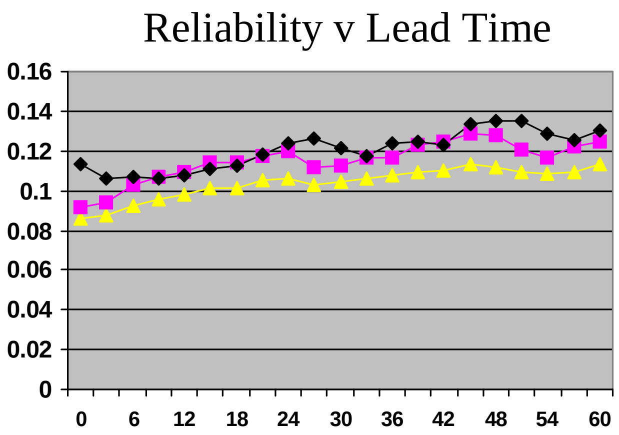


Figure 13: CRPS Reliability for all lead times. The integrated forecast is better at all lead times.

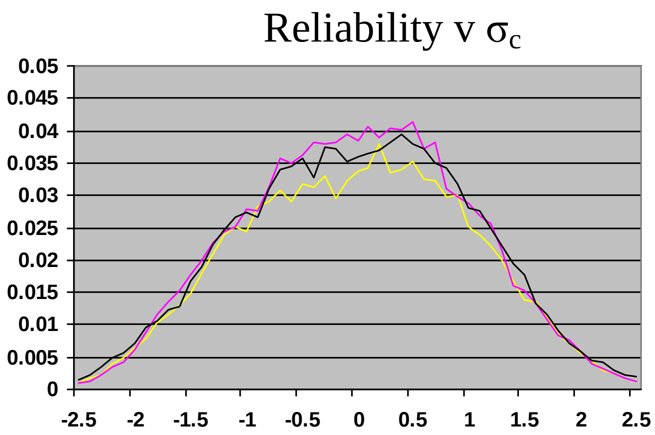


Figure 14: CRPS Reliability shown at hour 12 for all variance thresholds. The integrated forecast is better at most lead times.

References

Myers, B, Petty, M and Cowie, J, “An Automated Road Weather Prediction System for Road Maintenance Decision Support”, 18th Annual IIPS Conference 3.5, 2002 AMS Conference, Orlando, FL

Neilley, P, Myers, W and Young, G, “Ensemble Dynamic MOS”, 16th Conference on Probability and Statistics in the Atmospheric Sciences 3.6, 2002 AMS Conference, Orlando, FL

Gerding, S, and Myers, B, “Adaptive Data Fusion of Meteorological Forecast Modules”, 3rd Conference on Artificial Intelligence Applications to the Environmental Science 4.8, 2003 AMS Conference, Long Beach, CA

Clemen, Robert and Winkler, Robert, “Combining Probability Distributions from Experts in Risk Analysis” Risk Analysis, 19, (1999), 187-203