Ensemble forecast and verification of low level wind shear by the NCEP SREF system

Binbin Zhou*, Jeff McQueen, Jun Du,
Geoff DiMego, Zoltan Toth, Yuejian Zhu

NOAA/NWS/NCEP/Environment Model Center

1. Introduction

Ensemble forecasting is a new modeling technique to deal with the uncertainties and errors, in either initial conditions or models. The National Centers for Environmental Prediction (NCEP) has developed ensemble forecast systems at both global (since 1996) and short range scales (since 2001). In recent years, NCEP has been making efforts to apply its Short Range Ensemble Forecast (SREF) System to aviation weather forecast (Zhou et al, 2004) and has completed the primary framework including system configuration, post-processing, preliminary aviation-related ensemble products (mean, spread and probabilities for turbulence, icing, jet stream at different flight levels, surface visibility, wind speed/direction, cloud sky type, flight condition category, low level wind shear, convection, precipitation type, tropopause, frozen height, etc), and a web site to display these products. The SREF system runs twice a day (09Z and 21Z) out to 63 forecast hours with output in every 3 hours. The domain is CONUS. There are two basic models in the SREF system, the 32km ETA and 40km Regional Spectral Model (RSM). Perturbed initial conditions (breeding method) as well as multiple convection schemes are used in both models to generate a total of 15 ensemble members. The results are stored in GRIB-extension format files (Grid data) and BUFR format files (station data). Unfortunately, all these aviation ensemble products have not been verified yet since the ensemble forecast verification are not set up in NCEP Forecast Verification System (FVS). The verification of low level wind shear (LLWS) is our first attempt at verifying or evaluating aviation ensemble products. We are trying to put this verification and some other aviation ensemble products into NCEP's FVS. If this is successful, such verification can be performed routinely in the future.

This report has examined the SREF LLWS forecast with the NCEP EDAS re-analysis data, trying to answer the following questions: (1) What is the error/bias in SREF LLWS products? (2) How is SREF system performance in terms of the LLWS forecast? (3) What is the skill level for SREF LLWS forecast compared to NCEP current operational ETA (now NAM) and how does it change with forecast time? (4) What are the uncertainties involved in the SREF LLWS forecast? (5)How do we select the forecast probability threshold when using probability information in forecasting LLWS? Readers will find many differences between ensemble forecast verification and that of regular deterministic forecasts. In this report, the general principles of ensemble forecast verification are reviewed and will be used as a reference for the verification of other ensemble aviation products.

There are 4 sections in this report. Section 1 is the introduction, Section 2 is the forecast of LLWS in the SREF system, Section 3 first introduces the verification methods, including usual skill scores used in the ensemble forecast verification, and then discusses the results of each verification/evaluation scores, and Section 4 is the summery.

Corresponding author address: Binbin Zhou, NCEP/EMC, 5200 Auth Rd. Camp Springs MD 20746, Email: Binbin.Zhou@noaa.gov

2. Ensemble forecast of LLWS in SREF

(2.1) LLWS definition

Low level wind shear is hazardous to airplane landing operations and management and is a major concern of aviation weather forecasters at airports, particularly in the cold season (Cole, et al 2000).

According to *NWS Instruction 10-813, Terminal Aerodrome Forecasts* (2004), the definitions of LLWS are as follows:

LLWS – Wind vector difference between surface and 2000-foot-level (knots/2000 feet)

Severe LLWS – (1) If LLWS > 20 knots/2000feet, or (2) If LLWS > 20 knots/200 feet within any 200-layer bellow 2000 feet

(2.2) LLWS Forecast in the SREF system

In the SREF system, the above definitions are used to compute ensemble LLWS products including the mean (i.e. average among the 15 members), spread (i.e. standard deviation with its mean as reference) and probability for certain thresholds (i.e. severe LLWS) in the post processor.

It should be noted that, the SREF system is post-processing. That is, the 15 ensemble models run first to produce general output for each ensemble member. The output data only include basic meteorological parameters such as winds, temperature, humidity, pressure, etc. The aviation weather parameters such as LLWS are not included in the model output and need an additional post processing, applying certain algorithms, to generate the aviation-related parameters for each ensemble member. Then, statistical formulas are applied to produce their ensemble products.

In SREF system, each model (ensemble member) outputs are stored in GRIB-212 format which covers the domain of CONUS. Winds (u and v components) in the GRIB file are defined on the pressure levels beginning at 1000mb going up to 50mb, with fixed intervals of 25mb (about 200 meters, or 600 feet). However, the surface level and 2000-foot-level usually do not exactly match one of these pressure levels. Therefore, the surface height and 2000-foot-level must be determined first. Since surface wind is represented by the 10m wind, LLWS is calculated from the vector difference between the surface wind and the wind at 2000 feet plus 10 meters (2030 feet). The surface height is one of model outputs in the GRIB file, while the 2030-foot-level is obtained by searching from the surface upward, calculating the height of every pressure level until hits a level that just above 2000 feet. That means, the 2030-foot-level is located between this level (marked as L) and the level just below (marked as L-1). Linear interpolation between the L-1 and L level wind speeds is used to get the 2030 feet wind speed (See Figure 1).

Please note that the GRIB's vertical level's interval is 25mb or 200 m (about 600 feet), which is much larger than 200 feet. That means, SREF can not precisely estimate the second case of severe LLWS . We only compute the wind shear for 25mb layers to represent the second category of severe LLWS which, as expected, will underestimate the second occurrence of the severe LLWS.

After the LLWS for each ensemble member is computed, it is then used to generate LLWS ensemble products (mean, spread and probability). The products are still saved in GRIB-212 format files. The 2 cycles of SREF LLWS mean, spread and probability are routinely displayed at NCEP SREF web page at
http://www.emc.ncep.noaa.gov/mmb/SREF_avia/FCST/AVN/web_site/wshr/shr_09z_2000ft.htm

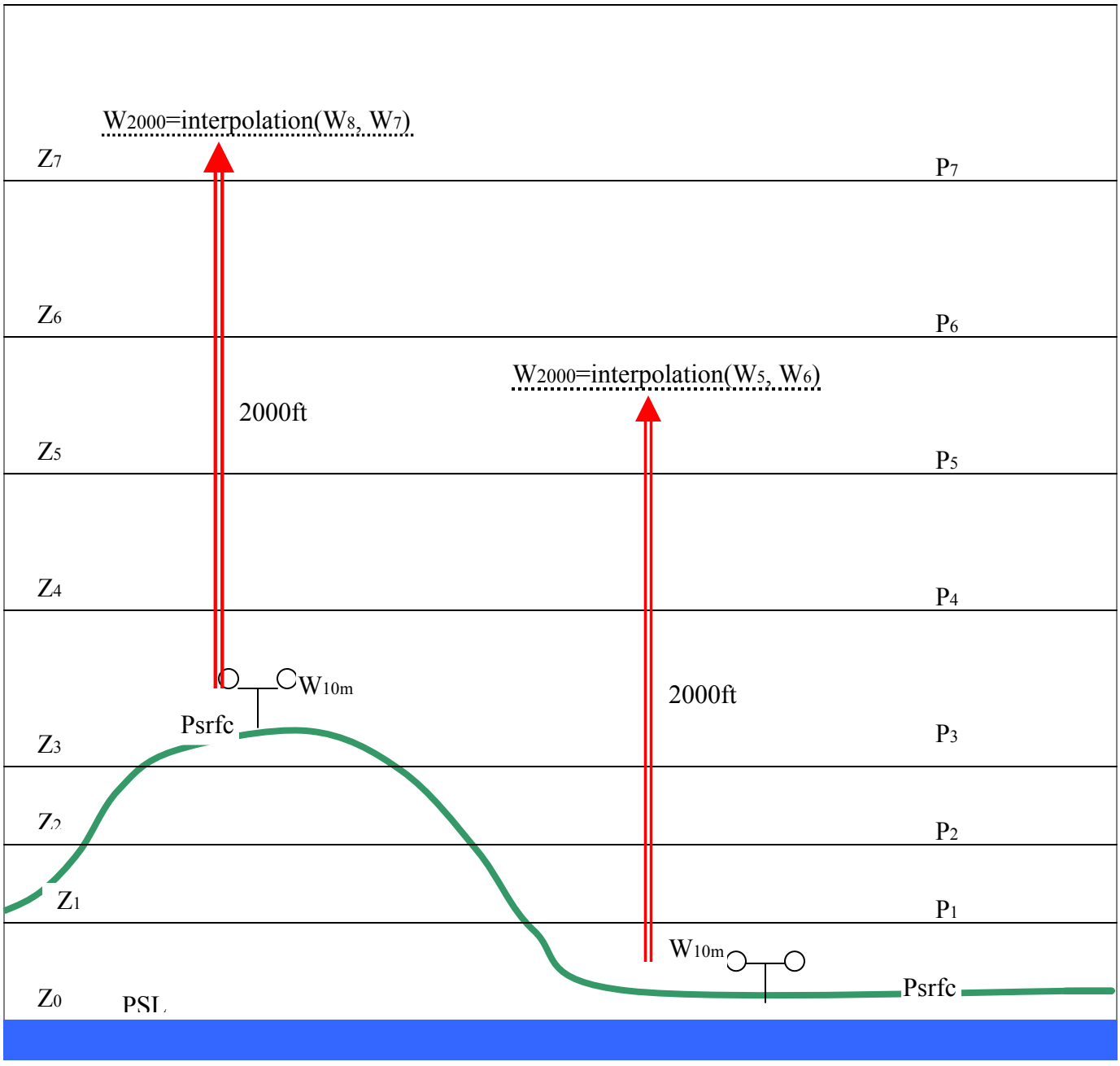


Figure 1: LLWS computation plot

Figures 2 and 3 show examples of SREF LLWS mean-spread and probability distribution 09Z forecast (September 20, 2004) over CONUS at forecast time = 6 hour, respectively.

In Figure 2, the mean value is indicated by line contours while the spread by the color shades. The red colored areas show the most uncertain LLWS forecasts, and the blue areas show the most certain forecast LLWS.

Figure 3 shows the severe LLWS (> 20knots/2000feet) probability distributions. The red areas indicate the most possible (> 90%) regions where LLWS is over 20knots/2000 feet, which is consistent with the mean values shown in Figure 3.

The usage of mean and spread (Figure 2) is straight forward. The mean value at a location represents the average of all 15 ensemble members' forecasts while the spread is the variability range for these forecasts, indicating the forecast variation range and uncertainty. But how to use the forecast probability plot (Figure 3) is not straight forward like selecting 90% or 100% as a threshold. If the probability larger than such high threshold, the forecast will lead to a very lower hit rate and a high missing rate. We will discuss this issue in the later section in section 3.

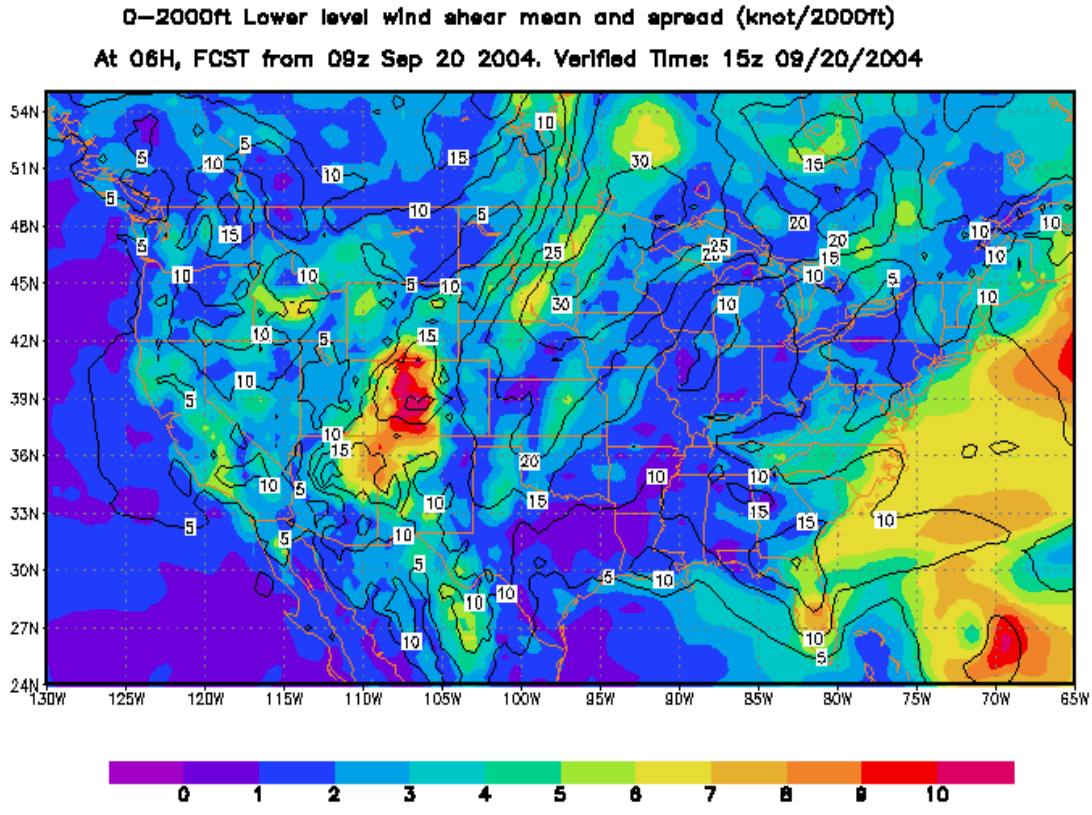


Figure 2. Mean and spread distributions

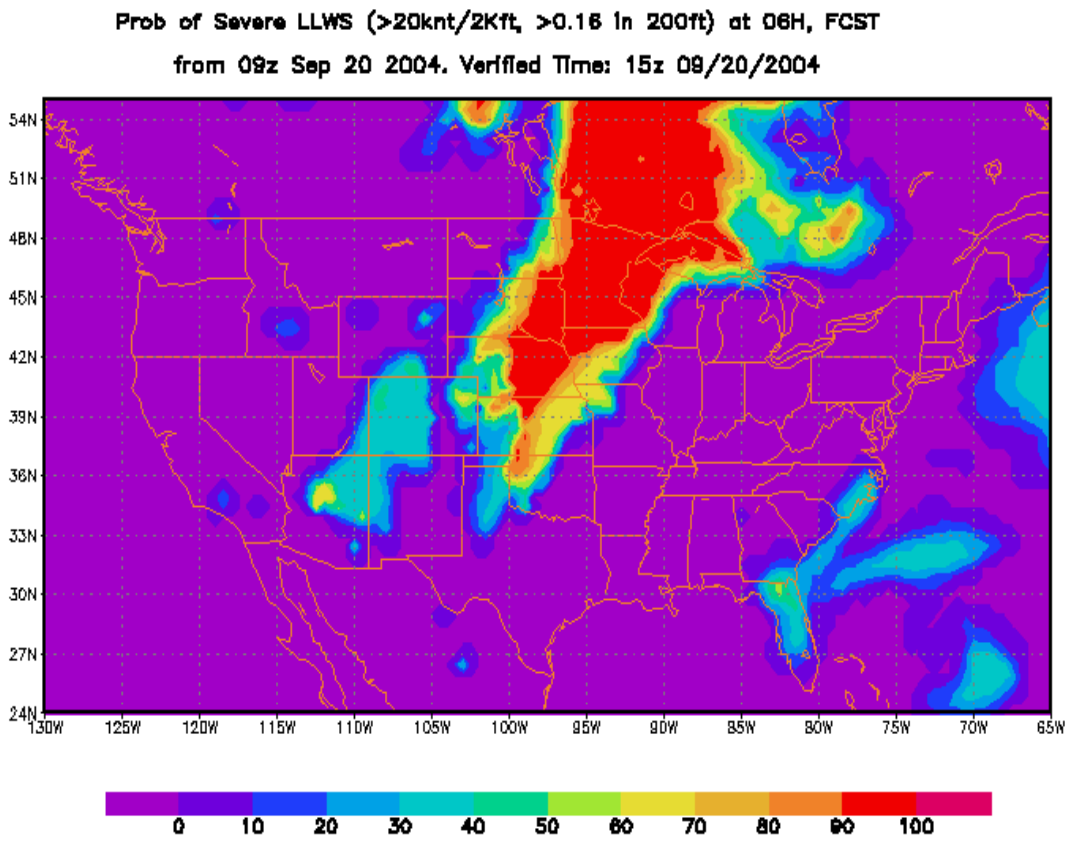


Figure 3, forecast probability distributions

3. Verification

(3.1) Verification data

NCEP ETA Data Assimilation System (EDAS) reanalysis are used as verified data. EDAS data are generated every 6 hours (00Z,06Z,12Z, and 18Z). Since SREF runs twice (09Z and 21Z) per day out to 63 hours, every SREF run can be verified by the 11 later on EDAS data files. For example, if the forecast is launched at 20040805 09Z, which creates 63 hours forecast up to 00Z of 20040808, then the verification can use 20040805's 12Z, 18Z EDAS data, 20040806's 00Z,06Z,12Z,18Z EDAS data, 20040807's 00Z,06Z,12Z,18Z EDAS data, and 20040808's 00Z EDAS data (See Figure 4).

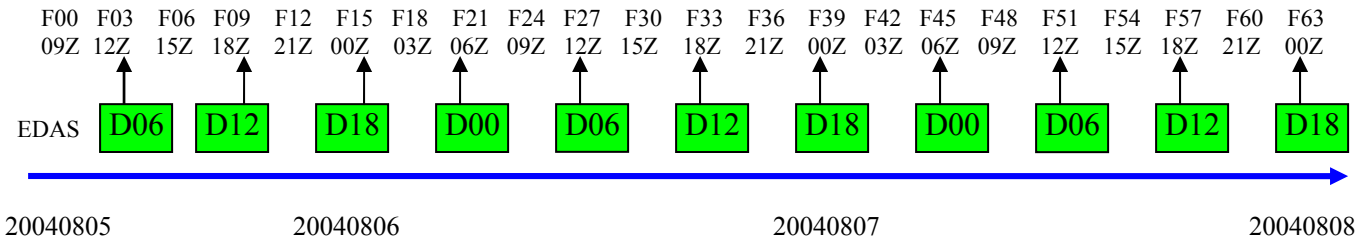


Figure 4: Verification configuration

The verification is grid-to-grid. In this report, we will show the verification results using 18 days of EDAS data, from August 18 to September 5, totally 18 day, representing a summer season case.

(*3.2) Verification/evaluation methods, results and discussions*

The verification methods for ensemble forecasts are broader than traditional deterministic forecasts since verification of ensemble forecast not only includes error, bias, skills like in the deterministic forecast, but also includes the evaluation of individual ensemble members as well as integration effects of ensemble members. The member evaluation is called system evaluation, such as performance equality, the distribution of observed data among the ensemble members. If all member behaviors are equal for a variable forecast, we say it is good. If most of the observed data can be captured by one of ensemble members, we say it is a good ensemble system. The skills of an ensemble forecast system are also evaluated by a set of scores or parameters, such as Brier Skill Scores (BSS), and Ranked Probability Skill Score(RPSS). The BSS can further be decomposed into reliability, resolution and uncertainty.

The following paragraphs will present each measures at first and then the results in detail.

(3.2.1) System evaluation 1: The ability of capture observed data

Such ability is usually evaluated with so-called Talagrand rank histogram. That, is, for each location, the forecast values of all ensemble members are first sorted by increase order, Second, M+1 bins are made (where M is the total number of ensemble members), with each bin representing the range of nearby members. The leftmost and rightmost represent the values < smallest member and > largest member, respectively. Last, examine the bins to see which observed data falls into which bin.

For example, there are 5 members in an ensemble. They gives the wind speeds as 2, 3, 1, 5, 6 m/s. The sorted values are 1,2,3,5,6 m/s, and the 6 bins are 0-1 m/s, 1-2 m/s, 2-3 m/s, 3-5 m/s, 5-6 m/s, >6 m/s. If the observed wind is 4 m/s, then it falls in the 4^{th} bin, if observed wind is 0.5 m/s, it falls in the first bin, and if observed wind is 7 m/s, it falls in the 6^{th} bin.

The SREF system has 15 members, so has a total of 16 bins. LLWS of the 15 members are sorted to create 16 bins. Then we see which of bins contains the observed LLWS. After

accumulating LLWS over all locations (grids) in the domain for each forecast time, we have a Talagrand histogram, as shown in Figure 5. In Figure 5, the 6 plots are for 6 forecast times. Each bin represents the percentage of observed data falling in the corresponding bin. We see that all plots show a U-shape. This is the typical Talagrand histogram for an ensemble forecast system, indicating that, there are about 20 % of LLWS observed data (leftmost bin + rightmost bin) are not captured by the current SREF LLWS forecast (also called an outlier rate of 20%). The perfect case would be 100% data being captured (an Outlier rate of 0%).

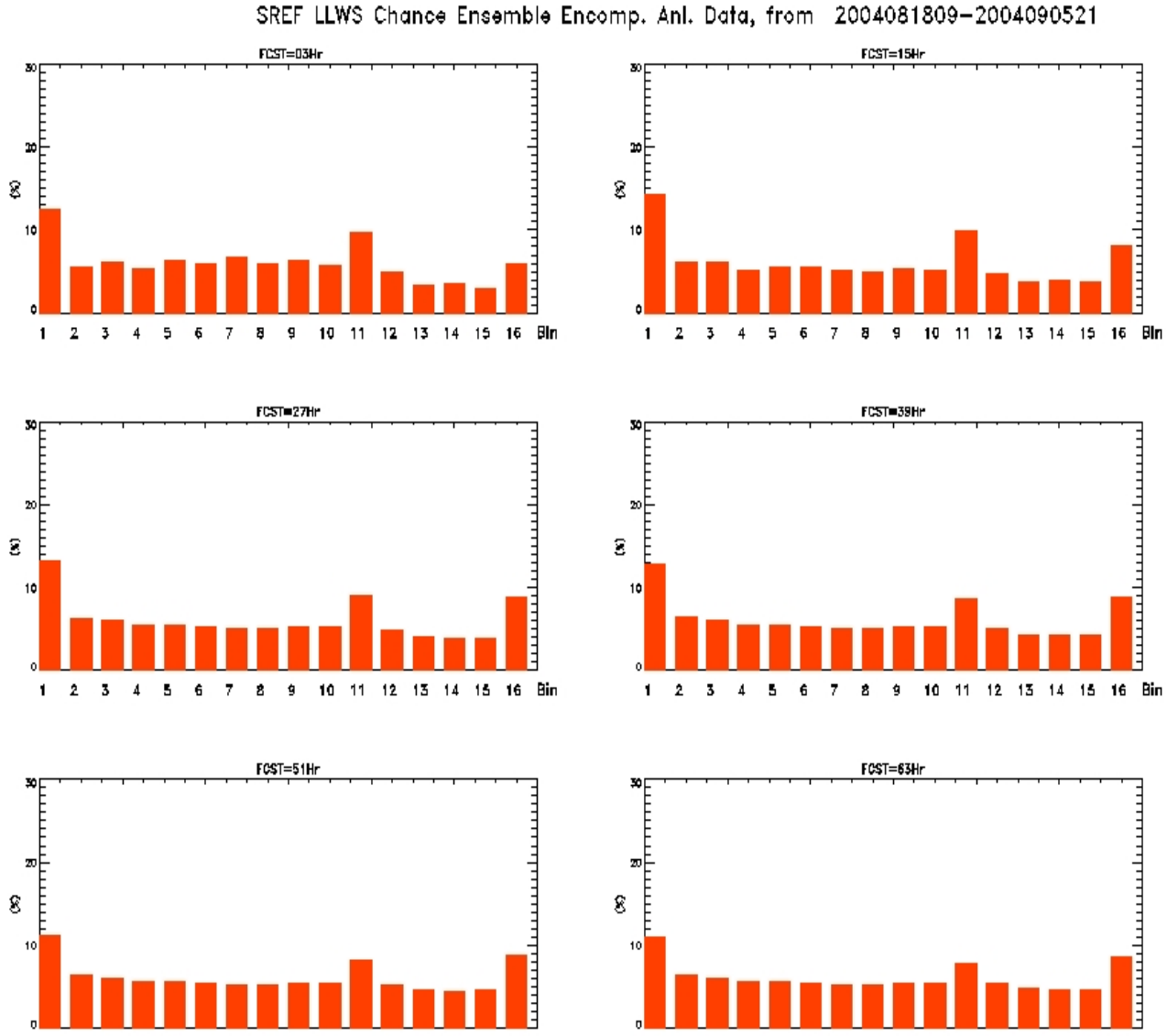


Figure 5: Talagrand Rank Histogram

The fact that not all observed data are captured has statistical significance: In general, if the larger the diversity among the ensemble members, the more possible is it that the observed data can be captured. The forecast diversity is expressed by forecast spread. This is good aspect of the spread. However, if the spread is too large, the forecast uncertainty is also large, and the confidence is then small. This is the bad aspect of the spread. In the SREF system, the spread is generated from a combination of randomly perturbed initial conditions and different convection schemes employed in the different members. We always hope for a smaller spread but lower outlier rate. In both low cases, the Talagrand histogram will have a ∩-shape instead of a U-shape. Unfortunately, we can not achieve both low low spread and outlier rate in the current SREF system for LLWS

(3.2.2) System evaluation 2: The member equal-likelihood

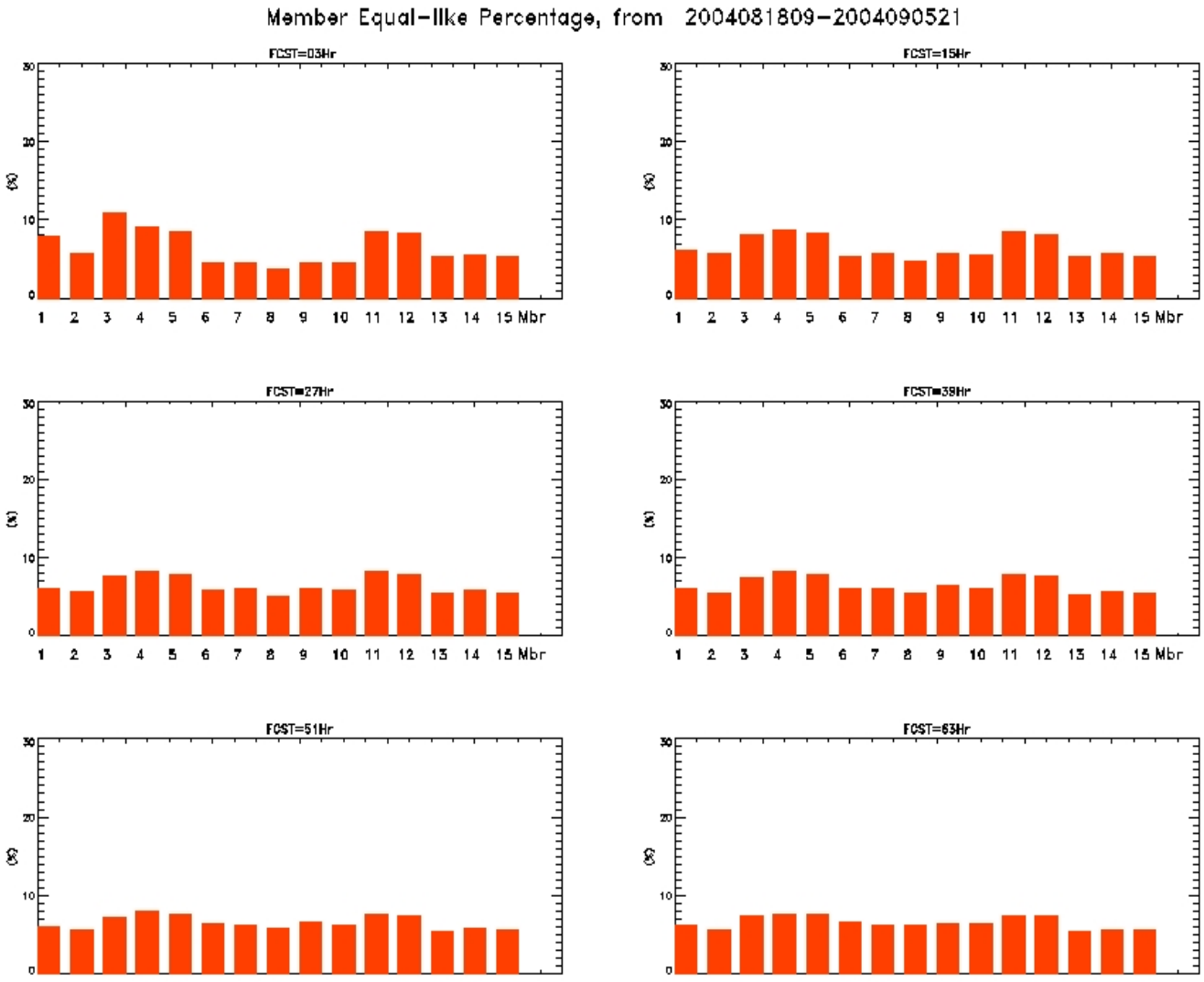


Figure 6. Member Equal-likelihood plots

The performance of an ensemble system can also be evaluated by the measurement called *member equal-likelihood*. Generally, in a good ensemble forecast system, all members should have equal ability to capture the observations. In other words, the observed data set should uniformly distributed among the ensemble members. This can be tested by using equal-likelihood plots. In this method, a bin is set for each ensemble member, and then the result is checked to see which member's forecast is closest to the observed data. Figure 6 is a set of statistical plots for LLWS at 6 forecast time over all LLWS forecasts.

The bin order is arranged in such way that ETA members sit in the first 10 bins while RSM members in the other 5 bins. It shows that, the equality among the 15 members is generally the same, except for the initial time. No member gets particular high percentage or particular low percentage. This test indicates that, all of the ensemble members in the SREF system have similar ability/performance to capture the observed LLWS. This is what we expected.

(3.2.3) RMSE, Bias, Spread, Correlation-coefficient, ROC, and ETS

RMSE (root mean square error), Bias, Correlation-efficient, ETS (equitable threat score, or Gilbert skill score) and ROC (relative operating characteristic) are the parameters often used in the traditional (deterministic) forecast verification. The difference here is that they are obtained from the ensemble mean instead of from single model and ETS and ROC are from the accumulated hit rate and false alarm rate The spread represents the diversity of forecasts among the members as mentioned before. Before showing the results, the definitions for these parameters are listed below since different researchers use different definitions at times.

$$RMSE = \sqrt{\frac{1}{N}\sum(\overline{F}i - Oi)^2}$$

where N is the sample size, $\overline{F}i$ is mean value at location i, Oi is the observed value at location i. RMSE indicates the average magnitude of forecast errors.

$$Bias = (\frac{1}{N}\sum Fi)/(\frac{1}{N}\sum Oi)$$

Bias indicates the average forecast magnitude compared to the average observed magnitude.

$$Corr = \frac{\sum(Fi - \overline{F})(Oi - \overline{O})}{\sqrt{\sum(Fi - \overline{F})^2}\sqrt{\sum(Oi - \overline{O})^2}}$$

which indicates how well the forecast values correspond to the observed data. $\overline{F}$ and $\overline{O}$ are averaged forecast and averaged observation over whole domain, respectively.

$$ETS = \frac{hits - hits_{random}}{hits + misses + falseAlarms - hits_{random}}$$

where *hits* is the number of counts where both observed LLWS and forecast mean LLWS are larger than the severe LLWS threshold, *misses* is the number of counts where observed LLWS is severe, but forecast mean is not severe, *falseAlarms* is the number of counts where observed LLWs is not severe, but forecast mean is. The hits due to random change are

$$hits_{random} = \frac{(hits + misses)(hits + falseAlarms)}{N},$$

Where N is total number of forecasts.

The relationship among hits, misses, and false alarms can be expressed in a 2 by 2 contingency table:

Table 1: 2 by 2 contingency table

		Observation	
		Not severe	Severe
Forecast mean	Not severe	**Correct negative**	**misses**
	Severe	**False alarms**	**hits**

A ROC plot is a plot in which the hit rate is drawn against the false alarm rate (FAR). In probabilistic forecast, the hat rate is defined by the following integrations:

Hit rate: $$H(pt) = \frac{1}{s}\int_{pt}^{1} f_{obsv}(p) f_{fcst}(p)dp$$

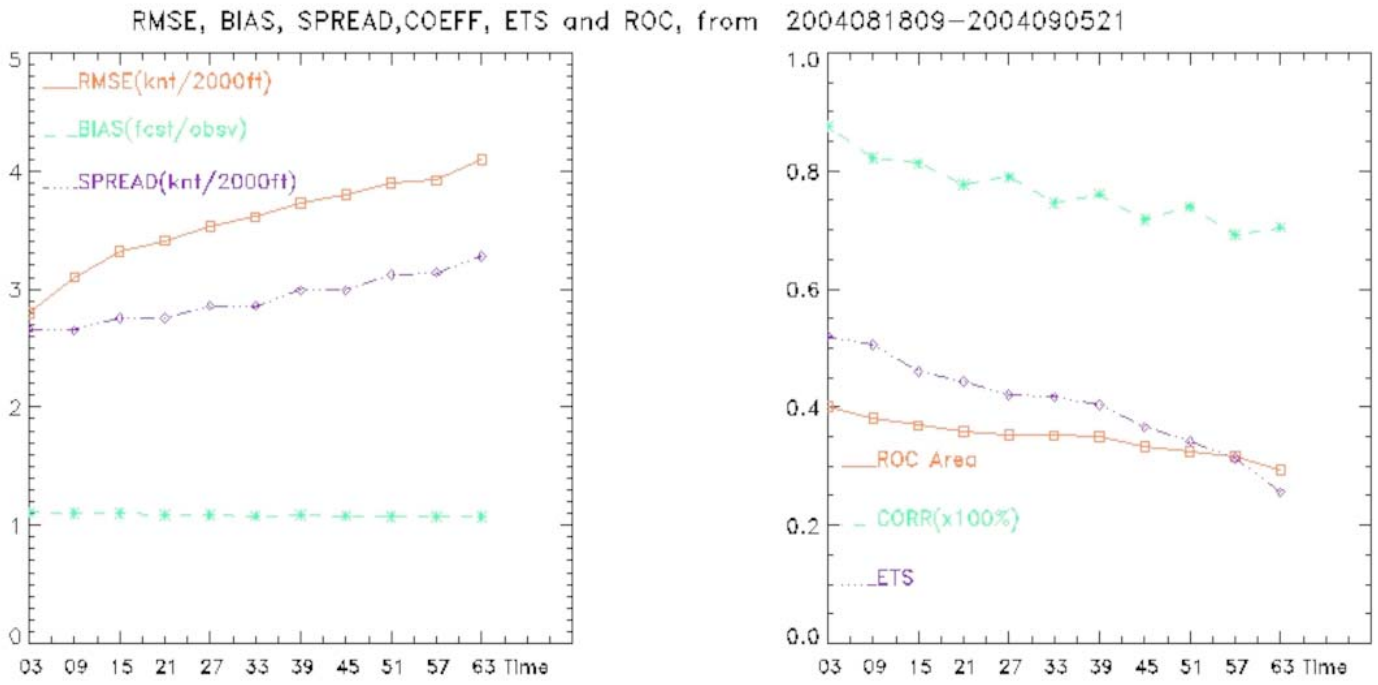
FAR:

$$F(pt) = \frac{1}{1-s}\int_{pt}^{1} [1 - f_{obsv}(p)] f_{fcst}(p)dp$$

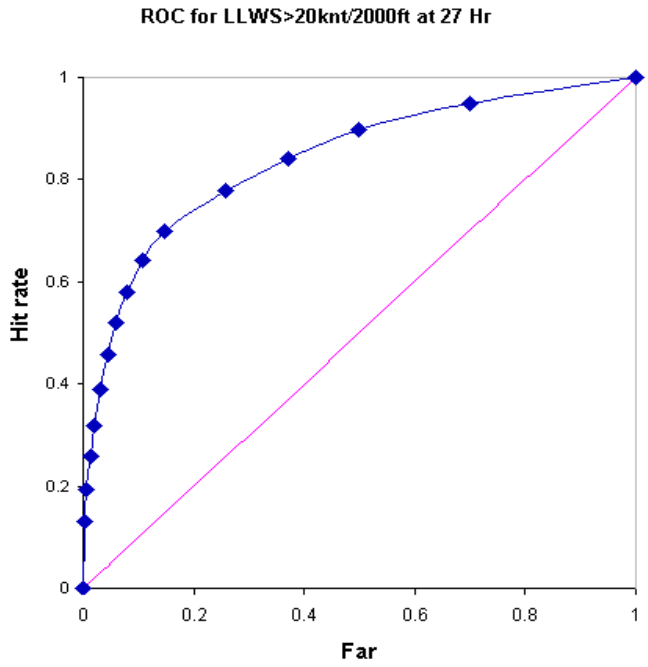
where $s = \int_{0}^{1} f_{obsv}(p) f_{fcst}(p)dp$, *H(pt),* also called the POD (probability of detection), is the hit rate under conditions such that only if $p > pt$, then a warning forecast is issued. *F(pt)* is the false warning rate under such conditions. $f_{obsv}(p)$ is the observed frequency for probability p, and $f_{fcst}(p)$ is the forecast frequency for probability p. Drawing the hit rate against the FAR can obtain the so-called ROC plot, which indicates the skill of the probabilistic forecast in terms of hit rate and FAR, if hit rate > FAR, has skill, otherwise no skill.

After examining the above definitions, let us now see the results shown in Figure 7(a),(b) and (c)

In Figure 7 (a) are the RMSE, Bias and Spread, showing that the forecast error and spread increases with the forecast time. It is can be expected that the model becomes less and less inaccurate over time, and the forecast uncertainty, which can be expressed by the forecast spread, becomes larger and larger after forecast begins. This is consistent with the Figure 7 (b), which shows that the skill of the SREF system becomes smaller and smaller with the time. The ROC area is the area below the curve and above the diagonal line shown in Figure 7 (c). The area of the best case is 0.5, as long as ROC area > 0.0, the forecast has skill. We can see that although the ROC area size decreases with the forecast time, after 63 hours, the SREF LLWS forecast still has relative larger skill (0.3). The same thing is true for ETS which also decreases with the forecast time, indicating the loss of skill with time. Figure 7 (c) only depicts ROC plot for one of the forecast times here.



(a) (b)



(c)

Figure 7: Error and Skills

The bias in Figure 7 (a) shows that the averaged forecast of LLWS is a little bit larger than observed LLWS with ratio = 1.05, but this ratio does not increase with the forecast time.

We should pay more attention to the behavior of spread in Figure 7 (a). As we said that magnitude of spread is an indicator of the forecast uncertainty. The larger the spread is, the larger uncertain the forecast is. In the SREF system, 15 ensemble members are generated from both initial condition perturbation (breeding) and multiple convective schemes. That implies that the spread of forecast LLWS spread comes from either errors in the initial conditions or uncertainty in the model convective physics or both. In general, the error in the initial conditions is small but will grow gradually with the forecast time while the diversity of the model physics is fixed and does not changed with forecast time. This implies that, the big jump in spread at the initial time is caused by the model diversity, and growth of error in the initial condition data is responsible for the spread's gradual increase and skill loss with forecast time.

Spread, most of people think, is an indicator of ensemble forecast error (This statement is still an open issue). For a good ensemble system, RMSE and spread should consistent in both in values and variations over time. In Fig 7 (a), at the early time, RMSE and the spread are close, while with forecast time, they go away gradually.

(3.2.4) Probabilistic measures -- Reliability, Resolution, Uncertainty, Brier Score and Brier Skill Score

The evaluation of the probabilistic performance of an ensemble forecast is through following measures:

Reliability – the difference between a forecast probability distribution and observed probability distribution. It is related to bias and can be improved by bias-correction techniques. The best value = 0.0

Resolution – the ability to distinguish forecast from averaged observed data or climate data, or back-
ground noise (uncertainty). Can not be improved by bias-correction techniques. The worst value = 0.0

Uncertainty – the error or variability in the observed (climatological) data which is used in either the initial conditions or in the comparison. Always>0.0. Note that, the forecast uncertainty is indicated by forecast spread

Brier score (BS) – the quadratic scoring measure for a probabilistic binary forecast defined as

$$B = \frac{1}{n}\sum_{j=1}^{n}(p_j - o_j)^2$$

where p_j is forecast probability, o_j is observed data (1 for the event happening , 0 for the event not happening). For a deterministic forecast, p_j = 1 (event happens) or 0 (event doesn't not happen). But for a probabilistic forecast, p_j's value is between 0.0 – 1.0 since ensemble forecast usually give uncertain forecast (probability is neither 0 nor 100%). The best value for a Brier score is 0 (perfect forecast system).

Brier skill score (BSS) – compares the BS to that for a reference forecast system, defined as

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

where the reference BS (BS ref) can be either climate data, observed data or BS for another forecast system. A BSS > 0.0 shows skill in

comparison to the reference (*BS* > *BSref*), otherwise, there is no skill.

It can be theoretically proved that, BSS can be decomposed into a BSS reliability part and a BSS resolution part, as follows

$$BSS = \frac{\text{Re}\, s - \text{Re}\, l}{Uncerta\ \text{int}\ y} = 1 - BSSrel - BSSres$$

since only BSS > 0 shows skill, the ensemble system with smaller uncertainty and resolution > reliability will be skillful, otherwise, such a system has no skill.

There are some other measures such as ranked probability score (RPS) and ranked probability skill score (RPSS) which are measures for multiple category forecasts. For the LLWS case, the severe LLWS forecast is a binary forecast (severe or not severe). For a binary forecast, it can be proved that RPS = BS, and RPSS = BSS. So we won't compute RPS and RPSS here.

The BS , and BSS referenced by observed data and the BSS referenced by operational ETA are shown in Figures 8, 9 and 10.

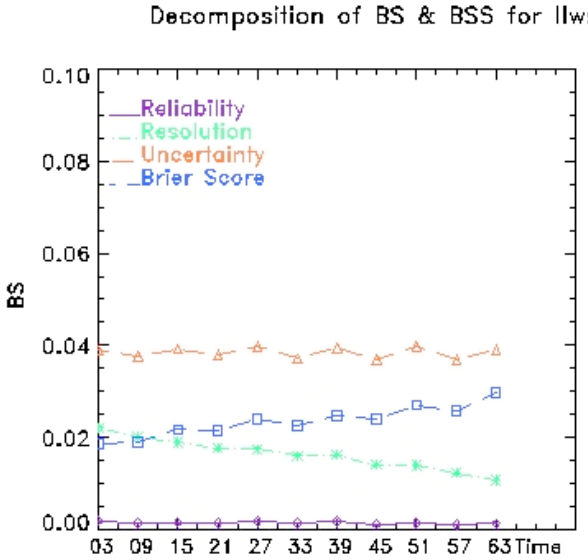


Figure 8: Brier Score

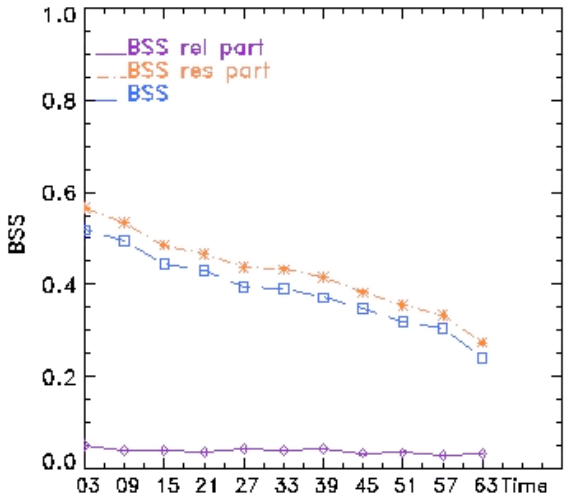


Figure 9: BSS referenced by observation

Figure 8 shows that, the reliability for SREF LLWS is very small, indicating a good reliability in general. Its value is also much smaller than the resolution, and the data which does not vary with the forecast time as. The BS value increases with the forecast time, indicating the forecast skill decreases with time which obviously is due to the resolution decreases seen in the same figure.

Figure 9 shows the decomposition of BSS into reliability and resolution parts, presenting the similar information on reliability and resolution as Figure 8. BSS is still larger than 0.0 after 63 hours in Figure 9 indicating that the SREF LLWS forecast is still skillful after 63 forecast hours compared to data uncertainty.

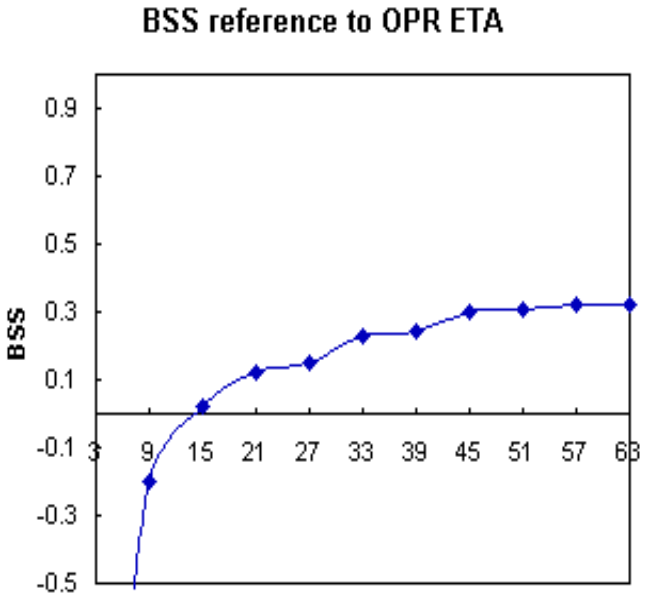


Figure 10: Ensemble mean BSS referenced by operational ETA

Besides using data uncertainty as a reference to compute the Brier skill score, we also use the operational 12km-resolution ETA model forecast to see if ensemble mean is better than the operational ETA. First we compute the Brier score for operational ETA, then the Brier score for the ensemble mean. Using above equation, we can obtain the Brier skill score shown in Figure 10. If the Brier skill score is less than zero, the ensemble mean LLWS is worse than the operational ETA, otherwise, it is better than the operational model.

It can be seen that, at very early times (before 9 hours), the operational ETA has better skill in LLWS forecast mean than the SREF, but after 15 hours, the SREF LLWS forecast mean is increasingly better than the operational ETA. This plot illustrates that, the ensemble forecast mean is advantageous over a single model forecast system, particularly for a little bit longer range forecasts, although the operational ETA has a higher horizontal resolution (12km) than the SREF models (32km for ETA and 40km for RSM). This result can be found from verifications of other SREF variables, such as 500mb u, v, T, RH, etc. From this result, we again confirmed our assumption that, even for a precise weather forecast model, the errors and uncertainties, which always exist in the initial conditions or in the model, will grow rapidly with the forecast time. Over very short time, they might be not important so that a precise model can still give a good forecast. However, with forecast time, growing errors will lead to more and more forecast error. That is one of reasons behind the motivation for using ensemble forecasts.

(3.2.5) Reliability plot

We have defined reliability and resolution. Their relationship can be expressed by so-called *reliability plot*, see Figure 11.

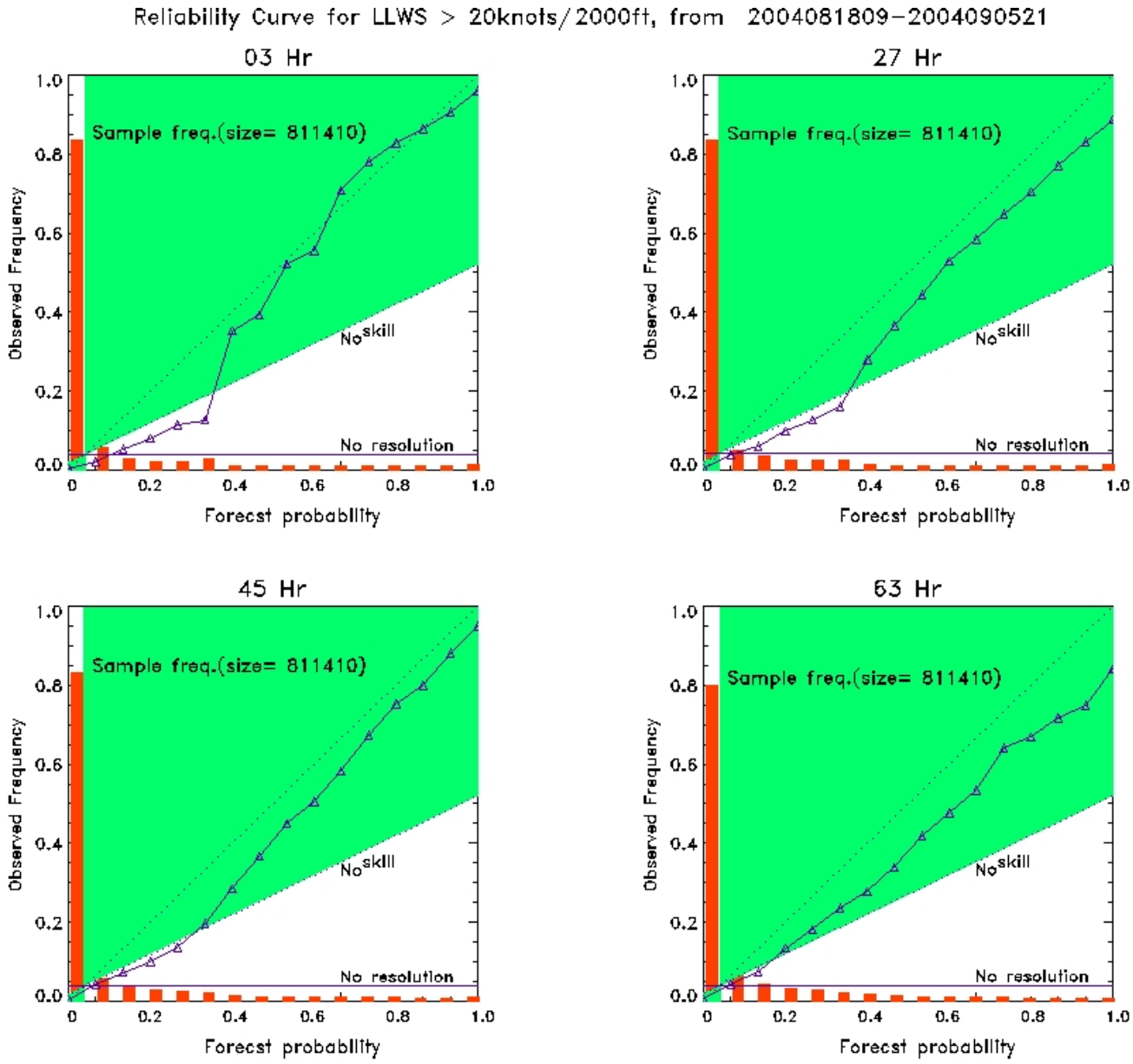


Figure 11. Reliability Plots

Figure 11 has reliability plots for 4 forecast times, 03 , 27 , 45 and 63 hours respectively. The following information can be found from the reliability plot:

(a) Reliability curve: also called probability bias, is the relationship between forecast probability and observed frequency. The best curve is a diagonal line. At early forecast times, the reliability curve is close to the diagonal line, which means good reliability. With increasing forecast time, the reliability curve goes away from the best line, and forecast probability in the SREF system is larger than observed frequency, indicating that, the reliability has decreased with forecast time. This is also consistent with the larger-than-one bias shown in Figure 7(a). The reliability is

actually the area (distance) between the diagonal line and the reliability curve. At the initial time, such an area is small, but later on, becomes larger.

(b) Data uncertainty: the variance in observation data which is not related to the forecast probability, so it is drawn as a constant line near the x-axis.

(c) Resolution: the distance (area) between the data-uncertainty line to the reliability curve. Therefore, the data-uncertainty line also can be seen as a no-resolution level. That is, when the reliability curve lowers down to the data-uncertainty line, the forecast has no resolution, or the forecast can not be distinguished from data uncertainty. At early times, the resolution is bigger, but later on, the reliability curve lowers down, and so does the resolution, but there is still pretty good resolution after 63 hours.

(d) Skill/no skill areas: the green area is the skillful area where the forecast has skill, and the blank no skill area is where the forecast has no skill. If the forecast probability is lower than 30%, all reliability curves are within the blank area, indicating the forecasts of severe LLWS will be not skillful. Only when forecast probability > 30%, the forecasts have skill.

(e) The red bins in the Figure 11 are the sample frequency for each forecast category. For example, for the case of forecast probability = 0 case, the sample frequency is about 83% for all the times. That means, of all the sample (or grid regions), in 83% of the regions, no any ensemble member got LLWS > 20knots/2000feet - all members predicted no severe LLWS. When the forecast probability increases, ito 100% , for instance, the sample frequency is about 1 %, which means only 1% of the region, all members forecasted severe LLWS forecasts.

Please note that, as mentioned before, the reliability can be improved by so-called bias correction method so that the reliability curve is closer to the diagonal line.

(3.2.6) Forecast probability threshold selection

How to select a forecast probability threshold (only above which, should LLWS warning be issued) is the remaining issue, if set it too high, forecast confidence is high but will lead to a high missing rate. On the other hand, if set it too low, the missing rate can be decreased, but the FAR is increased. For example, In a case of severe LLWS probability forecast, 3 forecast categories, 10%, 50% and 100% regions are identified. Thus, which region should be issued severe LLWS warning? A common sense argues selecting 100% (red color area in Fig. 3) since all ensemble members give a severe LLWS forecast and the confidence is highest with this selection. But one concern is that if the 100% region is selected, we might miss some regions where severe LLWS actually happens but not all members gave a severe LLWS forecast. Another choice is to select 50% as threshold. In this case, the missing rate can be lowered, but the false alarm rate (FAR) might be increased and confidence level is not as high since only 50% of the members predicted severe LLWS. So what forecast probability threshold should be used is a very practical problem for forecasters who wish to use the probability information. Here we suggest a rule, without considering economic factors, to select a probability with which, both the missing rate and the FAR are lowest. We have already shown that, ETS is related to both the missing rate and the FAR. If both the missing rate and the FAR are lowest, ETS will be at its largest. To confirm this statement, we computed the missing rate, the hit rate, and the FAR as well as ETS for different forecast probabilities, and

depicted them together in one plot (see Figure 12).

From Figure 12, it can be seen that, the best probability threshold is around 45 %, where both the missing rate and the FAR are lowest and ETS is highest (the maximum point of the ETS curve). That means we should select 45% as the forecast probability threshold for LLWS instead of the100% that common sense would say. That is, as long as an area has severe LLWS probability > 45 %, we can be relatively confident that severe LLWS will happen and be most unlikely to miss severe LLWS, but have the lowest false alarm rate. If we select 100% as the threshold, the missing rate will be very high, as high as 80% at early times, and it becomes almost 100% at 63hr (means no any severe LLWS was captured). The hit rate in this case is lowest. So back to Figure 3, those regions where the forecast probability > 45% should issue a severe LLWS warning.

It is interesting that the value of 45% is quite stable for all forecast times in Figure 12, although ETS decreases with the forecast time in corresponding to the increase in the missing rate and FAR, and the decrease in hit rate with the forecast time. As we expect, for perfect forecast system, the FAR and missing rate for all probabilities should be zero, and hit rate is always equal to 1. In this case, any forecast probability can be selected as the threshold. Unfortunately, an ensemble system is far from perfect . The errors and uncertainties in the forecast system always exist. However, we still don't know why highest ETS for LLWS is at 45% instead of a relative higher value such as 70 or 80% even for very early forecast stages. We don't know if this threshold can be increased after the bias is corrected. This will be confirmed in a later study.

Please note that, the above probability threshold analysis does not consider the economic factors which have an impact on the probability threshold selection policy. For example, if the economic loss of a FAR forecast is larger than a missing forecast, then the weight should have a higher weight, thus the threshold will be increased so that the FAR is reduced even though the missing rate is increased. Issuing LLWS warning will decrease the airport landing rate and affect the airport landing management. Thus, selecting a probability threshold not only reflects the reduction of forecast errors, but also depends on an economic cost-benefit or cost-loss analysis.

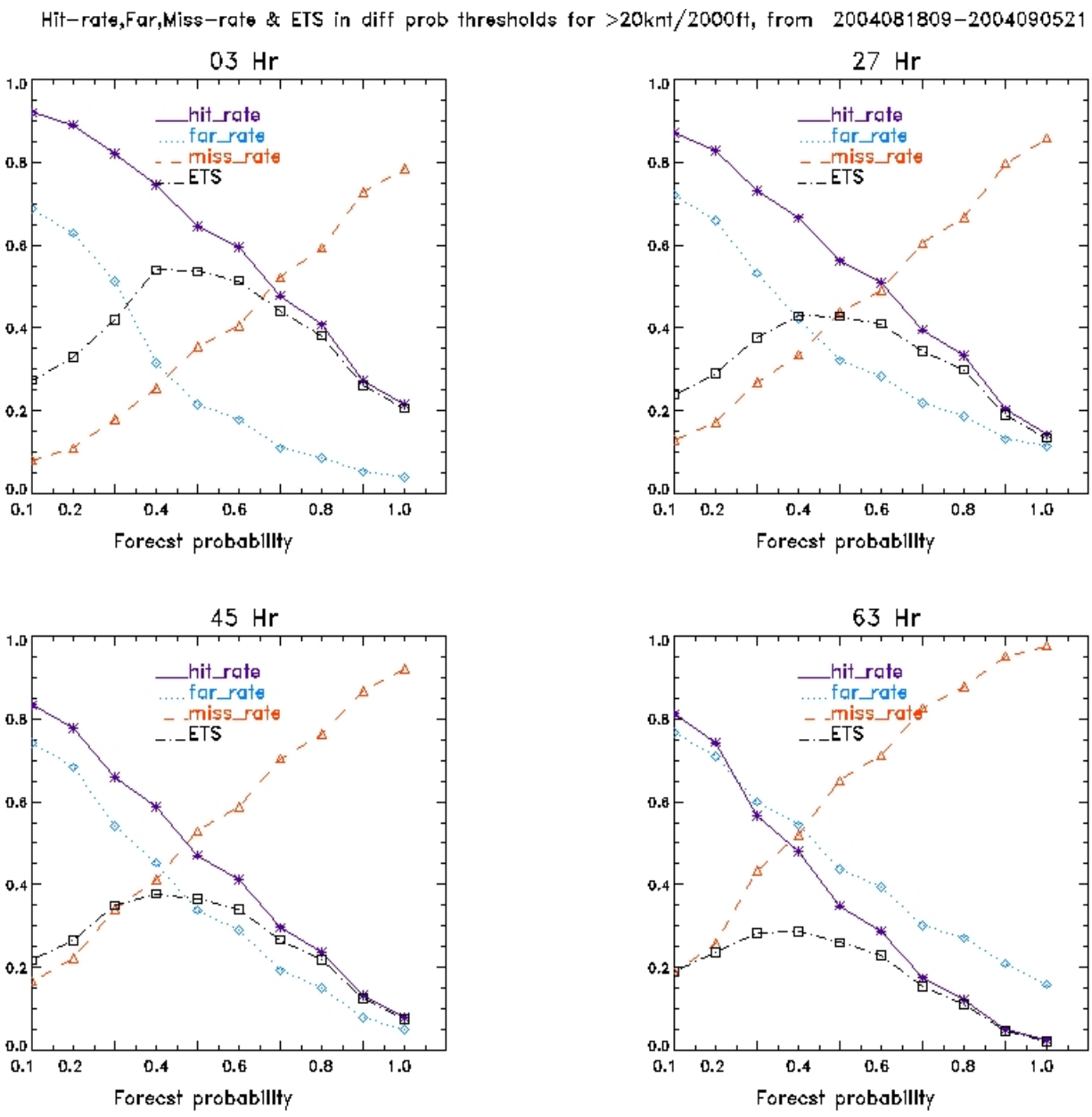


Figure 12, The best probability threshold for issuing sever LLWS

4. Summary

In this report, 18 day SREF LLWS forecast mean, spread and probability results were evaluated and verified against EDAS analysis data. The ensemble forecast verification/evaluation methods were simply reviewed, then the verification results were presented, showing that,

(i) SREF systematic performance for LLWS forecasts is promising in terms of member equal-likelihood and the ability to capture observations.

(ii) The forecast error and uncertainty (spread) increase while skills (ROC, ETS, etc) decrease with the forecast time.

(iii) The reliability, resolution and the uncertainty in the SREF LLWS forecast data are also captured by decomposing the Brier skill score, showing that, the Brier skill score decreases in response to the decrease in resolution, although the data uncertainty is stable for all forecast times.

(iv) The probability threshold for issuing severe LLWS warning was investigated. If not consider the economic factor, the value for this threshold is about 45% where both the missing rate and FAR are the lowest and ETS value is highest. This number is still mysterious and need to further investigate.

From this report, we have already seen the features and advantages of ensemble forecasts through the evaluation of probabilistic measures. With ensemble forecasting, the probabilistic information such as forecast range, diversity, probability distribution, uncertainty in both data and model, the ability to distinguish from the reference observations, etc., are numerically quantified, which could not be done by traditional deterministic forecasts.

This work can be considered as an example for verifying/evaluating other aviation products in the SREF system. Currently, those observation data for the other products are not available to the SREF system's verification package. If these data are available, the same methods or measures employed in this report can be applied.

We are working on the bias-correction technique which will be applied for the SREF system's post-processing. Such technique is based on the ensemble information from the SREF forecast outputs. The primary results have shown that it can reduce the both system error and random error significantly. If the bias-corrected winds can be further used in LLWS computation, we believe that the skill will be better than the current SREF LLWS forecast.

Acknowledgement

Thank Dr. Steve Silverberg of NCEP Aviation Weather Center (AWC) for giving us the instruction of developing SREF aviation ensemble products. Also hank Mike Graf and his group in NOAA Aviation Service Branch, for helping us to understand the requirements and definition for LLWS in aviation community.

Reference:

Cole, R. E., S. S. Allan, and D. W. Miller, 2000: *Vertical Wind Shear Near Airports as an Aviation Hazard*, 9th Conference on Aviation Range and Aerospace, Sept 11-15, Orlando, FL, Amer. Meteor. Soc.

Du, Jun, et al, 2004, *The NOAA/NWS/NCEP short range ensemble forecast(SREF) system: evaluation of an initial condition vs multi-model physics ensemble approach*. 16th Conference on Numerical Weather Prediction. Seattle, WA, Amer. Meteor Soc.

Operations and Services, Aviation Weather Services, NWSPD, 10-8, 2004, *NWS Instruction 10-813, Terminal Aerodrome Forecasts*, Feb. 1, 2004

Toth, Z., et al: 2003: Chapter 7, Probability and ensemble forecast, Environmental Forecast Verification: *A Practitioner's Guide in Atmospheric Science*, Edited by I. T. Jolliffe and D. B. Stephenson, John Willey & Sons.

Wilks D. S., 1995: *Statistical Methods in the Atmospheric Science*, Academic Press, 467 pp.

Zhou, Binbin et al. 2004, *An Introduction to NCEP SREF Aviation Project*, 11th Conference on Aviation Range and Aerospace, Oct 4-8, Hyannis, MA, Amer. Meteor. Soc.

Zhu, Yuejian, et al, 2002, The *economic value of ensemble-based weather forecasts*, Bull. Of Amer. Meteor. Soc. January, 73-83.