Ensemble Based Probabilistic Forecast Verification

Yuejian Zhu and Zoltan Toth
Environmental Modeling Center, NCEP, NWS/NOAA
Washington DC 20233
E-Mail: Yuejian.Zhu@noaa.gov

1. INTRODUCTION:

The NCEP ensemble verification system was developed to evaluate ensemble based probabilistic forecast in the 90s (Zhu et al., 1996). This system mainly focuses on two attributes: the ***reliability*** and ***resolution*** (Toth et al., 2003, 2006) of the NCEP ensemble based probabilistic forecast, in addition to the traditional verification measures such as Pattern Anomaly Correlation (PAC) and Root Mean Square (RMS) error for the ensemble mean, rank histogram, and outliers (Zhu, 2004; Toth et al., 2003), and Perturbation versus Error Correlation Analysis (PECA) (Wei and Toth, 2003), etc. For precipitation verification, Equitable Threat Score (ETS), True Skill Statistics (TSS) and Bias (BI) have been used to measure the ensemble mean (Zhu, 2007). In this ensemble based probabilistic verification system, the definitions of events are based on 1) user defined thresholds, 2) climatological percentiles, and 3) the ensemble members. In practice at NCEP, the climatological percentiles (10 climatologically-equally-likely bins) have been used for NCEP/GEFS (Global Ensemble Forecast System) daily verification. Therefore, the probabilistic skill scores for current NCEP/GEFS forecasts are based on the NCEP/NCAR 40-year reanalysis climatology (references). On a routine basis, this system generates a Brier Score (BS), Brier Skill Score (BSS) with its decomposition of reliability and resolution, Ranked Probability Skill Score (RPSS), Continuous Ranked Probability Skill Score (CRPSS), Relative Operational Characteristics (ROC) area score, Relative Economic Value (REV) score for selected cost/loss ratios to apply to upper atmospheric variables such as 500hPa geopotential height, and 850hPa temperature and near surface variables such as 1000hPa geopotential height, 2-meter temperature, and 10-meter wind (u and v). In terms of the ensemble mean, as in a deterministic forecast, ensemble spread and RMS error have been introduced; histogram (or Talagrand) distributions and outliers have been generated to measure the ensemble's reliability and consistency. This system was recently upgraded and applied to the Northern American Ensemble Forecast System (NAEFS), which combines the NCEP and CMC ensemble forecasts. This article mainly summarizes this verification system.

2. METHODOLOGY OF VERIFICATION:

a. RMS error and SPRD (ensemble spread):

RMS errors of the ensemble mean measure the distance between forecasts and analyses (or observations). SPRD (ensemble spread) is calculated by measuring the deviation of ensemble forecasts from their mean (Zhu, 2005). Figure 1 is an example of a display of RMS errors and ensemble spread (SPRD) for a 15-day lead-time forecast. Usually, SPRD is defined as:

$$SPRD = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\overline{f} - f(n))^2}$$

Where $\overline{f} = \frac{1}{N}\sum_{n=1}^{N} f(n)$ is for the ensemble mean and f is for the ensemble forecast.

In general, an ideal ensemble forecast will be expected to have the same size of ensemble spread as their RMS error at the same lead time in order to represent full forecast uncertainty (Zhu, 2005, Buzza et al., 2005). But most of the ensemble systems are underdispersed (less spread) for longer lead times due to an imperfect model system (or physical parameterizations) and other things. Therefore, a stochastic process will be introduced to increase ensemble spread for longer lead-time

forecasts (Hou et al., 2008). On the other hand, the ensemble mean consistently performs better than the high resolution deterministic forecast GFS (T382L64) after a 2-day lead time, while the high resolution GFS uses similar (or more) resources than the global ensembles (20 members at T126L28 resolution).

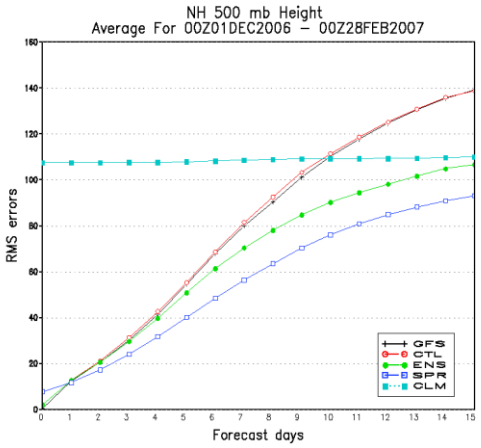

Fig. 1. RMS error for ensemble mean (blue) and ensemble spread (green) for NH ex-tropical 500hPa geopotential height of the 2006-2007 winter season, compared to the GFS (black) and ensemble control (CTL, red) RMS errors. The top curve (cyan) is for RMS error of climatology.

b. Histogram distribution:

A Histogram (or Talagrand) distribution is a simple measurement used to verify an ensemble system and its forecast distribution. The calculation formula of the Histogram Distribution (*HD*) for one grid point, at time *t*, analysis or observation (*a*), *N* ensemble forecasts *f(1,2,...N)* after re-ordering from low to high according to their values could be written as:

$$HD(n)=\begin{cases}\dfrac{1}{N}, n=1, a\le f(n)\\ \dfrac{1}{N}, n=2,...,N, f(n-1)<a\le f(n)\\ \dfrac{1}{N}, n=N+1, a>f(N)\end{cases}$$

There are a few resulting common shapes such as a U-shape, L-shape and A-shape. The U-shape represents an under-dispersed ensemble (less spread), A-shape means the ensemble is over-dispersed (more spread), and the L-shape represents a typically biased forecast. The best ensemble system will be expected to have a constant (or flat line) HD. Figure 2 shows an example of a NCEP GEFS (10-member) forecast for the period December 1st 2004 – February 28th 2005, for 1-, 3-, 5-, 8-, and 10-day NH 500hPa geopotential height. The histogram distribution for the raw forecast is over-dispersed at the short lead-time at that time period. There is a little cold bias for longer lead-times as you see the high bars move right as lead time increases.

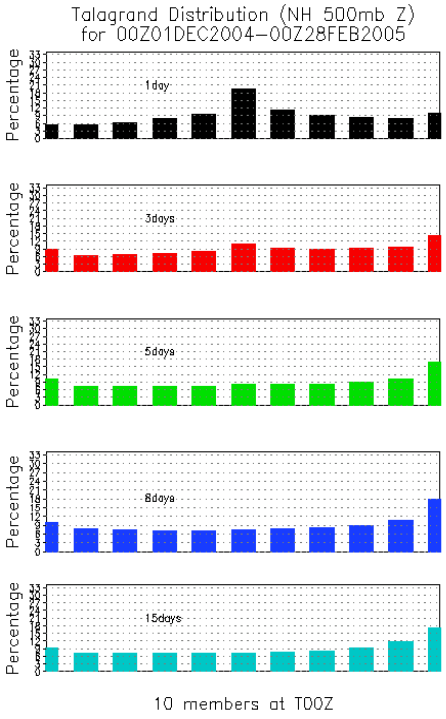

Fig. 2. NCEP global ensemble (10 member) histogram (Talagrand) distribution for NH ex-tropical 500hPa geopotential height for 1-, 3-, 5-, 8-, 15-day forecasts of the 2004-2005 winter.

c. CRPS and RPS

Continuous Ranked Probability Skill Score (CRPSS) and Ranked Probability Skill Score (RPSS) measure the reliability and resolution. The formulas can be written as follows:

$$CRPS=\int_{-\infty}^{+\infty}\left[F(x)-H(x-x_0)\right]^2 dx$$

Where the Heaviside Function H is

$$H(x-x_0)=\begin{cases}0, x \le x_0\\ 1, x > x_0\end{cases} \quad \text{and}$$

$$CRPSS = \frac{CRPS_r - CRPS_f}{CRPS_r}$$

Where *r* is for a reference and *f* is for a forecast.

$$RPS = 1 - \frac{1}{k-1}\left[\sum_{i=1}^{k}(\sum_{n=1}^{i} p_n - \sum_{n=1}^{i} o_n)^2\right]$$

And $$RPSS = \frac{RPS_f - RPS_r}{1 - RPS_r}$$

Where *P* is a forecast probability, and *O* is for an observation or analysis.

For statistics over a long period, CRPS is very similar to RPSS. Therefore, we consider it possible to use either one of these two measures, whichever is more convenient. There is a very good example of this in Figures 5 and 6 for NH extra-tropical 850hPa temperature.

d. Brier score and decomposition

There are many classical references that discuss the Brier score (BS), and its decomposition for reliability and resolution. (Wilks 1995; Toth et al., 2003; 2006). In general, BS can be expressed as the summation of reliability, resolution and uncertainty (Wilks, 1995). CRP or RPS can be considered as a total integration of all probabilities. Users can review all the references to understand BS, reliability, resolution and uncertainty. And here is the final formula for decomposition:

BS = Reliability – Resolution + Uncertainty

e. Hitting rate, false alarm rate, economic value

There is a traditional consideration for the hitting rate and false alarm rate. The typical application for this is the Relative Operational Characteristics (ROC) curve (Toth et al., 2003), or sometimes called the ROC area. Another application is the Relative Economic Value (REV), used when evaluating the cost and loss (Zhu and etc. 2002) which is very useful for decision makers.

3. APPLICATIONS:

The NCEP/GEFS and NAEFS unified verification system will focus on probabilistic forecast verification for mainly short- and medium-range ensemble forecasts. Currently, it is available for the global ensemble forecast only, but it will be soon applied to the short-range ensemble forecast system as well. The discussion in Section 2 (Methodology of ensemble verification) describes the main characteristics of a probabilistic forecast which are more completely measured in terms of reliability and resolution. The NCEP/GEFS and NAEFS product verification statistics have been generated for the seasonal average and the skill scores have been posted at: http://wwwt.emc.ncep.noaa.gov/gmb/yzhu/html/opr/naefs.html since June 2006.

4. VERIFICATION STATISTICS:

Upgrades to the NCEP/GEFS and NAEFS forecast systems are planned for implementation every half year or at least yearly. The verification statistics are basic measurements of model performance to allow the system developers to make a decision on whether to adapt a new method or not. The following selective statistics are part of the NCEP/GEFS and NAEFS verification. All the verified truth comes from the best available analysis; for the NCEP/GEFS the NCEP GDAS analysis will be used as truth, and CMC's analysis will be used to evaluate their forecast. For the joint ensemble (NAEFS) forecast, the mean of the two analyses (NCEP and CMC) will be used to verify joint ensemble forecasts. Figure 3 is for NCEP/GEFS only and compares raw forecasts to bias corrected forecasts. The histogram distributions show the model has a cold bias and less spread with different lead times with the NCEP/GEFS raw forecasts (black). However, the bias is mostly removed after bias correction (red) (Cui et al.,

2006), which was implemented at NCEP and CMC on May 30th 2006. Meanwhile, there are many differences between Figures 2 and 3 (histogram distributions) for the raw forecasts (black curves in Fig. 3) because Figure 2 presents the 2004 ensemble forecast model, while Figure 3 shows the 2006 ensemble forecast model. Figure 3 also has better forecasts than Figure 2 after bias correction (red curves) for all lead-times.

The Figures 4-9 are all skill scores for the 2006-2007 winter season which compare NCEP/GEFS 14-member raw forecasts (black, E14s), CMC/GEFS 16-member raw forecasts (red, E16m) and the combined NAEFS 30-member (NCEP(14) + CMC(16)) raw forecasts (green, E30n). Figure 4 shows CRPSS for NH extra-tropical 500hPa heights. NAEFS (green) raw forecasts are significantly improved in skill for all lead times, especially for longer lead times. In Section 2, CRPSS and RPSS (Figures 5 and 6) were discussed for NH extra-tropical 850hPa temperature. They are very similar to each other which suggest that either score could be used to verify the ranked probability.

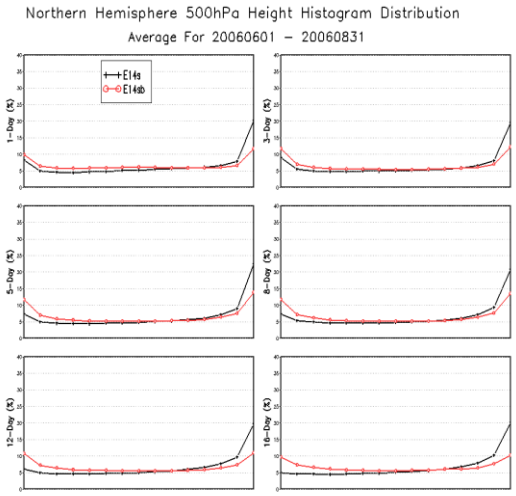

Fig. 3. NCEP global ensemble (14 member) histogram (Talagrand) distribution for NH ex-tropical 500hPa geopotential height for 1-, 3-, 5-, 8-, 13-, 16-d forecasts before (black) / after (red) bias correction of 2006-2007 winter season.

BSS and its decomposition (reliability and resolution) are shown in Figure 7 for NH extra-tropical 1000hPa height. The results are very similar to 500hPa height and 850hPa temperature. According to the formula in Section 2.d, BSS is equal to zero when resolution (going down with time from high to low) equals reliability (which goes up with time). The Reliability diagram (Figure 8) is more popular with many users. The diagonal line is the perfect line for a reliable forecast and the line is the perfect line for a reliable forecast and the further you get from this lines the worse the forecast. Apparently a bias corrected forecast has more reliability (Figure 8, red line).

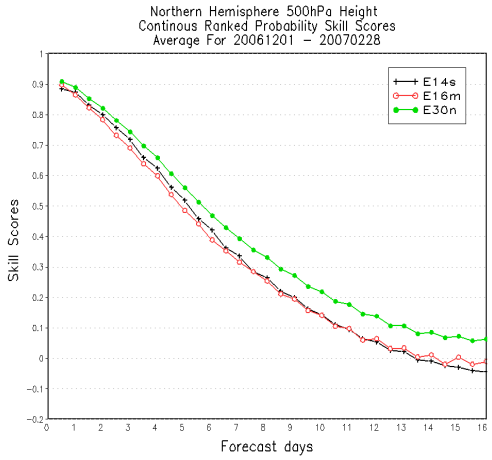

Fig. 4. CRPSS for the NCEP 14 global ensemble raw forecast (black) compared to the CMC 16 global raw forecast (red) and the combined NCEP and CMC ensembles (green) for NH extra-tropical 500hPa geopotential height for the winter 2006-2007.

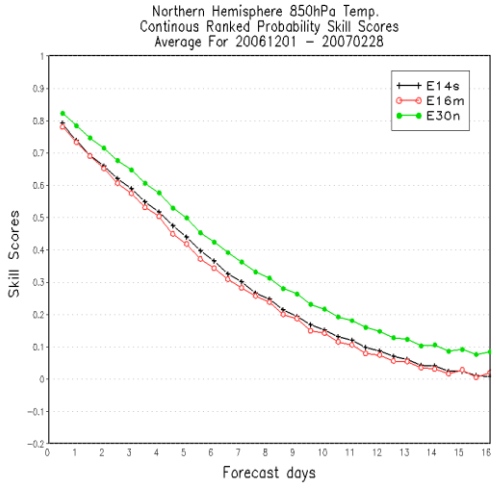

Fig. 5. CRPSS for the NCEP 14 global ensemble raw forecast (black), compared to the CMC 16 global raw forecast (red) and the combined NCEP and CMC ensemble (green) for NH extra-tropical 850hPa temperature for the winter of 2006-2007.

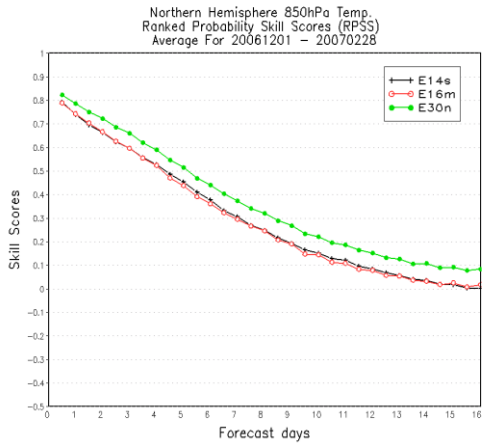

Fig. 6. RPSS for the NCEP 14 global ensemble raw forecast (black) compared to the CMC 16 global raw forecast (red) and the combined NCEP and CMC ensemble (green) for NH extra-tropical 850hPa temperature for the winter of 2006-2007.

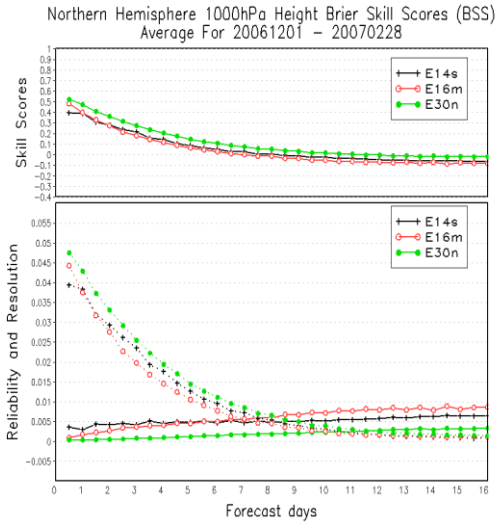

Fig. 7. BSS (top plot), Reliability (bottom plot, solid) and Resolution (bottom, dotted) for the NCEP 14 global ensemble raw forecast (black), compared to the CMC 16 global raw forecast (red) and the combined NCEP and CMC ensemble (green) for NH extra-tropical 1000hPa geopotential height for the winter of 2006-2007.

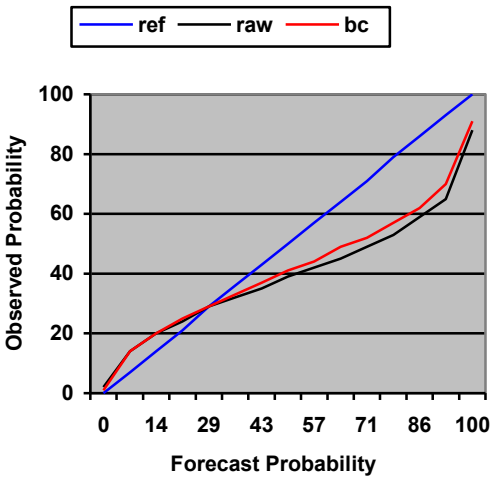

Fig. 8. Reliability diagram of the NCEP 14 global ensemble raw forecast (black) compared to the bias corrected forecast (red) for a 48 hour forecast of NH extra-tropical 1000hPa height for the winter of 2006-2007.

There are some differences between NCEP/GEFS and CMC/GEFS raw forecasts at the near surface (1000hPa geopotential height) when considering the ROC area (Figure 9). There is a similar result for 2-meter temperature for the 2006 CMC/GEFS model. There is much improvement after the CMC/GEFS implementation in July 2007 (personal communication). However, NAEFS still shows a significant improvement at all lead times.

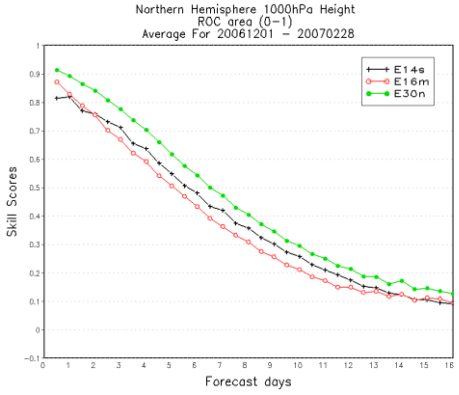

Fig. 9. ROC areas (from 0 to 1) for the NCEP 14 global ensemble raw forecast (black) compared to the CMC 16 global raw forecast (red) and the combined NCEP and CMC ensemble (green) for NH extra-tropical 1000hPa geopotential height for the winter of 2006-2007.

Economic values (Figure 10) really depend on the cost/loss ratio. These values vary when the cost/loss ratio changes (Zhu et al., 2002). In most cases, a 10:1 cost/loss ratio shows (near) maximum economic value.

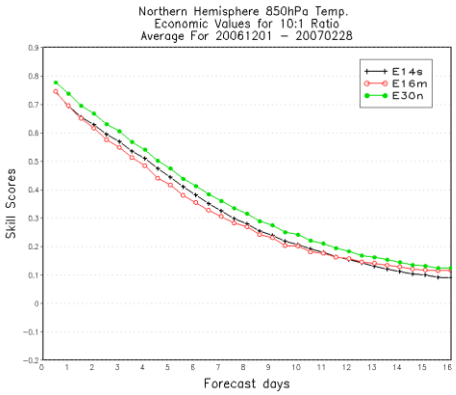

Fig.10. REV (10:1 cost/loss ratio) for the NCEP 14 global ensemble raw forecast (black) compared to the CMC 16 global raw forecast (red) and the combined NCEP and CMC ensemble (green) for NH extra-tropical 850hPa temperature for the winter of 2006-2007.

5. CONCLUSIONS:

In conclusion, authors believe that the verification results from various methods we proposed in Section 2 should be very similar, except for some of the reliability measurements only, such as the histogram distribution.

The Bias correction method which NAEFS introduced is very effective with the ensemble forecast system. There is only one example shown in this article as a demonstration. There is a full article that discusses the bias correction for NAEFS (Cui et al., 2006).

Both NCEP and CMC have very comparable ensemble forecast systems, which means the probabilistic skill scores are very similar to each other in terms of all measurements. Therefore, more value/skill is expected to be added by NAEFS, which is a combination of these two systems.

Acknowledgments: *We greatly appreciate the helps offered by Mary Hart before we finalize this manuscript.*

6. REFERENCES:

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, Y. Zhu, 2005: A comparison of the ECMWF, MSC and NCEP global ensemble prediction systems, *Monthly Weather Review,* Vol. 133, 1076-1097

Candille, G. and O. Talagrand, 2005: Evaluation of Probabilistic Prediction Systems for a Scalar Variable. *Quart. J. Roy. Meteor. Soc.,* 131, 2131-2150

Cui, B., Z. Toth, Y. Zhu and D. Hou, 2008: Statistical Downscaling Approach and its Application. Preprints, *19th Conference on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Sco.

Cui, B., Toth, Z., Zhu, Y., Hou, D., Unger, D., Beauregard, S., 2006: The Trade-off in Bias Correction between Using the Latest Analysis/Modeling System with a Short, versus an Older System with a Long Archive. *The First THORPEX International Science Symposium.* Deccember 6-10, 2004, Montréal, Canada, World Meteorological Organization, P281-284.

Hamill, T., 1997: Reliability Diagrams for Multicategory Probabilistic Forecasts. *Weather Forecast.,* 12, 736-741

Hou, D., Z. Toth, Y. Zhu and W. Yang, 2008: Evaluation of the Impact of the Stochastic Perturbation Schemes on Global Ensemble Forecast. Preprints, *19th Conference on Probability and Statistics,* New Orleans, LA, Amer. Meteor. Soc.

Murphy, A. and H. Daan, 1985: Forecast evaluation. In *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* ed. A. H. Murphy and R. W. Katz, pp. 379-437. Westview Press.

Wei, M. and Z. Toth, 2003: A New measure of ensemble performance: Perturbation versus Error Correlation Analysis (PECA). *Mon. Wea. Rev., 131,* 1549-1565

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

Toth, Z., O. Talagrand, and Y. Zhu, 2006: The attributes of forecast system, In book of: *Predictability of Weather and Climate,* Ed.: T. N. Palmer and R. Hagedorn, Cambridge University Press, 584-595

Toth, Z., Talagrand, O., Candille, G. and Zhu, Y. 2003. Probability and ensemble forecasts. In: *Forecast Verification: A Practitioner's Guid in Atmospheric Science* (eds Ian T. Jolliffe and David B. Stephenson). John Wiley & Sons Ltd., England, 137–163.

Zhu, Y., 2007: Objective evaluation of global precipitation forecast, In special collection of: *International Symposium on Advances in Atmospheric Science and Information Technology,* Beijing, China, p3-8

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability, *Advance in Atmospheric Sciences, Vol. 22,* No. 6, 781-788

Zhu, Y., 2004: Probabilistic forecasts and evaluations based on a global ensemble prediction system, *World Scientific Series on Meteorology of East Asia, Vol. 3* - Observation, Theory, and Modeling of Atmospheric Variability, 277-287

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne 2002: On the economic value of ensemble based weather forecasts, *Bulletin of American Meteorological Society, Vol. 83,* 73-83

Zhu, Y., G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., J79–J82.