Tethys, 3, 59–65, 2006
www.tethys.cat
ISSN-1139-3394
DOI:10.3369/tethys.2006.3.08

Journal edited by **ACAM**
(Associació Catalana de Meteorologia)

Involving human forecasters in numerical prediction systems

Víctor Homar[1] **and David Stensrud**[2]

[1]Grup de Meteorologia. Departament de Física. Universitat de les Illes Balears
[2]National Severe Storms Laboratory. National Oceanic and Atmospheric Administration. USA

Received: 25-VII-2006 – Accepted: 19-X-2006 – **Original version**

Correspondence to: victor.homar@uib.es

Abstract

Human forecasters routinely improve upon the output from numerical weather prediction models and often have keen insight to model biases and shortcomings. This wealth of knowledge about model performance is largely untapped, however, as it is used only at the end point in the forecast process to interpret the model-predicted fields. Yet there is no reason why human forecasters cannot intervene at other earlier times in the numerical weather prediction process, especially when an ensemble forecasting system is in use. Human intervention in ensemble creation may be particularly helpful for rare events, such as severe weather events, that are not predicted well by numerical models. The USA/NOAA SPC/NSSL Spring Program 2003 tested an ensemble generation method in which human forecasters were involved in the ensemble creation process. The forecaster highlighted structures of interest and, using an adjoint model, a set of perturbations were obtained and used to generate a 32-member ensemble. The results show that this experimental ensemble improves upon the operational numerical forecasts of severe weather. The human-generated ensemble is able to provide improved guidance on high-impact weather events, but lacks global dispersion and produces unreliable forecasts for non-hazardous weather events. Further results from an ensemble constructed by combining the operational ensemble perturbations with the human-generated perturbations shows promising skill for the forecast of severe weather while avoiding the problem of limited global dispersion. The value of human beings in the creation of ensembles designed to target specific high- impact weather events is potentially large. Further investigation of the value of forecasters being part of the ensemble creation process is strongly recommended. There remains a lot to learn about how to create ensembles for short-range forecasts of severe weather, and we need to make better use of the skill and experience of human forecasters in this learning process.

1 Introduction

The numerical forecasting of mesoscale phenomena and severe convective weather poses one of the most challenging problems faced today in the atmospheric community. Model physics, resolution and data assimilation techniques are continuously improving and examples of promising simulations of severe convective systems can be found in the literature. However models still do not provide consistently reliable guidance for operations about important aspects of severe weather such as initiation, mode, intensity and evolution of convection. Admittedly, short-range mesoscale numerical forecasts are hampered by the largely unknown observational sampling errors at the meso- and small-scales, as well as by the deficiencies in the models from such sources as physical parameterization schemes. Additionally, little is known about the limits of predictability at the spatial and temporal scales of intermittent weather systems responsible for producing severe weather. The perception that multiple sources of uncertainty may largely degrade the forecast decreases the confidence of the forecaster in the output produced by mesoscale numerical models, even when they provide highly realistic looking forecasts. Inevitably, observational dataset errors and model deficiencies, as well as the predictability concerns, introduce inherent uncertainties that always are present in the forecast.

Ensemble techniques are one method that can be used to explicitly account for uncertainties in the numerical forecasting system and their use may assist forecasters in assessing appropriate levels of confidence. However, identifying, quantifying and representing these uncertainties in the forecast system is a complex task. It is well known that combining the solutions of a number of slightly different numerical simulations not only produces a forecast that is more skillful than each individual simulation when examined over many cases (Leith, 1974), but also provides a quantitative indication of forecast uncertainty. How these realizations (i.e. ensemble members) are constructed is currently the subject of significant attention in the weather research community (Shapiro and Thorpe, 2004).

Multiple methods to choose an optimum ensemble of realizations that accounts for the analysis errors have been proposed. For forecasts in the medium-range, two well established strategies have been adopted by the major operational centers in the United States and Europe. The breeding (Toth and Kalnay, 1993) and singular vector (Buizza and Palmer, 1995) techniques have provided notable improvements in the skill of the medium-range forecasts, even without considering model deficiencies. Unfortunately, accounting for the initial conditions errors for applications on the mesoscale becomes more complex due to the larger and less known analysis error, the large role that physical process parameterization schemes play in model forecasts of sensible weather, and the end user's more sensitive dependence upon reliable forecasts.

Xu et al. (2001) propose a method to generate members for a short-range ensemble that benefits from forecaster's guidance in identifying areas where threatening weather is likely in the forecast and the atmospheric features that can influence the development of the threatening weather. The approach of Xu01 assumes that the experience and skill of human weather forecasters is a useful addition to the process of creating ensemble systems. It is well known that forecasters routinely improve upon numerical guidance, as clearly seen in skill scores for precipitation (Olson et al., 1995, e.g.). In addition, forecasters at the USA/NOAA/Storm Prediction Center regularly identify mesoscale-sized regions of significant severe weather threat through the issuance of outlooks and severe weather watches with high level of skill. There is no reason to assume that this human knowledge and experience, although subjective, cannot be made useful in the creation of ensemble members and thereby benefit the operational forecast process particularly for rare and significant events.

With the aim of assessing the value of short-range numerical forecast ensembles to assist in the operational forecasting of severe weather, the Storm Prediction Center and the National Severe Storms Laboratory, two USA/NOAA centers, conducted the 2003 Spring Program (SP03) experiment focused primarily on the generation and interpretation of mesoscale short- range ensembles. Encouraged by the promising conclusions of Xu et al. (2001), the SP03 included a subexperiment to test their method for a larger number of cases using operational forecasters as the main drivers of the system. The underlying idea was to create a daily, customized ensemble to provide guidance on the severe weather threat over the following 48 hours. Essentially, the ensemble dispersion was intended to be generated in specific areas, and focused upon specific fields of interest as opposed to everywhere in the domain, or following fast growing modes under global generic norms. This paper presents an overview of the verification results of this SP03 test ensemble.

2 Ensemble generation

The forecaster-generated ensemble consists of 32 members produced using the nonhydrostatic Pennsylvania State University-National Center for Atmospheric Research (PSUNCAR) Fifth-generation Mesoscale Model (MM5V3, www.mmm.ucar.edu/mm5/). This test ensemble of the SP03 experiment (MM5ADJ) ran weekdays from April 28 to June 6 (SP03 did not operate on weekends). To generate the set of different ICs for the ensemble, the method detailed in Xu et al. (2001) was followed: each day an experienced human severe weather forecaster was asked to identify 16 features of interest in the control run that were, in the forecaster's opinion, important to the potential development and/or evolution of severe weather on the following day (12 UTC to 12 UTC day 2). The forecaster was able to select atmospheric structures at any time (in 6 h intervals) from the 48 h Eta control forecast. Figure 1a show examples of human-drawn features of interest for the forecast cycle of May 5 2003.

Each day, for each of the 16 selected features of interest, an adjoint model integration (Errico, 1997) was correspondingly initialized and the sensitive areas of each forecaster-specified feature to the IC were derived. The adjoint model used is the MM5 Adjoint Modeling System developed by National Center for Atmospheric Research. Once the sensitivity fields were obtained from the adjoint, the horizontal wind components and temperature sensitivities were rescaled to a maximum amplitude of 8.0 m s-1 and 4.0 K, respectively. Finally, two MM5 simulations were run for each highlighted feature, each one using perturbations in both directions (positive and negative). Figure 1b shows an example of such perturbations for the temperature field at 700 hPa. Since the forecaster was requested to highlight 16 features each day, 32 perturbed simulations were produced to form the MM5ADJ ensemble.

Although the adjoint model is tangent linear, and hence the perturbations were defined strictly to change the forecaster-selected feature in a linear sense, the nonlinear evolution of the perturbation can be interpreted as a stochastic perturbation to the initial model state trajectory. However, this stochastic component of the perturbation will likely be confined about the area of concern in the forecast at the forecast time selected. In essence, by using both positive and negative perturbations the feature of interest likely is both enhanced and reduced equally in the linear sense. The nonlin-

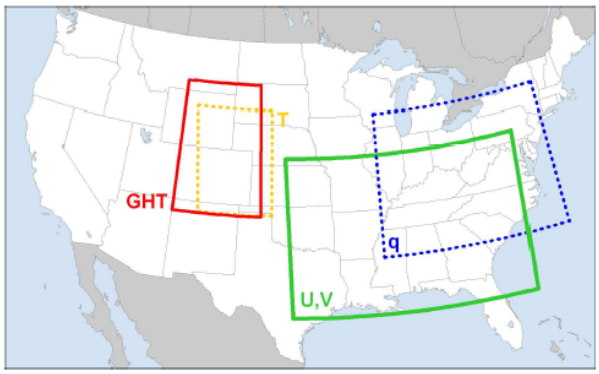

Fig. 1a

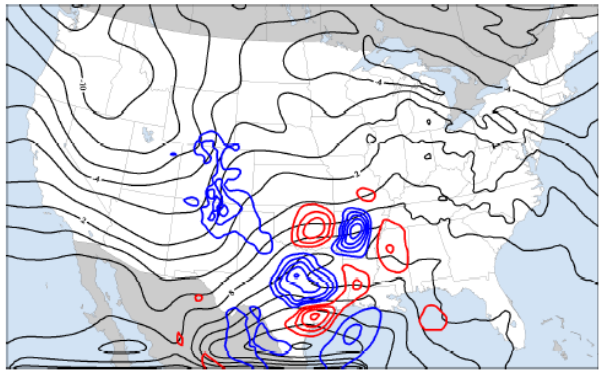

Fig. 1b

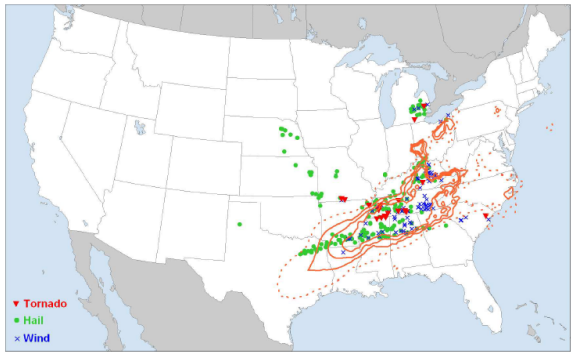

Fig 1c

Figure 1. Example steps from the proposed ensemble generation process: a) Areas and fields selected by the forecaster (here, geopotential height (GHT), temperature (T), wind components (U,V) and specific humidity (q)); b) Example of initial condition temperature perturbation at 700 hPa (black contours depict unperturbed temperature, red and blue show positive and negative (i.e. warm and cold) perturbations) used to generate a single ensemble member and derived from an adjoint model applied to the fields shown in a); and c) Storm reports and probability of severe weather (orange lines in 25% intervals) as forecast by the test ensemble for the 24 h period beginning on 12 UTC May 5. Dotted lines depict the 5% probabilities.

ear evolution of the positive and negative perturbations, however, may yield unexpected results since the specified feature of interest likely is not enhanced and reduced symmetrically in the two nonlinear forecasts. This nonlinear behavior is viewed as a positive attribute of the system, ensuring a rich diversity of solutions among the ensemble members over the forecaster defined regions of concern as opposed to the trivial effects of the purely linear evolution of the linearly-derived perturbations.

3 Verification datasets

The evaluation of the MM5ADJ is based on observations of severe weather over the continental United States (CONUS), east of the Rockies. The observational dataset used for verification is the SPC severe weather reports. This database contains a real-time list of tornado, large hail (larger than 20 mm) and convective wind (stronger than 50 knots) damage reports in the United States with information about the intensity of the event and its location in space and time. Figure 1c shows an example of the reports in the SPC database.

In addition to the objective verification against the observational dataset, the relative value of the MM5ADJ is assessed by comparing it against the operational short-range forecasts available for the same period:

- Subjective Day 2 Outlooks: The SP03 forecaster issued an experimental severe weather outlook for Day 2, following the same guidelines used for the routine operational SPC outlooks. Since the SPC outlooks are issued using 5 discrete probability categories: 0.00, 0.05, 0.15, 0.25 and 0.35, and in order to ensure fair comparison among the considered predictions, all forecasts considered in this study are truncated onto these categories.
- Operational Eta: The operational 12 UTC daily run from the NCEP Eta is included to add a reference from a deterministic model into the comparison.
- NCEP SREF System: The NCEP ensemble for short range forecasting during SP03 consisted of 10 members, five Eta and five Regional Spectral Model (RSM) members. Unfortunately, owing to problems with the data archive, only 11 days are available for comparison

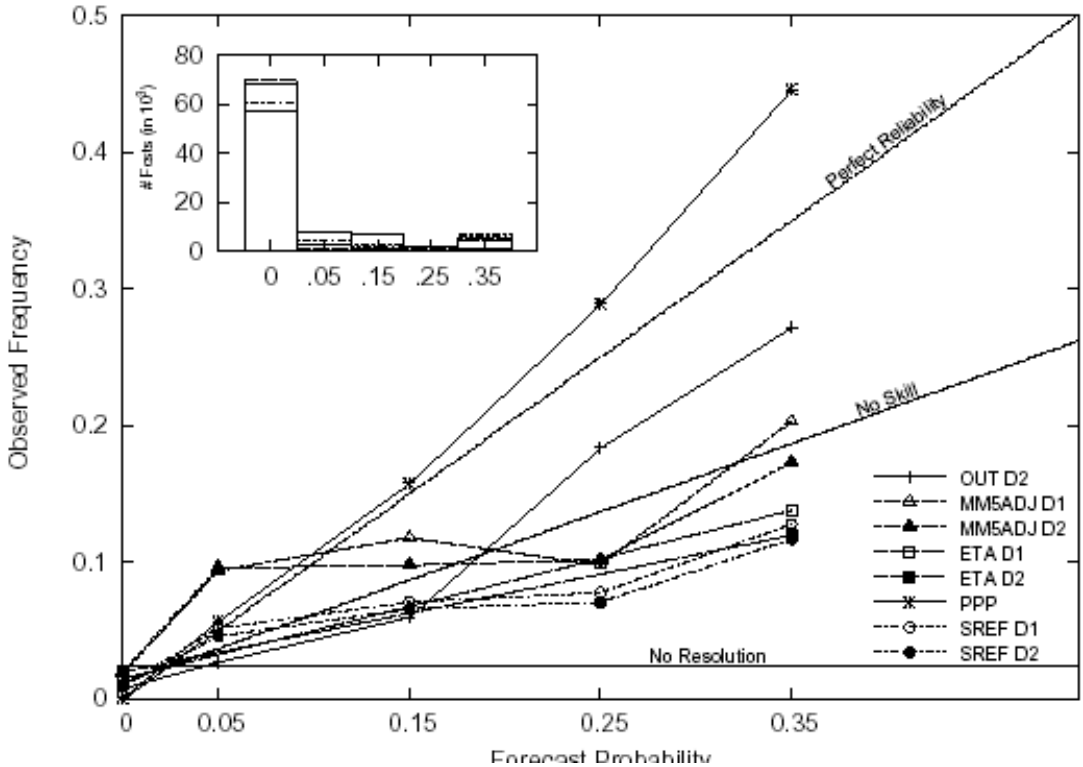

Figure 2. Attributes diagrams for the probability of severe weather as obtained from: SP03 preliminary Day 2 outlooks, T+24 and T+48 h MM5ADJ, Eta and SREF.

during the period that the SP03 lasted. All results obtained from such small sample of 11 days are complemented with a statistical significance test.

- Practically Perfect Prog: Brooks et al. (2003) discuss the concept of the practically perfect progs (PPP) and present the main characteristics. This hypothetical forecast is as accurate as could be expected for a forecaster already aware of the reports, given the limitations of real-world forecasting.

4 Verification of severe weather forecasts

Unlike the SPC human-rendered outlooks, current models do not explicitly forecast severe weather. The diagnosis of severe weather from analysis or models that do not explicitly resolve convection can be inferred, at least in part, through indices that characterize the environment and may allow some basic discrimination of the type or intensity of convective phenomena supported. In this study, severe weather is defined to occur within a grid box when both the Supercell Composite Parameter (SCP, Thompson et al. (2002)) > 1 and the triggering of the model's convective scheme occur simultaneously. Together, these two quantities specify regions in which the model jointly predicts an environment that is favorable for supercell thunderstorms, and in which convection develops.

Hence, the probability of occurrence of severe weather during a 24 h period at every grid point is simply defined as the number of ensemble members having a SCP larger than 1 and simultaneous convective precipitation at that grid point anytime during that 24 h period, divided by the total number of ensemble members. Figure 1c shows an example of this probability field from the MM5ADJ.

Verification of the probabilistic forecasts for all cases is done by using the attributes diagram. This diagram shows the observed frequency of an event as a function of the forecast category and allows an interpretation of skill for each forecast category separately. Figure 2 shows the attributes diagram for all the forecasts compared in this study. The sample climatological frequency is 0.016 severe reports per gridpoint during 24 h. Not surprisingly for the prediction of unlikely events, all forecasts in the comparison show good skill at predicting no occurrence of severe events (0.00 probs), with the human outlooks showing the highest reliability in this category. For low (0.05) and moderate (0.15) probabilities, the MM5ADJ and SREF are the only forecasts showing some skill, with especially good reliability at the low category for the SREF.

For higher probabilities (when a majority of the ensemble members agree), the MM5ADJ is the only model showing some skill at the 0.35 probability category and some resolution still exists between the 0.25 and 0.35

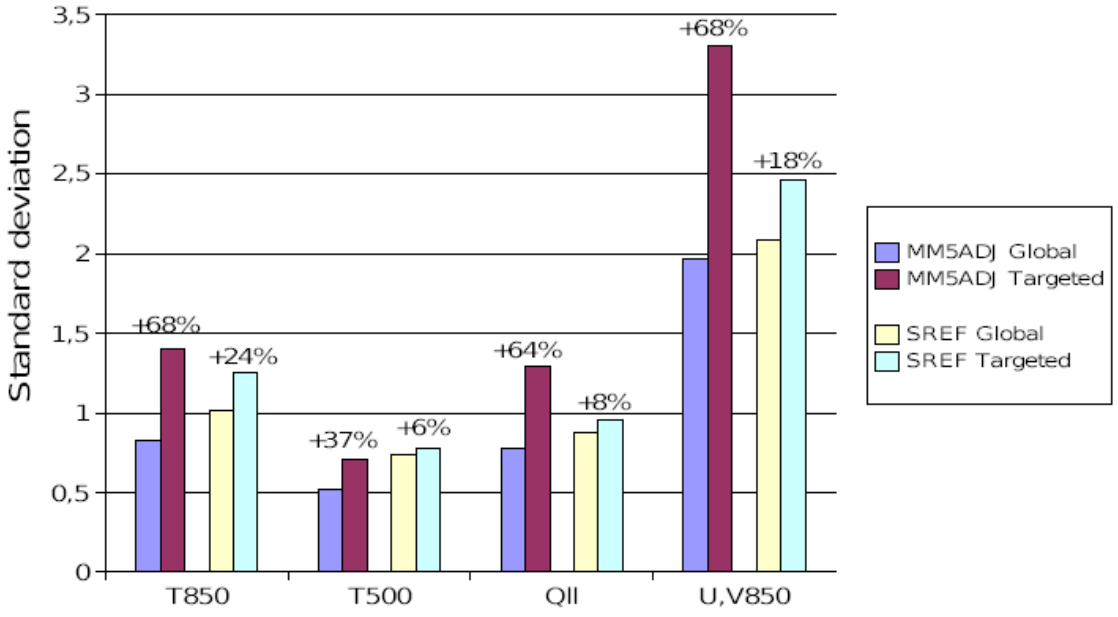

Figure 3. Mean of the standard deviation of the 36 h forecast, computed at sounding sites. Global values are averaged over the CONUS east of the Rockies, and the Targeted include only the areas delineated by the forecaster when defining the perturbations. Values above the bars indicate the percent change from the global to the targeted standard verification. Qll refers to the average of the standard deviation of Q at 1000, 850 and 700 hPa.

forecasts. The human outlooks, however, show skill at the high probability categories, revealing the skill of the forecasters when they show high confidence in the intensity of the situation of the day and decide to use high probabilities in the outlook. On the other hand, Eta forecasts are clearly hampered in this type of probabilistic verification, showing a clear overforecast of severe weather. Although, the SREF results show almost perfect reliability for the low category it has no skill for higher probability categories. The significance of the differences between the MM5ADJ and SREF results is assessed using a bootstrap non-parametric test with 10000 samples. Most of the differences between MM5ADJ and SREF visible in Fig. 2 are significant to the 99% confidence level. This result clearly shows the advantage of the MM5ADJ over the SREF in forecasting probabilities of severe weather at and above 0.15, usually associated with the more intense and damaging episodes. This is most likely a consequence of the customized design of the MM5ADJ to focus on the areas of severe weather threat, whereas the SREF system is designed to cover a wide range of mesoscale forecast aspects and shows its strength at the low probability range.

4.1 Targeted spread

To better understand the differences between the MM5ADJ and SREF systems in forecasting higher probability (0.25) episodes of severe weather, we analyze the ability of the MM5ADJ to generate spread specifically over the areas of concern defined by the forecaster. Two versions of the spread for each model are computed (Figure 3): the Global spread is the mean of the spread calculated at each sounding site within the CONUS, east of the Rocky Mountains; the Targeted spread is computed considering only the forecast at sounding sites within the areas of concern and times designated by the forecaster in constructing the ensemble. The relative increase of spread from the global to the targeted spread is much larger in the MM5ADJ than in the SREF, especially in low levels where increases in spread ranging 65 to 70% are obtained. Therefore, the breeding vectors technique (as well as the model diversity) in the SREF system produces larger dispersion in a global sense, whereas the customized MM5ADJ successfully targets ensemble dispersion both spatially and temporally over the region selected by the forecaster.

5 Mixed ensemble test

In order to test the effect of adding members to the MM5ADJ system that provide spread across the entire domain, we evaluate the forecast skill of an ensemble generated by combining the 32 MM5ADJ and 10 SREF members to produce a 42 member ensemble (42 ENS). This ensemble not only will benefit from a large number of members but also from being multimodel and including two initial conditions perturbation techniques. This ensemble is still primarily focused on targeting severe weather but may

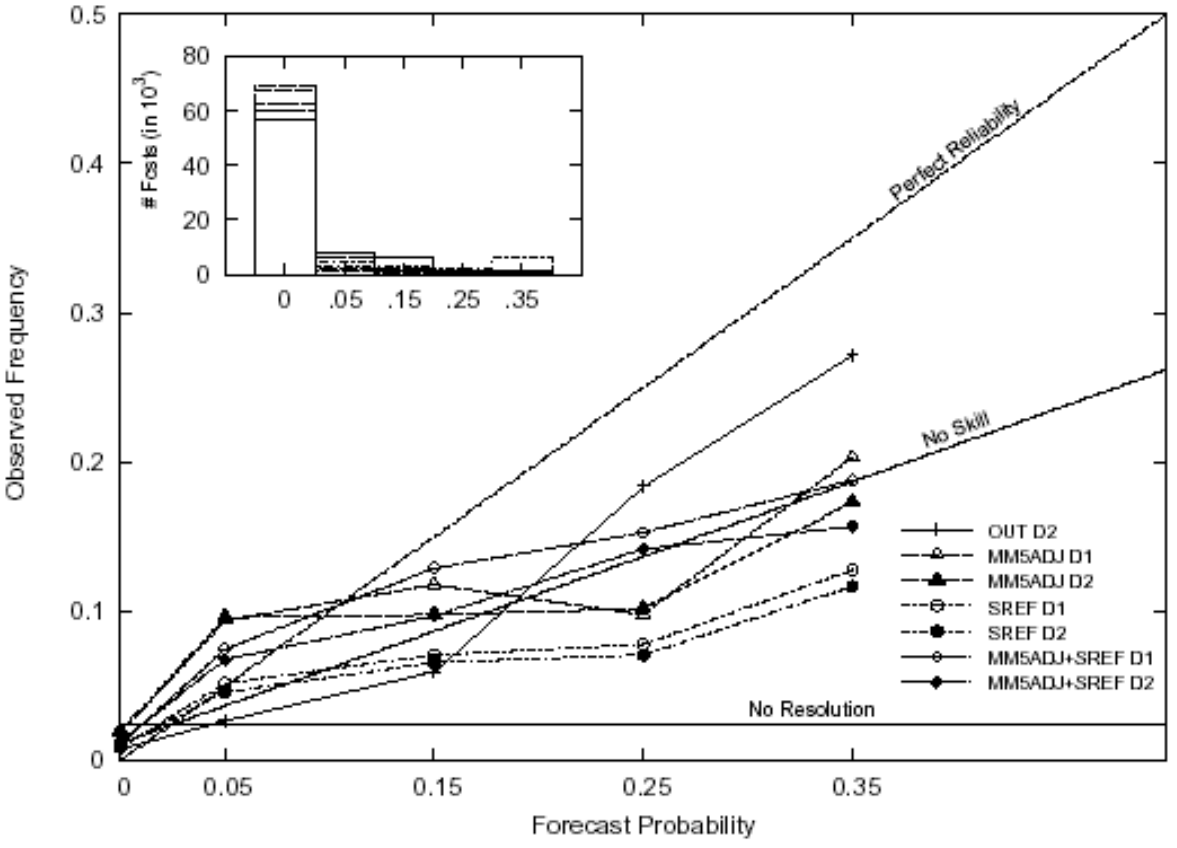

Figure 4. As in Fig 2 but for the 42 ENS system.

also benefit from the globally better scores of the 10 SREF members.

Severe weather forecasts are produced for the 42 ENS following the same method presented in the previous section. The bootstrap non-parametric test is also used to assess the significance of the differences between 42 ENS and MM5ADJ results. The attributes diagram curve for the 42 ENS forecasts shows the superior skill of this configuration as compared to the MM5ADJ for almost all probabilities (Fig. 4). Only for the 15 and 30% categories, the 42 ENS does not produce results significantly better than MM5ADJ for both Day 1 and 2.

6 Conclusions

The SPC/NSSL Spring Program 2003 included an experimental ensemble aimed at testing for an extended period of time the ensemble generation method of Xu et al. (2001) who proposed using human forecasters to identify atmospheric features they believed to be important to the development and evolution of severe weather during the 24-48 h forecast period. Using an adjoint model, perturbations to the forecast model initial conditions that would influence these forecaster selected atmospheric features are identified and used to create an ensemble of model forecasts. The performance of this experimental ensemble is evaluated by using severe weather reports.

The experiment was designed to run in real-time, with the initial hope that forecasters would have time to examine and verify the ensemble forecasts and gain experience in selecting the perturbation fields, vertical levels, and areas. Unfortunately, computer limitations did not allow for this learning experience to happen as the forecasts were available too late in the day. Thus, the forecasters were only given basic guidance on how to generate the perturbations. Many other aspects of the experiment also are imperfect and should be improved upon in future experiments. Yet the initial results are suggestive and warrant careful consideration.

Verification results show value in the experimental ensemble forecasts compared to the operational SREF system, despite the multiple improvements still possible to the experimental system. A single model is used in the experiment, with the human-selected perturbations the only source of dispersion in the ensemble system. Although basic

training was provided at the beginning of each experimental week of SP03 covering the selection of fields, levels, sizes, and time of the targeted structures, no definitive rules were made available to the forecasters on the construction of perturbations, because this had never before been conducted as a real- time experiment. Additionally, the forecasters had no previously experience with this type of ensemble creation and no quantitative feedback was provided to them during the experiment. Further research might indicate whether certain sizes, fields and levels are more appropriate to define the perturbations for specific types of predicted weather.

Despite the lack of previous knowledge and experience using this technique, the experimental ensemble is shown to improve the numerical forecasts of severe weather, arguably because it successfully generates dispersion over the areas of concern selected by the forecaster. However, the experimental ensemble forecasts of low probability severe weather have less skill than those of the SREF and operational Eta. A clear conclusion from these results is that this ensemble, customized to exclusively focus on high-intensity and damaging weather, lacks global dispersion and produces unreliable forecasts for non-hazardous weather events. Results from an ensemble constructed by combining globally perturbed members (from SREF) and humanly perturbed members (from MM5ADJ) show promising skill for the forecast of severe weather. While the experimental set up was not perfect, the results indicate that the value of human beings in the creation of ensembles designed to target specic weather threats is potentially large.

Further investigation of the potential value of humans being part of the ensemble process is strongly recommended, even if the end result is to learn how forecasters can provide real- time input into an automated ensemble generation system. We still have a lot to learn about how to create ensembles for short-range forecasts of high impact weather, and we need to make better use of the skill and experience of human forecasters in this learning process.

References

Brooks, H., Doswell-III, C. A., and Kay, M. P., 2003: *Climatological estimates of local daily tornado probability for the united states*, Weather Forecast., **18**, 626.640.

Buizza, R. and Palmer, T. N., 1995: *The singular vector structure of the atmospheric general circulation*, J. Atmos. Sci., **52**, 1434–1456.

Errico, R. M., 1997: *What is an adjoint model?*, Bull. Amer. Meteorol. Soc., **78**, 2577–2591.

Leith, C. E., 1974: *Theoretical skill of monte carlo forecasts*, Mon. Weather Rev., **102**, 401–418.

Olson, D. A., Junker, N. W., and Korty, B., 1995: *Evaluation of 33 years of quantitative precipitation forecasting at the NMC*, Weather Forecast., **10**, 498–511.

Shapiro, M. A. and Thorpe, A. J., 2004: *Thorpex international science plan*, Tech. rep. World Meteorological Organization, **Version 3**.

Thompson, R. L., Edwards, R., and Hart, J. A., 2002: *Evaluation and interpretation of the supercell composite and significant tornado parameters at the storm rediction center*, 21st Conference on Severe Local Storms, Amer. Meteorol. Soc., San Antonio, TX, USA, pp. J11–J14.

Toth, Z. and Kalnay, E., 1993: *Ensemble forecasting at NMC: The generation of perturbations*, Bull. Amer. Meteorol. Soc., **74**, 2317–2330.

Xu, M., Stensrud, D. J., Bao, J.-W., and Warner, T. T., 2001: *Applications of the adjoint technique to short-range ensemble forecasting of mesoscale convective systems*, Mon. Weather Rev., **129**, 1395–1418.