The Effects of High Resolution Model Output on Severe Weather Forecasts as Evaluated in the SPC/NSSL Spring Program 2004

Adam French
National Weather Center Research Experience for Undergraduates
Valparaiso University

Steven J. Weiss
NWS Storm Prediction Center, Norman, Oklahoma

John S. Kain
Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma and
NOAA National Severe Storms Laboratory

Research Experience for Undergraduates Final Project

Last Revision: 29 July 2004

Corresponding Author:

Adam French
8 Woodside St
Manchester, CT 06040
Ph. (860) 646-6660
Email: Adam.French@valpo.edu

This material is based on work supported by the National Science Foundation under Grant No. 0097651

Abstract

The goal of this study is to determine the usefulness of high resolution model output when it comes to severe weather forecasts and what the value of this usefulness might be in relation to the cost of running such models. This was accomplished by looking at data gathered during the 2004 SPC/NSSL Spring Program comparing a forecast made using current operational models, and then a forecast made for the same forecast period using the 4km WRF output. Further analysis was done using verification data for the four models examined. The results of these analyses determined that the 4km WRF models improve upon current abilities when it comes to forecasting severe weather, however these improvements were small. The small improvements combined with the high cost (in computer time and money) of running the high resolution models makes their value hard to determine. However, given some of the potential evident in cases such as 28 May 2004 it makes sense to continue work on development of 4km WRF models as future versions may provide a vast improvement over current forecast models.

1. Introduction

Every spring since 2000, a collaborative effort between researchers and forecasters has taken place at the Storm Prediction Center (SPC) in Norman, OK. Known as the Spring Program, this collaboration is geared towards developing and evaluating tools to improve severe weather forecasts, at the SPC (Weiss et al, 2004). The main focus for this development in previous years has been on the use of numerical weather prediction models in applications that would benefit the SPC (Janish et al; 2001, Weiss et al, 2003). This trend continued in 2004 when the overarching goal of the program was to determine the usefulness of high resolution forecast models in improving severe weather forecasts.

The impetus for this undertaking stems from the ongoing desire for increased lead time on convective watches, both on the part of those issuing the watches and those who use them, including Weather Forecast Offices (WFOs) and the public (Weiss et al, 2004). While this undertaking necessitates the use of current observed data to the fullest extent possible, increasing watch lead time requires a look several hours into the future to see what the atmosphere is going to do and how the situation is going to evolve. This view into the near-term future is something that can be provided by forecast models, and when it comes to predicting severe weather on the scale of thunderstorms and convective complexes, it makes sense to look at model output that is on a similar scale. This can be done through the use of high resolution (small grid spacing) models that can depict these smaller scale features. However, the amount of computing time it takes to run the model increases greatly as a function of increasing resolution (Lilly, 1990). Consequently one has to find a grid spacing that is small enough to resolve mesoscale features, yet not so

small that it takes prohibitively long to run the model. According to Weisman et al (1997) a grid resolution of 4km provides just such a medium, being the largest grid scale that can still handle and properly resolve mesoscale features. In light of this, three Weather Research and Forecasting (WRF) models run on a 4km grid spacing with explicit precipitation processes (no parameterized convection) were evaluated in the 2004 Spring Program to test their utility for forecasting severe weather in comparison with a pair of currently operational, and several experimental, mesoscale models.

The primary goal of the 2004 Spring Program was to evaluate how useful these higher resolution forecast models are in forecasting severe weather events. A related secondary objective was to assess not just if they are useful, but if they are useful enough to warrant devoting the resources (both in money and computing power) to run them operationally (Weiss et al 2004). For a higher resolution model to be run operationally it would require a large investment in both money and computing power, and it is necessary to determine whether or not such an investment would really be worthwhile. Thus, the goal of this study is not just to determine whether or not higher resolution model output provides forecasters with a useful tool that can aid in improving severe weather forecasts, but also determine if the usefulness and improvement are significant enough to warrant dedicating the resources needed to make such a modeling system.

2. Methodology

To determine the usefulness of higher resolution models in the forecasting of severe weather, three WRF models run at 00z on a 4km grid spacing were evaluated regarding their handling of convective initiation, convective evolution and convective

mode. These three models included a WRF Mass Core run by the University of Oklahoma Center for Analysis and Prediction of Storms (OU/CAPS) that used the ARPS Data Assimilation System (ADAS) and Level II radar data for initialization (hereafter referred to as the WRF-CAPS), a WRF Mass Core run by the National Center for Atmospheric Research (NCAR) using 40km Eta background fields for initialization (hereafter WRF-NCAR), and the WRF/NMM from the Environmental Modeling Center (EMC) using EMC's Nonhydrostatic Mesoscale Model (NMM) core, also initialized with 40km Eta background fields (hereafter WRF-NMM) (Weiss et al, 2004). Output from these three models was used to update a preliminary human forecast that was based on observational data as well as output from the operational 12z 12km Eta, 12z 20km RUC models and 09z EMC Short Range Ensemble Forecast (SREF). This procedure was designed so it could be determined how the higher resolution models affected the forecast, and whether or not they contributed to an improved forecast.

The daily operations of the 2004 Spring Program began Monday through Friday at 8AM. The forecast/evaluation teams met and discussed the preliminary and final forecasts made the previous day, using radar data and severe storm reports to verify each forecast and subjectively assess how well it predicted the day's events. Following this verification, the focus shifted to the preliminary forecast for severe weather valid during a six hour period (usually 18-00z) for the current day. For this, the forecasters looked at observed weather data (satellite imagery, surface data, upper air data, radar, etc) as well as output from the operational 12z 12km Eta, 12z 20km RUC, and 09z SREF. Output from the SREF was examined, however it was not evaluated in the same in depth manner as the mesoscale and storm scale models, and thus is not included in this paper. From

these data, as well as the current SPC Day One outlook, the forecasters picked a focus region and then proceeded to delineate specific severe storm threat areas within the region using probability lines similar to those found in SPC outlooks. After this forecast was issued, the team completed the web-based Real Time Mesoscale Model Evaluation (RTMSME), which included specific questions designed to measure forecaster confidence in the 12km Eta and 20km RUC solutions for the day's severe weather threats.

Once the preliminary forecast and RTMSME were complete, the forecast teams shut down any real-time data feeds that were coming into the SPC's Science Support Area (the SSA, where the Spring Program took place) and began examining experimental output from the three 00z 4km WRF models in order to make the final forecast. The incoming real-time data was shut down in order to ensure that any changes to the preliminary forecast were based only on WRF model output. After studying this output the forecasters made any changes they felt necessary to the preliminary forecast, creating a final forecast for the same six-hour forecast period. They then completed the Real-Time Storm Scale Model Evaluation (RTSSME) forms which, like the RTMSME, provided an expression of the forecaster's confidence in how the higher resolution models were handling the expected weather situation.

Following the daily SPC/NSSL map discussion regarding the day's forecast challenges, the team evaluated the performance of the forecast models for the previous day's forecast period. In this evaluation, rather than assessing forecaster confidence in the models, the team used radar data to verify how well the various models actually did with regards to convective initiation, convective evolution, and convective mode. In this

manner all five of the aforementioned models were evaluated. In addition, this assessment also included a 12z run of the WRF-CAPS model, and two mesoscale WRF models run with parameterized convection, a WRF-NMM with 8km grid spacing and a WRF-NCAR at 10km. These additional models are not included in the evaluation discussed in this paper as they were only verified the next day, and were not used in the creation of either the preliminary or final forecasts. As such they played no part in showing the effects of the high resolution models on the forecast.

It is important to stress that both the real time and next day evaluations are subjective by design. Rather than using an objective verification metric, such as the Equitable Threat Score, as is common when evaluating the performance of forecast models. The Spring Program utilizes a more subjective format based on discussion between the forecasters and researchers that leads to a consensus conclusion regarding the score for each specific model. This consensus is based on the opinions and impressions of the forecasters. While this may appear less accurate, as it can include the biases of the forecasters involved, it reflects the needs and requirements of forecasters who are actually using the model, so that the model and model display systems can be developed in a way that will be most useful to those using it. Through its use in previous Spring Programs, this type of subjective verification, compliments objective measurements and has been deemed to be a viable and useful method for evaluating forecast models (Kain et al 2003).

3. Data

The 2004 Spring Program was conducted from 19 April to 4 June 2004, during which 35 days of model and forecast evaluation data. These data consisted of subjectively determined numerical scores and written comments on a variety of topics related to the models being evaluated. While in theory this should result in a fairly sizeable dataset, issues arose as model availability and applicability varied on a day to day basis, meaning that for any given day there may be data missing for one or more of the models being evaluated. This created a problem as far as constructing datasets that maximized both the number of days in the sample as well as the number of models available on each day.

Since the focus of this paper is to determine the usefulness of high resolution model data in improving severe weather forecasts, it was deemed necessary to first and foremost include the three 4km WRFs in the analysis. The original intention was to compare these three models with both the 12km Eta and the 20km RUC, in order to compare the experimental high resolution models with both of the mesoscale models currently used in operational forecasting. However, due to the limited number of days of complete RUC data[*], the three 4km WRFs were compared solely to the 12km Eta. This resulted in a 15 day dataset of model verification data (from the Yesterday's Model Verification dataset or YMV), allowing for comparisons of model performance between the operational Eta and 4km WRF models. Also, a dataset of human forecast verification data (looking at the accuracy of both the preliminary and final human forecasts, and any

[*] While the RUC was available very often during the Spring Program, it only ran out 12 hours from 12z, so if the forecast period extended past 00z the RUC would be unusable for some if not all of the evaluation fields. Also, there were some cases where the RUC's three-hourly output made it impossible to compare to the other models, and thus it was not evaluated. For these reasons it was decided to discount the RUC data in this analysis in favor of a larger dataset.

improvement or decline from the first to second) spanning 29 days was utilized to examine how use of the higher resolution models impacted severe weather forecasts. For the model verification data as well as the forecast verification data, both numerical scores as well as written comments were used in order to determine the skill levels of the models and forecasts, and to see why they ended up scored the way that they were.

4. Results

4.1 Model Verification

After analyzing the aforementioned datasets, it was concluded that higher resolution forecast models do provide improvement when it comes to forecasting severe weather. This result was evident from both the model evaluation/verification data as well as from the preliminary and final forecast verification data. After compiling mean scores for convective initiation, convective evolution and convective mode from the next day verification data for the 12km Eta, 4km WRF-CAPS, 4km WRF-NCAR and 4km WRF-NMM, it was determined that the 4km WRF-NMM scored the highest for all of the evaluation criteria. As was seen in Fig. 1, the WRF-NMM scored significantly higher than the next model in the rankings, the 12km Eta, in all categories except for convective evolution. Here, significance refers to statistical significance determined using a t test.

4.2 Preliminary and Final Forecast Verification

Upon examining these data and determining the daily change between the preliminary forecast score (i.e. how good the preliminary forecast was) and the final

forecast score (i.e. how good the final forecast was) it was found that the majority of the time a positive change between the two forecasts was recorded (Fig 2). This meant that typically the final forecast, which incorporated output from the higher resolution WRF models, was an improvement over the preliminary forecast, i.e., that on average the use of higher resolution models did result in a more accurate forecast for severe weather.

4.3 Case of 28 May 2004

A fine example illustrating the performance of the 4km WRF models and the resultant improved final forecast can be found by looking at 28 May 2004. The preliminary severe weather forecast for the 28th outlined a rather large area in the northern plains with a 15% probability of severe storms (Fig. 3). Forecasters were expecting some isolated cells and supercells early but mainly expected the storms to evolve into a multicell cluster through the remainder of the forecast period. After looking at the output from the 4km WRF-NCAR and WRF-NMM models, which were developing cells in northeast Nebraska/southern South Dakota into Iowa, a substantial change was made to the forecast. The 15% probability area was reduced in size, and an area of 25% probability was added to the southeast South Dakota area (Fig. 4). During the next day of operations at the Spring Program (31 May) these forecasts were verified using severe storm reports and it was found that the majority of the reports were concentrated in the 25% area created based on the information provided by the high resolution models (Fig. 4). Furthermore, both the WRF-NMM and WRF-NCAR model forecasts showed a strong correspondence with observed radar (Fig. 5). This case provides a "best case" example as to how well the 4km grid spaced models on occasion helped improve a forecast for severe weather made using larger scale models.

5. Discussion

It is clear that the use of higher resolution model output has potential to lead to improvements in severe weather forecasting. Not only did the final forecasts made using these experimental models generally improve over preliminary forecasts that incorporated output from lower resolution operational models, but one of the 4km models scored the highest overall in all aspects of the convective forecasting evaluation. However, one has to consider the value of this improvement when it comes to considering the continued development of the 4km WRF models and any possible future operational usage. While on average the final forecast scores improved over the scores given to the preliminary forecast, 80% of the time this improvement was small (by only one point (Fig. 6)). So while there was improvement, it was only incremental. Furthermore, while the 4km WRF-NMM was the highest scoring model in all evaluations, the next highest scoring model was the 12km Eta. This suggests that while one of the 4km WRFs outperformed the mesoscale model, the other two higher resolution models failed to do so in a consistent manner. This can call into question the usefulness of high resolution models when a currently available model is doing as well or better than two of the three high resolution examples.

This question of just how much more useful the 4km models are is an important one to consider mainly because to run a large domain storm scale model in an operational environment would take a great deal of computing power and money. Lilly (1990) used the example of doubling the spatial resolution of a forecast causing a 16-fold increase in computer capacity in his discussion of using numerical models to forecast thunderstorms.

In the current situation, implementation of a 4km model would not just double model resolution, but triple it, so the increase in required computer capacity would be even greater than 16-fold. This creates a critical issue when it comes to working with these higher resolution models, be it in development or eventually operations, as the computers they run on would have to not only be able to handle the large increase in required capacity/computing power, but also do so in a timely enough manner so the model output would be available in time to be used in the forecasting process. This would take considerable computer resources and money, which may seem like a lot to do for a one point increase in forecast score, especially when the current Eta model, that takes nowhere near the resources of a high resolution WRF, was able to outscore two of the three high resolution models.

However, as in the case of 28 May, having such a tool at their disposal may give forecasters more insight into what is going to happen than they otherwise would have, and thus be able to properly forecast a severe weather event they may have mishandled or even missed otherwise. While it is true that in this case, and in others, the forecasters had the correct general area for severe weather in their preliminary forecast, these forecasts were typically overdone, as was definitely the case on 28 May. Having too broad of an area included in the severe weather outlook makes it difficult for forecasters to zero in on where the severe weather is going to occur and provide the best warning for that area. In the 28 May case, the preliminary forecast had no fewer than 8 County Warning Areas (CWAs) included within the bounds the outlook. In reality, almost all of the severe reports, and all of the tornado reports occurred in the Sioux Falls CWA. The 4km WRF-based final forecast indicated this quite well as the 15% outlook area was drawn in closer

to the bounds of the Sioux Falls CWA, and the 25% outlook was centered right in it. Being able to make such a move operationally would have made it possible to both keep the WFOs in the unaffected CWAs from having to devote time to worrying about a situation that was not going to develop and in doing so possibly cause undo alarm among the public. At the same time, and perhaps more importantly, the WFO for the CWA that was under the higher probability would be more aware of the situation and be able to pay closer attention to their specific area rather than responding to a more general threat across the region. Both of these outcomes would be better for those being affected by severe weather as the public would be better warned when there is a severe weather threat while not being desensitized to such threats by repeated false alarms due to over done outlooks/forecasts.

Thus the answer to the question of whether or not making continued development of high resolution models and perhaps making them operational is worth the cost to do so is not something that is clear. The costs in time and computing power to do so are very large and very tangible. At the same time, trying to put a value on the benefits of such forecasts is very hard to do at this time, especially as on average the improvement was not very large. However, given that the three WRF models are still under development, one has to consider their potential. While in current configurations the 12km Eta outperforms two of the models, and all three generally only result in modest improvements over the mesoscale model-based forecasts, days such as the 28 May case show what could be possible from these models. With further development and changes, forecast improvements such as those seen on 28 May could become more common place and the true ability of higher resolution models could be realized. Also, while these

models continue with development computer power is increasing as well. Perhaps by the time a 4km WRF is fully developed, running a large domain 4km grid spaced model operationally will be more feasible and less cost prohibitive, and the question of its value could be answered more definitively and confidently.

6. Conclusion

This paper explained what, if any, improvements might be possible in severe weather forecasting through the use of high resolution, 4km grid spacing model output. Three versions of the WRF run at 00z (WRF-CAPS, WRF-NCAR and WRF-NMM), as well as the 12km operational Eta were examined during the 2004 Spring Program. It was determined that the 4km WRF-NMM outperformed the other models in the categories of convective initiation, evolution and mode and it was significantly better, in every category except evolution where its score was close to that of the 12km Eta. Also, a majority of the time, the final human produced severe weather forecast made incorporating the WRF models scored higher than a similar preliminary forecast based on output from operational mesoscale models. However, for the vast majority of these cases the forecast improvement was only incremental. This, combined with the findings that the operational Eta outscored two of the three high resolution models calls into question just how much of an improvement these models provide. While these findings may raise some questions as to the value of high resolution model output one has to keep in mind that the models examined are still very much experimental and have a lot of development ahead of them. What the models do show, through days such as the 28 May case, is that

they have the potential to be a very useful and effective tool for forecasting severe weather. As such, further development of these high resolution models is recommended as with more refined models it may become clearer just how valuable and useful they can be.

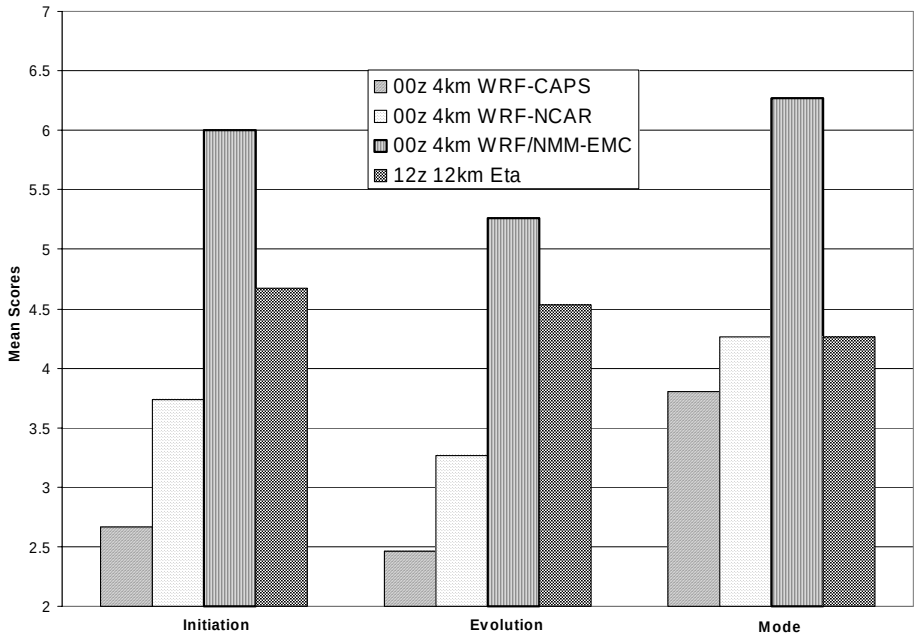

Figure 1: Average verification scores for 12z 12km Eta, 00z 4km WRF-CAPS, 00z 4km WRF-NCAR and 00z 4km WRF-NMM for convective initiation, convective evolution and convective mode. Based on a sample of 15 days

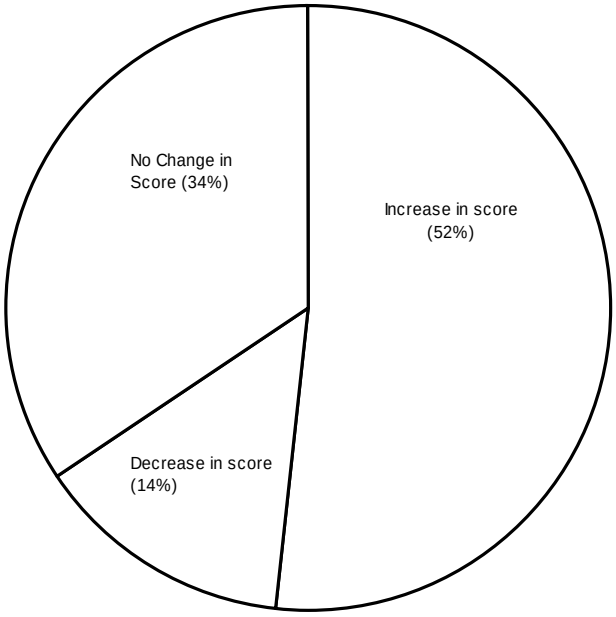

Figure 2: Percentages of total days where the final forecast verification score increased, decreased or stayed the same in comparison to the preliminary forecast verification score. Based on a sample of 29 days

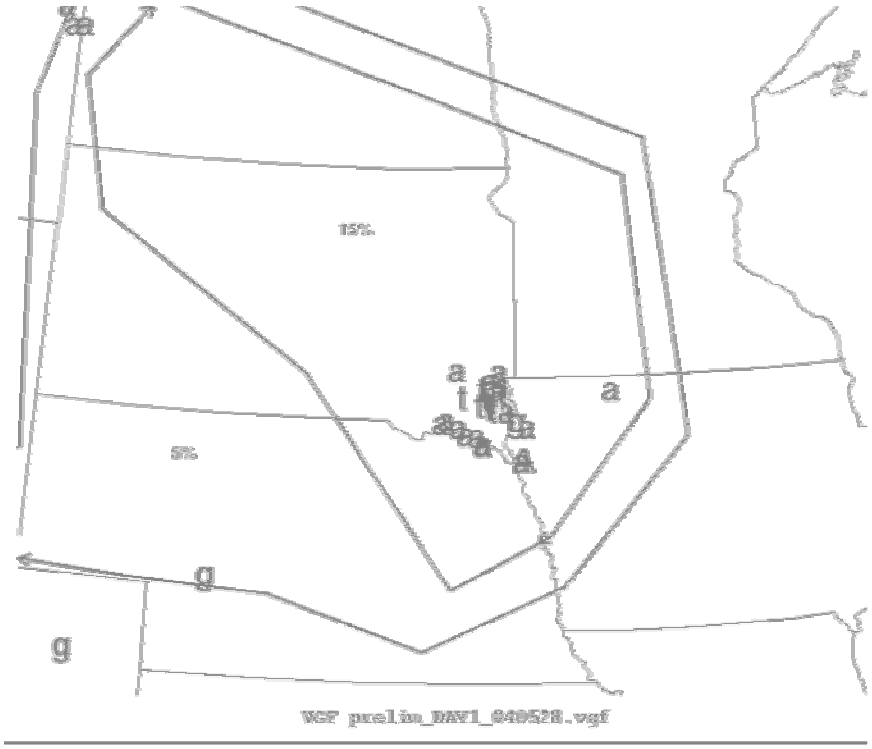

Figure 3: Preliminary forecast from 28 May with storm reports. Notice the broad area covered by the 15% probability zone and that the storm reports are mainly clustered in one part of this area.

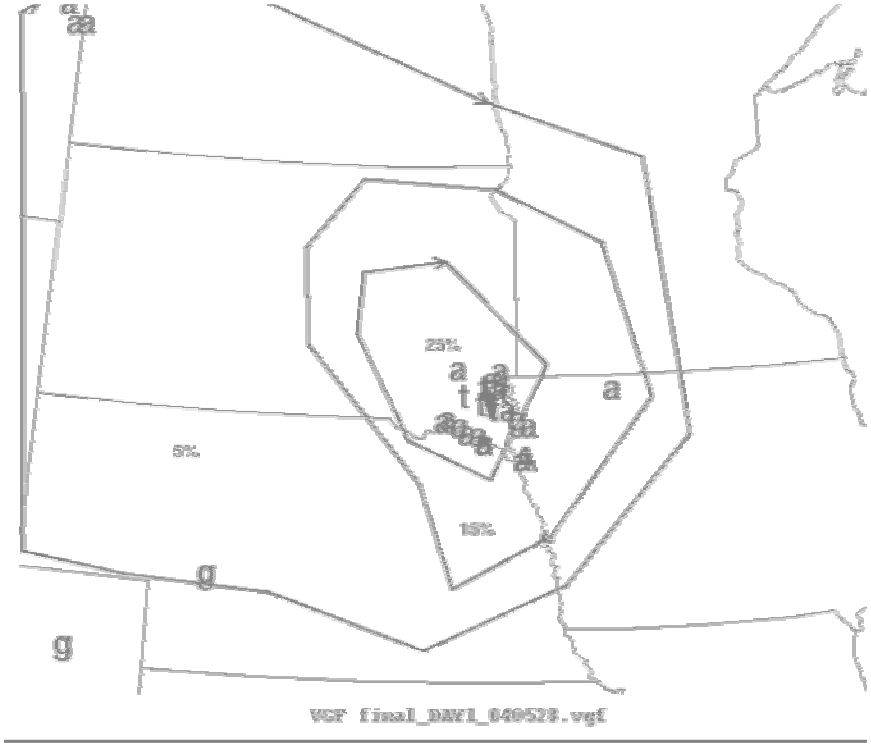

Figure 4: Final forecast for 28 May 2004 with storm reports. Notice the 15% probability area has shrunk and now focuses on the area of severe reports and that the added 25% probability area includes most of the severe reports, including all of the tornado reports.

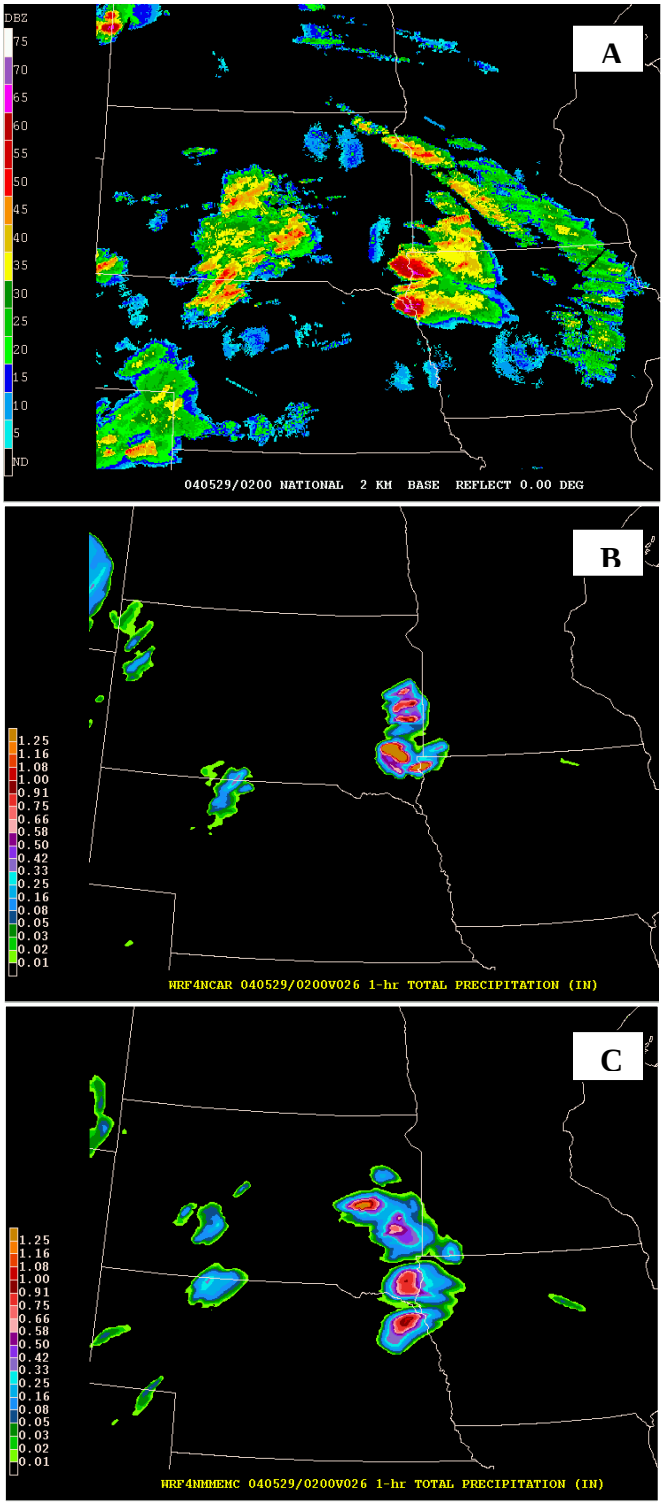

Figure 5: A) Base reflectivity one hour precip totals for 02z 29 May 2004. B) 1 hour precip output from WRF-NCAR for 02z. C) 1 hour Precip output from WRF-NMM for 02z.

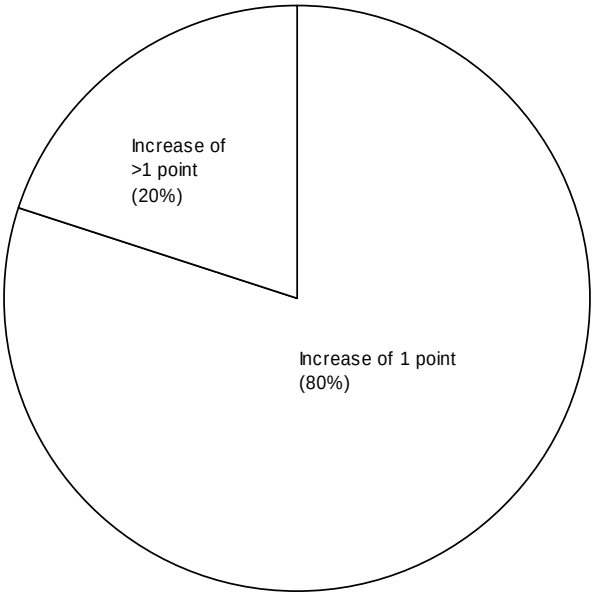

Figure 6: Percentages of days when the final forecast score increased by 1 point and by more than one point.

Acknowledgements: The following are gratefully acknowledged for their contributions to this work: Daphne Zaras, Lance Maxwell OU CAPS, Jason Levit SPC, David Schultz, NOAA/NSSL, National Science Foundation under Grant No. 0097651

References

Janish, P.R., Weiss, S.J., Kain, J.S., & Baldwin, M.E. (2001). Spring Program 2001 Program Overview and Operations Plan. Norman, OK: Storm Prediction Center/National Severe Storms Laboratory. (available online at http://www.spc.noaa.gov/exper/Spring_2001)

Kain, J.S., Baldwin, M.E., Janish, P.R., Weiss, S.J., Kay, M.P., Carbin, G.W. (2003). Subjective Verification of Numerical Models as a Component of a Broader Interaction Between Research and Operations. Weather and Forecasting, 18, 847-860.

Lilly, D.K. (1990). Numerical Prediction of Thunderstorms – Has its Time Come? Quarterly Journal of the Royal Meteorological Society, 116, 779-798.

Weisman, M.L., Skamarock, W.C. & Kemp, J.B. (1997). The Resolution Dependence of Explicitly Modeled Convective Systems. Monthly Weather Review, 125, 527-548.

Weiss, S., Bright, D., Levit, J., Schneider, R., Stensrud, D., & Kain, J. (2003). SPC/NSSL Spring Program 2003 Program Overview and Operations Plan. Norman, OK: Storm Prediction Center/National Severe Storms Laboratory. (available online at http://www.spc.noaa.gov/exper/Spring_2003).

Weiss, S., Levit, J., Bright, D., Schneider, R., Kain, J., & Baldwin, M. (2004). SPC/NSSL Spring Program 2004 Program Overview and Operations Plan. Norman, OK: Storm Prediction Center/National Severe Storms Laboratory. (available online at http://www.spc.noaa.gov/exper/Spring_2004).