Evaluation of Experimental Forecasts from the 2009 NOAA Hazardous Weather Testbed Spring Experiment Using Both Traditional and Spatial Methods

Tara Jensen[1]*, B. Brown[1], M. Coniglio[2], J. S. Kain[2], S. J. Weiss[3], L. Nance[1], and Tressa Fowler[1]

[1] NCAR/Research Applications Laboratory, Boulder, Colorado, USA
[2] NOAA/National Severe Storms Laboratory, Norman, Oklahoma, USA
[3] NOAA/Storm Prediction Center, Norman, Oklahoma, USA

Introduction

NOAA Storm Prediction Center (SPC) and National Severe Storms Laboratory (NSSL) have a collaborative testbed facility called the Hazardous Weather Testbed. The NOAA Hazardous Weather Testbed (HWT) has conducted Spring Experiments since 2000. The 2009 Spring Experiment took place over a 5 week period from 4 May – 5 June. A main focus of recent Spring Experiments is to gain an understanding of how to better use the output of near-cloud resolving configurations of numerical models to predict convective storms. HWT Spring Experiment participants have found through subjective evaluation that high resolution convective storm predictions are at times difficult for operational forecasters to reconcile, in part because many solutions appear to be plausible for a given mesoscale environment. The Development Testbed Center (DTC) collaborated with HWT in 2009 to help evaluate performance of three models during the Spring Experiment. The goal of the 2009 objective evaluation was to assess the impact of radar assimilation on the forecasts of strong convection. Generally speaking, the objective evaluation provided by MET supported the subjective evaluation performed by the forecasters. The results from both the traditional and spatial methods will be presented in this paper.

Methodology

a. Model configuration and initialization procedures

CAPS ran a 4-km resolution ensemble prediction system and provided the output to the Spring Experiment in both 2008 and 2009 (Xue et al. 2008; 2009). Each year, two members of the ensemble were configured identically but initialized differently. Specifically, the initialization of one of these members included assimilation of radar and other observational data, while the other one did not. This study focuses exclusively on output from these two members.

The these two members are both based on the WRF-ARW core, with 4km grid spacing, and 51 vertical levels. Microphysical parameterizations used include: Thompson cloud microphysics, MYJ PBL/Turbulence, and Goddard/RRTM (SW/LW) radiation schemes. The forecast domain is shown in Fig. 1. Initial and boundary conditions were generated by interpolating the North American Mesoscale (NAM) 12 km model (Rogers et al. 2009) to the 4 km high-resolution grid. One of these members (hereafter "C0") used this background directly as initial condition while the other member (hereafter "CN") incorporated additional observational datasets in a storm-scale analysis, including assimilation of radar reflectivity and velocity data in the initial condition.

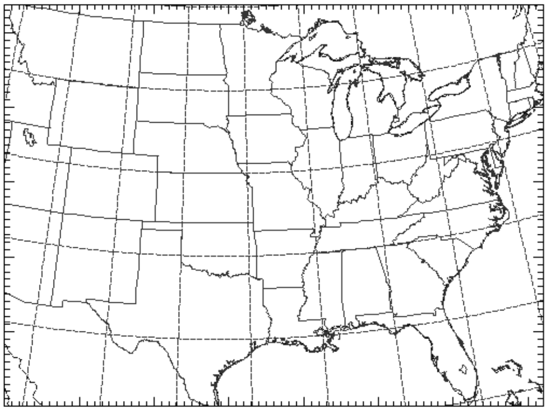

Figure 1. Model domain for the 4 km 2009 CAPS forecasts

Specifically, the unique assimilation process in the CN run ingested data from the national network of WSR-88D radars, typically using the Level II dataset, but occasionally using the Level III data when Level II was not available. Information from conventional rawinsonde, wind profiler, METAR (Meteorological Aviation Report) surface observations and Oklahoma Mesonet observations was also included. Furthermore, visible and infrared channel-4 data from GOES satellites were used in the cloud analysis package. Details about this complex assimilation process can be found in Xue et al. (2008 and 2009).

Both members were integrated out to the 30 h forecast time, but the focus here will be primarily on the 0-12 h – the period when the assimilated radar data is expected to have the most impact (Zhang et al. 2007).

b. Verification data sets

Simulated reflectivity (SR) from convection allowing models has proven to be a very useful diagnostic output field because it provides important clues about a variety of circulations and processes in a model forecast (e.g., Xue et al. 2003; Koch et al. 2005). The SR was computed from the three-dimensional hydrometeor field as described in Kain et al. (2008), with all relevant parameters (such as those describing particle size distributions) set to the values used by the Thompson microphysical parameterization that was used during model integration.

For this study, composite SR was used, meaning that gridded values represent the largest computed simulated reflectivity at any level in each vertical column. The observed reflectivity (OR) data came from the National Severe Storms Laboratory (NSSL) national 1 km radar mosaic (Vasiloff et al. 2007). Model SR and precipitation fields were extracted from the experimental model datastream at the HWT and transferred to the DTC, along with verifying reflectivity and precipitation fields from the NSSL national radar reflectivity and National Multisensor quantitative precipitation estimate (QPE) mosaics (also refered to as NMQ Q2). These datasets were ingested at the DTC and several different types of verification statistics were computed. This paper focuses on a specified (moveable) regional domain where active weather was expected at model initialization time (0000 UTC) each day.

c. Subjective Evaluation

Graphical displays of the statistical results were posted to an internal web page along with selected output fields such as simulated reflectivity in time for subjective assessments and critical examination by forecast teams during the SE2009 daily activities. This group evaluation was led by a DTC scientist each day, as the DTC rotated several scientists through SE2009 on a weekly basis. The group was instructed to focus on assessing 1) the degree to which objective verification metrics corroborated subjective impressions and 2) the potential utility of the various objective metrics in an operational environment.

d. Objective Evaluation

Verification procedures at the DTC used the Meteorological Evaluation Tools (MET) software package (Brown et al. 2007). Traditional verification metrics, such as the Gilbert Skill Score (GSS) and Critical Success Index (CSI), bias, and false alarm ratio, were computed using the MET Grid-Stat package. Additional verification statistics designed to quantify the correspondence between objects, or features, in forecasts and observations were computed using the MET Method of Object-based Diagnostic Evaluation (MODE) package (Davis et al. 2006; Brown et al. 2007). MODE relies on user-specified parameters to identify coherent objects and match them to identify similar features (such as precipitation elements) in forecasts and observations. In this study, the matched features were overlaid to enhance visual comparison of forecasts and observations. Matched object pairs were also used to calculate but traditional statistics as well as attributes such as centroid distance, difference between object angles, ratio between object areas, and median (50th percentile) and near-peak (90th percentile) intensity.

Results

Output from this forecasting system incorporating radar assimilation was visually compared to output from forecasts with no assimilation of radar data during the HWT 2009 Spring Experiment. The differences were scrutinized in order to assess the impact of the radar-data assimilation. In addition, the impact was measured after the close of the experimental periods using various objective verification metrics.

The GSS start out at a high level in the CN runs, with the magnitude depending on the reflectivity threshold. But GSSs from the CN and C0 runs converge at about the 10 h time (Fig. 2). Although the convergence point may be a few hours earlier than seen with the previous (i.e., 2008) dataset (which was verified on the basis of accumulated precipitation), it is not clear whether there is any real significance to this difference. In a broad sense, traditional verification metrics indicate a decrease in the impact of the radar assimilation after 6-12 h.

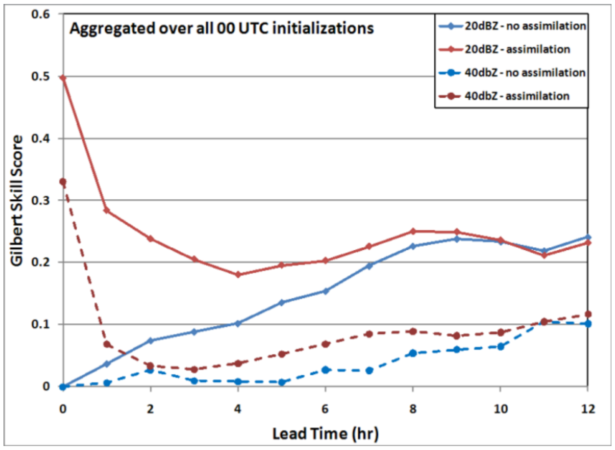

Figure 2. Gilbert Skill Score for simulated reflectivity forecasts as a function of time for all forecasts initialized at 00z during SE2009. The red (blue) curves are derived from the run with (without) assimilation of radar data, solid (dashed) lines represent scores using a 20 (40) dBz reflectivity threshold.

A visual inspection of the radar (CN) runs compared with no-radar (C0) runs indicated an initial positive impact from the CAPS assimilation system during the first 6 h of integration. As this time period being during which precipitation features were spinning up in the C0 runs, this is not too surprising. After the 6 h time the evaluation teams focusing on visual side-by-side comparisons often found it difficult to discern which forecast was better.

However, a simple overlay techniques introduced through the use of MODE output (see Figure 3) seemed to indicate a small yet systematic phase lag in the C0 forecasts, similar to the phase shift identified in MCS simulations by Dawson and Xue (2006). Therefore, not only did precipitation systems require 3-6 h of integration time to spin up when radar and other observational data were not assimilated, these features were often displaced slightly upstream compared to corresponding features in observations and the CN forecasts.

This phase lag is also indicated by examining the mean distance between the centroids of the matched forecast and observed objects identified by the MODE tool (see Figure 4). At the 1hr lead time, the CN model on average had a 10 grid-point (40 km) difference between the centroids of the 20dBZ objects. In contrast, the C0 model has a 22 grid-point (88 km) mean distance between centroids of similarly thresholded objects. The difference between mean distances decreases significantly during lead time 2-4 hr but still represents a 2-3 grid-point (8-12 km) slow bias in the C0 model. The sample size for this plot is on the order of 30-40 matched objects and no statistical significance can be attributed at this time. The implied phase lag in the C0 model may help to explain why the CN forecasts earned slightly higher aggregate objective verification scores after approximately 6 h when any advantage in overall convective evolution and morphology past this time was not readily discernible in side-by-side subjective assessments.

Summary

In a general sense, the objective evaluation performed during the HWT 2009 Spring Experiment supported the subjective conclusions drawns from the comparing no-radar (C0) and radar (CN) runs. In some cases, the objective evaluation helped identify and potentially explain difference in the two fields.

Traditional evaluation methods indicate there is an improvement in skill scores on short-term forecasts (0-6 hr) of Simulated composite Reflectivity (SR) when radar assimilation is included in the initial conditions for a convective allowing model. The magnitude of increased skill score depends on the reflectivity threshold. However, skill scores for the no-radar and radar assimilation methods appear to converge by the 10hr forecast. Although the convergence point may be a few hours earlier than seen it is not clear whether there is any real significance to this difference. In a broad sense, traditional verification metrics indicate a decrease in the impact of the radar assimilation after 6-12 h.

Object-based verification metrics, such as those available in MODE, are capable of diagnosing systematic biases, such as the upstream lag proposed above. Time series of attributes derived from MODE-defined objects may help with quantifying the model response. However, generation of statistically significant inferences from these metrics is quite challenging due to sample size constraints (i.e. number of identified objects is small over a 4-6 week time period) and is the subject of ongoing work.

Finally, application of traditional verification metrics to reflectivity fields rather than the more traditional accumulated precipitation fields provides an example of how simulated forecast fields may provide meaningful evaluation metrics to help forecasters synthesize the model output.

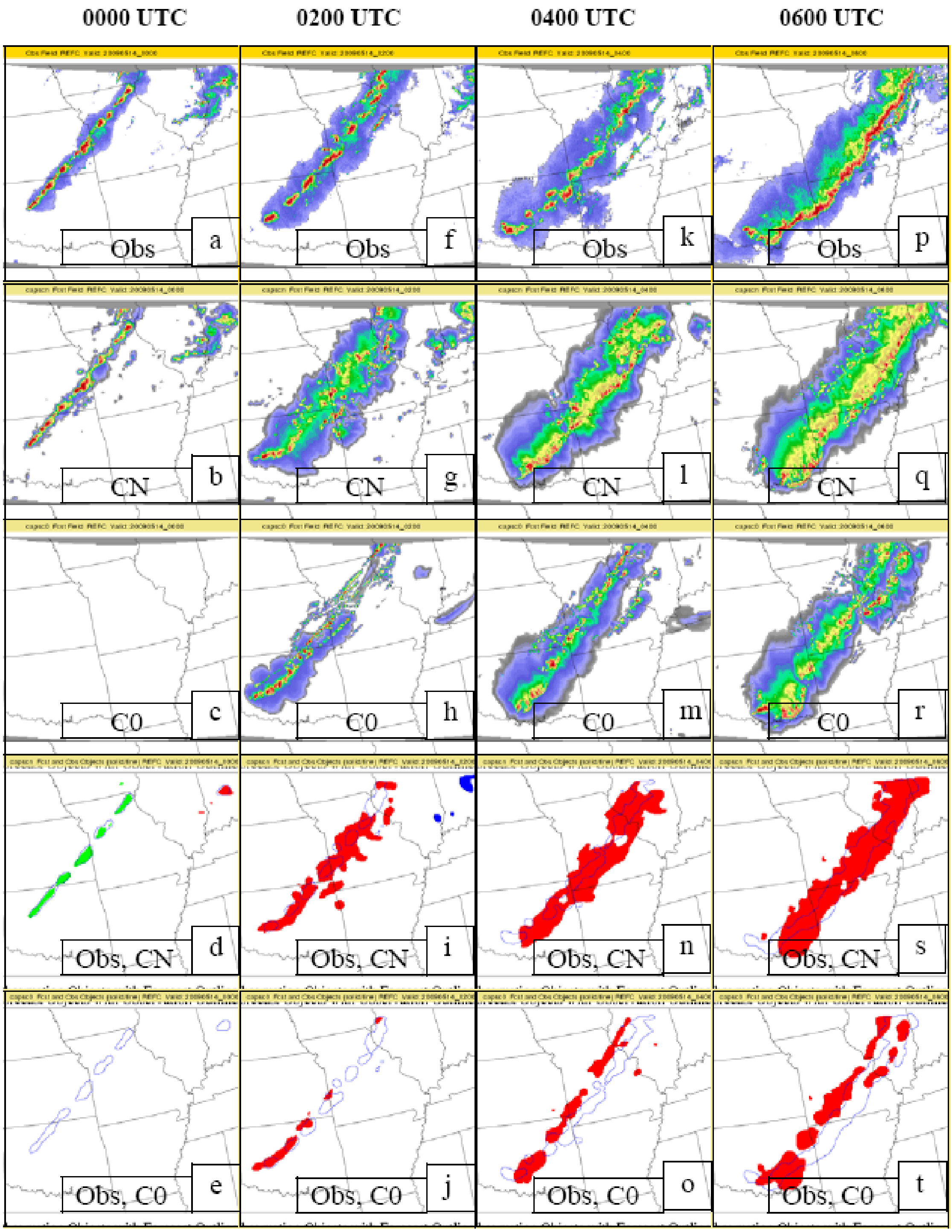

Figure 3. Comparison of observed and simulated composite reflectivity for forecast hours 0, 2, 4, and 6, beginning at 0000 UTC 14 May 2009. Observations are in the top row, with the CN and C0 forecasts in the second and third rows, respectively. In the bottom two rows, the location of observed features is outlined in blue and the features predicted by the CN (fourth row) and C0 (fifth row) are overlaid and color filled. Objects filled with a single color have been grouped (or merged) by the MODE software as a collection of related objects. Each time period is analyzed independently by MODE.

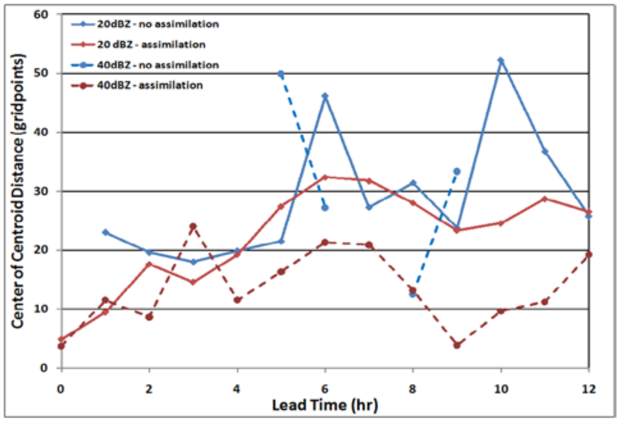

Figure 4. Distance between centroids of matched simulated reflectivity forecast/observed reflectivity object pairs as a function of time for all forecasts initialized at 00z during SE2009. The red (blue) curves are derived from the run with (without) assimilation of radar data, solid (dashed) lines represent scores using a 20 (40) dBz reflectivity threshold.

Acknowledgements

This analysis would not have been possible without the contribution of the CAPS Storm Scale Ensemble Forecasts to the HWT Spring Experiment. The CAPS forecasts were partially supported by the NOAA/CSTAR program and were performed at the Pittsburgh Supercomputing Center supported by the National Science Foundation. Supplementary support was provided by NSF ITR project LEAD led by Kelvin Droegemeier (ATM-0331594), and by NSF grant ATM-0802888. David O'Neal of PSC is thanked for his computing system support

Dedicated work by many individuals led to the success of SE2009. At the SPC, HWT operations were made possible by technical support from Jay Liang, Gregg Grosshans, David Bright, Greg Carbin, and Joe Byerly. At the NSSL, Brett Morrow, Brad Sagowitz, Steve Fletcher, and Jeff Horn also provided valuable technical support. Steve Sullivan of the DTC helped with data retrieval. David Ahijevych, Tressa Fowler, and Jamie Wolff, also from the DTC, led discussions related to forecast verification during SE2009 daily activities. Ryan Sobash of NSSL and OU developed critical web displays used at HWT for SE2009.

References

Brown, B. G., R. G. Bullock, J. H. Gotway, D. Ahijevych, C. A. Davis, E. Gilleland, and L. Holland, 2007: Application of the MODE object-based verification tool for the evaluation of model precipitation fields. *Preprints, 22nd Conference on Weather Analysis and Forecasting/18th Conference on Numerical Weather Prediction*, Amer. Meteor. Soc., Park City, UT. paper 10A.2

Dawson, D. T., II and M. Xue, 2006: Numerical forecasts of the 15-16 June 2002 Southern Plains severe MCS: Impact of mesoscale data and cloud analysis. *Mon. Wea. Rev.*, **134,** 1607-1629.

Kain, J. S., S. J. Weiss, D. R. Bright, M. E. Baldwin, J. J. Levit, G. W. Carbin, C. S. Schwartz, M. Weisman, K. K. Droegemeier, D. Weber, and K. W. Thomas, 2008: Some practical considerations for the first generation of operational convection-allowing NWP: How much resolution is enough? *Wea. Forecasting*, **23**, 931-952.

Koch, S. E., B. S. Ferrier, M. Stolinga, E. Szoke, S. J. Weiss, and J. S. Kain, 2005: The use of simulated radar reflectivity fields in the diagnosis of mesoscale phenomena from high-resolution WRF model forecasts. *Preprints, 11th Conference on Mesoscale Processes*, Albuquerque, NM, Amer. Meteor. Soc., paper J4J.7

Rogers, E., G. J. DiMego, T. L. Black, M. B. Ek, B. S. Ferrier, G. A. Gayno, Z. Janjic, Y. Lin, M. E. Pyle, V. C. Wong, W.-S. Wu, and J. Carley, 2009: The NCEP North American mesoscale modeling system: Recent changes and future plans. *Preprints, 23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, Amer. Meteor. Soc., Omaha, NE. paper 2A.4

Vasiloff, S.V., D. J. Seo, K.W. Howard, J. Zhang, D. H. Kitzmiller, M. G. Mullusky, W. F. Krajewski, E. A. Brandes, R. M. Rabin, D. S. Berkowitz, H. E. Brooks, J. A. McGinley, R. J. Kuligowski, and B. G. Brown, 2007: Improving QPE and very short term QPF: An initiative for a community-wide integrated approach. *Bull. Amer. Meteor. Soc.*, **88**, 1899–1911.

Xue, M., D.-H. Wang, J.-D. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Physics*, **82**, 139-170.

Xue, M., F. Kong, K. W. Thomas, J. Gao, Y. Wang, K. Brewster, K. K. Droegemeier, J. S. Kain, S. J. Weiss, D. R. Bright, M. C. Coniglio, and J. Du, 2008: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2008 Spring Experiment. *Preprints, 24th Conf. Severe Local Storms,* Savannah, GA, Amer. Meteor. Soc., Paper 12.2.

Xue, M., F. Kong, K. Thomas, J. Gao, Y. Wang, K. Brewster, K. Droegemeier, J. S. Kain, S. J. Weiss, D. R. Bright, M. C. Coniglio, and J. Du, 2009: CAPS realtime 4 km multi-model convection-allowing ensemble and 1 km convection-resolving forecasts for the NOAA Hazardous Weather Testbed 2009 Spring Experiment. *Preprints, 23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, Amer. Meteor. Soc., Omaha, NE. paper 16A.2

Zhang, F., N. Bei, R. Rotunno, C. Snyder, and C.C. Epifanio, 2007: Mesoscale predictability of moist baroclinic waves: Convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594.