

The Developmental Testbed Center (DTC) Objective Evaluation Performed During the Hazardous Weather Testbed (HWT) 2010 Spring Experiment.

Tara Jensen¹, Steve Weiss², Jack Kain³, Michelle Harrold¹, Ming Xue⁴, Fanyou Kong⁴, Barb Brown¹, Patrick Marsh³, Adam Clark³, Kevin Thomas⁴, Mike Coniglio³, and Russ Schneider²

¹ NCAR/Research Applications Laboratory (RAL), Boulder, Colorado

² NOAA/Storm Prediction Center (SPC), Norman, Oklahoma

³ NOAA/National Severe Storms Laboratory (NSSL), Norman, Oklahoma

⁴ Center for Analysis and Prediction of Storms (CAPS), University of Oklahoma, Norman, Oklahoma

Abstract

Every spring approximately 60-70 researchers and forecasters participate in the Hazardous Weather Testbed (HWT) Spring Experiment to investigate the use of convection-allowing model forecasts as guidance for the prediction of severe convective weather. The collaboration between HWT and Developmental Testbed Center (DTC) for the HWT Experimental Forecast Program (EFP) Spring Experiment has strengthened and grown over the past three years. In addition to the severe convective weather component, the 2010 Spring Experiment objective evaluation plan includes evaluation of forecasts from WRF convection-allowing models for extreme precipitation events as well as aviation related thunderstorm indicators.

The DTC objective evaluation during the 2010 HWT Spring Experiment complements the subjective evaluation that has traditionally taken place. With the addition of probabilistic verification capabilities in the DTC's Model Evaluation Tools (MET), both probabilistic products and deterministic forecasts will be evaluated this year. It is anticipated that both the subjective and objective evaluations will support the use of latest scientific developments and forecast strategies by the SPC, NOAA/NWS, Hydrometeorological Prediction Center (HPC) and NOAA/NWS Aviation Weather Center (AWC).

This year's Spring Experiment ran from May 17 – June 18, 2010 and overlapped the VORTEX-2 field campaign. This presentation paper provides preliminary results from the 2010 Spring Experiment.

Methodology

The NOAA HWT Spring Experiment is a yearly experiment that, in recent years, has investigated the use of convection-allowing model forecasts as guidance for the prediction of severe convective weather. A variety of model output has been examined and evaluated daily during the experiment and experimental severe weather forecasts have been created and verified. The variety of models available to the Spring Experiment has allowed the HWT to explore different types of guidance, including products derived from both ensembles and deterministic forecasts. This year, the models available to the Spring Experiment include: NOAA National Severe Storms Laboratory (NSSL) 4km WRF-ARW, NOAA National Centers for Environmental Prediction (NCEP) Environmental Modeling Center (EMC) 4km WRF-NMM, National Center for Atmospheric Research (NCAR) 3km WRF-ARW, University of Oklahoma Center for Analysis and Prediction of Storms (CAPS) 4km Storm Scale Ensemble Forecast (SSEF) 26 member multi-model ensemble, and the NOAA Earth System Research Laboratory (ESRL) High Resolution Rapid Refresh (HRRR) 3km WRF-ARW.

The overarching goals of the 2010 HWT-DTC Spring Experiment evaluation are to 1) provide objective evaluations of the experimental forecasts; 2) supplement and compare to subjective assessments of performance; and 3) familiarize forecasters and researchers with both new and traditional approaches for evaluating forecasts. The objectives for the 2010 Spring Experiment include: 1) augmenting the samples available to evaluate the impact of radar assimilation on short-term forecasts; 2) providing evaluation of probabilistic quantitative precipitation forecasts (QPF) of extreme precipitation events; and 3) providing evaluation of predicted radar echo top heights, defined by the height of the 18 dBZ reflectivity.

This year DTC is evaluating (in near real-time) all 26 members of the CAPS SSEF ensemble as deterministic models, some of the CAPS SSEF ensemble products, and the HRRR deterministic model. DTC is also bringing in two operational models, the EMC North American Mesoscale (NAM) 12 km WRF-NMM deterministic model, and the EMC Short Range Ensemble Forecast (SREF) 32-35 km 21 member multi-model ensemble. Table 1 indicates the forecast variables, observation fields, and evaluation metrics applied to these models. A full description of the each model contributed to HWT may be found on their website at:

http://hwt.nssl.noaa.gov/Spring_2010/

Table 1. List of variables (and thresholds) to be evaluated during SE 2010.

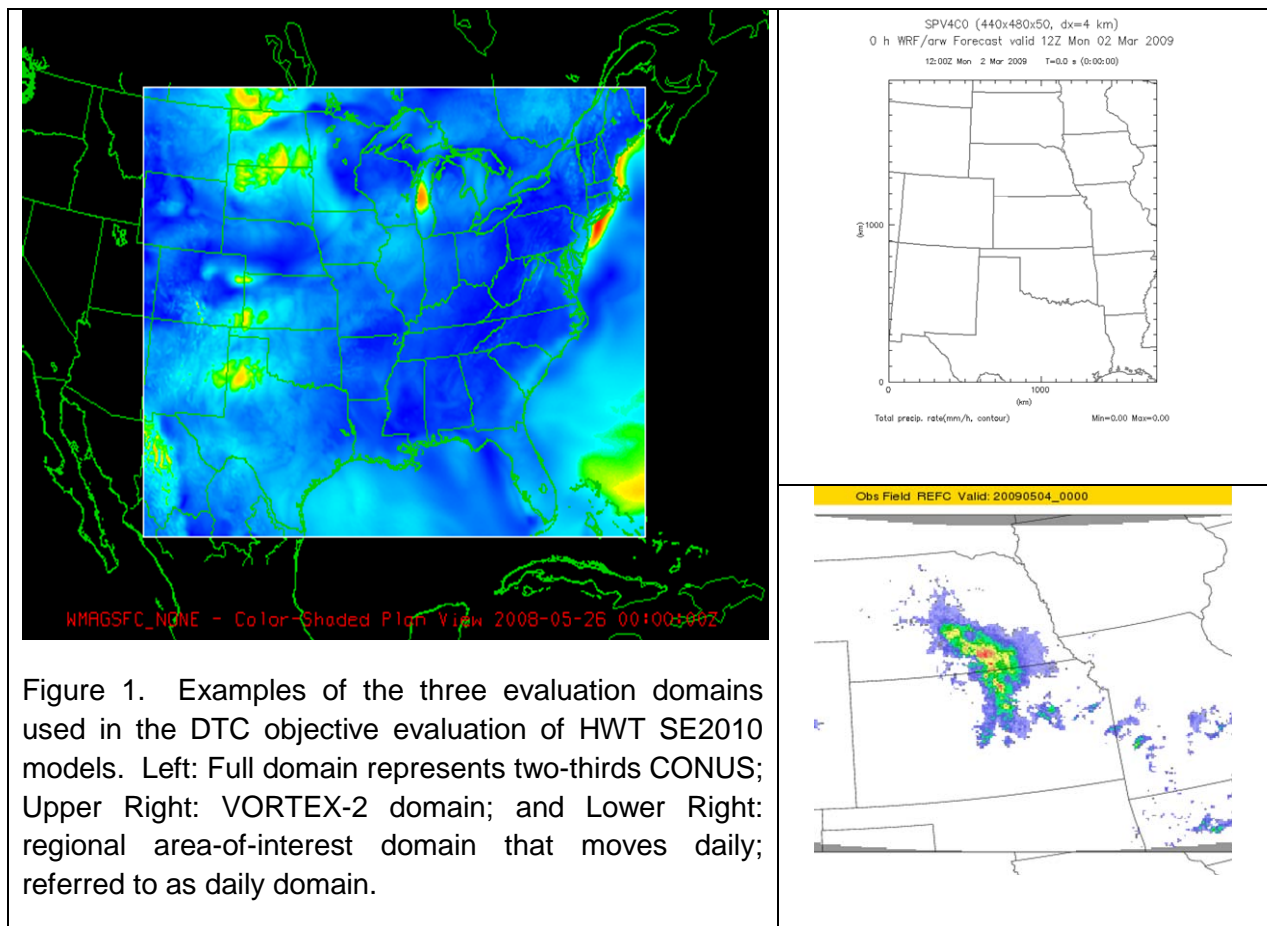
FCST Field	Observation	Grid-Stat	MODE	Models
Prob of Exceed (0.5", 1", 2" over 3 and 6 hrs)	0.5", 1", 2" QPE over 3 and 6 hrs	Brier Score, Decomp of Briar score, Area under ROC	None	Ensemble products from CAPS and SREF
50% Prob of Exceed (0.5", 1", 2" over 3 and 6 hrs)	0.5", 1", 2" QPE over 3 and 6 hrs	None	MMI, Intersection Area, Area Ratio, Centroid Distance, Angle Difference, % Objects and Area Matched, 50 th and 90 th percentile	Ensemble products from CAPS and SREF
QPF (0.25", 0.5", 1.0", 2" over 3 and 6 hrs)	0.25", 0.5", 1.0", 2" QPE over 3 and 6 hrs	GSS, CSI, FAR, PODY, FBIAS	Same as above	CAPS members, CAPS ens products, SREF ens mean, HRRR, NAM
Sim. Comp. Refl (20,30,40,50 dBZ)	Q2 Composite refl (20,30,40,50 dBZ)	GSS, CSI, FAR, PODY, FBIAS	Same as above	CAPS members, CAPS ensemble products, HRRR, NAM
18 dBZ Echo Top (18, 25, 30, 35, 40, 45 kft)	Q2 18dBZ Echo Top (18, 25, 30, 35, 40, 45 kft)	GSS, CSI, FAR, PODY, FBIAS	Same as above	CAPS members, CAPS ens mean, HRRR
Prob of 40dBZ echos	Q2 Composite reflectivity (40dBZ)	GSS, CSI, FAR, PODY, FBIAS	None	Ensemble products from CAPS and SREF
50% Prob of 40dBZ echos	Q2 Composite reflectivity (40dBZ)	None	Same as above	Ensemble products from CAPS

ROC (Receiver Operator Characteristics Curve), GSS (Gilbert Skill Score), CSI (Critical Success Index), FAR (False Alarm Ratio), PODY (Probability Of Detection Yes), FBIAS (Frequency Bias), MMI (Median of Maximum Interest).

Objective evaluation was performed using the DTC's MET tool. Traditional statistics for categorical (thresholded) forecasts were computed using the Grid-Stat tool in MET. Specific statistics provided include: Gilbert Skill Score (GSS), Critical Success Index (CSI), Probability of Detection Yes (PODY), False Alarm Ratio (FAR), and Frequency Bias (FBIAS). Evaluation using a spatial method was performed using the Method for Object-Based Diagnostic Evaluation (MODE) tool in MET. Metrics provided include: Median of Maximum Intensity (MMI), Intersection Area, Area Ratio, Centroid Distance, Angle Difference, % Objects and Area Matched, and 50th and 90th percentile within the object. Davis *et al.* (2006) provides an overview of the application of MODE, including many of these attributes to the QPF verification problem. A complete description of the MET tools may be found at:

<http://www.dtcenter.org/met/users/docs/overview.php>.

Included in the evaluation were the 00 UTC initialization of each model and the 12 UTC initialization of the HRRR, NAM, and two deterministic update members provided by CAPS in support of the VORTEX-2 experiment. All models were re-gridded to the 4 km Stage IV grid configuration. The 00 UTC models were evaluated over three regions: the entire domain, the static VORTEX-2 domain provided by CAPS at the 12 UTC initialization time, and a regional, movable area-of-interest domain selected by HWT Spring Experiment participants each day. Figure 1 depicts examples of these domains.



Ensemble Product Description

For this paper, the Ensemble Simple Mean represents the arithmetic mean at each gridpoint. Ensemble Simple Frequency means the number of ensemble members meeting or exceeding the threshold at a give gridpoint. Probability Neighborhood method (PN) applies a neighborhood method using a Gaussian filter with radius of interpolation of 40 km and prior to computing the ensemble relative frequency. The Probability Matching method (PM) orders the PDF of the maximum intensity of all members at each grid point (Ebert, 2001). It assigns the maximum intensity value to the location of the maximum in the simple ensemble mean, the second highest intensity value to the location of the second maximum value in the simple mean and so on. This allows the shape of the ensemble mean to be maintained but the intensities to be more dynamic.

Preliminary Results

The goal of the objective evaluation is to help the forecasters and researchers understand the skill of these storm scale models. A comparison of subjective evaluation with the objective results was performed during the Experiment daily. The following section depicts three subjective evaluations with corresponding objective evaluation. The results in this section should be viewed as preliminary as they are aggregated over initial data available in real time during the Spring Experiment, and the data samples are not homogeneous.

a. Simulated Radar Reflectivity

Simulated radar reflectivity was used by Spring Experiment participants taking part in the Severe Weather and Aviation Weather forecast exercises. The SE2009 participants subjectively found that radar data assimilation methods provide somewhat more realistic forecasts of simulated radar reflectivity fields during 00-06hr lead times but small scale features of the simulated reflectivity appeared to lose coherence more rapidly. There also appeared to be better spatial over-lap of the simulated field when radar assimilation methods were employed. A more detailed description of this subjective impression may be found in Kain et. al (2010) or Jensen et. al (2010).

The SE2010 objective evaluation supports this subjective impression and agrees with the findings in Kain et. al. Figure 2 shows GSS (also known as Equitable Threat Score) and FBIAS for simulated reflectivity greater than 20 dBZ aggregated over all available runs during the Spring Experiment. During lead times 0 - 3 hours, there is a marked increase in GSS for the three control members with radar assimilation (red lines) as compared with control members with no radar assimilation (red lines starting at a GSS of 0 at lead-time 0), and this increase is maintained up to 5-6 hours. However, there is a marked decrease in GSS during that same time and diminishes quickly out to 15 hr lead time, implying the benefits gained from radar assimilation are washed out fairly quickly. We note there that positive impacts of radar data in hourly precipitation GSS were reported by Xue et al. (2008) and Kong et al (2009) to last between 6 to 12 hours for the 2008 and 2009 spring seasons of CAPS forecasts, depending on the threshold.

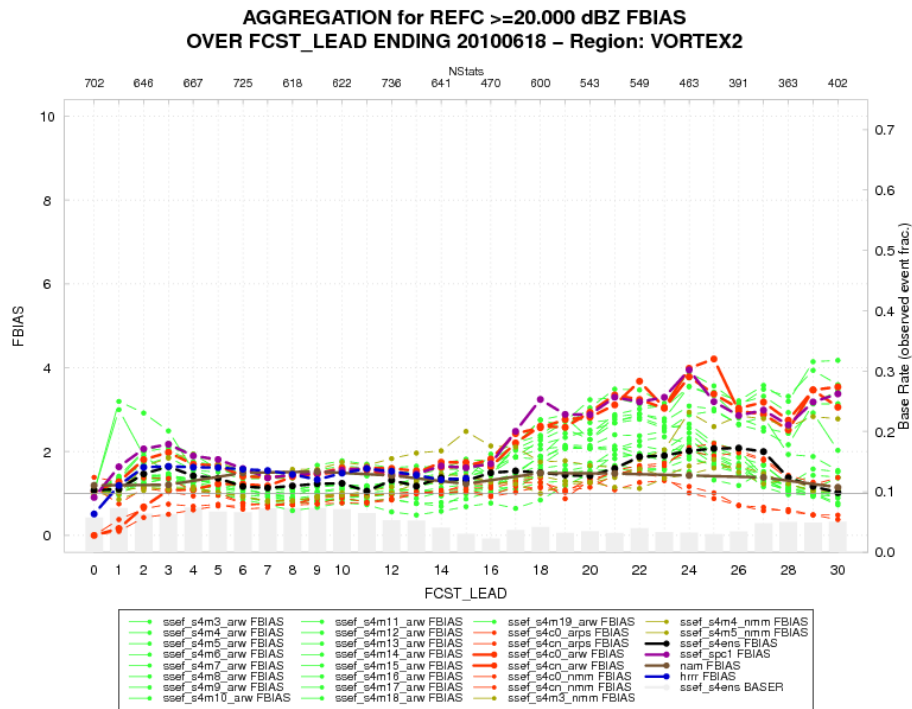
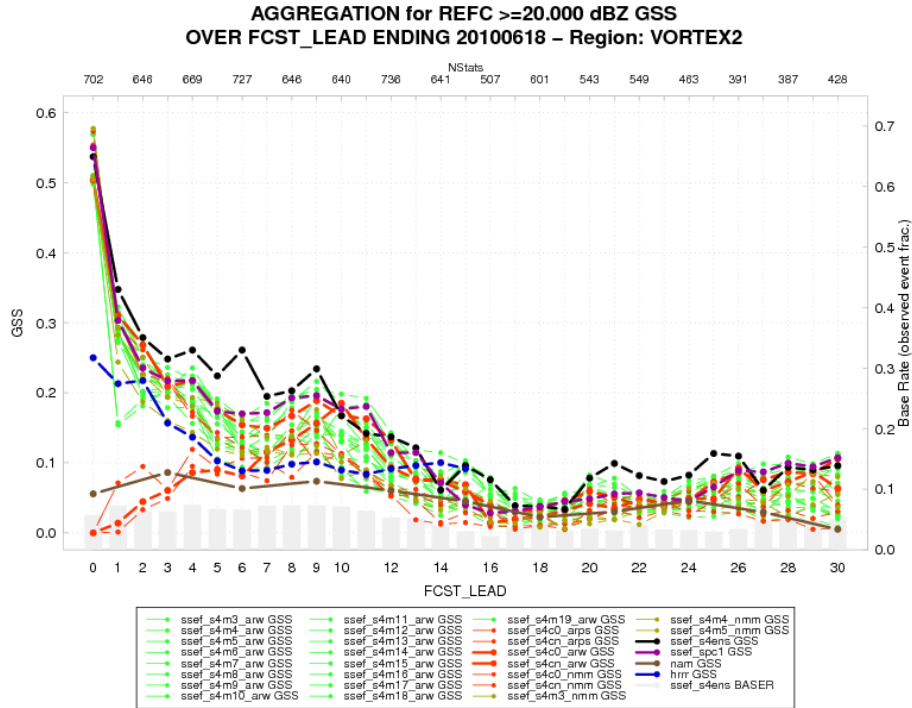


Figure 2. GSS (top) and FBIAS (bottom) for simulated reflectivity greater than 20 dBZ aggregated over four weeks of the HWT Spring Experiment. Red lines: control members of the CAPS SSEF representing models with (ssef_s4cn) and without (ssef_s4c0) radar assimilation. Green lines: WRF-ARW members of the CAPS SSEF. Gold lines: WRF-NMM members of CAPS SSEF. Black: CAPS SSEF simple ensemble mean. Purple: CAPS SSEF 1 km WRF-ARW simulation. Brown: NCEP/EMC NAM 12 km WRF-NMM simulation. Blue: NOAA/ESRL HRRR 3 km WRF-ARW simulation. Gray bars: Base Rate (or observed event fraction) and scaled on right axis.

The CAPS Probability Matched mean (black), and CAPS 1 km simulation (purple) appear to outperform HRRR (blue) and NAM (brown), so do most of the radar-assimilating 4 km members of CAPS, especially in the first 12 hours. When comparing convective allowing models, it should be noted that the HRRR 00 UTC run does not include upper air data assimilation and hence skill may be reduced due to this deficiency. The HRRR 01 UTC run has the upper air data assimilation and may make a better benchmark for future comparison. GSS increases at later lead times which may be explained by an increase in convection during the afternoon and evening hours. This increase is reflected in Base Rate (depicted by gray bars). The period of lower GSS scores between 14 and 18 hours correspond to the morning minimum of convective activities.

The skill demonstrated by GSS is sensitive to hits and therefore should be interpreted in light of FBIAS. Frequency bias, or the ratio of forecast area to observed area, can be used to provide a sense of over or under-prediction of a particular variable. Frequency Bias aggregated over the experiment is shown in Figure 5. The optimal value of frequency bias is 1. Scores larger than this indicate an over-forecast of the given threshold. Less than 1 indicates an under-forecast. During 0-6 hr lead times, some of the skill demonstrated by convective allowing models with respect to the NAM baseline may be due to the higher FBIAS exhibited in Figure 2.

b. Probabilistic Quantitative Precipitation Forecasts

The QPF and probabilistic QPF fields were used primarily by the participants in the Hydrometeorological forecasting exercise. Subjectively, it appeared that the simple ensemble relative frequency probability product derived from the CAPS SSEF ensemble is a step-forward from using SREF and NAM for the QPF problem, but two other ensemble products appear to be equally promising. One method is the neighborhood method (PN QPF). The other is a probability matching method (PM QPF).

Figure 3 is a screenshot of one of the displays available on the DTC/HWT Objective Evaluation Website (<http://verif.rap.ucar.edu/eval/hwt/2010/>) and provides a sample of how the subjective impression was formed. The top row shows the observed field, CAPS simple ensemble relative frequency probability field, SREF simple ensemble relative frequency probability field, CAPS PN field, NAM deterministic QPF field, and CAPS PM QPF field. The bottom row shows the respective objects that were identified by the MET spatial verification package called MODE.

Traditional statistics and spatial attributes for these fields are provided in Table 2 and Table 3 respectively. Based on the Brier Score, listed in Table 2 and the optimal value is 0, the CAPS simple PQPF appears to have more skill than the other two products. However, the CAPS PN PQPF shows more resolution and has a higher Area under the Receiver Operating Characteristic Curve and could potentially be deemed more skillful for this model run.

Evaluation using a spatial method is provided in Table 3 using MODE Forecast-Observed cluster attributes. Centriod distance is the distance between the centriods of the “rubber band” that connects the maximum extension of the object cluster. Smaller is better for this attribute. Area ratio is defined as the ratio of the forecast area to the observed area. Once again, smaller indicates a better match. Numbers less than 1.0 indicate an under-forecast in area while numbers greater than 1.0 indicate an over-forecast in area. Intersection area indicates the the area of the two object clusters that intersect. A larger value would indicate better object agreement. The median (50th Percentile) and near-peak (90th percentile) accumulations within the forecast cluster objects and observed cluster objects were differenced to provide the final two metrics listed in Table 3. Based on these spatial results, it appears that the convective allowing models in the CAPS SSEF provided a better agreement with observations, with the

CAPS simple arithmetic mean edging out the other two ensemble products (PN and PM). Follow-on work will include aggregating these metrics.

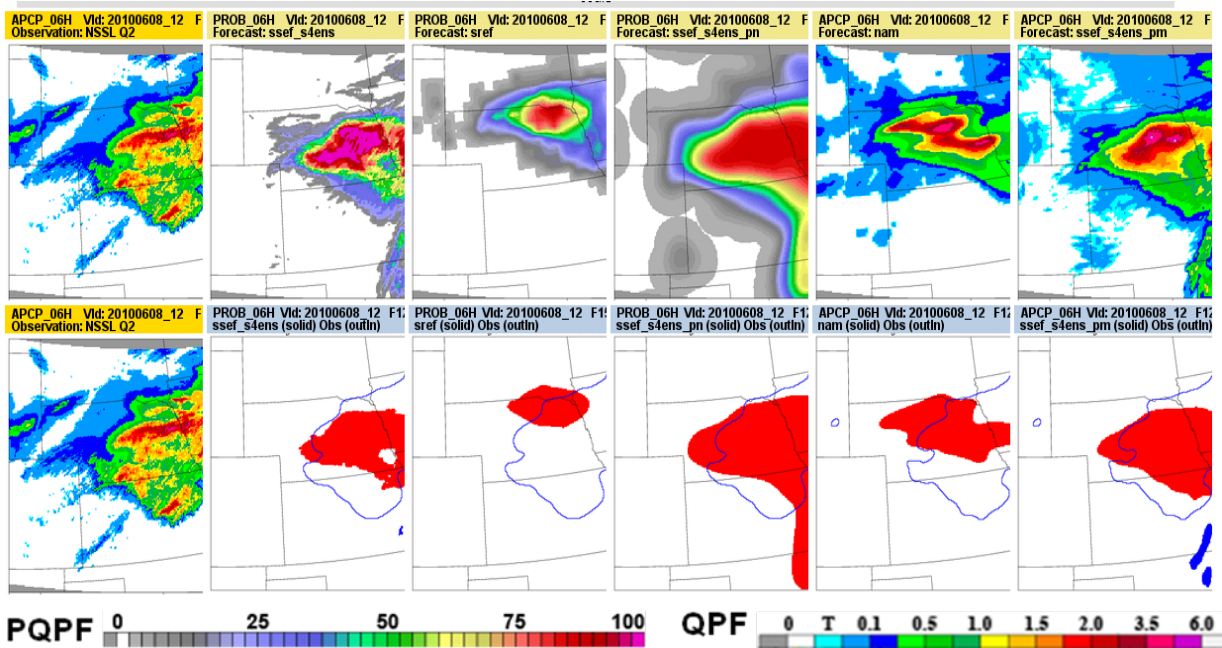


Figure 3. Screenshot of a Probability QPF spatial verification display available on DTC/HWT objective evaluation website. Plots are the 12hr forecast valid at 8 June 2010. The top row: observed field; CAPS Simple PQQF field; SREF Simple PQQF field; CAPS PN PQQF field, NAM QPF field, and CAPS PM PQQF field. The bottom row shows forecast (solid) and observed (blue line) objects identified by MODE for probability >50% and accumulated precipitation > 0.5 inch. PQQF units – percent. QPF units – inches.

Table 2. Traditional statistics for 12hr forecast valid on 8 June 2010 at 12 UTC. Statistics calculated by the DTC MET Grid-Stat tool. SREF data is from 15hr forecast valid at same time.

Model	CAPS Simple PQQF	SREF Simple PQQF	CAPS PN PQQF	NAM QPF	CAPS PM QPF	CAPS Simple QPF
Brier Score	0.081	0.134	0.088	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>
Reliability	0.004	0.020	0.013	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>
Resolution	0.068	0.032	0.071	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>
Uncertainty	0.145	0.145	0.145	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>
Area Under ROC	0.867	0.720	0.924	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>

Table 3. Spatial attributes for 12hr forecast valid on 8 June 2010 at 12 UTC. Statistics calculated by the DTC MET MODE tool. SREF data is from 15hr forecast valid at same time. F denotes forecast and O denotes observation object clusters defined by MODE interest maps.

Model	CAPS Simple PQPF	SREF Simple PQPF	CAPS PN PQPF	NAM QPF	CAPS PM QPF	CAPS Simple QPF
Centroid Distance	33.0 km	186.2 km	69.8 km	152.2 km	53.0 km	51.8 km
Area Ratio (F/O)	0.56	0.20	1.15	0.45	0.71	0.66
Intersection Area	6539 gridsq	1530 gridsq	9798 gridsq	4178 gridsq	8112 gridsq	7680 gridsq
50% Accum Diff (F-O)	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>	0.007 in	0.07 in	-0.08
90% Accum Diff (F-O)	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>	0.70 in	0.54 in	0.02

c. Radar Echo Top Product

The radar echo top product was used primarily by participants in the Aviation Weather forecast exercise. The Spring Experiment participants examined individual members of the CAPS SSEF system. One comparison was how different the simulated radar reflectivity field was for different cloud microphysics packages. One subjective impression that occurred on several days was that the Thompson scheme in WRF-ARW model tended to over develop the stratiform region. This over-prediction impacted the simulated reflectivity and hence 18 dBZ radar echo top height, and potentially accumulated precipitation products.

Figure 4 is a screenshot of another display available on the DTC/HWT Objective Evaluation Website (<http://verif.rap.ucar.edu/eval/hwt/2010/>) and provides a sample of how the subjective impression was formed. The top row shows the observed field, CAPS PM mean, CAPS cn_arw member with Thompson microphysics scheme, CAPS m15_arw member with WRF double moment microphysics scheme, CAPS m16_arw member with WRF single moment microphysics scheme, and CAPS m17_arw member with Morrison microphysics scheme. The bottom row shows the respective objects that were identified by the MET spatial verification package called MODE. There is a significantly larger stratiform cloud shield in the WRF-ARW Thompson simulation than the other microphysics perturbations. This large shield appears to dominate the simple ensemble mean, likely reflecting the presence of 9 SSEF members that are configured with the Thompson scheme.

The CAPS PM mean appears to have a sizable over-prediction of radar echo top > 25,000 ft. This may be due to the over-prediction of stratiform region of the convection in several members, including the Thompson scheme members. It is interesting to note that the other three simulations with large FBIAIS use the Thompson scheme for cloud microphysics. It

appears using the PM method may exacerbate the over-representation of the stratiform region. These results are preliminary and it is recommended the reader uses them with caution. A full analysis will be performed retrospectively.

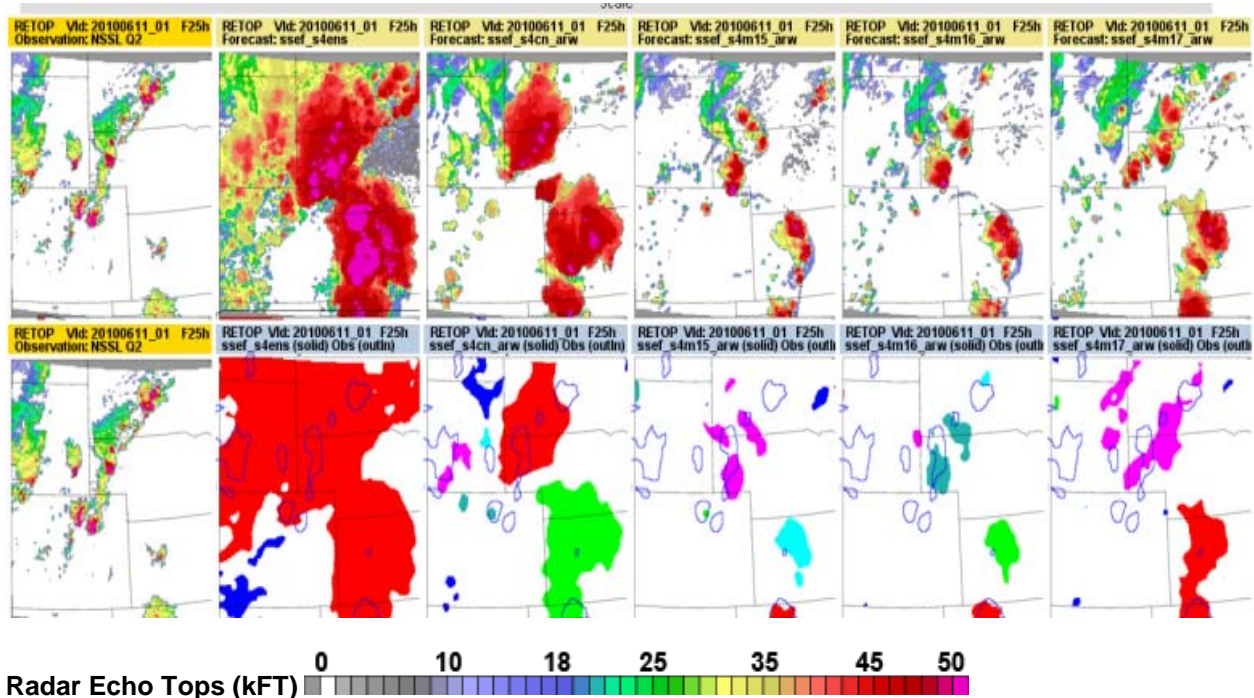


Figure 4. Screenshot of a 18dBZ radar echo top height and spatial verification display. Plots are the 12hr forecast valid at 8 June 2010 12UTC. The top row: Q2 observed field; CAPS Simple Probability field; SREF Simple Probability field; CAPS Probability Neighborhood field, NAM deterministic QPF field, and CAPS Probability Matched QPF field. The bottom row shows forecast (solid) and observed (blue line) objects identified by MODE for 18dBZ echo top height > 25000 ft. Different colors indicate forecast cluster of objects matched with underlying observation objects.

Summary

In recent years, the NOAA Hazardous Weather Testbed Spring Experiment has presented the research community with a unique opportunity to have their storm-scale simulations used and evaluated by the forecasting community. DTC has been collaborating with HWT by providing objective evaluation during the Spring Experiment for the past three years. Each year this collaboration has grown. This year, 29 deterministic models and four probabilistic products from two ensemble systems were evaluated in near real-time. Evaluation of individual runs, as well as aggregations of the experiment, were available on-line for use each day in the Spring Experiment at <http://verif.rap.ucar.edu/eval/hwt/2010/>. DTC also had a staff member in attendance throughout the Spring Experiment to lead a discussion on objective verification.

This paper shows examples of preliminary results from the experiment along with some sample interpretation. Initial results support the findings on impact of radar assimilation on the 0-12 hr forecast presented in Kain et. al (2010). Additionally, evaluation of CAPS ensemble products indicate the neighborhood probability QPF (PQPF) and probability matched QPF products show potential for utility in a forecast setting. Finally, this paper provides one example of how the

DTC/HWT objective evaluation may be able to provide feedback to model developers for model physics improvement. The Thompson microphysics scheme in CAPS SSEF WRF-ARW simulations appeared to over-develop the stratiform region of some convective systems. This over-prediction was reflected in the large frequency bias of the CAPS simple ensemble mean of 18dBZ echo top height. A more thorough analysis of the results will be conducted retrospectively.

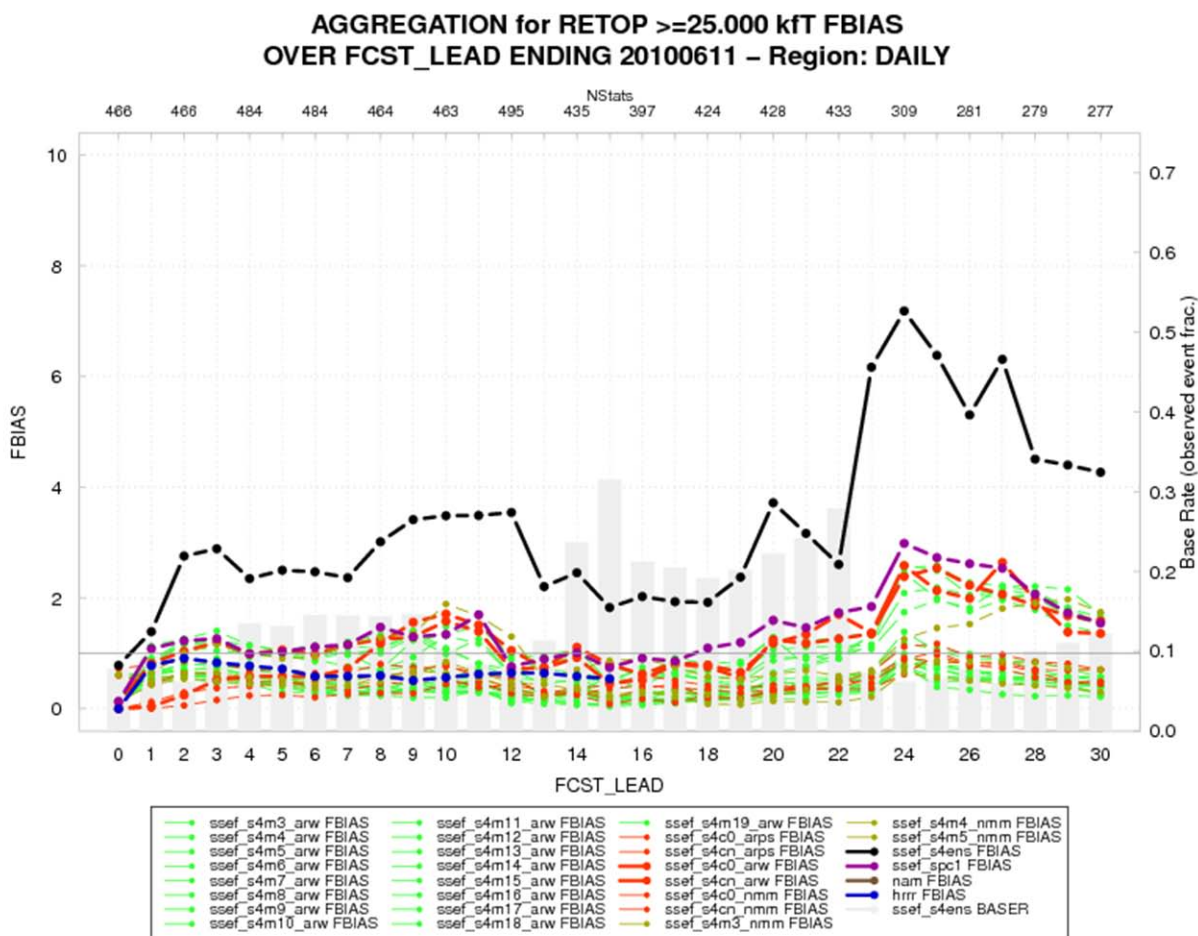


Figure 5. Frequency Bias for 18dBZ echo top height >25,000 ft aggregated over four weeks of the HWT Spring Experiment. Red lines: control members of the CAPS SSEF representing models with (ssef_s4cn) and without (ssef_s4c0) radar assimilation. Green lines: WRF-ARW members of the CAPS SSEF. Gold lines: WRF-NMM members of CAPS SSEF. Black: CAPS SSEF simple ensemble mean. Purple: CAPS SSEF 1 km WRF-ARW simulation. Brown: NCEP/EMC NAM 12 km WRF-NMM simulation. Blue: NOAA/ESRL HRRR 3 km WRF-ARW simulation. Gray bars: Base Rate (or observed event fraction) and scaled on right axis.

References

Davis C.A., B.G. Brown, and R.G. Bullock, 2006. Object-based verification of precipitation forecasts, Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.* **134**:1785--1795.

Ebert, E.E., 2001. Ability of a Poor Man's Ensemble to Predict the Probability and Distribution of Precipitation. *Mon. Wea. Rev.* **129**: 2461-2480.

Jensen, T., B. Brown, M. Coniglio, J.S. Kain, S.J. Weiss, and L. Nance, 2010. Evaluation of Experimental Forecasts from the 2009 NOAA Hazardous Weather Testbed Spring Experiment Using Both Traditional and Spatial Methods. *20th Conf. on Prob. and Stats in the Atmos. Sci.*, Atlanta, GA USA, American Meteorological Society

Kain, J.S., M. Xue, M.C. Coniglio, S.J. Weiss, F. Kong, T.L. Jensen, B.G. Brown, J. Gao, K. Brewster, K.W. Thomas, Y. Wang, C.S. Schwartz, and J.J. Levit, 2010. Assessing Advances in the Assimilation of Radar Data within a Collaborative Forecasting-Research Environment. *Wea. and Forecasting*, in press.

Kong, F., M. Xue, M. Xue, K. K. Droegemeier, K. W. Thomas, Y. Wang, J. S. Kain, S. J. Weiss, D. Bright, and J. Du, 2008: Real-time storm-scale ensemble forecast experiment - Analysis of 2008 spring experiment data. *Preprints, 24th Conf. Several Local Storms*, Amer. Meteor. Soc., Savannah, GA, , paper 12.3.

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.

Acknowledgements

The Developmental Testbed Center is funded by the National Oceanic and Atmospheric Administration, Air Force Weather Agency and the National Center for Atmospheric Research. The CAPS research was supported by an allocation of advanced computing resources provided by the National Science Foundation. The computations were performed on Athena (a Cray XT4) at the National Institute for Computational Science (NICS; <http://www.nics.tennessee.edu/>)

Dedicated work by many individuals led to the success of SE2010. At DTC, Paul Oldenburg, John Halley Gotway, Randy Bullock, and Nancy Rehak developed the evaluation system. David Ahijevych, Jamie Wolff, and Isidora Jankov also from the DTC, led discussions related to forecast verification during SE2010 daily activities. At the SPC, HWT operations were made possible by technical support from Israel Jirak, Chris Melick, and Andy Dean. At the NSSL, Ryan Sobash provided valuable technical support. David Novak and Faye Barthold from NOAA/Hydrometeorological Prediction Center, and Jason Levit from NOAA/Aviation Weather Center provided much needed guidance on meaningful evaluation criteria. Curtis Alexander from NOAA/Earth System Research Laboratory provided some insightful comments about the High Resolution Rapid Refresh performance.