P.48

The Developmental Testbed Center Objective Evaluation Performed During the Hazardous Weather Testbed 2010 Spring Experiment.

Tara Jensen[1*], Steve Weiss[2], Jack Kain[3], Michelle Harrold[1], Ming Xue[4], Fanyou Kong[4], Barb Brown[1], Patrick Marsh[3], Adam Clark[3], Kevin Thomas[4], Mike Coniglio[3], and Russ Schneider[2]

[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
*Presenting Author E-mail: jensen@ucar.edu

Abstract: 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. 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. This year's Spring Experiment ran from May 17 – June 18, 2010.

Approach

This year DTC evaluated (in near real-time):

•CAPS SSEF 4 km ensemble members as deterministic run (26 multi-model)
•CAPS 1km deterministic
•CAPS SSEF 4 km ensemble products (15 radar assimilation models used)
•NOAA/ESRL HRRR 3 km deterministic model
•NOAA/EMC NAM 12 km deterministic model *(baseline)*
•NOAA/EMC SREF 32-35 km ensemble products (21 member multi-model) *(baseline)*

•Observations: NSSL NMQ Q2 QPF and Reflectivity products

A full description of the each model contributed to HWT may be found on their website at: http://hwt.nssl.noaa.gov/Spring_2010/. Objective evaluation results are available at http://verif.rap.ucar.edu/eval/hwt/2010/.

Variables Evaluated :

FCST Field	Observation (NSSL Q2 fields)	Traditional (MET/Grid-Stat)	Spatial (MET/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, 50th and 90th 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	See above	CAPS members, CAPS ens mean, SREF ens mean, HRRR, NAM
Sim. Comp. Refl (20,30,40,50 dBZ)	Composite refl (20,30,40,50 dBZ)	GSS, CSI, FAR, PODY, FBIAS	See above	CAPS members, CAPS ens products, HRRR, NAM
18 dBZ Echo Top (18, 25, 30, 35, 40, 45 kft)	18dBZ Echo Top (18, 25, 30, 35, 40, 45 kft)	GSS, CSI, FAR, PODY, FBIAS	See above	CAPS members, CAPS ens products, HRRR
Prob of 40dBZ echos	Composite reflectivity (40dBZ)	Brier Score, Decomp of Briar score, Area under ROC	None	Ensemble products from CAPS and SREF
50% Prob of 40dBZ echos	Composite reflectivity (40dBZ)	None	See above	Ensemble products from CAPS

Grid: All models were re-gridded to the 4 km Stage IV grid configuration.

Domains: 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.

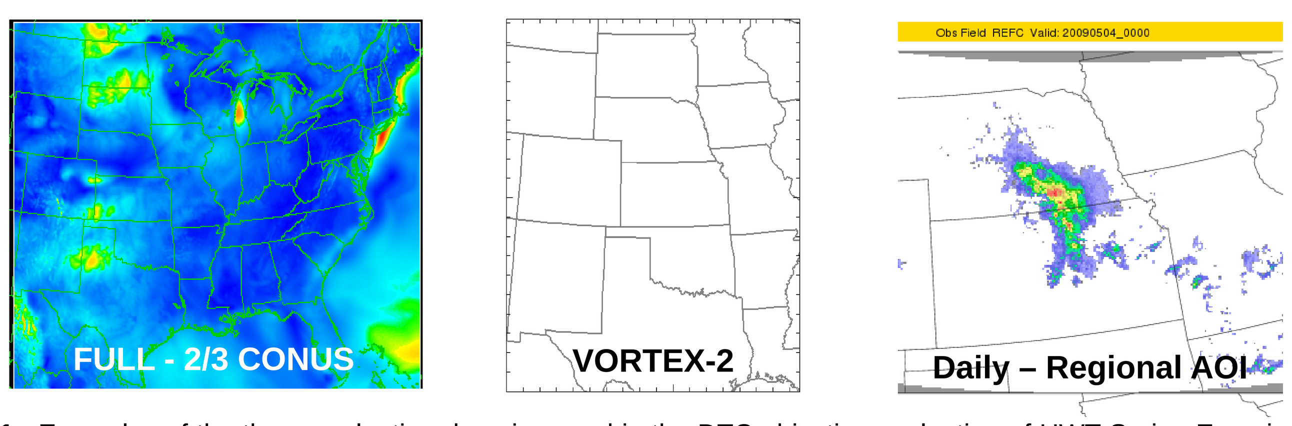


Figure 1. Examples of the three evaluation domains used in the DTC objective evaluation of HWT Spring Experiment models. Left: Full domain represents two-thirds CONUS; Middle: VORTEX-2 domain; and Lower Right: regional area-of-interest domain that moves daily; referred to as daily domain.

Preliminary Results: Severe Storms

Details:

DOMAIN: VORTEX-2

Field: 20 dBZ reflectivity (REFC)

CAPS SSEF Ensemble PM Mean
CAPS SSEF 1 km Model
CAPS SSEF ARW-CN (control w/o radar assimilation)
CAPS SSEF ARW-C0 (control w/o radar assimilation)
HRRR
NAM

PM Mean = Probability Matching used to compute mean REFC [reference: Ebert (2000), MWR]

Statistics aggregated over entire 5 week experiment

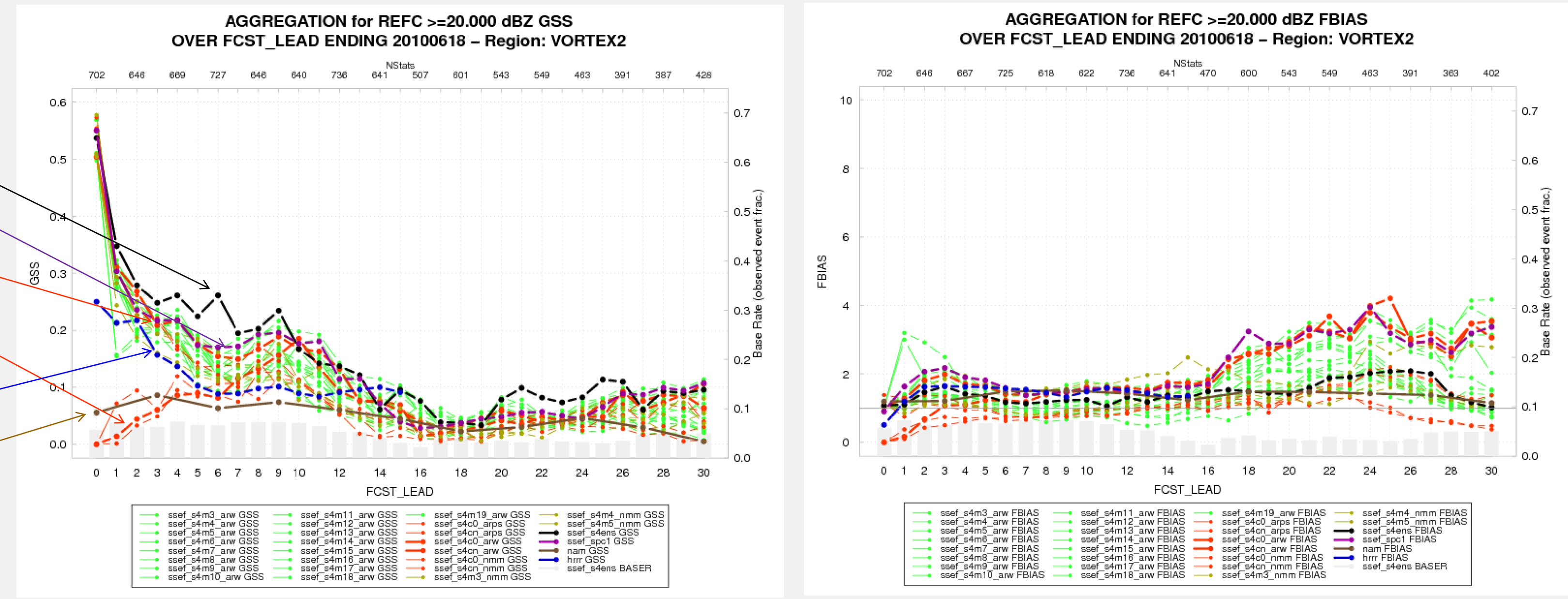


Figure 2. GSS (left) and FBIAS (right) for simulated reflectivity greater than 20 dBZ aggregated over five weeks of the HWT Spring Experiment. Red lines: control members of the CAPS SSEF representing models with (ssef_s4cn) and without (ssef_s4c0) radar assimilation. Green lines: WRF-ARW members of the CAPS SSEF. Gold lines: WRF-NMM members of CAPS SSEF. Black: CAPS SSEF Prob Match 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.

Discussion:

Based on Gilbert Skill Score,
(an assessment of how well the forecast "yes" events correspond to observed "yes" events while taking into account possibility of a random chance "yes")

CAPS SSEF PM Mean and 1 km Deterministic runs appear to provide advantage over baselines HRRR and NAM.

Radar assimilation appears to provide advantage over no assimilation during 0-6hr fcst

However some of that advantage may be due to a positive FBIAS, which is

$$\frac{\text{forecast events}}{\text{observed events}}$$

Increased Forecast Events improves potential for a Hit

Preliminary Results: Hydro/QPF

Details:

DOMAIN: DAILY

Fields:
Quant. Precip Est. (QPE)
Quant. Precip. Fcst. (QPF)
Probability QPF (PQPF)
MODE Objects for >0.5 inch

Simple PQPF = Ensemble Relative Frequency at Gridpnt
PN PQPF = Neighborhood method applied before computing Ens. Rel. Freq.
PM QPF = Probability Matching used to compute QPF intensity

Statistics for Single Valid Time 8 Jun 2010 – 12 UTC

Probabilistic Scores (from MET/Grid-Stat)

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.

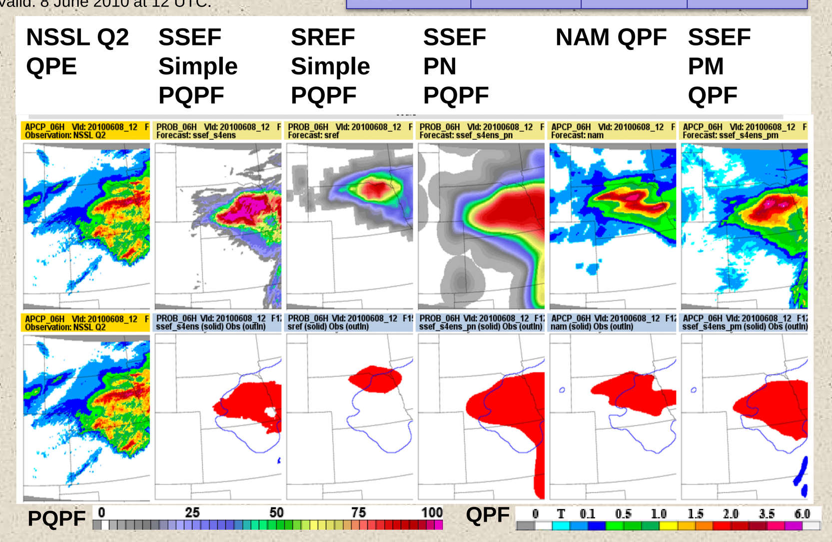
Model	CAPS Simple PQPF	SREF Simple PQPF	CAPS PN PQPF
Brier Score	0.081	0.134	0.088
Reliability	0.004	0.020	0.013
Resolution	0.068	0.032	0.071
Uncertainty	0.145	0.145	0.145
Area Under ROC	0.867	0.720	0.924

$$BS=\frac{1}{T}\sum(p_i-o_i)^2=\frac{1}{T}\sum(p_i-\overline{o}_i)^2-\frac{1}{T}\sum n_i(p_i-\overline{o}_i)^2+\overline{o}(1-\overline{o})$$

Brier Score *(i.e. Mean Sqare Prob Error)*; **Reliability** *(i.e. Conditional Bias)*; **Resolution** *(i.e. Event Discrimination)*; **Uncertainty** *(i.e. Observation Variance)*

Where: T = Total Sample; p_i = forecast probability; o_i = corresponding obs
This score is sensitive to the base rate or climatological frequency of the event.
Forecasts of rare events can have a good BS without having any actual skill.
Since Brier score is a measure of error, smaller values are better.

Figure 3. Example Screen Shot from DTC/HWT Website for PQPF/QPF evaluation. MODE objects for >0.5 inches. Valid: 8 June 2010 at 12 UTC.



Spatial Attributes (from MET/MODE)

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
Intersect. Area	6539 gridsq	1530 gridsq	9798 gridsq	4178 gridsq	8112 gridsq	7680 gridsq
50% Accum Diff (F-O)	n/a	n/a	n/a	0.007 in	0.07 in	-0.08 in
90% Accum Diff (F-O)	n/a	n/a	n/a	0.70 in	0.54 in	0.02 in

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.

Discussion:

Based on the Brier Score
SSEF Simple and PN PQPF appear to have more skill than than SREF Simple PQPF.
Based on Reliability and Resolution
The SSEF PN PQPF shows slightly more ability to resolve events but has slightly less reliability.
Based on Area Under the Curve
With a higher Area under the Receiver Operating Characteristic Curve SSEF PN QPN may be deemed the slightly more skillful for this application.

Based on Centroid Distance, Area Ratio, Intersection Area
SSEF Ensemble Products appear to provide better QPF location guidance than SREF and NAM for this case.
Based on 50th Percentile and 90th Percentile Intensity Differences
SSEF Simple Ensemble Mean provided the best guidance for this case.

Aggregations of these individual case products will become available on the website near the end of the summer.

Preliminary Results: Aviation Weather

Details:

DOMAIN: DAILY

Fields:
Observed Reflectivity (REFC)
Forecasted Simulated REFC
Observed 18 dBZ Radar Echo Top (RETOP) height
Forecasted RETOP height
MODE Objects for >25,000 ft

Microphysics Schemes:
Thompson
WRF Single Moment (WSM)
WRF Double Moment (WDM)
Morrison

All Models in example have radar assimilation in Initial Conditions

PM RETOP: = Probability Matching used to compute QPF intensity

Statistics aggregated over entire 5 week experiment

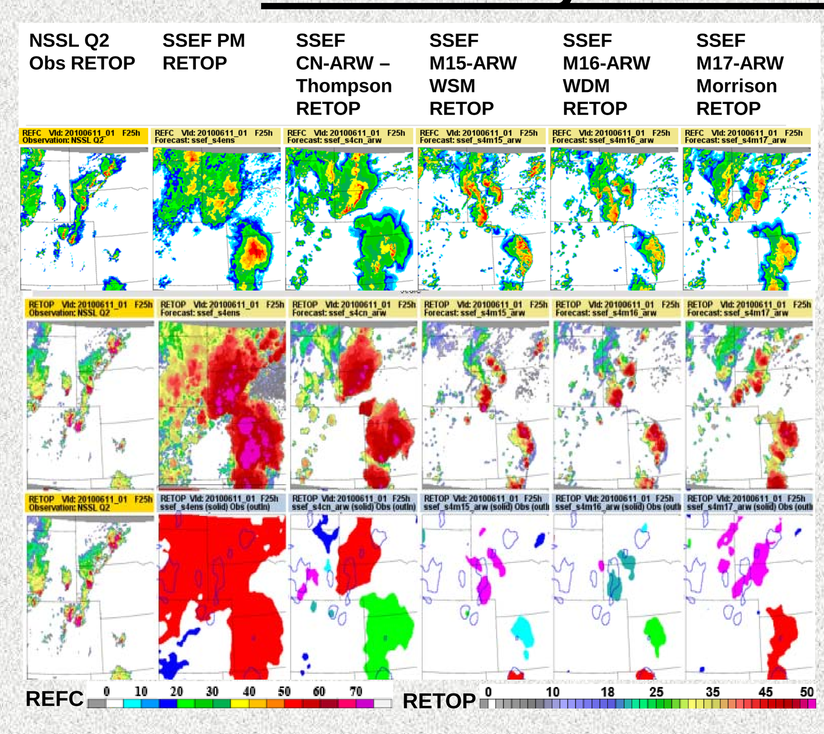


Figure 4. Example Screen Shot from DTC/HWT Website for REFC/RETOP evaluation. MODE objects for >25kFT Valid: 11 June 2010 at 01 UTC.

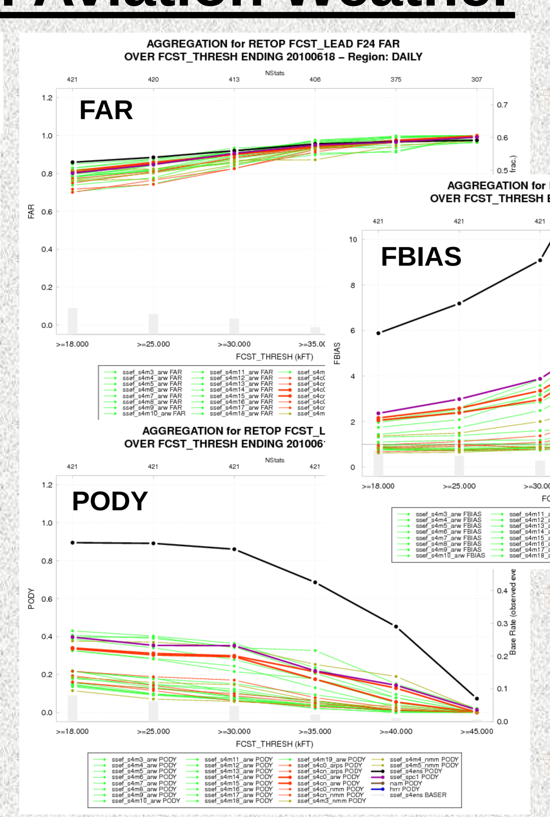


Figure 5. FAR(top left), PODY (bottom left) and FBIAS (right) for radar echo top heights aggregated over five weeks and plotted as a function of threshold. Colors are as in Fig 2.

Discussion:
Based on Visual Inspection
(of Fields and MODE Objects to left)

SSEF Probability Matched Mean appears to over-estimate RETOP spatial coverage

Thompson scheme members appears to be significantly over-predicting the stratiform cloud shield (green in REFC over predicted leading to much larger area for RETOP also).

Thompson members appear to dominate Probability Matching (example not shown here – see attached plot)

Other Members seem to have better Frequency Bias but are over developing southern convective region. This may be a timing error.
Based On FAR (top middle)
SSEF Probability Matched Mean appears to have high FAR
Based On FBIAS (bottom middle)
SSEF Probability Matched Mean appears to have significantly Probability of Detection but this is most likely due to high FBIAS
Based On FBIAS (right)
SSEF Probability Matched Mean appears to over-estimate RETOP spatial coverage

Summary

The DTC/HWT Objective Evaluation goals include are to 1) to provide objective evaluations of the experimental forecasts; 2) to supplement and compare to subjective assessments of performance; and 3) to expose the forecasters and researchers to both new and traditional approaches for evaluating forecasts.

In total, 30 models and 3 ensemble product methods were evaluated in near real-time for the 5 week HWT 2010 Spring Experiment. Three forecast challenges were addressed: Severe Weather, QPF, and Aviation Weather. The DTC evaluated 1-2 variables for each area.

Some preliminary objective results:
•Models using radar assimilation methods exhibited improved skill during 0-6 hr lead times
•CAPS Convective Allowing Models with radar assimilation generally exhibited improved skill over operational baselines (such as NAM, HRRR, and SREF) but this improvements may be a by-product of increased frequency bias and hence may result in increase false alarms.

Preliminary observations about Ensemble Products:
•Simple arithmetic mean field only available for QPF field – it showed improved skill over other methods on some days but not all. Further investigation is recommended.

•Probability Matched mean product exhibits higher skill for reflectivity (REFC) and quantitative precipitation forecast (QPF) fields but appears to produce an sizable frequency bias (FBIAS) for radar echo top (RETOP), and hence increased false alarm ratio (FAR)s that may make the field unusable

•Probability Neighborhood product was evaluated for QPF only. While the product exhibits a smoother Probability QPF field, which in general appears to increase traditional skill, the tendency for a sizable frequency bias may decrease its utility. The true test is which product was found to be the most useful. Further investigation of subjective logs is needed.

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. Paul Oldenburg, John Halley Gotway, Randy Bullock, and Nancy Rehak developed the DTC 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 scientific and technical support. David Novak, Faye Barthold (from NOAA/Hydrometeorological Prediction Center), and Jason Levit (from NOAA/Aviation Weather Center) provided much needed guidance on meaningful evaluation criteria.