V. Developmental Testbed Center Objective Evaluation Background

New Objective Verification Approaches

Subjective verification of model forecasts has been a cornerstone to HWT activities in previous years. This approach has provided valuable insights into how forecasters use numerical models, and facilitates the gathering of information about the value of new guidance tools from the perspective of a forecaster. In addition, traditional verification measures (e.g., Equitable Threat Score or ETS) used for synoptic scale and mesoscale model forecasts of discontinuous variables such as precipitation typically provide less useful information (and even misleading information) about forecast accuracy as the scale of the phenomena being evaluated decreases. This is because the ETS is proportional to the degree of grid scale overlap in space and time between the forecasts and observations, and there is typically low predictability on convective scales. Despite these limits, operational severe weather forecasters have often found value in higher resolution forecasts of thunderstorms and convective systems, since they can provide unique information about convective mode, coverage, and evolution that is not resolved by mesoscale models using parameterized convection. In recent years, we have found that subjective evaluation has great potential to serve as a comparative benchmark for assessing new objective verification techniques designed for high resolution NWP from convection-allowing models (CAMs), and has had a significant positive impact on model development strategies.

In order to better utilize subjective and objective verification techniques in a complementary manner, simulated composite reflectivity and 6-hr QPF output from several model runs will be evaluated using subjective visual comparisons and objective statistical measures produced by the Developmental Testbed Center's (DTC) Model Evaluation Tools (MET). The focus this year will be on probabilistic predictions, particularly of extreme precipitation events and strong convection as it relates to convective initiation. All members of the Center for Analysis and Prediction of Storms (CAPS) Storm Scale Ensemble Forecast (SSEF) system will be evaluated for select variables (see Table 1). Ensemble products from the twenty-four or twenty-five member (ssef_s4ens), fifteen member (ssef_s4ens15), and five member (ssef_s4ens5) ensembles selected by the NOAA Storm Prediction Center (SPC) will also be evaluated. Operational (or near-operational) models will be used as a baseline for comparison. Probabilistic baselines include ensemble products from the Short Range Ensemble Forecast (SREF) and Hybrid Regional Ensemble Forecast (HREF), and High-Resolution Model Output Statistics (HRMOS) systems. The deterministic models include the 12 km operational North American Mesoscale Model (NAM), the 3 km High Resolution Rapid Refresh (HRRR), both 4km east HiResWindows (NMM and ARW), the 12 km NMM-B parent domain (NMMB_12) and 4 km CONUS nest (NMMB_4). Other contributing models, such as the Storm Scale Ensemble of Opportunity generated by SPC and the NOAA/GSD LAPS short-range deterministic and ensemble member will be brought in and archived for retrospective studies.

MET is designed to be a highly-configurable, state-of-the-art suite of verification tools. We will focus on the use of the object-based verification called Method for Object-based Diagnostic Evaluation (MODE) that compares gridded model data to gridded observations for the QPF and simulated reflectivity forecasts. MODE output including plots of the objects (see Figure 1) and the attributes associated with the objects will be used to evaluate the CAMs to diagnose different types of convective modes considered important in forecasting convective weather. We will also be providing plots of the smoothed fields for calculating neighborhood statistics (see Figure 1) along with aggregation of statistics such as Fraction Skill Score (FSS). Traditional categorical verification statistics for both probabilistic and single-value (deterministic) fields will be computed. Some of these scores will be plotted and many of them will be available in the DTC database and displayed using the web-based METViewer interface. Details about the DTC MET system can be found at http://www.dtcenter.org/met/users/ . A description of the statistics and MODE attributes provided during the experiment can be found in the MET Users Guide (in Grid Stat and MODE sections as well as Appendix C) and can be downloaded from: http://www.dtcenter.org/met/users/docs/overview.php

Verification "truth" will be provided by NSSL National Mosaic and Multi-Sensor QPE (NMQ) multi-sensor Quantitative Precipitation Estimates (QPE) and three-dimensional radar reflectivity datasets. See http://www.nssl.noaa.gov/projects/q2/ for more information about the NMQ.

Table 1. List of variables (and thresholds) to be evaluated and ready during the subjective evaluation portion of the 2011 Spring Experiment. Evaluation of the Storm Scale Ensemble of Opportunity, several sub-ensembles representing physics experiments, and the LAPS short-range ensemble will be performed by DTC retrospectively. Additional variables, such as 1-hr Accumulated Precipitation and 1 km reflectivity and corresponding probability fields will also be evaluated as resources allow retrospectively.

Members	**REFC (20,30,35,40,50,60 dBZ)**	**APCP_06 (0.5,1.0,2.0")**	**Prob_APCP_06 (0.5,1.0,2.0")**
SSEF Ens (Mean, Max, Prob-Match, Prob-Neigh)	GSS, CSI, FBIAS, FSS MODE Attrib. Rank Histograms	GSS, CSI, FBIAS MODE Attrib. Rank Histograms	Brier Score, ROC, AUC, Reliability Dia. MODE Attrib.
SSEF Ens5,15 SSEF Ens Bias Corr. (Mean, Max, Prob-Match, Prob-Neigh)		GSS, CSI, FBIAS MODE Attrib. Rank Histograms	Brier Score, ROC, AUC, Reliability Dia. MODE Attrib.
Prob. Baselines SREF HREF HRMOS	GSS, CSI, FBIAS, FSS MODE Attrib.	GSS, CSI, FBIAS MODE Attrib.	Brier Score, ROC, AUC, Reliability Dia. MODE Attrib.
Det. Baselines HRRR EastNMM (HiRes) EastARW (HiRes) NAM_Ops NMMB_12 NMMB_4	GSS, CSI, FBIAS, FSS MODE Attrib.	GSS, CSI, FBIAS MODE Attrib.	
SSEF members	GSS, CSI, FBIAS, FSS MODE Attrib.	GSS, CSI, FBIAS MODE Attrib.	

Key: Traditional statistics include Gilbert Skill Score (GSS), Critical Success Index (CSI), Frequency Bias (FBIAS), Brier Score, Receiver-Operator Characteristic Curve (ROC), Area under ROC Curve (AUC), and Reliability Diagrams. The statistic calculated for the neighborhood is Fraction Skill Score (FSS). MODE attributes include centroid distance, intersection area, symmetric difference and more. Rank Histograms (or Talagrand Diagrams) and Spread will be provided for SSEF and SSEO ensembles.

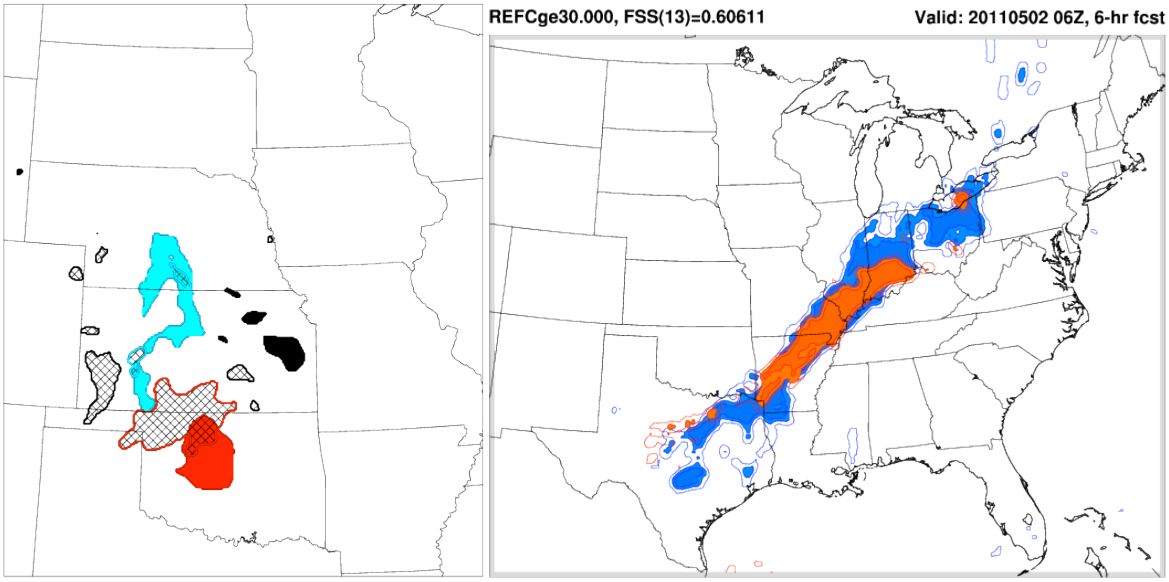

Figure 1. Example of MODE object plot (left) and Neighborhood plot (right). MODE objects: observation (hatched) and forecast (solid) objects/clusters of objects. Colors correspond to matched clusters. Black are unmatched objects/clusters. Neighborhood: fractional coverage of observation (orange) and forecast (blue) fields used in calculating Fraction Skill Score (FSS). Neighborhood reflected by number after FSS.