8B.6 QUANTITATIVE PRECIPTIATION FORECAST (QPF) VERIFICATION COMPARISON BETWEEN THE GLOBAL FORECAST SYSTEM (GFS) AND NORTH AMERICAN MESOSCALE (NAM) OPERATIONAL MODELS

Jamie Wolff*, Barbara Brown, John Halley Gotway, Michelle Harrold, Louisa Nance, Paul Oldenburg and Zach Trabold

National Center for Atmospheric Research/Research Applications Laboratory (NCAR/RAL) and Developmental Testbed Center (DTC), Boulder, CO USA

1. Introduction

Numerical weather prediction (NWP) models continue to move toward higher resolution, which, in turn, provides both a finer level of detail and a more realistic structure in the resulting forecast. It is widely acknowledged, however, that using traditional verification metrics for evaluation may unfairly penalize these high-resolution forecasts (e.g., Davis et al. 2006; Roberts and Lean 2008). Traditional verification requires near-perfect spatial and temporal placement for a forecast to be considered good; this approach favors smoother forecast fields of coarser resolution models and offers no meaningful insight regarding why a forecast is considered good or bad. In contrast, more advanced spatial verification techniques, such as object-based methods, can provide information on differences between forecast and observed objects in terms of displacement, orientation, intensity and coverage areas; neighborhood methods can provide information on the spatial scale at which a forecast becomes skillful.

The Developmental Testbed Center (DTC) performed an extensive evaluation of the Global Forecast System (GFS) and the North American Mesoscale (NAM) operational models to quantify the differences in the performance of Quantitative Precipitation Forecasts (QPF) produced by two modeling systems that vary significantly in horizontal resolution. Traditional verification metrics computed for this test included frequency bias and Gilbert Skill Score (GSS). Two advanced spatial techniques were also examined - the Method for Object-based Diagnostic Evaluation (MODE) and the Fraction Skill Score (FSS) - in an attempt to better associate precipitation forecast differences with different model horizontal scales.

2. Data

2.1 Model Data

QPF output at 3-h intervals from the GFS and NAM models initialized at 00 UTC daily were retrieved from the U.S. National Centers for Environmental Prediction (NCEP) for 18 December 2008 through 15 December 2009. The NAM native output employs an E-grid domain with approximately 12-km grid spacing, while the GFS native output is a global Gaussian grid with 0.5 x 0.5 degree resolution. The ***copygb*** program, developed by NCEP, was used to regrid the native model output to 4-, 15- and 60-km contiguous U.S. (CONUS) grids on a Lambert-Conformal map projection. The budget interpolation option in ***copygb*** was utilized (described in Accadia et al. 2003). This approach attempts to conserve the total area-average precipitation amounts.

*Corresponding author address: Jamie Wolff, NCAR/RAL, P.O. Box 3000, Boulder, CO 80307, email: jwolff@ucar.edu

2.2 Precipitation Analyses

For this evaluation, precipitation accumulation periods of 3 h and 24 h were assessed. The observational datasets included the 4-km NCEP Stage II analysis (Lin and Mitchell 2005) for the 3-h accumulations and the 1/8-degree NCEP Climate Prediction Center daily gauge analysis (Higgins et al. 2000) for the 24-h accumulations (valid at 12 UTC). The observational datasets were also interpolated, using the budget option, to the same 4-km, 15-km and 60-km domains as the model output, prior to the comparison with the forecasts.

3. Verification Method

Objective model verification statistics were generated using version 3.0 of the Model Evaluation Tools (MET) software package, which offers a wide variety of state-of-the-art verification methods (Fowler et al. 2010). Grid-to-grid comparisons were performed to verify QPF using traditional metrics (Wilks 1995), including frequency bias, which measures the ratio of the frequency of forecast events to the frequency of observed events and indicates whether the forecast system has a tendency to under-forecast (<1) or over-forecast (>1) events, and GSS, which measures the fraction of observed and/or forecast events that were correctly predicted and is adjusted for hits associated with random chance, where zero indicates no skill and one is a perfect score. In addition, an object-based verification approach, MODE, and a neighborhood verification method, FSS, were applied. The process of identifying and verifying objects with MODE is fully described in Davis et al. (2006). Briefly, this approach consists of the following steps: (i) resolve objects within the raw forecast and observation fields, (ii) determine if objects within each field should be *merged*, (iii) determine which objects should be *matched* between the forecast and observation fields and determine if additional *merging* in either field should occur, (iv) calculate various quantities of interest (attributes) to assess forecast quality. For FSS, the purpose of the method is to obtain a measure of how forecast skill varies with spatial scale (Roberts and Lean 2008), and includes the steps: (i) convert all forecast and observed fields into binary fields, for each threshold of interest (ii) generate fractions within a square of length *n* that have exceeded the threshold, (iii) compute the mean squared error relative to a low-skill reference forecast (i.e., FSS).

Both the NAM and GFS QPF fields were interpolated to the same 4-, 15- and 60-km domains as the observations. Verification results were computed for select spatial (CONUS, CONUS-East) and temporal (all cases, seasonal) aggregations; however, only the CONUS domain over the entire set of cases will be

discussed here. For the 3-h QPF, all verification scores were evaluated every 12 h out to 84 h. The 24-h QPF verification scores were evaluated at the 36-, 60- and 84-h lead times. A subset of these results will be discussed in Section 4.

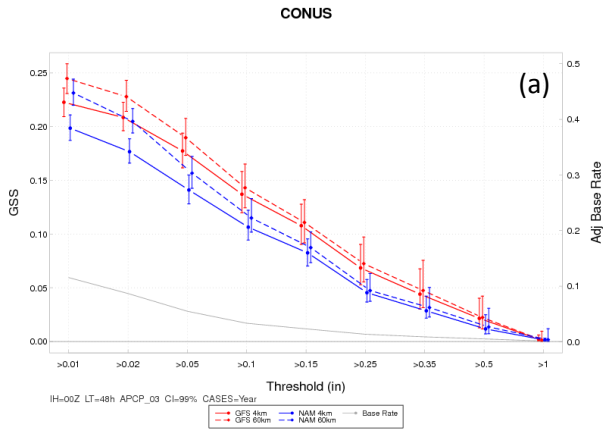
Verification statistics generated by MET for each retrospective case were loaded into a MySQL database, from which data was then retrieved to compute and plot requested aggregated statistics using routines developed by the DTC in the statistical programming language, R. The traditional verification metrics are accompanied by confidence intervals (CIs), at the 99% level, computed using a bootstrapping technique. When comparing the models, a conservative estimation of statistically significant (SS) differences was employed based solely on whether the aggregate statistics with the accompanying CIs overlapped. If no overlap was noted for a particular threshold, the differences between the models were considered SS.

4. Results

4.1 Traditional Verification Metrics

4.1.1 Gilbert Skill Score

For the 3-h QPF, both of the models show an exponential decrease in GSS values with increasing threshold, regardless of forecast lead time (Fig. 1a). Also, as expected, the base rate decreases with increasing threshold, with very few observations associated with the highest accumulation values. When examining lead times valid at 00 UTC (i.e., 24-, 48-, 72-hr), the NAM forecast interpolated to the 4-km domain is SS lower than that interpolated to the 60-km domains at the lowest thresholds. In general, most other thresholds for both the NAM and GFS forecasts valid at 00 UTC and all other lead times and thresholds valid at 12 UTC (not shown) reveal no SS differences. The latter result also holds for the 24-h QPF (Fig. 1b).



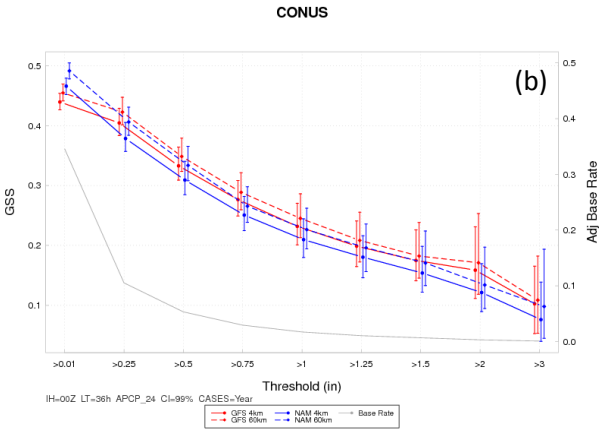


Figure 1. Threshold series plots of (a) 48-h lead time for 3-h QPF (in) and (b) 36-h lead time for 24-h QPF (in) for median GSS aggregated across all model initializations run. The GFS is shown in red, NAM in blue, 4-km domain in solid lines and 60-km domain in dashed lines. The vertical bars represent the 99% CIs. Associated with the second y-axis, the light grey line is the adjusted base rate, or the ratio of observed grid box events to the total number of grid boxes in the domain, by threshold.

4.1.2 Frequency Bias

The sign of the bias for both models depends strongly on the threshold and lead time. For forecast lead times valid at 00 UTC, there is a SS high bias for the 3-h QPF for thresholds at and below 0.02" for the NAM and 0.05" for the GFS (Fig 2a). At and above the 0.15" threshold, both models exhibit a SS low bias. Neither of these results depends on the resolution of the interpolated verification domain. However, for the forecast lead times valid at 12 UTC (i.e. 12-, 36-, 60 and 84-hr) the interpolated verification domain also has an impact on the results. For the 36-, 60- and 84-hr lead times the NAM exhibits a SS high bias for thresholds less than or equal to the 0.15" threshold (Fig 2b). For the GFS, the SS high bias is for thresholds at or below 0.10". A transition to a SS low bias is noted for both models at and above the 0.25" thresholds, excluding the NAM forecasts interpolated to the 4km domain, where the CIs more often encompass one, indicating the estimated bias is not statistically different from the value for an unbiased forecast. The CIs associated with the bias for the NAM 24-h QPF interpolated to both the 4-km and 60-km domains encompass the value of one for all thresholds above 0.75" while the lower thresholds have a SS low bias (Fig. 2c). However, the results for the GFS reveal a general SS high bias for the lowest and highest thresholds, while CIs for mid-range thresholds also encompass a bias value of one.

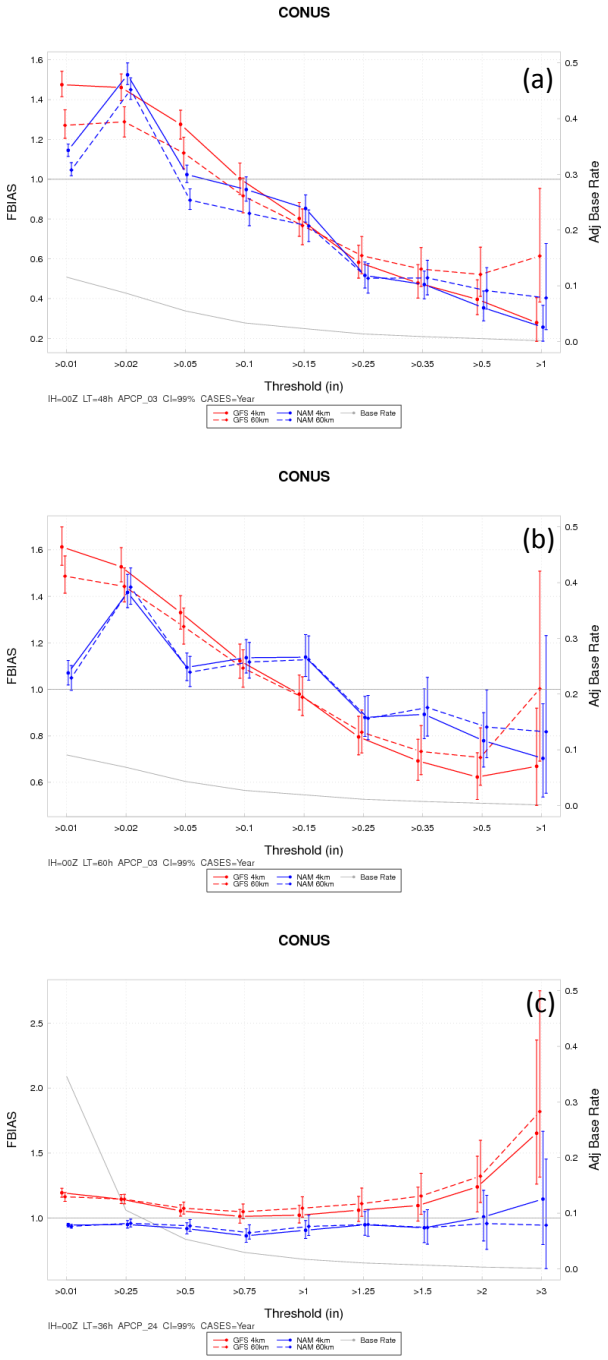


Figure 2. Threshold series plots of (a) 48-h lead time for 3-h QPF (in) (b) 60-h lead time for 3-h QPF (in) and (c) 36-h lead time for 24-h QPF (in) for median frequency bias aggregated across all model initializations run. The GFS is shown in red, NAM in blue, 4-km domain in solid lines and 60-km domain in dashed lines. The vertical bars represent the 99% CIs. Associated with the second y-axis, the light grey line is the adjusted base rate, or the ratio of observed grid box events to the total number of grid boxes in the domain, by threshold.

4.2 Spatial Verification Techniques

4.2.1 Method for Object-based Diagnostic Evaluation (MODE)

For this test, a convolving disk of 2 gridpoints for the 60-km domain, 8 gridpoints for the 15-km domain, and 15 gridpoints for the 4-km domain was used. A threshold of 0.01″ for the 3-h and 0.2″ for the 24-h precipitation accumulation fields was then applied to define discrete rain objects of the approximate size of interest for this study. A second, lower threshold was then applied to each individual field and merging of objects in the same field was allowed if the original objects defined became one object after the lower threshold was applied. Objects were matched between the forecast field and observed field if the total interest value between a forecast observation object pair (a weighted sum of interest values for centroid distance, boundary distance, angle difference, area ratio and intersection area ratio) was greater than 0.7. Additional merging was allowed if two or more objects in one field matched the same object in the other field. An example of the objects created from the forecast and observation fields for one particular valid time are shown in Fig. 3. Many unique attributes can be examined when looking at MODE output; however, only a few will be discussed in this paper.

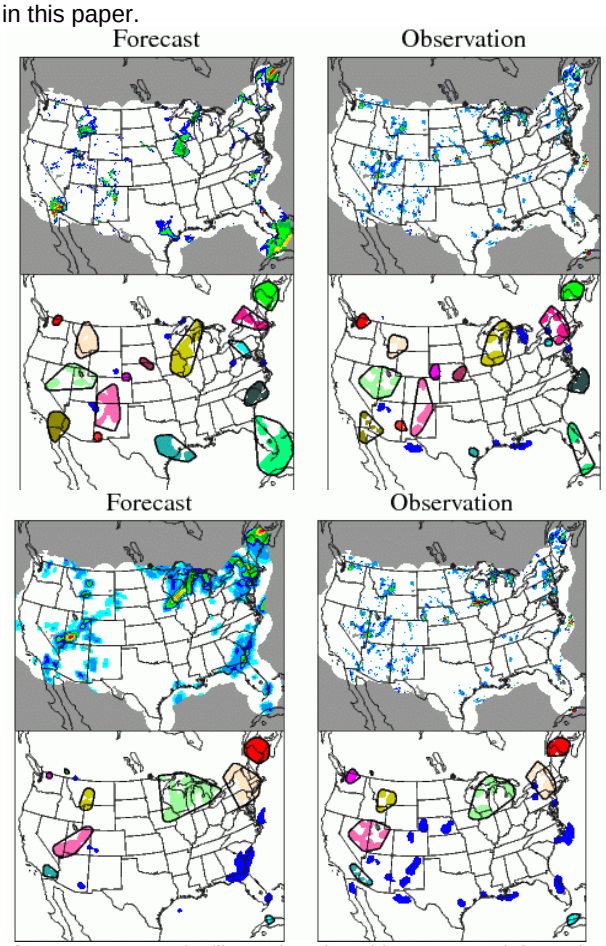


Figure 3. Example illustrating the objects created from the NAM (top) and GFS (bottom) 3-h QPF fields (left panel) and the associated fields from the Stage II observation field (right panel). Both the forecast and observations are on the 4-km domain.

Fig. 4 shows the total counts of objects from the two models and the observation field on the 4-km domain by forecast lead time. The count represented here is the total number of (simple matched and unmatched) objects in each field summed over the entire set of valid times run. From this plot, it is seen that the total object counts from the NAM4 more closely reproduces that of the observations, while there are too few overall objects within the GFS4. In addition, box plots of the distribution of object size by lead time are shown in Fig. 5. Though both models produce objects that are generally too large, the NAM4 better matches the distribution of object sizes in the observation field than the GFS4. Even though the GFS has too few objects, those that are produced are too large; for all valid times except between 18-03 UTC, the GFS has a SS higher total area coverage when compared to NAM4 though both models have SS higher total area coverage for all forecast lead times when compared to the observations (Fig. 6). The box plots for the GFS indicate a significant bias towards larger objects, and, even though there are too few number of total objects defined in the GFS forecast, the size of those objects are larger than those from the NAM, covering a larger total area.

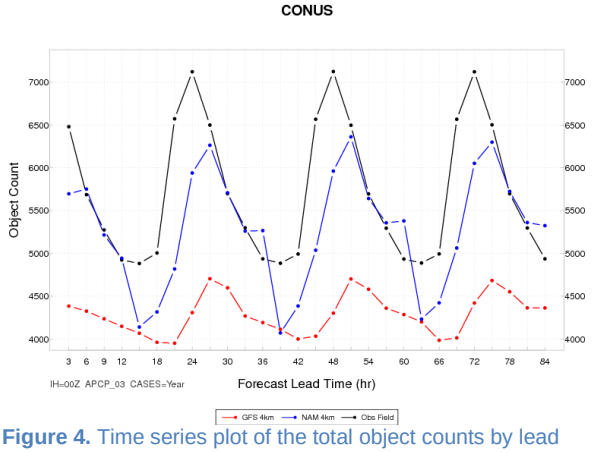


Figure 4. Time series plot of the total object counts by lead time for the GFS4, NAM4 and observations on the 4-km domain aggregated across all model initializations run. The GFS object counts are shown in red, NAM in blue and the observations in black.

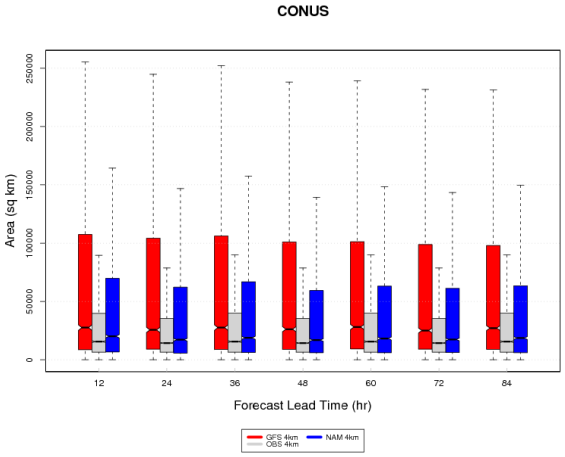


Figure 5. Box plots by lead time of the size distribution for objects defined within the GFS4 (red), NAM4 (blue) and observation field (grey) for the 4-km domain. The top and bottom of the box correspond to the 25th and 75th percentile, respectively; the black line at the "waist" is the median value.

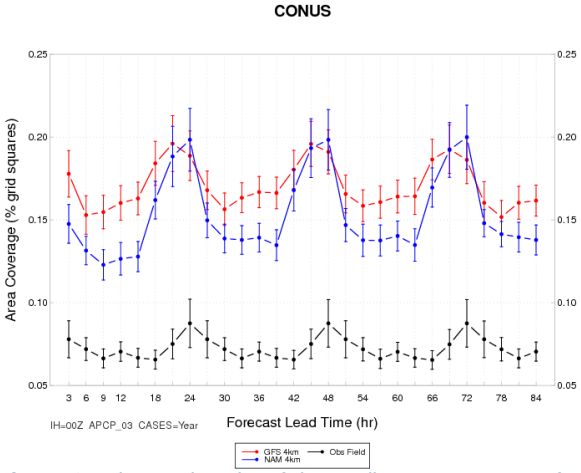


Figure 6. Time series plot of the median area coverage from the GFS4 (red), NAM4 (blue) and observation field for the 4-km domain and aggregated across all model initializations run. The vertical bars represent the 99% CIs.

The area ratio of all forecast objects to all observed objects is a way for MODE to compute its own version of frequency bias. Similar to the results seen from the traditional metric, for the 3-h accumulations, the GFS on the 4-km grid has a higher bias at nearly all forecast lead times; the only exception is for 00 UTC valid times (Fig. 7).

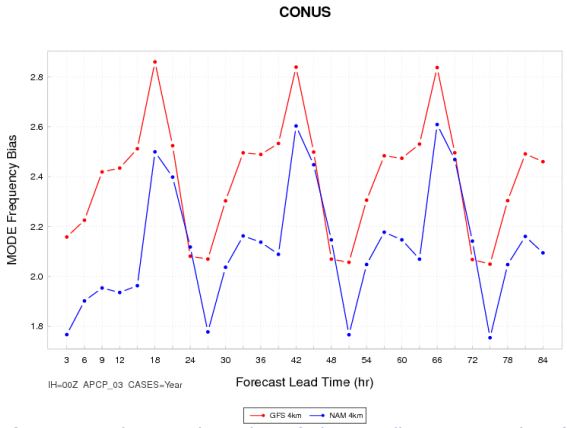


Figure 7. Time series plot of the median area ratio of all forecast objects to all observed objects for the 4-km domain and aggregated across all model initializations run.

4.2.2 Fractional Skill Score (FSS)

For this evaluation, neighborhood sizes (in terms of grid squares) of *n*=3, 5, 7, 9, 11, 13 were computed for each model. For the NAM15, additional neighborhood sizes of *n*=19, 21, 27, 29, 35, 37, 43, 45, 51, 53, 59, 61 were also examined to represent the similar horizontal scales as those used for the 60-km neighborhood sizes. Because the neighborhood size is required to be an odd integer and the resolution difference between the two models is a factor of four, an exact match between the neighborhood sizes for the GFS60 and NAM15 was not possible.

As expected, FSS increases with neighborhood size and decreases with lead time (Fig. 8). A diurnal cycle is also evident with the largest FSS values seen during the evening and overnight hours (00-12 UTC) and the smallest values during the daytime (15-21 UTC). For the 0.1″ threshold, the FSS for the NAM15 (e.g., 405-km neighborhood) is consistently larger than the corresponding value for the GFS60 (e.g., 420-km neighborhood). When examining multiple precipitation accumulation thresholds for a single lead time, it is clear that there is an increase in FSS with decreasing threshold (Fig. 9); however, the result holds that the NAM15 FSS is always higher than the GFS60 at similar neighborhood sizes.

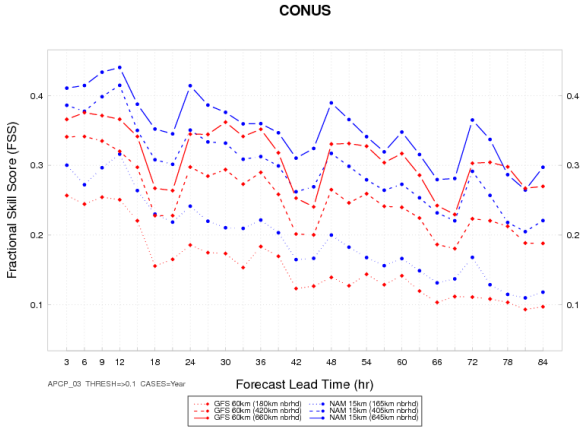


Figure 8. Time series plot of FSS using a threshold of 0.1" aggregated across all model initializations run. The GFS60 is shown in red, for n=3 (dotted), 5 (dashed) and 7 (solid) and the NAM15 in black for n=11 (dotted), 19 (dashed) and 27 (solid).

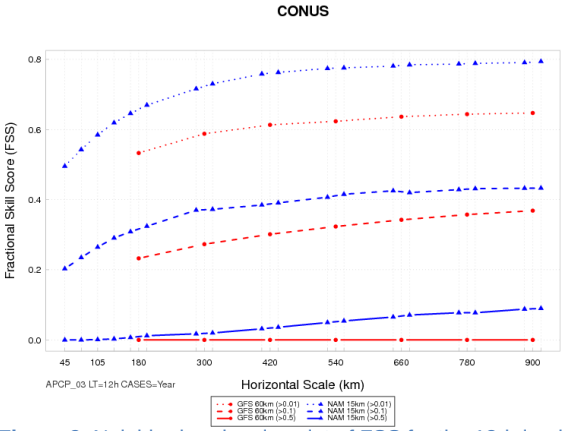


Figure 9. Neighborhood series plot of FSS for the 12-h lead time aggregated across all model initializations run. The GFS60 is shown in red and NAM15 in black for thresholds of 0.01" (dotted), 0.1" (dashed) and 0.5" (solid).

5. Summary

Results comparing two NWP models which vary significantly in horizontal resolution were described. Using traditional verification metrics, there is no notable, consistent forecast improvement with the higher resolution NAM model even though, subjectively, the finer detail more closely matches the observations in many cases. However, when looking at MODE and FSS, additional attributes can be examined that provide further information about the skill of the forecast. The objects created by MODE reveal additional information on the area and correspondence of objects, between the forecasts and observations. A clear high bias is noted for the GFS object areas. The FSS evaluation clearly shows that the higher-resolution NAM has comparable skill to the GFS at considerably smaller neighborhood sizes and larger FSS values at comparable neighborhood sizes.

Acknowledgements

The authors would like to thank Ying Lin at NCEP/EMC for her assistance in acquiring the model and observation data used for this evaluation. This work was funded by the National Oceanic and Atmospheric Administration. NCAR is sponsored by the National Science Foundation.

References

Accadia, C., S. Mariani, M. Casaioli, and A. Lavagnini, 2003: Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids. *Wea. and Forecasting*, **18**, 918-932

Davis, C., B. Brown, R. Bullock, 2006: Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas. *Mon. Wea. Rev.*, **134**, 1772-1784.

Fowler, T. L., T. Jensen, E. I. Tollerud, J. Halley Gotway, P. Oldenburg, R. Bullock, 2010: New Model Evaluation Tools (MET) Software Capabilities for QPF Verification. *Preprints*, 3rd Intl. Conf. on QPE, QPF and Hydrology, Nanjing, China, 18-22 October 2010.

Higgins, R. W., W. Shi, E. Yarosh, and R. Joyce, 2000: Improved United States precipitation quality control system and analysis. *NCEP/Climate Prediction Center Atlas 7,* National Weather Service, NOAA, U.S. Department of Commerce, 40 pp.

Lin, Y. and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and applications. *Preprints*, 19th Conf. on Hydrology, American Meteorological Society, San Diego, CA, 9-13 January 2005.

Roberts, N. M., H. W. Lean, 2008: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Mon. Wea. Rev.*, **136**, 78-97.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.