# 24.6 A COMPARISON OF MM5, WRF, RUC, AND ETA PERFORMANCE FOR GREAT PLAINS HEAVY PRECIPITATION EVENTS DURING THE SPRING OF 2003

Peter J. Sousounis*, Todd A. Hutchinson, Stephen F. Marshall

Weather Services International Corporation
Andover, MA

## 1. INTRODUCTION

An effort was completed recently at WSI Corporation to evaluate the performance of existing numerical weather prediction (NWP) mesoscale models with an emphasis on evaluating the precision as well as the accuracy of 0-12 hour precipitation forecasts. For example, one of the goals of the research effort was to evaluate model accuracy of onset and cessation of heavy (convective) precipitation on hourly time scales. Towards that end, a suite of four different existing numerical weather prediction mesoscale models was recently evaluated.

An active 2003 Spring severe weather season provided a good opportunity to evaluate a series of nearly consecutive periods of very active weather. The active weather was concentrated over the Great Plains where the Storm Prediction Center (SPC) logged nearly 1300 reports of hail, nearly 3200 reports of damaging high winds, and almost 500 reports of tornadoes between April 15 and May 11 2003.

More than one hundred different model configurations (spanning four different models) combined with 39 different initialization times resulted in almost 4000 simulations being performed within a two month work period. The model output was verified using traditional techniques (e.g., threat scores) as well as using a newly developed technique called acuity-fidelity (Marshall et al. 2004). The objective of this acuity-fidelity (AF) technique is to account for temporal and intensity errors as well as spatial errors and then to cast the result in terms of a unidimensional result. Hence the utility of this method is to evaluate model forecasts more accurately and fairly compared to traditional methods.

## 2. METHODOLOGY

### *a. model simulations*

The four models evaluated were MM5 (v3.5, Grell et al, 1993), WRF (v1.3, Michalakes et al. 2001, Skamarock et al. 2001), ARPS (v5.0.0, Xue et al. 2000), and the workstation Eta (Listemaa, 2002). Different configurations were run as existing options for existing physical parameterizations for each model allowed. The bulk of the configurations were provided by MM5 and WRF (48 each) because of the many user specified options available with these models. The ARPS model allowed some choices for physics options, which resulted in 16 configurations that could be evaluated. The workstation Eta was evaluated using only two configurations – one with Betts-Miller-Janjic (BMJ) convection and one with Eta-Kain-Fritsch convection (cf. Black, 1994). While more configurations were certainly possible, especially with MM5 and WRF, the total number was limited by computing power and the time frame within which results were needed.

The models were run out to 12 hours with one-way nesting at 36 and 12 km horizontal resolutions. Because neither the workstation Eta nor the WRF currently has the capability to run simultaneous nests, the 12 km simulations for these models were run using a sequential nesting technique. This technique required the 36 km output to be saved at very high temporal resolution for boundary conditions to the 12 km domain. The 36 km domain covered much of the continental US, while the 12 km domain covered a section of the Great Plains (see Fig. 1). Initial and boundary conditions for the 36 km domain were obtained from the NCEP Eta model.

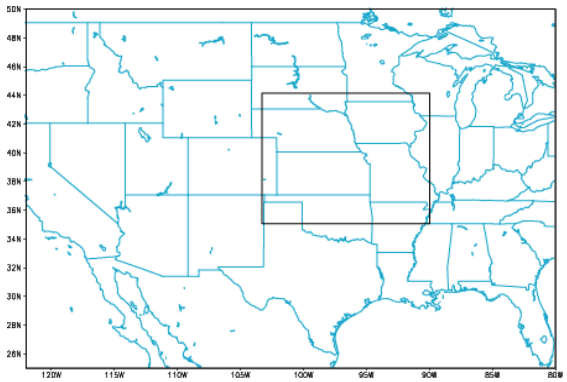


*FIG. 1. Domains used for 36 km (large rectangle) and 12 km (small rectangle) resolution simulations.*

The simulations were performed on six Dell dual processor PCs. The processor speeds ranged from 2.4-3.08 GHz. Not all the configurations ran successfully for all the dates. Not all simulations were completed within the needed time frame. The MM5 completed all configurations and all dates successfully, while the WRF had problems with about half of the 48 configurations. The ARPS model ran so slowly that only one configuration was completed in enough time to be evaluated. The reason for the slowness was linked to the fact that the model could not be configured to run stably with a time-step greater than 20 sec – even with a

*Corresponding Author Address: Dr. Peter J. Sousounis, WSI Corporation, 400 Minuteman Road, Andover, MA 01810. Email: psousounis@wsi.com.

36 km horizontal grid point separation. One configuration was completed but the others were abandoned owing to the realization that it would take about 200 hours (> 1 week) of computing time to complete one configuration. In contrast, the MM5 ran very stably at 36 km with a time-step of 90 sec. Each MM5 configuration took about 10 hours to complete all the 36 km simulations for the 39 dates.

*b. verification techniques*

A standard verification technique of equitable threat score (ETS) was used to evaluate model performance. The ETS is defined as:

$$ETS = (a-e)/(a+b+c-e) \quad (1)$$

where a = the number of grid points with correct "yes" forecasts,
b = the number of grid points where the phenomenon was forecast but not observed,
c = the number of grid points where the phenomenon was observed but not forecast,
e = the number of grid points with a correct forecast that would be expected from random chance.

More information on ETS is available in Wilks (1995) and Colle et al. (2000). Because of the emphasis on precipitation with respect to the research effort at WSI, attention was paid not only to 12 hourly accumulated precipitation totals, but also to the hourly forecasted amounts (e.g., 0-1 h, 1-2 h, etc.) Towards that end, while a traditional measure of skill such as ETS may be appropriate for evaluating 0-12 h amounts, it may not be so appropriate for evaluating hourly amounts because slight errors in timing of precipitation (e.g., too early or too late by 1-2 hours) would be reflected in dramatically lower ETSs. Additionally, it is a well known consequence that high resolution precipitation output can appear worse than low resolution output from a verification standpoint because of added spatial structure (Mass et al. 2002). Thus, there was motivation to develop a verification technique that not only evaluated the spatial characteristics of a precipitation forecast, but also its temporal and intensity aspects.

A technique referred to acuity-fidelity was developed and implemented to quantify the skill of a forecast using the three metrics of space, time, and intensity. Acuity represents the model's skill at detecting the features of the observed data. The acuity of a forecast is calculated for each observed data point by finding the best matching forecast for that observation. Instead of automatically associating an observation with the forecast that shares its location and time, the best match is obtained by minimizing a cost function calculated between the target observation and many candidate forecast data. The candidate forecast datum that produces the smallest penalty is deemed the best match, and is therefore associated with the observation. Fidelity represents the faithfulness of the model's predictions to the observed data. The fidelity of a forecast is calculated much like the acuity, except the roles of the observations and forecasts are reversed. Thus for each target forecast datum, the best matching observation is found within a multidimensional field of candidate observations.

The cost function is defined in terms of four components: one each for errors in distance, time, and intensity, and a fourth term to account for missed events.

$$J = J_s + J_t + J_i + J_e \quad (2)$$

In this study, intensity refers to one hour accumulated precipitation but in general it could be any dependent variable. To calculate a total acuity or fidelity penalty, all the cost function components must be converted into common units. Here the choice was made to convert the time, intensity and event penalties into equivalent distances using the following component definitions:

$$J_s = \Delta x \quad (3)$$
$$J_t = U_e \, \Delta t \quad (4)$$
$$J_i = D_I \, \Delta I \quad (5)$$
$$J_e = f(J_{miss}, \text{Intensity regimes}) \quad (6)$$

In these equations, the variables $\Delta x$, $\Delta t$, and $\Delta I$ represent the absolute differences in position, time, and intensity, respectively, between an observed datum and a forecast datum. The coefficient $U_e$ is the characteristic event velocity used to relate temporal and spatial errors. The coefficient $D_I$ is the distance-intensity ratio used to relate intensity and spatial errors. $J_{miss}$ is the maximum value of J, and represents the penalty associated with a complete miss. The intensity regimes are a list of intensity values that define categories within the intensity continuum. More information on acuity-fidelity is available in Marshall et al. (2004).

## 3. RESULTS

The precipitation output from each of the simulations was verified against Stage 4 hourly precipitation amounts (Baldwin and Mitchell, 1997).

*a. equitable threat score analysis*

The ETSs were computed for 00-12 h accumulated precipitation over an area bounded by 35N-44N and 103W-90W, which corresponded very nearly to the domain used for the 12 km simulations (see Fig. 1) for each model configuration run for all 39 dates. Table 1 summarizes the results for the configurations evaluated at 36 km at five different precipitation thresholds as indicated. The summary values indicated in Table 1 for each model at each threshold were obtained as lumped values. That is, the a, b, c, and e quantities in (1) were summed over all 39 dates and then the ETS was computed. As a result, the simulations from those dates on which heavy precipitation was either observed and/or forecasted had a correspondingly greater impact. Additionally. because MM5 and WRF were run at many different configurations, the values shown in Table 1 for these models are those from the best configuration,

| *model* | *p > 0.01* | *p > 0.05* | *p > 0.25* | *p > 0.50* | *p > 1.00* |
|---|---|---|---|---|---|
| ARPS | 0.287 | 0.287 | 0.238 | 0.188 | 0.130 |
| MM5 | 0.387 | 0.404 | 0.342 | 0.244 | 0.146 |
| WRF | 0.360 | 0.385 | 0.347 | 0.247 | 0.129 |
| WSETA | 0.374 | 0.360 | 0.301 | 0.219 | 0.104 |
| NCEPETA | 0.352 | 0.371 | 0.307 | 0.196 | 0.098 |
| NCEPRUC | 0.357 | 0.367 | 0.275 | 0.167 | 0.040 |

*Table 1. Mean values of equitable threat score for 00-12 h accumulated precipitation at five different precipitation thresholds (inches) for the four models evaluated as well as for two NCEP models.*

defined as that with the highest ETS averaged over all five thresholds.

A comparison of the results at 36 km suggests that the best MM5 configuration had a statistically significant higher ETS than the best WRF configuration for all but one precipitation threshold. The ETS scores from both models were higher than those from either the workstation Eta or ARPS. One characteristic result from the ARPS simulations, which explains partially its poorer performance, was that on some situations it performed dramatically worse than the other models, although on about half of the simulated periods it performed better than either MM5 or WRF. Note also that both the best WRF and MM5 configurations typically yielded higher ETSs than either the NCEP-RUC or NCEP-Eta at all thresholds.

A breakdown of ETSs by parameterization option revealed that for the 36 km MM5 and WRF configurations evaluated, certain parameterization options performed consistently better over certain ranges of precipitation thresholds. For example, in MM5, the best options for convection, microphysics, boundary layer, and surface layer when considering all precipitation amounts (p > 0.01 in) were GRELL, SCHULTZ, BLACKADAR, and 5-LAYER, respectively. For heavy precipitation (p > 1.00 in), the best options for convection, microphysics, boundary layer, and surface layer in MM5 were KUO, SHULTZ, BURKE-THOMPSON, and NONE, respectively. Both MM5 and WRF illustrated distinct crossover points where one option performed better at a lower threshold and a different option performed better at a higher threshold. Interestingly, for both the MM5 and WRF models and for all thresholds, the configurations that contained the set of options that yielded the highest ETSs by parameterization category were indeed the best configurations, a result that highlights the linear benefits provided by each parameterization option.

While ETS may be a reliable indicator of skill when evaluating models with coarse resolution on 12 hour time scales, Fig. 2 presents three examples where the reliability breaks down when evaluating model performance at higher spatial or temporal resolution. Example 1 in the top row of Fig. 2 shows 00-12 h forecast precipitation accumulation valid at 00 UTC 05 May 2003 from a particular WRF configuration from the 36 and 12 km domains. Note that the 12 km version (right panel) appears to be the more accurate but that from a ETS perspective the 36 km version (middle panel) has a higher score. Three features support this subjective assessment: (1) the precipitation over central Nebraska is forecast better in the right panel than in the middle panel; (2) the precipitation in northeast Iowa is forecast better in the right panel than in the middle panel; and (3) the bands of heavy precipitation across southern Missouri are forecast better in the right panel than in the middle panel. Example 2 in the middle row of Fig. 2 shows 00-12 h forecast precipitation accumulation valid at 00 UTC 07 May 2003 from the 12 km domain from two WRF configurations that differ only in terms of cumulus parameterization. Subjectively, note that the precipitation forecast from the configuration with the Betts-Miller-Janjic option (right panel) appears better than the one with the Eta-Kain-Fritsch option (middle panel) although the ETS for the Eta-Kain-Fritsch simulation is higher than that from the Betts-Miller-Janjic simulation. Specifically, the orientations of bands of precipitation are better represented in the right panel than in the left panel. Finally, Example 3 in the bottom row of Fig. 2 shows 11-12 h forecast precipitation accumulation valid at 00 UTC 05 May 2003 from the 12 km domain from two different WRF configurations. Once again, the configuration that appears to provide a better forecast (right panel) is the one with the lower ETS. In this example, note that the precipitation is over a one hour period (11-12hr) so not only are the precipitation amounts lower than in the first two examples but visual agreement between (both) forecasts and the observed amount is poorer and hence the ETSs are much lower. Still, from a subjective standpoint the forecast in the right panel shows precipitation areas that have better orientation in eastern Nebraska. Also, the precipitation in Missouri is at least of comparable intensity. The three examples in Fig. 2 highlight the inappropriateness of using ETS (or other traditional measures of forecast accuracy) to evaluate high resolution output.

*b. acuity-fidelity analysis*

Acuity-fidelity was performed on model output from the MM5 and WRF configurations at 36 km and on output from the WRF configurations at 12 km at 03, 06, 09, and 12 h. Values for the coefficients $U_e$ and $D_i$ were set at 10 $ms^{-1}$ and 20 $km(mm/hr)^{-1}$ respectively. The combined acuity-fidelity scores (acuity+fidelity) are shown in Table 2 for a threshold of 0.25 in/hr (6.35 mm/h). Again, for MM5 and WRF the results from the best configurations are shown. A comparison of MM5 and WRF results at 36 km demonstrates the slightly better performance of WRF over MM5. The relative performance of the other models is the same from an acuity-fidelity perspective or from an ETS perspective. Despite the fact that the ETSs in Table 1 and the acuity-fidelity values in Table 2 are different metrics and that they are computed over different time-slices of the 00-12 hr forecasts, it is still curious that MM5 performs better with respect to one metric and WRF performs

better with respect to the other metric. One possible explanation is that the WRF AF scores are associated with a higher percentage contribution from best matches at different times, which suggests that it does better with respect to the timing of precipitation on an hour-by-hour basis. The timing of precipitation is a feature that ETS can not capture.

| *model* | *02-03 hr* | *05-06 hr* | *08-09 hr* | *11-12 hr* | *avg* |
|---|---|---|---|---|---|
| ARPS | - | - | - | - | - |
| MM5 | 163 | 176 | 166 | 175 | 170 |
| WRF | 135 | 146 | 148 | 160 | 147 |
| WSETA | 172 | 196 | 188 | 169 | 181 |
| NCEPETA | 185 | 196 | 193 | 177 | 188 |
| NCEPRUC | 202 | 213 | 206 | 205 | 206 |

*Table 2. Values of acuity+fidelity for a precipitation intensity threshold of 0.25 in/hr at four different forecast times for three of the four models evaluated at 36 km resolution as well as for two NCEP models averaged over the 39 cases.*

A breakdown of acuity-fidelity contributions in fact reveals that timing errors account for 50 km of error in MM5, and only 30 km in WRF. The difference in timing errors almost explains the difference between the two average values.

A comparison of WRF results from the 36 km and 12 km domains in Table 3 illustrates how the 12 km results are more accurate from an A-F perspective, as one may hope. A primary reason for the lower numbers (better scores) from the 12 km domain is that the simulation is not penalized so severely for spatial errors as is the case with other metrics including root mean square error and threat score. While the 36 km results may still be penalized even less because of broad-brushing precipitation in areas where it is not observed, the 36 km results are penalized more than the 12 km results for intensity errors.

| *model* | *02-03 hr* | *05-06 hr* | *08-09 hr* | *11-12 hr* | *avg* |
|---|---|---|---|---|---|
| WRF-36 | 135 | 146 | 148 | 160 | 147 |
| WRF-12 | 202 | 213 | 206 | 205 | 206 |

*Table 3. Values of acuity+fidelity for a precipitation intensity threshold of 0.25 in/hr averaged over the 39 cases for WRF-36 and WRF-12.*

The acuity-fidelity scores also correlate more closely with subjective evaluation of which model configuration performs better in certain situations. Returning again to the three examples shown in Fig. 2, Table 4 indicates how the acuity-fidelity scores compare with the ETSs for the selected situations.

| *metric* | *ETS col 2* | *ETS col 3* | *A-F col 2* | *A-F col 3* |
|---|---|---|---|---|
| ROW 1 | 0.268 | 0.229 | 151 | 107 |
| ROW 2 | 0.332 | 0.280 | 174 | 107 |
| ROW 3 | 0.086 | 0.049 | 080 | 066 |

*Table 4. Values of ETS and acuity+fidelity for the three examples shown in Fig. 2. Second and third columns correspond to ETSs in middle and right columns of Fig. 2 respectively. Fourth and fifth columns correspond to mean acuity-fidelity scores (km) for middle and right columns of Fig. 2 respectively.*

Note that in each of the three examples, the better acuity-fidelity scores correspond to the right panels in Fig. 2 that were chosen subjectively as the better forecasts. Note also that in the first two examples the acuity-fidelity scores are weighted means from the 1-hr precipitation accumulation forecasts for the four different forecast hours evaluated (cf. Tables 2 and 3). While the better agreement from acuity-fidelity in Example 1 may be intuitive and anticipated because the metric does not unfairly penalize a forecast for spatially displaced areas of precipitation, the advantage of the metric from a temporal perspective can not be illustrated from the particular panels shown in Fig. 2 because they are effectively snapshots at selected times. The real advantage from acuity–fidelity is that it accounts for potentially better matches at adjacent times – before or after the snapshot time. Another difference between the ETS and acuity-fidelity scores is that errors in intensity count in acuity-fidelity, even when considering events within a particular threshold, whereas with ETS, all precipitation amounts above a particular threshold contribute equally to the score.

## 4. SUMMARY

A suite of existing numerical weather prediction models was evaluated recently at WSI Corporation for the purpose of identifying the most skillful models and configurations for forecasting heavy precipitation. Four different models were evaluated but by far the greatest number of model configurations was provided by MM5 and WRF. An approximate total of 80 different model configurations was used to make 36 km resolution - 12 h forecasts over the continental US for 39 different times. A subset of the 80 configurations was run at 12 km resolution over the Great Plains. In all, nearly 4000 12 h forecasts were performed and verified over a 3-month period.

Verification results of the 36 km domains using a traditional equitable threat score analysis show that MM5 had a statistically significant edge over WRF at most thresholds evaluated. Both were more skillful than the workstation Eta and ARPS models that were used, and both were more skillful that the NCEP ETA and RUC models for the dates that were examined.

A new technique called acuity-fidelity (A-F) was developed to evaluate model performance – taking into account model timing and intensity errors. The A-F

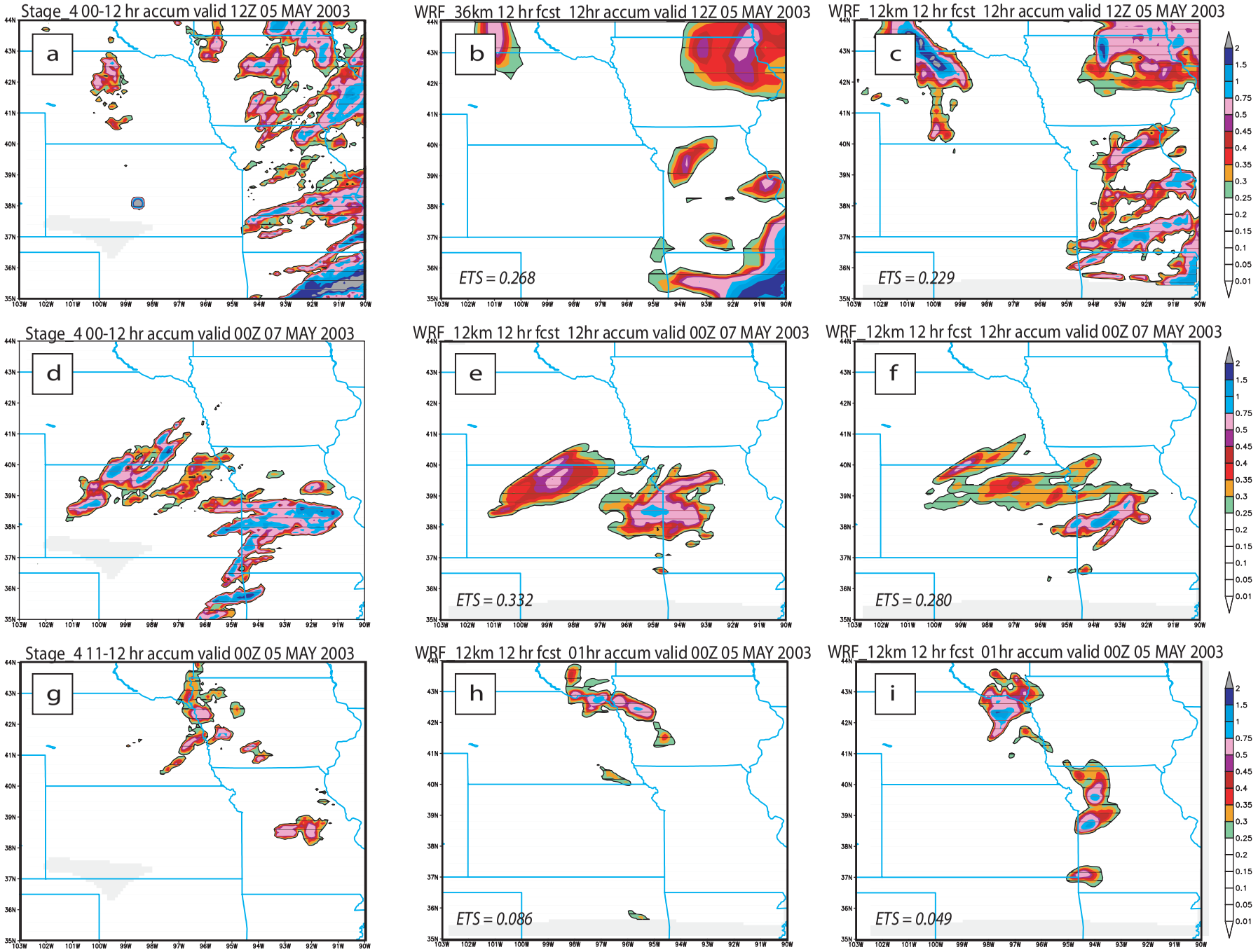


*Fig. 2. Precipitation forecasts from selected model configurations examined in the study. Top row: 00-12 hr accumulation valid at time shown for a) Stage 4 analysis, b) portion of 36 km domain from a particular WRF configuration, c) 12 km domain for same WRF configuration. Middle Row: 00-12 hr accumulation valid at time shown for d) Stage 4 analysis, e) 12 km domain from a particular WRF configuration, f) 12 km domain from a similar WRF configuration differing only in convective parameterization. Bottom Row: 11-12 hr accumulation for g) Stage 4 analysis, h) 12 km domain from a particular WRF configuration, i) 12 km domain from a different WRF configuration. Precipitation amounts in inches as indicated. Amounts less than 0.25 inches omitted. Equitable threat scores correspond to precipitation amounts greater than 0.25 inch threshold.*

results showed that WRF had more skill than MM5 at 36 km. It also showed that the WRF 12 km results had more skill than the WRF results at 36 km – a result that did not exist when using ETS analysis.

## REFERENCES


Baldwin, M. E., and K. E. Mitchell, 1997: The NCEP hourly multi-sensor U.S. precipitation analysis for operations and GCIP research. Preprints, *13th Conf. on Hydrology,* Long Beach, CA, Amer. Meteor. Soc., 54-55.

Black, T. L. 1994: The new NMC mesoscale Eta model: Description and forecast examples. *Wea. Forecast*. **9**, 265–284.

Colle, Brian A., Mass, Clifford F., Westrick, Kenneth J. 2000: MM5 Precipitation Verification over the Pacific Northwest during the 1997–99 Cool Seasons. *Wea. Forecast.* **15**, 730–744.

Grell, G. A., J. Dudhia and D. R. Stauffer, 1993: A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). NCAR Technical Note, NCAR/TN-398+ STR, 117 pp.[Available from NCAR Information Services, P.O. Box 3000, Boulder, CO 80307.]

Listemaa, S. A., 2002: Workstation ETA verification efforts at the lower Mississippi River Forecast Center. Preprints, *16th Conf. on Probability and Statistics,* Orlando, FL, Amer. Meteor. Soc., 54-56.

Marshall, S. F., P. J. Sousounis, and T. A. Hutchinson, 2004: verifying mesoscale model precipitation forecasts using an acuity-fidelity approach. *Preprints 20th Conference on Weather Analysis and Forecasting*, Seattle, Amer. Meteor. Soc., J13.3.

Mass, C. F., D. Ovens, K. J. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce better forecasts? The results of two years of real-time numerical weather prediction in the Pacific Northwest. *Bull. Amer. Meteor. Soc.*, **83**, 407-430.

Michalakes, J., S. Chen, J. Dudhia, L. Hart, J. Klemp, J. Middlecoff, and W. Skamarock, 2001: Development of a Next Generation Regional Weather Research and Forecast Model. *Developments in Teracomputing*: Proceedings of the Ninth ECMWF Workshop on the Use of High Performance Computing in Meteorology. Eds. Walter Zwieflhofer and Norbert Kreitz. World Scientific, Singapore. 269-276.

Skamarock, W. C., J. B. Klemp, and J. Dudhia, 2001: Prototypes for the WRF (Weather Research and Forecasting) Model. *Preprints, Ninth Conf. on Mesoscale Processes, Fort Lauderdale, FL, Amer. Meteor. Soc.,* J11-J15.

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

Xue, M., K. K. Droegemeier, and V. Wong, 2000: The Advanced Regional Prediction System (ARPS) - A multi-scale nonhydrostatic atmospheric simulation and prediction model. Part I: Model dynamics and verification. Meteorol. Atmos. Phys. **75**, 161-193.