J13.3 VERIFYING MESOSCALE MODEL PRECIPITATION FORECASTS USING AN ACUITY-FIDELITY APPROACH

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

WSI Corporation
Andover, MA

1. INTRODUCTION

Precipitation has been a target of verification for mesoscale numerical model forecasts since the early days of numerical weather prediction. For daily or longer time periods, a mean square error or threat score approach may be suitable. However, for verifying forecasts of hourly precipitation, particularly from models with very fine grid spacing, such traditional approaches can unfairly penalize one model while unfairly rewarding another.

A classic example is the decrease in skill of a model's precipitation forecast as its grid spacing is decreased. To the human eye, simulations on a finer mesh often appear to have better representation of mesoscale features than coarser simulations. However, traditional metrics often rank such fine mesh simulations as inferior to coarser runs from the same model (Mass et al. 2002). Subjectively, this effect appears to be caused by improper location or timing of the mesoscale features in the fine mesh model. The fine mesh model is penalized heavily both for a lack of precipitation at the locations and times where it was observed and for having too much precipitation at nearby times and locations where precipitation was not observed. By contrast a coarser forecast tends to predict smoother fields, e.g. weaker precipitation over greater areas, potentially leading to smaller point-by-point penalties and hence a superior skill score.

Recently, a verification strategy has been developed at WSI Corporation to more fairly evaluate forecasts with fine grid spacings and short temporal frequencies. This method differs from traditional techniques in the way it associates forecast and observational data to form forecast-observation pairs. In this scheme, the skill of the forecast is measured by two metrics called acuity and fidelity.

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.

*Corresponding Author Address: Stephen F. Marshall, WSI Corporation, 400 Minuteman Road, Andover, MA 01810. Email: smarshall@wsi.com.

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.

In this paper, we develop a cost function for verifying precipitation forecasts using the acuity-fidelity method and explore the sensitivities of this cost function's parameters. The validity of the verification scheme is demonstrated by visually comparing precipitation output from different models with corresponding acuity-fidelity scores. The utility of the acuity-fidelity technique is demonstrated by comparing its results to those from a more traditional threat score approach.

2. METHODOLOGY

To assess precipitation forecasts, a cost function was defined with four components: one each for errors in distance, time, and intensity, and a fourth term to account for missed events.

$$J = J_s + J_t + J_i + J_e \quad (1)$$

In this study, intensity refers to one-hour accumulated precipitation; 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. We chose to convert the time, intensity and event penalties into equivalent distances using the following component definitions:

$$J_s = \Delta x \quad (2)$$
$$J_t = U_e \, \Delta t \quad (3)$$
$$J_i = DI \, \Delta I \quad (4)$$
$$J_e = f(J_{miss}, \text{Intensity regimes}) \quad (5)$$

In these equations, the variables Δx, Δt, and ΔI represent the absolute difference in position, time, and intensity, respectively, between an observed datum and a forecast datum. U_e is the characteristic event velocity used to relate temporal and spatial errors. DI is the distance-intensity ratio used to relate intensity and spatial errors. J_{miss} is the maximum value of J, and represents the worst possible penalty; the minimum penalty is 0. The intensity regimes are a list of intensity values that define categories within the intensity continuum.

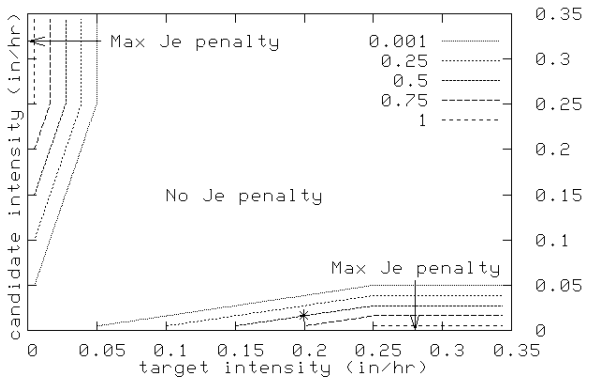

Fig. 1 - Ratio J_e/J_{miss} for intensity regimes defined by thresholds at 0.005, 0.05, and 0.25 in/hr.

The event term J_e is calculated by determining the intensity regimes for a forecast-observation pair. The intensity regime is a real number that represents a position within the defined ranges, e.g. a value midway between the first and second thresholds would have an intensity regime value of 1.5. J_e is set to 0 if the difference between the forecast and observed regimes is less than 1. J_e is set to J_{miss} if the pair differs by 2 or more intensity regimes. If the forecast and observation differ between 1 and 2 intensity regimes, J_e is calculated as a linear interpolation between 0 and J_{miss}. Figure 1 shows the ratio J_e/J_{miss} as a function of target and candidate intensities for regimes defined by thresholds at 0.005, 0.05, and 0.25 in/hr. For example, a target intensity of 0.2 in/hr and a candidate intensity of 0.016 in/hr (indicated on the plot by an asterisk) results in J_e = 0.5 * J_{miss}.

The intensity regimes are designed to represent important distinctions between data that have very similar intensity values, e.g. to discourage matches between effectively non-precipitating observations and lightly precipitating forecasts. This effect could also be represented using a non-linear definition for J_i that depends on the magnitudes of the intensity values, rather than just their differences. This effect was extracted into a separate event term to allow more insight into the matching algorithm. J_e values > 0 indicate some degree of disassociation between the target datum and its candidate field.

3. SENSITIVITY STUDY

Since the acuity-fidelity technique relies upon several configurable parameters, it is prudent to test the sensitivity of the results to their assumed values. Such a sensitivity study was performed on 3, 6, 9, and 12-hour forecasts of hourly precipitation from 39 cases of precipitating weather observed in the central U.S. during April and May of 2003. Results are presented for 4 model configurations, although many more were evaluated (See Sousounis et al. 2004 for further details).

The sensitivity study includes forecast data from two NCEP models, Eta and the Rapid Update Cycle model (RUC), as well as two different grid spacing configurations of the Weather Research and Forecasting model (WRF) (Skamarock 2001). The NCEP Stage IV precipitation analyses (Baldwin and Mitchell 1997) are used as the observation. The Stage IV data are available on 4.8 km grids with hourly temporal frequency, while the Eta and RUC are available on 40 and 20 km grids, respectively, and have three-hour temporal frequency. The WRF simulations were run on 12 and 36 km grids with an output frequency of twelve minutes. The WRF simulations were initialized using Eta analyses, and were run with the following parameterizations: Kain-Fritsch cumulus, OSU land-surface model, MRF planetary boundary layer, Kessler microphysics, and RRTM radiation.

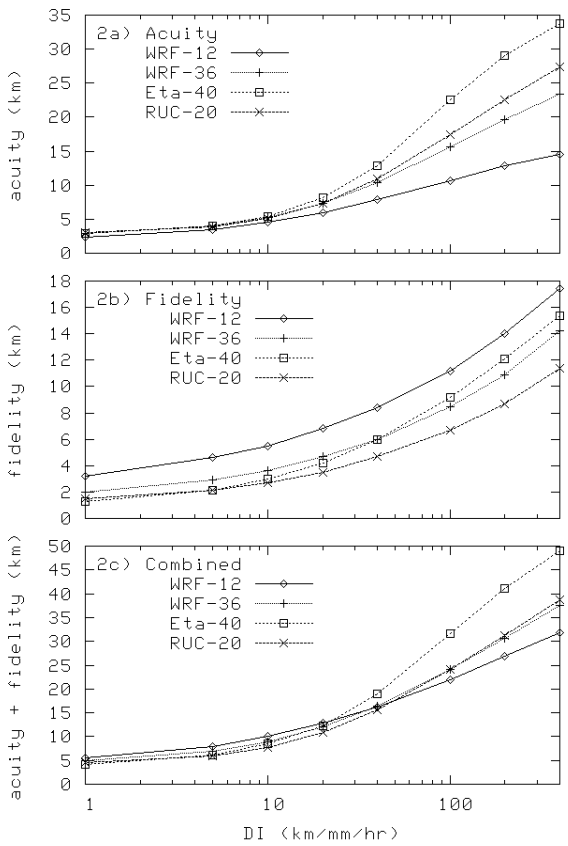

Fig. 2 - a) acuity, b) fidelity, and c) combined score in km for 4 models as a function of DI (km/mm/hr). Each point is an average over 156 times (39 initialization times and 4 forecast periods) and a variable number of grid point locations, depending on the model.

Acuity and fidelity were calculated for each observation and forecast datum within a region defined by a latitude-longitude box from 36.5° to 44° N and 103° to 92° W. However, the best matches were not constrained to fall within this region. The verification region contained 43,080 grid points from the Stage IV data, 487 from Eta-40, 1,973 from RUC-20, 546 from WRF-36, and 4,938 from WRF-12.

The base values for the cost function parameters were set to these values: J_{miss} = 1000 km, U_e = 10 m/s,

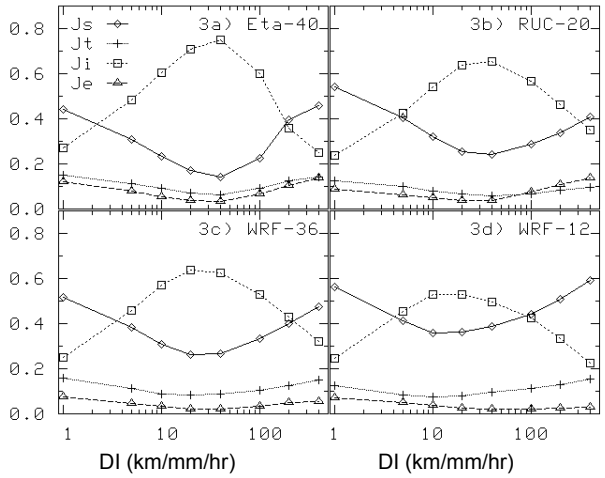

Fig. 3 – Mean relative contributions of the acuity components J_s, J_t, J_i, and J_e as a function of DI in km/mm/hr for a) Eta 40 km, b) RUC 20 km, c) WRF 36 km, d) WRF 12 km.

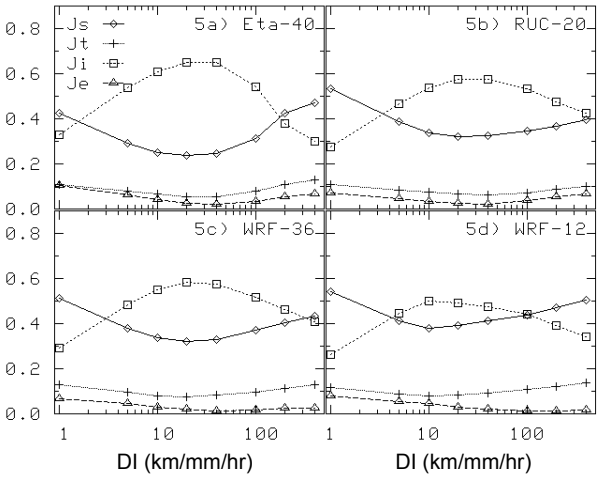

Fig. 5 - Mean relative contributions of the cost function components averaged between acuity and fidelity as a function of DI in km/mm/hr for the same models as Fig. 3

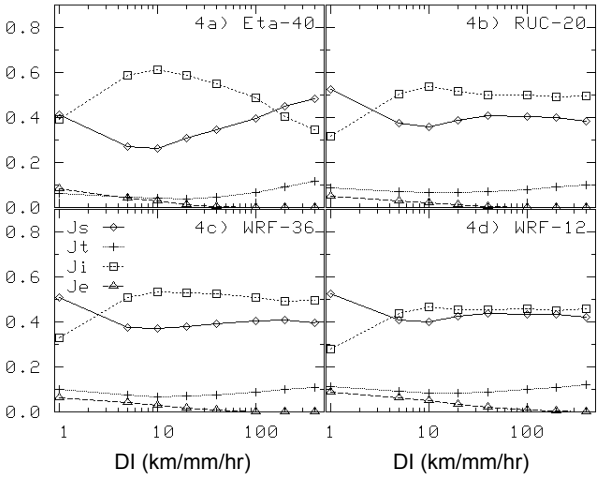

Fig. 4 – Mean relative contributions of the fidelity components J_s, J_t, J_i, and J_e as a function of DI in km/mm/hr for the same models as Fig. 3.

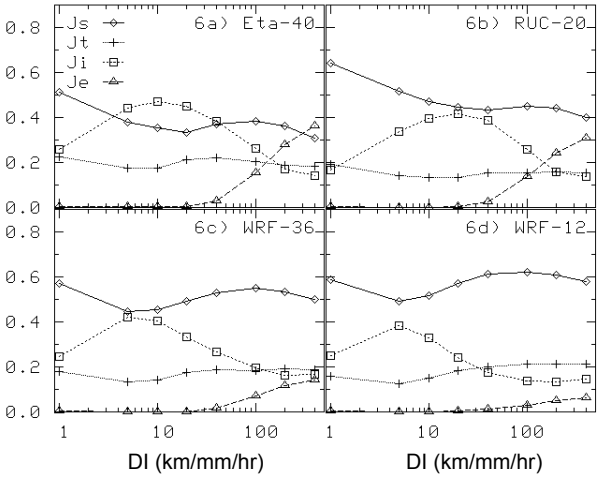

Fig. 6 – Stratified mean relative contributions of the cost function components averaged between acuity and fidelity as a function of DI for same models as Fig. 3. Values are stratified for target data with intensity ≥ 0.25 in/hr.

DI = 20 km/mm/hr, and intensity thresholds at 0.005, 0.05, and 0.25 in/hr. We justified the base values for DI and U_e by examining the effects of their variation, and the results of that study are presented in the paragraphs that follow. Variations in J_{miss} and the intensity regimes were not examined in detail, but their effects are most likely small, as they mainly affect the calculation of J_e. J_e generally has a small effect on the total cost unless the other parameters have extreme values, as will be seen in the sensitivity calculations for DI and U_e.

Figure 2 shows the changes in the mean acuity, mean fidelity, and their sum, called the combined score, as a function of DI. As expected, the cost function values for all models increase with DI. The WRF-12 has the best acuity scores and worst fidelity scores. The small acuity penalty represents a relatively high skill at predicting observed features, while the large fidelity penalty indicates the tendency of the fine mesh model to forecast features at the wrong place or time.

Figure 2 also shows that acuity is larger than fidelity for all values of DI. This is because acuity is dominated by error associated with the observations of intense precipitation, while fidelity can avoid these observations during its search for a best match. A graphical example of this can be seen in Fig. 9. Another way to think of the difference between acuity and fidelity is that acuity contains contributions from all the observations, but only a subset of the forecasts, while fidelity contains information from all the forecasts, but only a subset of the observations.

Figure 3 shows the mean relative contribution of each component of acuity for each model as a function of DI. All models show a similar trend, with a maximum percentage of the cost associated with intensity at

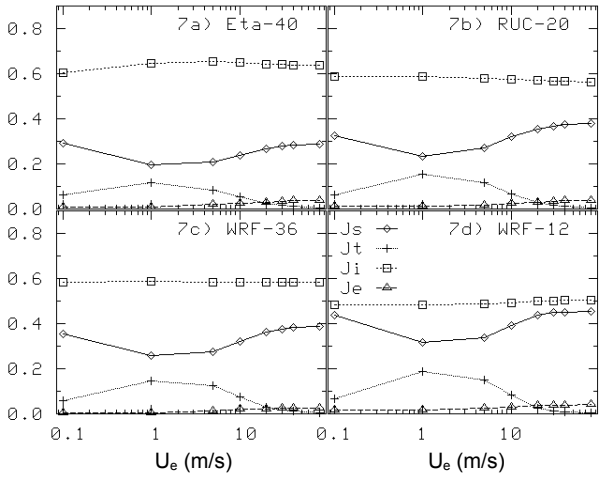

Fig. 7 - Mean relative contributions of the cost function components averaged between acuity and fidelity as a function of U_e (m/s) for the same models as in Fig. 3.

moderate values of DI and minima in the J_i contribution at the extreme values of DI. This pattern is caused by two competing effects. At small values of DI, the intensity penalty is so small that even large intensity differences contribute very little to the cost function. At large values of DI, the intensity errors are so strongly penalized that the best match is likely to be a datum with a very similar intensity to the target, even if it is separated widely from the target datum in space and time.

Figure 4 contains plots similar to Fig. 3, except for fidelity. The fidelity contributions look similar to acuity at small values of DI, but then become relatively insensitive at larger DI. The most notable effect at large DI is the decline in the importance of J_e to effectively 0.

The DI-sensitivities can be used to justify the choice of a DI value for future verifications. For that purpose, one may seek a value that minimizes the contribution of J_e, because J_e represents an inability to associate forecast and observed data, and it is desirable to find all reasonable associations between the two. Minimizing the J_e contribution to fidelity favors large values of DI, while minimizing the J_e contribution to acuity favors more moderate DI values.

To address this difference in the location of the minimum J_e contribution, an average of the acuity and fidelity mean relative contributions are shown in Fig. 5. A stratified version of this plot is shown in Fig. 6, where the acuity contribution includes only the observations where intensity $\geq$ 0.25 in/hr and the fidelity contribution includes only the forecasts where intensity $\geq$ 0.25 in/hr. The non-stratified plot shows a minimum J_e contribution at DI = 40 for all models except WRF-12, where DI = 100. However, the stratified plot shows a different and more consistent pattern, with the J_e contribution essentially 0 for DI $\leq$ 20, then increasing with DI. This represents the difficulty of making associations between forecasts and observations with large intensity values when DI $\geq$ 40. Since these moderate-to-heavy precipitation cases are typically of the greatest interest, we chose DI = 20 to minimize the J_e contribution in the stratified case, while still preserving a small J_e contribution in the non-stratified case.

For the U_e sensitivity study, the cost function component breakdown averaged over acuity and fidelity is shown in Fig. 7. In the U_e study, acuity and fidelity both exhibited similar patterns, and stratification on intensity had a much weaker effect than in the DI study, so for brevity, these plots are not shown.

The main feature of the U_e sensitivity is that J_t only contributes significantly to the cost function for moderate values of U_e. As U_e increases, J_t decreases mostly at the expense of increasing J_s, indicating that large temporal penalties result in a wider spatial search. At all U_e, J_e is negligible, making it difficult to use minimization of the J_e contribution as a criterion for picking the best U_e value. Instead, we focused on picking a U_e value that would allow temporal searching to play a significant role in the verification, i.e. $1 \leq U_e \leq 10$ m/s. We chose 10 m/s because it is the typical scale of horizontal atmospheric motions. However, a lower value would be more appropriate if one were interested in maximizing the potential for temporal matching.

4. COMPARISON TO SUBJECTIVE VERIFICATION

The acuity-fidelity technique is designed to measure the skill noticed in subjective verification of fine mesh forecasts that have not been captured in traditional skill metrics. To show that acuity-fidelity is consistent with subjective verification, we present in Fig. 8 the graphical data needed to subjectively assess several nine-hour forecasts of one-hour accumulated precipitation. The subjective assessment will be compared to corresponding acuity-fidelity metrics. Note that the authors have compared numerous subjective verifications to acuity-fidelity metrics, but for brevity, we present only one example. Readers are warned that this example is not necessarily representative of other cases we have studied.

The precipitation analysis and forecasts shown in Fig. 8 are valid at 21 UTC on May 4, 2003; data from the preceding and subsequent hours are presented to the left and right of the target data to give a temporal context. The top row shows the Stage IV analysis, which is taken as truth. The next 4 rows show the forecasts from WRF 12 km, WRF 36 km, Eta 40 km, and RUC 20 km, respectively.

Note that the Eta and RUC one-hour precipitation accumulations were derived from three-hour accumulations through simple division. A more fair assessment of RUC should include use of the one-hour precipitation accumulations that are available every three hours. This was not done because acuity-fidelity requires the candidate data set to be a time series with equal time increments and no temporal gaps.

It is also of note that the WRF accumulations were calculated from 12-minute output, which allowed more

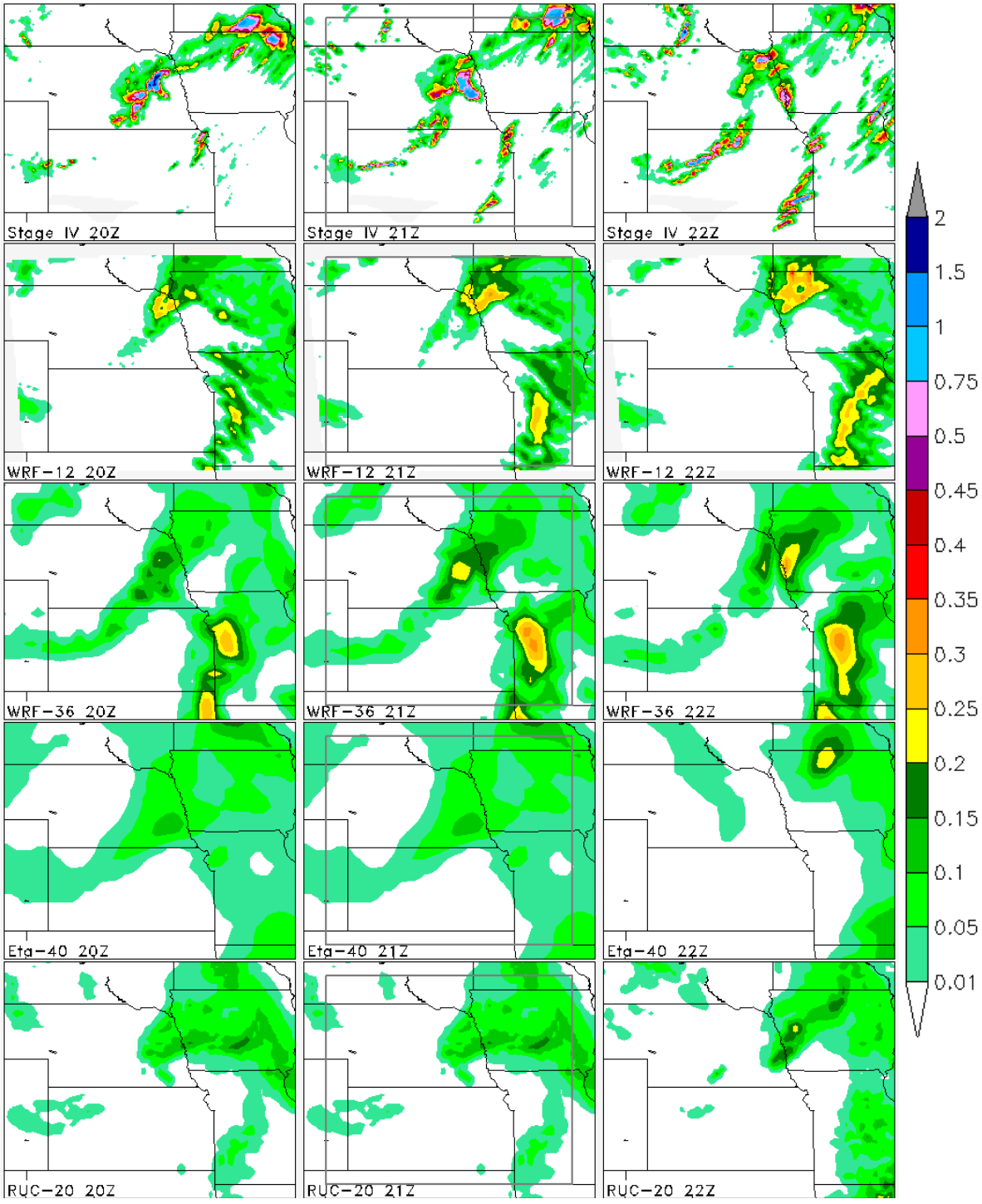

Fig. 8 – One hour accumulated precipitation in inches for periods ending 20, 21, and 22 UTC on May 4, 2003. The data sources are (from top row) Stage IV, WRF-12, WRF-36, Eta-40, and RUC-20, respectively. All forecast models were initialized on May 4 at 12 UTC. The gray boxes on the plots at 21 UTC indicate the region over which verification metrics were calculated (36.5° to 44° N and 103° to 92° W).

temporal precision when searching the WRF forecasts for a best match to observations during acuity calculations. There were four temporally-overlapping, one-hour precipitation accumulations available between each hour, e.g. for the periods ending at 20:12, 20:24, 20:36, and 20:48 UTC. For brevity, these are not shown, but they were part of the candidate field used during acuity calculations for the WRF simulations.

Model	Threshold	Acuity	Fidelity	Combined
Eta-40	0	21	10	31
RUC-20		19	**8**	**27**
WRF-12		**16**	12	28
WRF-36		20	14	35
Eta-40	0.005	58	**17**	**75**
RUC-20		63	21	83
WRF-12		57	35	92
WRF-36		**53**	29	83
Eta-40	0.05	101	**27**	128
RUC-20		108	34	142
WRF-12		94	47	142
WRF-36		**79**	42	**121**
Eta-40	0.25	233	*N/A*	*N/A*
RUC-20		239	*N/A*	*N/A*
WRF-12		206	69	275
WRF-36		**183**	**67**	**249**

Table 1- Mean acuity, fidelity, and combined score in km for the one hour precipitation forecasts shown in Fig. 8. Values are shown for 4 stratification thresholds of the target intensity: 0 (unstratified), 0.005, 0.05, and 0.25 in/hr. The best (minimum) scores for each category are bolded.

It is evident in Fig. 8 that all the forecast models failed to predict precipitation rates ≥ 0.5 in/hr. This is expected, because the models have from 3 to 10 times the grid spacing of the Stage IV analysis, and hence cannot typically generate the most intense precipitation that was observed. However, it is still instructive to be able to assess how close these models came to reproducing such observed high resolution features, even if they did not fully succeed. This approach considers coarseness of grid spacing or temporal frequency as a potential source of error, and thus allows an assessment of the effects on forecast skill of decreasing a model's grid spacing or increasing its temporal output frequency. In order to isolate scale issues from the analysis, all data would need to be filtered to common spatial and temporal scales before the acuity-fidelity technique is applied. Such a scale-neutral analysis is left as future work.

In the authors' subjective opinions, the WRF-36 seems to best represent the features of the Stage IV data at 21 UTC. It is the only model that predicted a precipitation maximum in eastern Nebraska, where it was observed. Both Eta-40 and WRF-36 predicted a broad trail of precipitation across northwestern Kansas that corresponds to a scattered line of observed precipitation. This line is spottier in the RUC data, and non-existent in the WRF-12. The WRF-12 had the most intense accumulations of all the models, but it located one of its main cells in an area with little observed precipitation; however this forecast feature was close to an observed cell in northeastern Nebraska. The WRF-12 placement of this cell looks better in the forecast valid at 20 UTC, indicating a possible timing error. The WRF-12 and WRF-36 had roughly equivalent forecasts of the narrow line of observed precipitation near the Kansas-Missouri border; both models predicted the feature too far to the south and east. However WRF-12 had more evidence of the observed cellular structure, particularly in the forecast valid at 20 UTC.

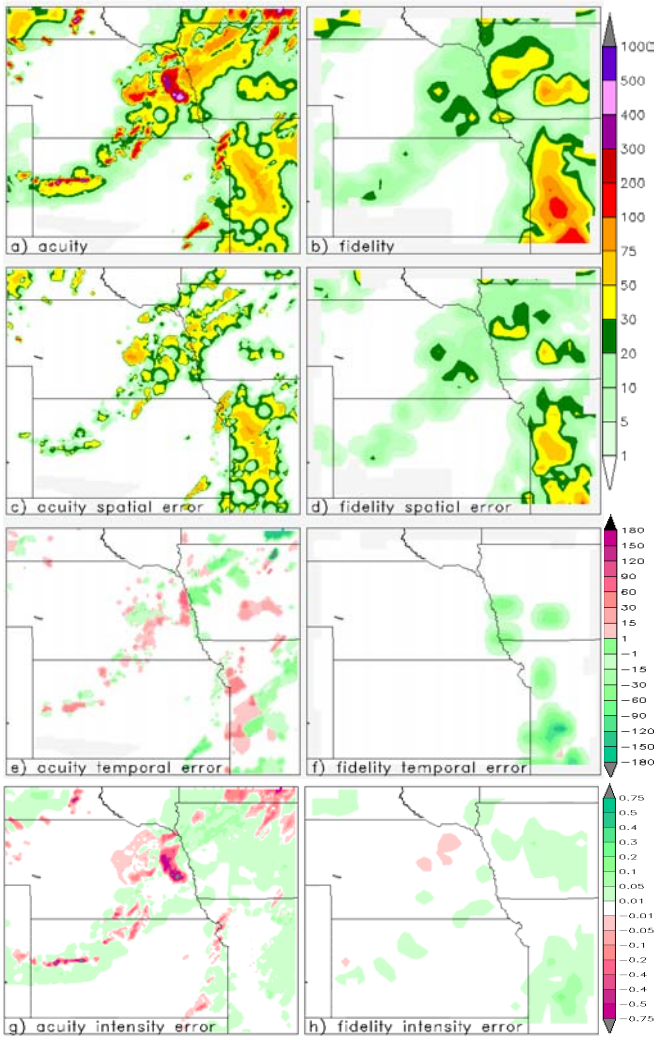

Fig. 9 – Spatial distribution of acuity and fidelity metrics for one-hour precipitation forecasts from the WRF 36 km model as shown in Fig. 8. Metrics are a) acuity, b) fidelity, c) acuity spatial error, d) fidelity spatial error, e) acuity temporal error, f) fidelity temporal error, g) acuity intensity error, h) fidelity intensity error. Units are km for a-d, minutes for e-f, and in/hr for g-h. In e and f, green colors indicate where the model forecasted features too early, and red colors indicate where the model was too late. In g and h, green and red indicate forecasts of precipitation that were too heavy or too light, respectively.

The RUC had a similar placement of this feature to the WRF simulations, but produced less precipitation.

For comparison, the mean acuity and fidelity scores for each forecast model at 21 UTC are shown in Table 1. The acuity and fidelity scores were calculated for 4 stratifications of target intensity: 0, 0.005, 0.05, and 0.25 in/hr. Since the stratification is always on the target intensity, the mean acuity values include only data where the observed precipitation rate was greater than or equal to the stratification threshold value; for fidelity, the stratification is on the forecast precipitation rate. Since neither RUC nor Eta forecast any points with a precipitation rate of 0.25 in/hr or greater, these

Model	Threshold	POD	FAR	TS
WRF-12	0.005	0.464	0.391	0.357
WRF-36		0.420	0.117	0.398
WRF-12	0.05	0.206	0.665	0.146
WRF-36		0.245	0.475	0.200
WRF-12	0.25	0.000	1.000	0.000
WRF-36		0.000	1.000	0.000

Table 2 – Threat score statistics for the forecasts shown in Fig. 8, using thresholds at 0.005, 0.05 and 0.25 in/hr.

models do not have a fidelity score for the 0.25 in/hr threshold.

The mean acuity-fidelity metrics support aspects of the authors' subjective verification, but also reveal some other patterns that were not as easily discernable. For example, WRF-36 has the best acuity score for all stratifications except the unstratified case, indicating that it was the best model at predicting the observed precipitation. The Eta-40 had the best fidelity scores for moderate thresholds, indicating that it made relatively few erroneous forecasts, once very light precipitation values were excluded.

The RUC-20 had the best combined score for the unstratified case, followed closely by the WRF-12. These models had the finest grid spacings and produced precipitation over smaller, more specific areas. This resulted in fewer erroneous forecasts (and hence better fidelity) as well as better detection of the non-precipitation observations (and hence better unstratified acuity).

It is also instructive to examine the spatial distribution of acuity and fidelity metrics, including components of the cost function. Figure 9 shows acuity and fidelity for the WRF-36 model, as well as the spatial, temporal, and intensity components of the error; the event component was negligible in this case. The spatial error is equal to J_s, while the temporal and intensity components are equivalent to J_t and J_i, respectively, except that they are signed values and use units of minutes and inches per hour, respectively. These units can be related to equivalent distances by multiplying by U_e or DI, respectively.

A comparison of Figs. 9a and 9b can explain the disparity between relatively large mean acuity penalties and small fidelity scores. Mean acuity is dominated by large acuity penalties in areas with intense observed precipitation. For example the acuity penalties in eastern Nebraska peak between 400 and 500 km, while fidelity penalties in the same area have a maximum of 50 km. Because the observed features are of limited spatial extent, the fidelity calculation algorithm is able to find good matches to the forecast field by searching small distances in space, except in the vicinity of observed cells, where intensity contributes slightly to the error. This can be seen by comparing Figs. 9b, d, f, and h: the fidelity in eastern Nebraska is almost completely characterized by spatial error.

In contrast, the acuity of the observed line of precipitation across northwestern Kansas is determined by a combination of small effects from spatial, temporal, and intensity errors. Slight adjustments in time and space explain much of the error, except in the vicinity of observed cells, where the intensity error dominates.

Model	Threshold	POD	FAR	TS
WRF-12	0.005	0.311	0.406	0.257
WRF-36		0.342	0.363	0.286
WRF-12	0.05	0.208	0.654	0.149
WRF-36		0.265	0.673	0.171
WRF-12	0.25	0.089	0.891	0.051
WRF-36		0.125	0.961	0.030

Table 3 – Threat score statistics for 39 test cases and 4 forecast periods used in sensitivity study.

Model	Threshold	Acuity	Fidelity	Combined
Eta-40	0	8	4	12
RUC-20		7	**4**	**11**
WRF-12		**6**	7	13
WRF-36		7	5	12
Eta-40	0.005	52	**16**	**68**
RUC-20		59	18	77
WRF-12		**39**	32	71
WRF-36		48	22	70
Eta-40	0.05	94	**33**	**127**
RUC-20		110	42	153
WRF-12		**68**	64	132
WRF-36		86	47	133
Eta-40	0.25	228	**65**	293
RUC-20		232	99	331
WRF-12		**130**	101	**231**
WRF-36		177	79	256

Table 4 – Acuity-fidelity statistics for 39 test cases and 4 forecast periods used in sensitivity study. Values are in km. The best scores in each threshold category are bolded.

5. COMPARISON TO OTHER METRICS

In this section we compare the skill rankings for our test cases as measured by the acuity-fidelity and threat score approaches. For this study, threat score is calculated by examining both the forecasts and observations of one hour precipitation at a common time and location, and determining if the accumulations are above or below a set threshold. A 2x2 truth table is constructed based on this data, and is used to calculate probability of detection (POD), false alarm rate (FAR), and threat score (TS) (Wilks 1995).

These metrics are tabulated in Table 2 for the one date examined in the previous section. For brevity, only the WRF-12 and WRF-36 models are shown. From a threat score perspective, the models have only a fair skill at the lowest threshold, and have no skill at a threshold of 0.25 in/hr. This result runs counter to both subjective analysis and acuity-fidelity, which both indicate at least some skill in the forecast fields.

To make the comparison more statistically meaningful, we compare the mean POD, FAR, and TS to mean acuity, fidelity, and combined score for all 39 test cases and all 4 forecast times used in the sensitivity study. The threat score statistics shown in Table 3 are

more moderate on average than in the May 4 case, showing less skill at the lowest threshold, but more at the highest threshold. The acuity-fidelity results shown in Table 4 are quite similar to the May 4 case, except that WRF-12 proved more skillful on average than WRF-36.

The threat score statistics show relatively little skill compared to either acuity-fidelity or subjective verification. This can be explained by the small amount of information from the forecast and observed fields that are used by these metrics. Threat score statistics make no use of the relative spatial or temporal distribution of forecast and observed values, and make very little use of the dependent variable itself, reducing its precision from a real number to a Boolean.

The acuity-fidelity method, stratified at 0.25 in/hr, awards the best mean acuity to the finest mesh model (WRF-12) and the best mean fidelity to the coarsest mesh model (Eta-40). This indicates that Eta-40 made the fewest erroneous forecasts with intensities greater than 0.25 in/hr, but that WRF-12 had the best representation of the observed precipitation greater than 0.25 in/hr.

In Tables 1 and 4, mean values of acuity and fidelity were used to represent forecast skill; however the statistical processing of raw acuity-fidelity data could be made more complex than a simple average. In fact, any technique that operates on forecast-observations pairs could be applied to the pairs generated by the acuity-fidelity technique. Exploration of the use of alternate statistical techniques with acuity-fidelity is left for future study.

6. CONCLUSIONS

In this paper, we introduced a new verification technique called acuity-fidelity. We applied this technique to the verification of precipitation forecasts and explored the sensitivities of the scheme's configurable parameters. While only precipitation forecasts were studied in this paper, acuity-fidelity could be applied to other phenomena, and is particularly well-suited to rare-event forecasting using numerical weather prediction models. Acuity-fidelity could be used to verify parameters other than precipitation either by modifying the configurable values of the cost function used in this study, or by developing a new cost function.

We demonstrated that acuity-fidelity measures the skill of precipitation forecasts in a way that is consistent with subjective verification based on visual inspection, particularly if the metrics are stratified by intensity. We also showed that visualization of the components of acuity and fidelity can be used as a tool for exploring and characterizing the skill of a forecast.

Finally, we showed that the acuity-fidelity technique provides a more fair assessment of forecasts than traditional metrics, such as threat score. This is particularly true if the forecasts have significant temporal or spatial errors. Acuity-fidelity may provide a way to objectively assess forecasts that previously have been amenable only to subjective verification.

Acknowledgements. Thanks are given to Peter Neilley for helpful conversations and suggestions that improved the quality of this work.

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.

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.

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.

Sousounis, P. J., S. F. Marshall, and T. A. Hutchinson, 2004: A Comparison of MM5, WRF, RUC, and Eta Performance for Great Plains Heavy Precipitation Events During the Spring of 2003. *Preprints 20th Conference on Weather Analysis and Forecasting*, Seattle, Amer. Meteor. Soc., 24.6.

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