16A.8

SCALE SENSITIVITIES IN MODEL PRECIPITATION SKILL SCORES DURING IHOP

Stephen S. Weygandt, Andrew F. Loughe[1], Stanley G. Benjamin, Jennifer L. Mahoney

NOAA Research - Forecast Systems Laboratory, Boulder, CO
[1]Cooperative Institute for Research in Environmental Sciences, Colorado State University, Ft. Collins, CO

1. INTRODUCTION

Traditional statistical measures used to evaluate precipitation forecast skill are affected by variations in the resolved scale of the features in both the forecasts and the observations. This scale-dependence complicates the comparison of precipitation fields that contain differing degrees of small-scale detail and is especially important for warm season precipitation, which is dominated by convective storms. These storms produce precipitation patterns with significant small-scale variability, which are extremely difficult to accurately predict. With the ever-increasing resolution of numerical models, forecast precipitation fields with a similarly large amount of small-scale detail can now be generated. Frequently, however, traditional scores (such as the equitable threat score) are worse for these detailed forecasts than for forecast fields with less small-scale detail. This is because the detailed forecasts often produce "near-misses" for precipitation maxima, even though they quite accurately depict the overall character of the precipitation. Despite a general recognition of this scale dependency and some assessments of it (Gallus 2002, Tustison et al. 2001), no systematic evaluation of the dependency has been completed. Recognition of the dependency has, however, led to a significant research effort aimed at developing more sophisticated verification metrics that more accurately quantify the realism of detailed precipitation forecasts.

In this study, we quantitatively document the scale-sensitivities in precipitation skill scores for four numerical model formulations run during the International H_2O Program (IHOP). IHOP was a field project run in the Southern Plains during the spring of 2002, with the goal of obtaining better observations of moisture and evaluating the observational requirements of atmospheric water vapor for modeling applications. The models compared include the operational 12- km Eta (ETA12), the operational 20-km RUC (RUC20), an experimental 10-km RUC (RUC10) and an experimental 12-km LAPS/MM5 (LMM12). The comparison of the equitable threat and bias scores for the models (verified against stage IV precipitation data) on different resolution grids is complemented by spectral analysis of the various forecast and verification fields. From this analysis we test the hypothesis that skill scores for models verified on different resolution grids are not directly comparable. Furthermore, we document how the skill-score sensitivity depends on the spectral characteristics of the precipitation field, as well as the bias of the field. Toward that goal, we have considered two sets of experiments, one in which in which both the forecast and verification precipitation fields are systematically upscaled to larger grid-resolution and one in which only the forecast fields are upscaled. This upscaling of high-resolution forecasts allows us to isolate the scale effects from effects due to variations in model skill for different resolutions, effectively producing for each model a series of equivalent forecasts, varying only in the degree of small-scale detail retained. Initial work has focused on detailed analysis of a single case. The case chosen well represents the key sensitivities to be evaluated. We are currently extending the single case study analysis to a multi-week period from IHOP.

2. EXPERIMENT METHODOLOGY

For both experiments, a fairly simple procedure for examining the scale sensitivity in the precipitation skill-scores is used as summarized in Fig. 1. For expt. 1, in which both the forecast and verification fields are upscaled, we first compare the equitable threat scores (ETS), bias scores, and spectra for the four models on domain-matched grid sub-sections extracted from each model's native grid. This is accomplished by determining the largest common domain among the native grids of the four models, excluding any model points directly impacted by the lateral boundary

* *Corresponding author address:* Stephen S. Weygandt, NOAA/FSL, R/FS1, 325 Broadway, Boulder, CO 80305, Stephen.Weygandt@noaa.gov

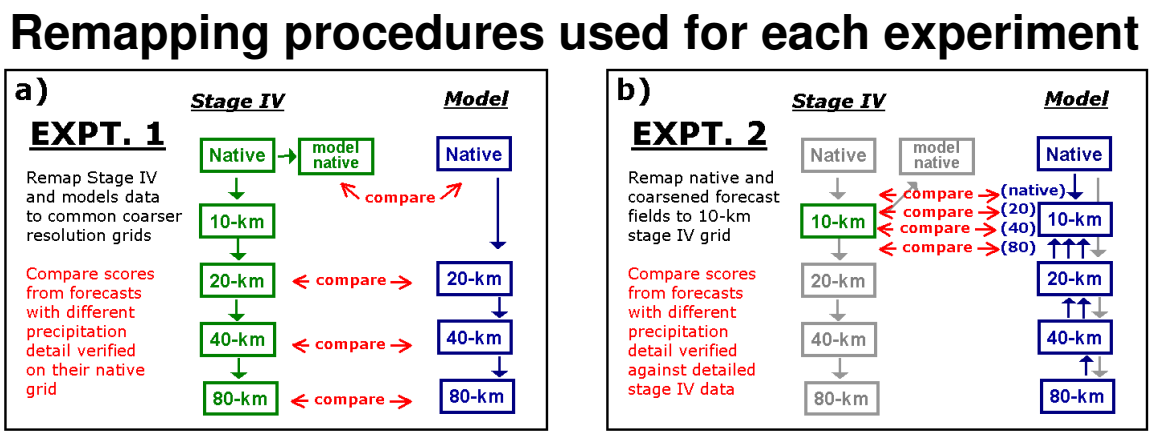

Fig. 1 Schematic diagram depicting the remapping procedures used for each experiment. a) For expt. 1, both forecast and verification fields are remapped to matched common 20-, 40-, and 80-km grids. b) For expt. 2, the native and coarsened forecast fields are remapped to the common 10-km grid

conditions. For each model, the corresponding rectangular subsection of the precipitation field is then isolated for comparison. Note that even though the exact gridpoints do not match between the different native grid subsections, it is imperative that the domain coverage of the models match. Otherwise the spectra and skill-scores cannot be directly compared. Skill scores for each model (on its native grid resolution) are then computed relative to stage IV data remapped to each model's native grid subsection. The stage IV precipitation data (Baldwin and Mitchell, 1998) are on a 4-km resolution national mosaic composited from gauge and radar data estimates supplied by each River Forecast Center. All remappings are accomplished using a standard procedure from the National Centers for Environmental Prediction (NCEP, Baldwin 2000), which approximately conserves the total precipitation volume, and introduces minimal smoothing. The remapping is accomplished by performing a nearest neighbor analysis from the original (input) grid to a 5x5 array of points encompassing each target (output) grid square. A simple average of these 25 target sub-grid values yields the remapped value at the target gridpoint.

Skill score computations are complemented by intercomparison of the precipitation spectra for the model forecast fields (on their native-grid sub-sections) and for the stage IV data interpolated to a neutral, domain-matched 10-km grid. The spectra are computed using the Errico (1985) technique, in which a 2-D Fourier transform is performed to determine spectral coefficients. Multiplication of the spectrum coefficients by their complex conjugate yields the 2-D variance spectrum, which is converted to 1-D by an annular average.

The grids (each forecast model and the 10-km stage IV verification) are then systematically upscaled to identically matched 20-, 40-, and 80-km grids, using the NCEP interpolation routine. Skill score and spectra comparisons can then be made at each of the three common grid resolutions. For each model, skill scores can be compared as a function of verification resolution and precipitation threshold. In this manner, we effectively create for each model a series of equivalent forecasts differing only in the degree of small-scale detail retained. This allows us to document the change in skill attributed solely to the upscaling of the forecast and verification fields from native through 80-km gridlengths. As such, this study represents an extension of the work by Gallus (2002), with a complementary comparison of the spectra, following Baldwin and Wandishin (2002). From the expt. 1 results we are able to address the question of how comparable are skill scores from different grid resolutions (containing different amounts of small-scale detail) verified on their native grid.

For expt. 2, in which only the forecast fields are upscaled, the upscaled forecast fields from the common 20, 40, and 80-km grids are remapped to a common 10-km grid as depicted in Fig. 1b. Forecast fields from the cutdown native grids are also remapped to the common 10-km domain. These forecast fields are then verified against stage IV data that has been remapped to the common 10-km domain. It is important to note that in contrast to the expt.1 remappings, in which the goal is to remove small-scale detail as the fields are upscaled, for expt. 2, the remappings merely transform the given field to a 10-km grid.

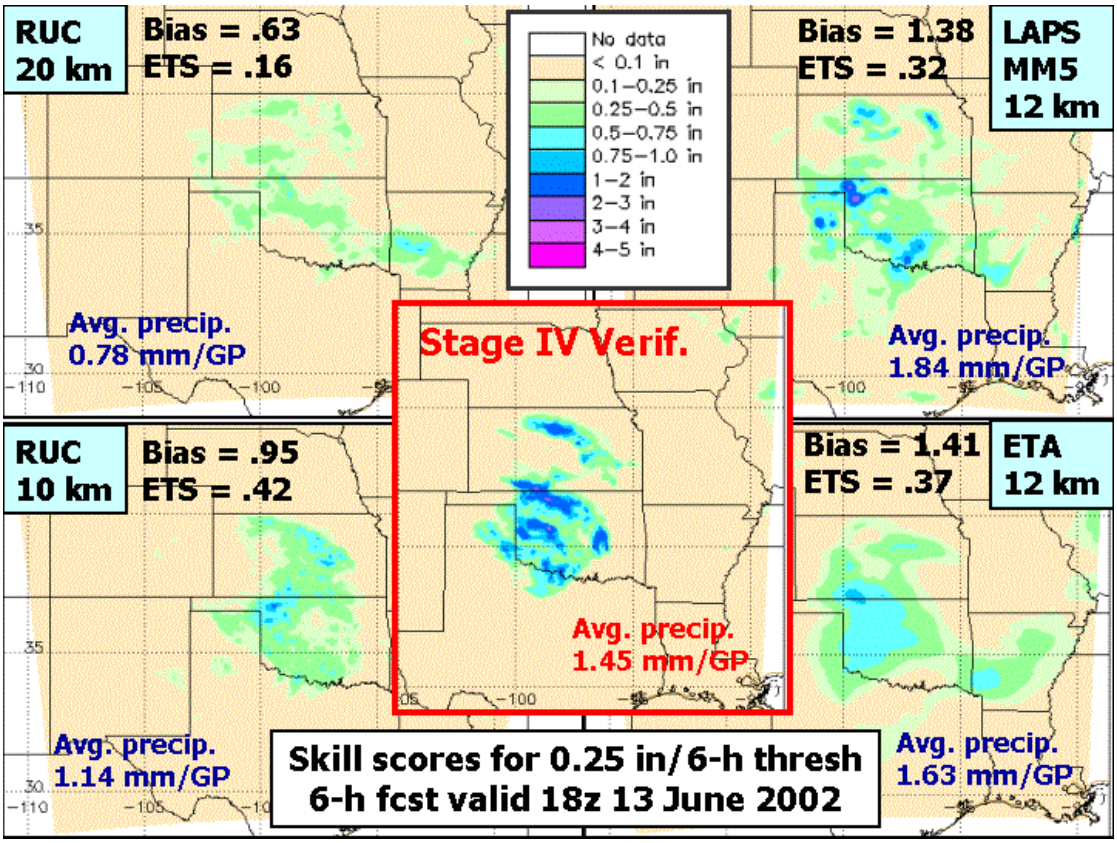

Fig. 2 6-h model predicted precipitation (in) for period ending 1800 UTC, 13 June 2002. Model forecasts shown include RUC 20-km (RUC20), RUC 10-km (RUC10), LAPS MM5 12-km (LMM12), and ETA 12-km (ETA12). Also shown is Stage IV precipitation verification. Contour bands are as indicated in the top center key. For each model, skill scores (ETS and bias) are shown for the 0.25" threshold as well as the average precipitation per gridpoint computed on the common 10-km grid

From the expt. 2 results, we are able to address the question of for a fixed, highly detailed verification field, how do skill scores change as small-scale detail is systematically removed from the forecast fields.

3. RESULTS

We present here a detailed analysis of a single 6-h precipitation forecast that highlights many of the scale-sensitivity issues in precipitation verification. Fig. 2 shows the ETA12, RUC20, RUC10, and LM12 6-h accumulated precipitation from the period 1200-1800 UTC, 13 June 2002, as well as the corresponding stage IV verification as depicted on the Real-Time Verification System (RTVS, Mahoney et al. 200X) webpage. This active IHOP period presents an ideal case, because all models show reasonable skill in the overall precipitation location, allowing the substantial variations in scale among the models to manifest themselves in the skill-score analysis. The stage IV verification is typical of most warm season precipitation fields, with many small-scale heavy precipitation areas. As indicated by the .25" threshold skill scores, the highly detailed RUC10 performs very well (high ETS, bias near 1). The ETA12 and LMM12 also perform quite well (slightly lower ETS, bias near 1.4), but the ETA12 contains substantially less small-scale detail than the LMM12 (or any of the other fields). The RUC20 precipitation location is also good, but the amounts are lighter, leading to lower ETS and bias scores. Also indicated on each panel is the average precipitation per gridpoint, computed for the common 10-km grid. Comparison of these values gives a measure of the total precipitation volume bias for each model.

Fig. 3 shows the spectra for the various models computed on the matched native sub-grids, as well as the spectrum for the stage IV verification interpolated to the common 10-km grid. Comparison of the various spectra confirms the

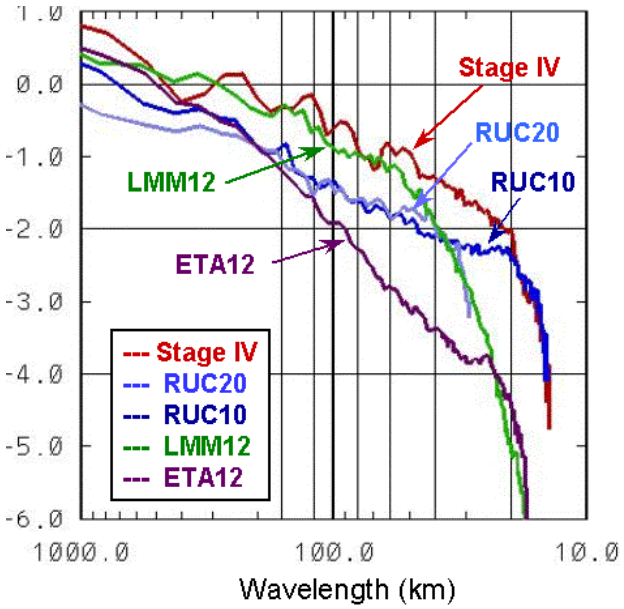

Fig. 3 Spectra computed for the model predicted and Stage IV verification 6-h accumulated precipitation fields. Models shown include RUC20, RUC10, LMM12, and ETA12.

qualitative assessment concerning the degree of small-scale detail in each of the model precipitation fields. In particular, the smoothness of the ETA12 precipitation field (see Fig. 1) is clearly evident in the reduced amplitude and steeper slope of the ETA12 spectral curve relative to the other spectra. The spectral slopes of the other model forecasts are in much better agreement with that of the stage IV verification data. The minimal numerical smoothing employed in the RUC model formulations is evident in the RUC10 and RUC20 spectral curves. The LMM12 curve shows the best match to the stage IV verification for wavelengths greater than 40 km, but indicates significant smoothing of the shortest wavelength features.

a) UPSCALING FORECASTS AND VERIFICATION

We begin our evaluation of the impact of upscaling both the model forecast and verification precipitation fields by assessing the changes in the spectra as the fields are upscaled. Fig. 4 shows the spectral changes as the fields are upscaled for the stage IV verification as well as for the RUC10 and ETA12 model fields. As expected, the changes are similar for the stage IV and RUC 10 models, with substantial reductions in the short wavelength spectral amplitude as the fields are upscaled. In contrast, the ETA12 spectral amplitude shows much less change as the ETA12 fields is upscaled. As expected and in accordance with Fig. 2 and 3,

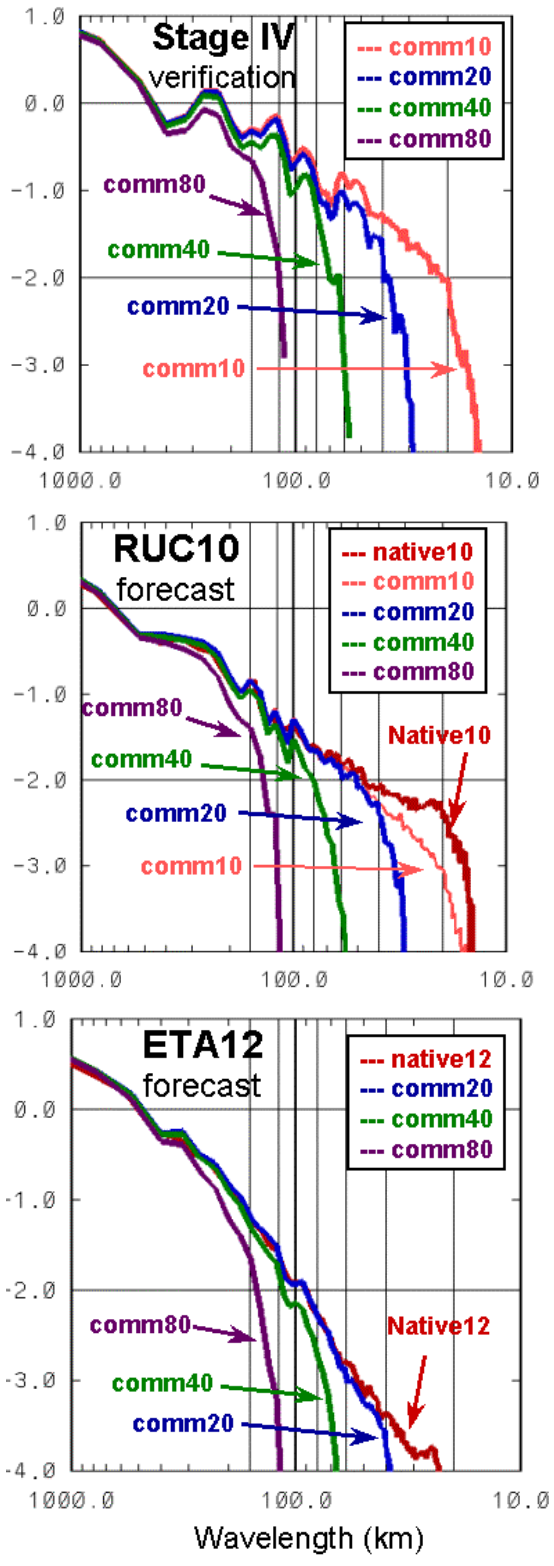

Fig. 4. Spectra computed for the 6-h accumulated precipitation fields on the matched native sub-grids, and common 20-, 40-, and 80-km grids. Shown are the Stage IV verification, RUC10 forecast, and ETA12 forecast.

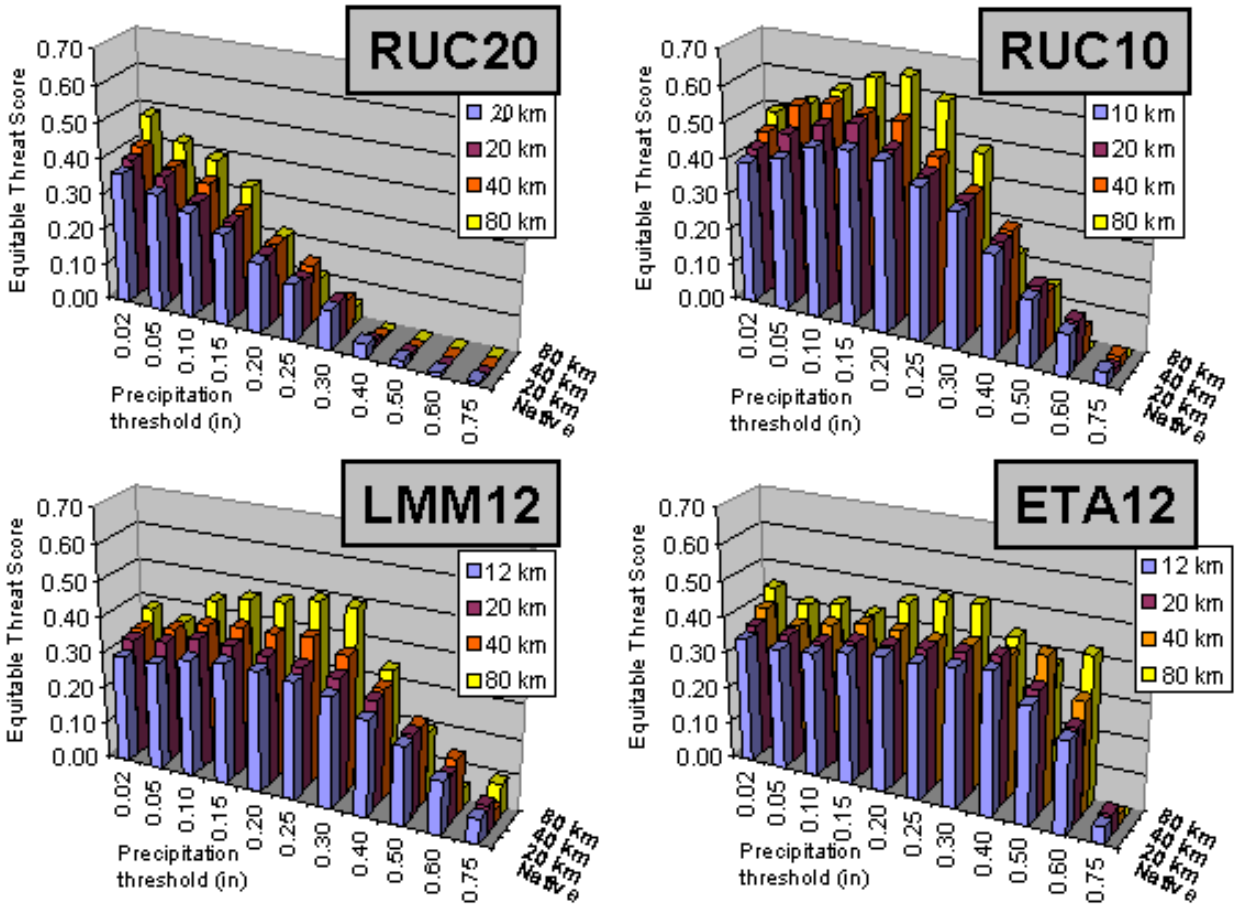

Fig. 5 Equitable threat score (ETS) values computed for a range of precipitation thresholds (horizontal axis) and a range of grid resolutions (native, 20-km, 40-km and 80-km). Models shown include RUC20, RUC10, LMM12, and ETA12.

the ETA12 field is less changed by the upscaling because the field is already quite smooth.

The spectral curve for the remapping of the RUC10 from its cutdown native grid to the common 10-km grid is shown to illustrate the spectral changes introduced by the NCEP remapping algorithm for a grid-resolution neutral transformation. This subject has been addressed previously by Accadia et al. (2003) who found that the NCEP algorithm was superior to simple bilinear interpolation, though it did improve ETS values and produce changes in the bias. Their results are consistent with the smoothing of very small-scale details as revealed by the spectra for the RUC10 remapped to common 10-km grid (shown in Fig. 3b). This small degree of smoothing inherent in the NCEP algorithm should not compromise our results, because we apply the technique consistently to remap all fields (forecast and verification) to coarser resolution grids. As confirmed by the spectra, this remapping removes small-scale detail commensurate with the grid coarsening.

Fig. 5 illustrates, for each model and over a range of precipitation thresholds, the change in the equitable threat scores associated with upscaling the forecast and verification precipitation fields from native through 80-km. The general improvement in the ETS values as the model and verification fields are upscaled largely confirms the hypothesis that skill scores for models verified on different resolution grids are not directly comparable and is consistent with the results of Gallus (2002). For several variations of the Eta model run at 10-km, he found that ETS values were generally higher when verification was performed on a 30-km grid rather than the native grid. Closer examination of Fig. 4 (and Table 1 of Gallus 2002),

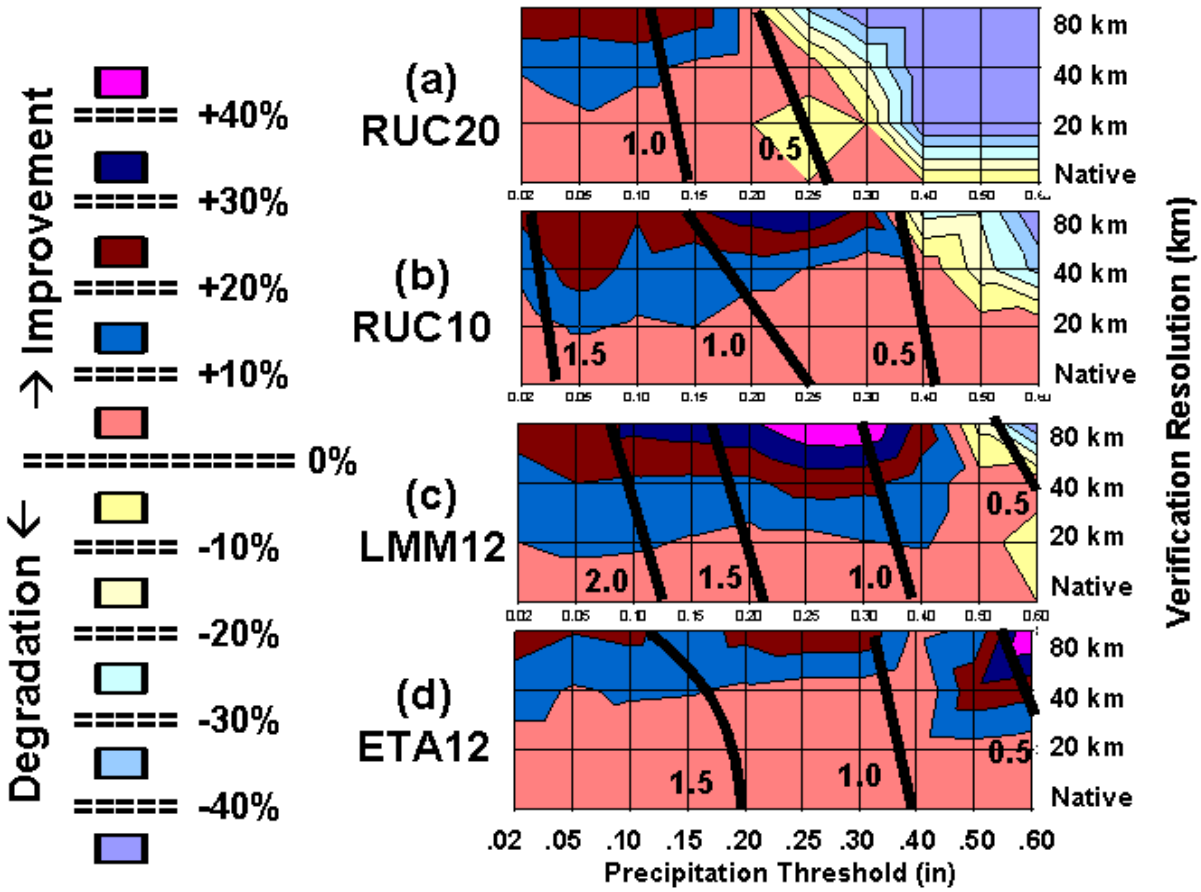

Fig. 6 Percent change in equitable skill score (ETS) computed for a range of precipitation thresholds (horizontal axis) and a range of grid resolutions (native, 20-km, 40-km and 80-km) for the case where both forecast and verification are upscaled. Models shown include (a) RUC20, (b) RUC10, (c) LMM12, and (d) ETA12. Color bands are as indicated in key on the left. Overlaid upon the percent change plots are approximate contours ford selected precipitation bias values (thick black lines).

reveals a more complicated pattern of ETS change as the verification grid is upscaled. This pattern is best revealed by normalizing the ETS values in Fig. 5 by the ETS for the native grid, yielding a percent change in the ETS due to the grid upscaling. Fig. 6 shows the result of this normalization, a contour plot of the percent change in the ETS relative to the native grid ETS as a function of precipitation threshold and verification grid resolution. Overlaid upon this plot are approximate locations of specific bias values in the same threshold/resolution space.

A number of interesting patterns are revealed in Fig. 6. For all three models with substantial mesoscale detail (RUC20, RUC10, LMM12), a cutoff precipitation threshold exists, with ETS improvement occurring below this threshold and ETS degradation occurring above this threshold. Furthermore, the cutoff threshold is a function of the verification resolution, decreasing as the precipitation fields are upscaled. These patterns are consistent with expectations for upscaling of detailed fields. The smoothing of fields (both forecast and verification) as they are upscaled causes a decrease in coverage for the higher precipitation thresholds, and a corresponding decrease in ETS values. This reduction in coverage and decrease in skill begins at the largest thresholds (amounts larger than those shown in Fig. 5) and progresses to successively smaller thresholds as the up-scaling proceeds to larger scales.

This cascade of precipitation from larger to smaller thresholds as the grids are upscaled is confirmed by calculation of the fractional area covered by precipitation exceeding each threshold for each grid. Fig. 7 shows the results of such a calculation, the fractional coverage of RUC10 predicted precipitation in excess of each threshold for the native and upscaled grids. Comparison of the fours curves for the different precipitation thresholds illustrates the overall decrease in

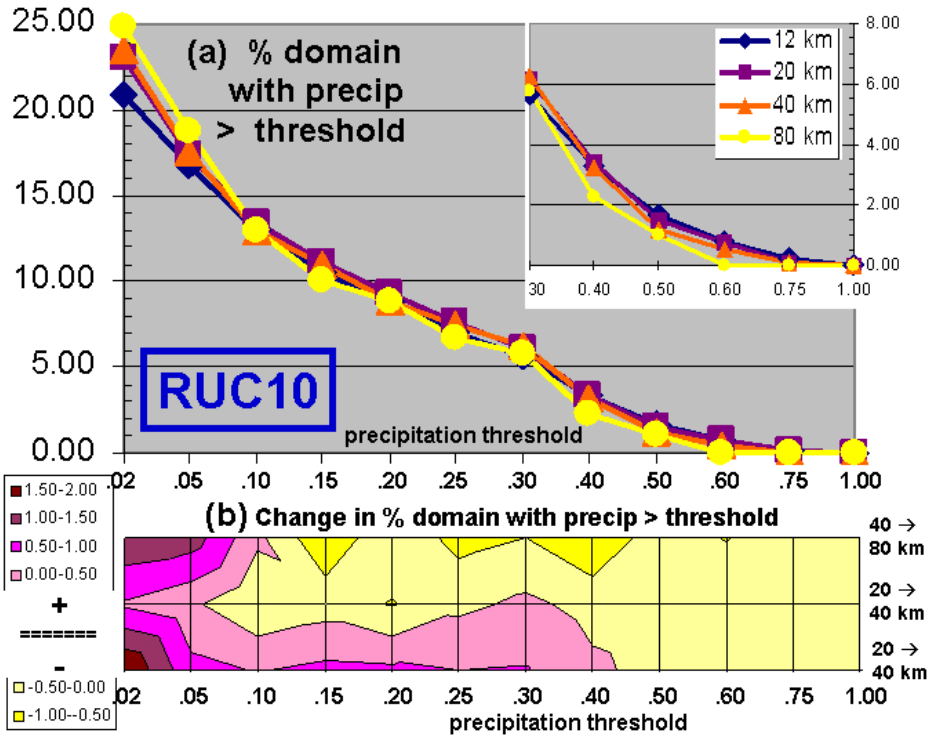

Fig. 7 a) Percent of the RUC10 domain with precipitation in excess of each threshold for each grid resolution. Comparison of the curves shows the cumulative change in fractional coverage as the field is upscaled. b) Change in percent coverage for each upscaling, with yellows denoting a decrease in percent coverage and pinks indicating an increase in percent coverage. Note that these changes are incremental (not cumulative) and must be summed to get the total change.

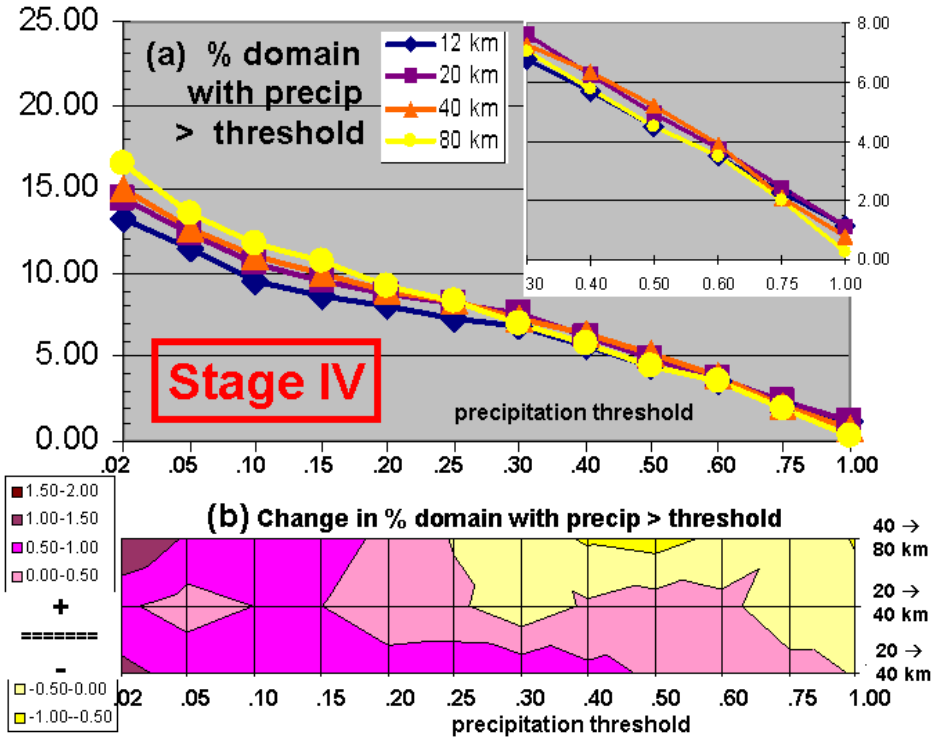

Fig. 8 Same as Fig. 7, but for Stage IV verification data.

fractional coverage for large precipitation thresholds and increase in fractional coverage for small thresholds. To better illustrate this, the change in fractional coverage for each coarsened grid relative to the next finer resolution grid is shown in Fig. 7b. This depiction clearly shows the fractional coverage decrease at larger thresholds progressing to smaller thresholds as the forecast is further coarsened.

Of course, a similar cascade occurs for the stage IV verification field as depicted in Fig. 8. For all resolutions, the stage IV fields have a larger fractional coverage for the large precipitation thresholds and a smaller fractional coverage for the smaller thresholds compared to the RUC10 forecast fields (more intense precipitation maxima, but smaller total areal coverage), accounting for the bias vs. threshold relationship seen in Fig. 6b.

Note that for any threshold and grid resolution, the bias is exactly specified by the ratio of the forecast and verification fractional coverage. Thus, we see that differences in the cascade between the forecast and verification fields directly explain the change of bias as the grids are upscaled. As an example, the strong RUC10 bias decrease with upscaling between the 0.15 and 0.25 thresholds occurs because the stage IV fractional coverage is increasing at these thresholds, while it is decreasing for the RUC10. Because the Stage IV has larger precipitation maxima, the stage IV cascade results in differential increase in fractional coverage for nearly all thresholds, yielding a general decrease in RUC10 bias for all thresholds.

For the ETA12, the cascade of precipitation from larger to smaller thresholds is less pronounced because the initially smooth precipitation field undergoes less modification in the upscaling process. Thus, ETS changes occur primarily from the smoothing of only the stage IV verification, resulting in increasing verification coverage for nearly all thresholds, but nearly constant forecast coverage. This leads to a markedly different pattern of ETS change for the ETA12 (Fig. 4d). Here, the percent improvement in ETS is less pronounced than all but the RUC20, but improvement occurs for all precipitation thresholds. Distinctly missing from the ETA12 model plot is any evidence of a cutoff threshold.

Understanding how differences in the precipitation cascade between the verification and the various models impact the bias and ETS values facilitates a better understanding of the sensitivities displayed in Fig. 6. All models overpredict the smallest and underpredict the largest thresholds, with a slight decrease in bias as the model and verification fields are upscaled. For the RUC20, RUC10, and LMM12, the transition from ETS improvement with upscaling to ETS degradation with upscaling shows some correlation with bias. For each of these models, the cutoff is reasonably well predicted by the 0.5 bias line. This is consistent with the cascade of precipitation from larger to smaller amounts as the verification grid is coarsened. For a given precipitation threshold, as the ratio of forecast to observed points falls below a certain ratio, the skill is measured by the ETS begins to fall. For the RUC20 model, which has the most significant underprediction (especially for larger thresholds) the ETS reduction is most severe and extends to the lowest thresholds. No such correlation between bias and ETS change is seen for the ETA12 plot (Fig. 6d), as large improvements occur for the largest threshold, which has a bias near 0.5. Again, this is consistent with the fact that for the ETA12, only the verification is being significantly smoothed by the upscaling.

Finally, we present a more explicit illustration of the improvement in ETS as small-scale features are removed from the forecast and verification fields. Fig. 9 shows a comparison of the ETS for the LMM12 and ETA12 on their native grids and the common 40-km grid. On the native grids, ETA12 scores are clearly better for all precipitation thresholds. For both models, improvement occurs for nearly all thresholds as the fields are upscaled from the native 12-km grid to the common 40-km. For all but the largest thresholds, however, the relative improvement is less for the ETA12 than the LMM12. The result is that for the common 40-km grid, the ETS values are nearly identical for the two models over a broad range of low and moderate precipitation thresholds.

b) UPSCALING ONLY FORECASTS

In the second set of experiments we examine the impact on the skill scores when only the forecast fields are upscaled. This expt. focuses on the important question of how does the smoothing of forecast fields affect their skill when verified against a fixed high-resolution stage IV data. As described in Sect. 2, the NCEP remapping algorithm was used to transform the coarsened fields (20, 40 and 80-km grids) to the common 10-km domain. This was accomplished in a series of steps with the grid resolution doubled each time. In addition, the precipitation field from each model's native grid was remapped to the common 10-km grid.

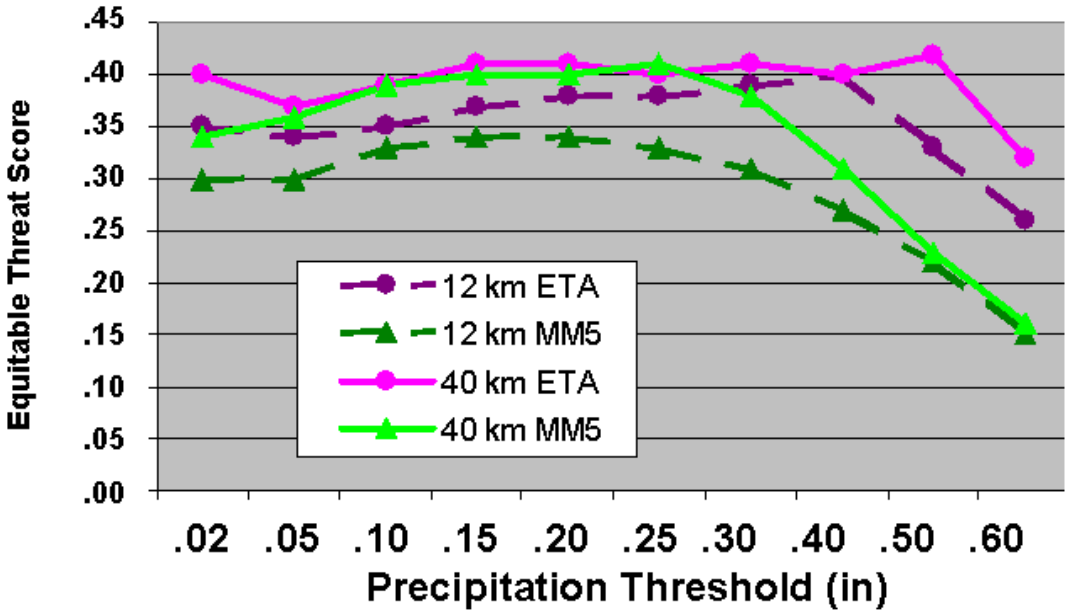

Fig. 9 Equitable skill score (ETS) values computed for a range of precipitation thresholds (horizontal axis) for two models (LMM12 and ETA12) and two grid resolutions (native 12-km and 40-km).

Visual examination of the 10-km remapped fields (not shown) indicates near perfect agreement with the coarser resolution input fields, however, spectra from the 10-km remapped fields indicate a slight amount of smoothing during the remapping process and the introduction of a small amount of noise. These imperfections in the remapping should have very little impact on the results presented.

As depicted in the schematic shown in Fig. 1b, we now illustrate for each model the change in ETS when only the forecast field is upscaled. Analogous to Fig. 6, Fig. 10 shows the percent change in the ETS (relative to the native grid forecast remapped to the common 10-km) as a function of precipitation threshold and smoothness of the forecast field. Although all fields are defined on the common 10-km grid, the 20-, 40-, and 80-km fields are progressively smoother than the 10-km field. This distinction between grid spacing of the field and scale of the features depicted in the field has sometimes been referred to as effective resolution and is discussed by Pielke (2001), Baldwin and Wandishin (2002) and others.

Overall, the pattern of ETS change in Fig. 10 is quite similar to that shown in Fig. 6. Moreover, the differences can be readily explained by noting that in Expt. 1 (Fig. 6), the cascade of precipitation from high to low thresholds is occurring for both the forecast and verification fields, but in expt. 2 (Fig. 10) the cascade is only occurring for the forecast fields. Thus in expt. 2, the biases decrease at the highest thresholds and increase at the lowest thresholds, as indicated by the slopes of the bias lines in Fig. 10.

The lack of skill change at any threshold for the ETA12, is consistent with the fact that little modification is occurring for either the forecast or verification field. The verification is specifically held constant and the forecast field is little changed because it as already quite smooth on the native grid. The expt. 2 percent improvements for the other models are somewhat reduced from expt. 1 because only one of the two fields (the forecast) is being smoothed, somewhat reducing the likelihood of increasing the fraction of hits at a given threshold. The LMM12 decrease in skill for small thresholds is related to the increase in bias beyond values optimal for the ETS.

4. DISCUSSION

The results from Expt. 1, where both the forecast and the verification are upscaled are well known and we merely document the sensitivity for a particular case. They do confirm the fact that forecasts verified on different resolution grids are not directly comparable using the ETS. The forecast verified on the coarser resolution grid will have an advantage due solely to the difference in the grid resolutions.

The results from the Expt. 2 are not surprising, and underscore the difficulty of showing

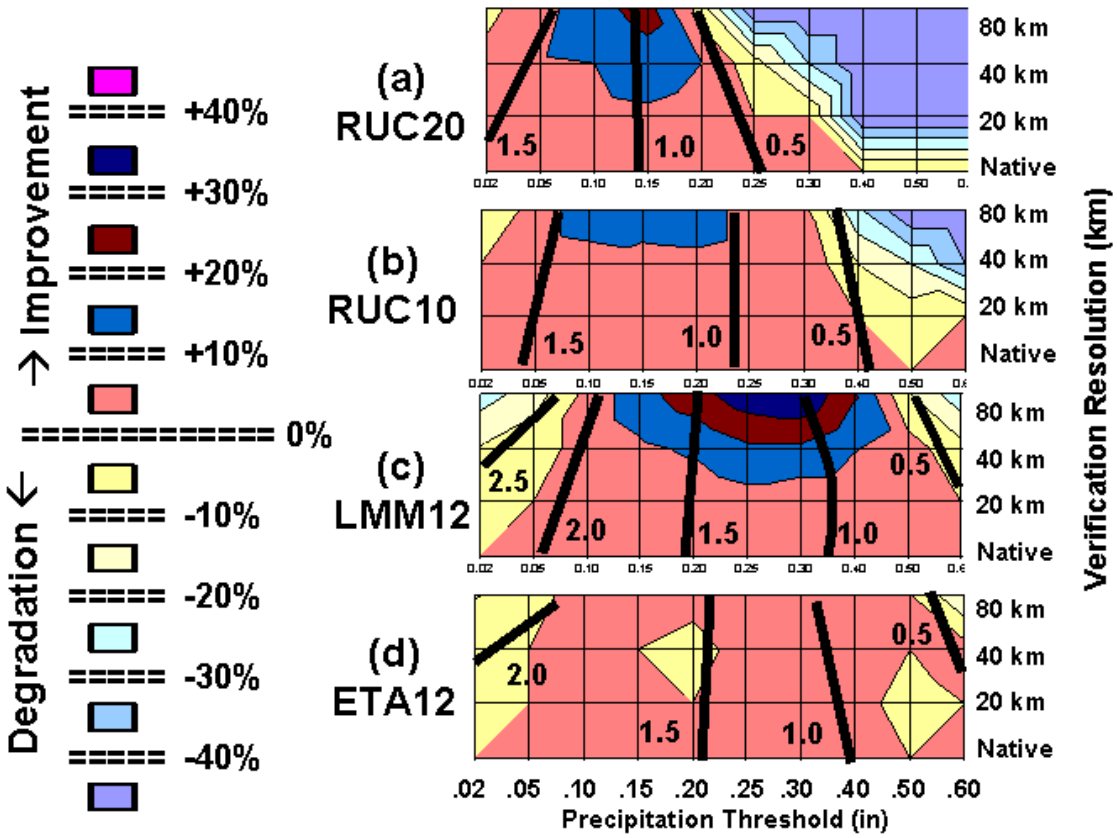

Fig. 10 Percent change in equitable skill score (ETS) computed for a range of precipitation thresholds (horizontal axis) and a range of effective resolutions (native, 20-km, 40-km and 80-km) for the case where only the forecast fields are smoothed. Models shown include (a) RUC20, (b) RUC10, (c) LMM12, and (d) ETA12. Color bands are as indicated in key on the left. Overlaid upon the percent change plots are approximate contours for selected precipitation bias values (thick black lines).

improvement, as measured by the ETS, for high-resolution forecasts. Because these forecasts frequently produce small-scale precipitation fields with phase errors, smoothing the forecast field almost always improves the forecast even when verified against highly detailed fields. The potential improvement for various thresholds from smoothing the forecast is modulated by a number of factors, including 1) the degree of smoothness in the initial field, 2) the overall bias of the field, and 3) the degree to which the initial field is affected by small phase errors in small-scale details.

With respect to the first factor, the ETA12 provides a clear example of an initially smooth field (even though it is on a high resolution grid). In effect, the benefit from smoothing the forecast has already been realized. With respect the second factor, the LMM12 and RUC20 are on the opposite extremes. For forecasts like the RUC20 with its pronounced dry bias, ETS reductions are almost inevitable for large thresholds, because the smoothing reduces bias to near zero. ETS improvements can still occur for low to moderate thresholds, as noted for the RUC20 in Fig. 10. With its large overall bias (as indicated by the large average precipitation per gridpoint value in Fig. 2) and abundant small-scale detail, the LMM12 is ideally suited to improve at medium to large thresholds, as noted in Fig. 10.

With respect to the third factor, two extreme cases can be considered. First, very poor forecasts exist that will not benefit from smoothing (eg: very large phase errors, completely missed or erroneously predicted precipitation area). Secend, some forecasts accurately predict small-scale features. These truly superior forecast profit little from smoothing. As an extreme example of such a forecast, consider the verification of progressively

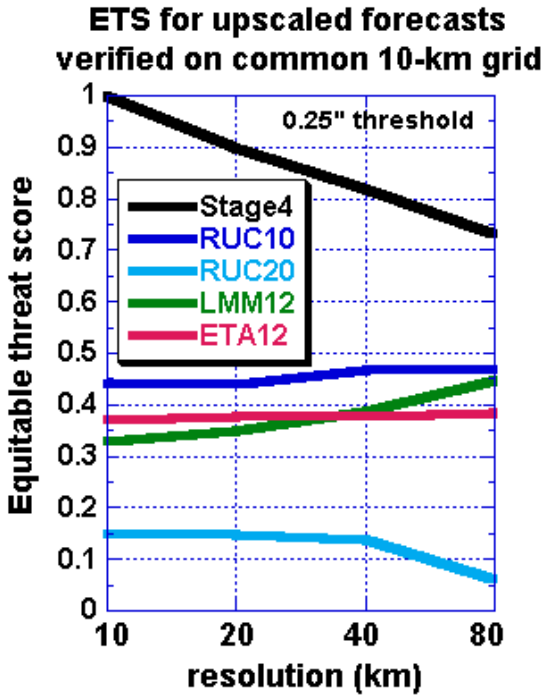

Fig. 11 Change in ETS value as the predicted precipitation field is smoothed for four models (RUC20, RUC10, LMM12, and ETA12) and the Stage IV data, verified against the Stage IV data on the common 10-km grid. Values shown are for the 0.25" threshold.

smoother versions of the stage IV against the stage IV data on the common 10-km grid. In this case, the 10-km stage IV "forecast" perfectly predicts all the small-scale features, and the ETS values would decrease (from 1) as the forecast is smoothed.

Fig. 11 illustrates this smoothing vs. ETS relationship (for the 0.25" theshold) for a perfect stage IV "forecast" as well as the various model forecasts. As expected, the ETS for the perfect stage IV forecast decreases with increasing smoothing. The LMM12 ETS value increases with smoothing, while the ETA12 and RUC10 ETS values are nearly constant. Because of its strong dry bias, the RUC20 ETS decreases as the field is smoothed.

The differing behaviors between the perfect and actual forecasts as they are smoothed gives some indication of the extreme demands that the ETS places on highly detailed forecasts. In some sense, the stage IV perfect forecast curve provides a practical upper-bound on the ETS score that can be attained for a given effective resolution forecast verified against the high-resolution data. Furthermore the slope of the stage IV perfect forecast curve provides a measure of the degree of small-scale detail in the verification field, with less negative slopes denoting smoother fields.

5. SUMMARY AND FUTURE WORK

Detailed analysis of a single case has yielded results that appear to confirm the initial hypothesis that precipitation skill-scores for models verified on different resolution grids should not be directly compared because: 1) ETSs generally increase as forecasts are verified on progressively coarser domains and 2) the improvement is greatest for fields that contain the largest amount of small-scale detail. While a general recognition of this sensitivity exists, with the exception of the work by Gallus (2002), the sensitivity has not been quantitatively documented. As discussed by Gallus (2002), this smoothness/skill relationship has significant implications for the downscale extension of mesoscale models. Our results support Gallus' conclusion that it may be difficult to show improvement in ETS values for models with increasingly fine resolution. The degree to which small-scale details should be retained in mesoscale models (and more sophisticated techniques used to verify the models) is the focus of some attention in the mesoscale modeling community. Our aim in this research is not to answer that question, but to provide a systematic documentation of the scale-sensitivities that do exist for traditional skill scores, such as the ETS.

For both exts. 1 and 2, we are currently extending the single case study analysis to include two one week periods (55 cases) encompassing the most convectively active periods during IHOP. Preliminary results for the first expt. indicate the trends documented in this single case study are also seen in the multi-case average. Further analysis of these multi-case results for both expts. is ongoing. In the future, we hope use this set of model forecasts and verification data as a testbed for evaluating more sophisticated scale-dependent verification techniques.

6. ACKNOWLEDGMENTS

This work was funded by a grant to NOAAA/OAR from the U.S. Weather Research Program. We would like to express our sincere appreciation to M. Baldwin and M. Wandishin for providing the spectral decomposition code to us. Discussions with M. Baldwin, Barry Schwartz, and Tom Hamill on a number of issues related to this work were quite helpful. We acknowledge Barry Schwartz and Susan Carsten for careful scientific and technical reviews, respectively.

7. References

Accadia, C., S. Mariani, M. Casaioli, and A. Lavagnini 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and simple neareast-neighbor average method on high-resolution verification grids. *Wea. Forecasting*. **18**, 918-832.

Baldwin, M.E 2000: Quantitative precipitation verification documentation. [Available online at: http://www.emc.ncep.noaa.goc/mmb/ylin/pcpverif/scores/docs/pptmethod.html

Baldwin, M.E., and K.E. Mitchell, 1998: Progress on the NCEP hourly multi-sensor U.S. precipitation analysis for operations and GCIP research. Preprints, *2nd Symp. on Integrated Observing Systems,* 78th AMS Annual Mtg., 10-11.

Baldwin, M.E., and M.S. Wandishin, 2002: Determining the resolved spatial scales of the Eta model precipitation forecasts. Preprints, *15th Conf. on Num. Wea. Pred.,* San Antonio, TX, Amer. Meteor. Soc., 85-88.

Errico, R.M., 1985: Spectra computed from a limited area grid. *Mon. Wea. Rev.* **113**, 1554-1562.

Gallus, W.A., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting*. **17**, 1296-1302.

Mahoney, J.L., J.K. Henderson, B.G. Brown, J.E. Hart, A.F. Loughe, C. Fischer, and B. Sigren, 2002: The real-time verification system (RTVS) and its application to aviation weather forecasts. Preprints, *10th Conf. OnAviation, Rnage, and Aerospace Meteorolog.,* Portland, OR, Amer. Meteor. Soc., 20-23.

Tustison, B., Harris, D. and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.* **106**, 11,775-78.