2002

Impact of verification grid-box size on warm-season QPF skill measures

William A. Gallus Jr.
Iowa State University, wgallus@iastate.edu

Impact of verification grid-box size on warm-season QPF skill measures

Abstract

A 10-km-grid-spacing version of NCEP's Eta Model was used to simulate 11 warm-season convective systems occurring over the U.S. upper midwest. Quantitative precipitation forecasts (QPFs) from the model valid for 6-h periods were verified using 4-km-grid-spacing stage-IV precipitation estimates. Verification first was performed on the model's 10-km grid by areally averaging the 4-km observations onto the model grid. To investigate and quantify the impact of the verification grid-box size on some standard skill scores, verification was also performed by averaging the 10-km model forecasts onto 30-km grid boxes and then areally averaging the observations onto the same 30-km grid. As a final test of the impact of the verifying grid-box size, the same 11 events were simulated with a 30-km version of the Eta Model, with verification then being performed on this 30-km grid. For all cases in both the 10- and 30-km versions of the model, 12 variations of the model were used, with variations involving either (i) modifications to the initial conditions to better represent mesoscale features present at the initialization time or (ii) changes in moist physics. Equitable threat scores (ETSs) increased when verification occurred on a coarser grid, whether the coarser grid was created by averaging the 10-km model results or was that used in the 30-km model runs. This result suggests that it may be difficult to show improved skill scores as model resolution improves if the verification is performed on the model's own increasingly fine grid. It should be noted, however, that the use of different verification resolutions does not change the general impacts on ETSs of variations in model physics or initial conditions. The sensitivity of ETSs to verifying grid-box size does, however, vary somewhat between model variants using different model moist-physics formulations or initialization procedures.

Keywords

computer simulation, mathematical models, precipitation (Meteorology), quantitative precipitation forecasts (QPF), weather forecasting, convective system, grid, precipitation (climatology), United States

Disciplines

Atmospheric Sciences | Geology

Comments

This article is from *Weather and Forecasting* 17 (2002): 1296, doi: 10.1175/1520-0434(2002)017<1296:IOVGBS>2.0.CO;2. Posted with permission.

Impact of Verification Grid-Box Size on Warm-Season QPF Skill Measures

WILLIAM A. GALLUS JR.

Department of Geological and Atmospheric Science, Iowa State University, Ames, Iowa

21 February 2002 and 31 July 2002

ABSTRACT

A 10-km-grid-spacing version of NCEP's Eta Model was used to simulate 11 warm-season convective systems occurring over the U.S. upper midwest. Quantitative precipitation forecasts (QPFs) from the model valid for 6-h periods were verified using 4-km-grid-spacing stage-IV precipitation estimates. Verification first was performed on the model's 10-km grid by areally averaging the 4-km observations onto the model grid. To investigate and quantify the impact of the verification grid-box size on some standard skill scores, verification was also performed by averaging the 10-km model forecasts onto 30-km grid boxes and then areally averaging the observations onto the same 30-km grid. As a final test of the impact of the verifying grid-box size, the same 11 events were simulated with a 30-km version of the Eta Model, with verification then being performed on this 30-km grid. For all cases in both the 10- and 30-km versions of the model, 12 variations of the model were used, with variations involving either (i) modifications to the initial conditions to better represent mesoscale features present at the initialization time or (ii) changes in moist physics. Equitable threat scores (ETSs) increased when verification occurred on a coarser grid, whether the coarser grid was created by averaging the 10-km model results or was that used in the 30-km model runs. This result suggests that it may be difficult to show improved skill scores as model resolution improves if the verification is performed on the model's own increasingly fine grid. It should be noted, however, that the use of different verification resolutions does not change the general impacts on ETSs of variations in model physics or initial conditions. The sensitivity of ETSs to verifying grid-box size does, however, vary somewhat between model variants using different model moist-physics formulations or initialization procedures.

1. Introduction

Skill scores for quantitative precipitation forecasts (QPFs) have generally improved over the years, in part because of the use of finer horizontal resolution (e.g., Black 1994). Unfortunately, these scores have tended to remain especially poor for warm-season convective rainfall (e.g., Olson et al. 1995; Stensrud et al. 2000), for which (i) a number of physical processes occur on scales too small to be adequately resolved by even the higher-resolution operational models used today, requiring use of various parameterization schemes (convection, boundary layer processes, radiation, evapotranspiration, and cloud microphysical processes), and (ii) much of the convection is forced by mesoscale features also poorly resolved in the models (Kain and Fritsch 1992; Stensrud and Fritsch 1994).

The important role of mesoscale dynamics in producing warm-season rainfall might suggest that large improvements are likely in skill scores as horizontal resolution of the models becomes sufficiently fine to accurately represent these small-scale circulations. Brooks et al. (1992) have cautioned, however, that increased horizontal resolution might produce little if any improvement in convective forecasts because of limits on the predictability of some phenomena and incomplete data sampling for model initialization. Recently, Gallus and Segal (2001) showed that in a 10-km-grid-spacing version of the Eta Model, it was difficult to consistently improve equitable threat scores (ETSs) by a significant amount when three different techniques were used to improve the initialization of mesoscale features presumed to be important in the formation of later convective systems. In addition, recent verification statistics for National Centers for Environmental Prediction (NCEP) models [e.g., McDonald and Horel (1998); M. Baldwin, National Severe Storms Laboratory (NSSL), 2001, personal communication] suggest that improvements in ETSs are leveling off as horizontal resolution becomes increasingly fine. Mass et al. (2002) note that although improved grid resolution may not improve objectively scored accuracy measures, the realism of simulated features may be better.

Corresponding author address: Dr. William A. Gallus Jr., Dept. of Geological and Atmospheric Science, Iowa State University, 3025 Agronomy Hall, Ames, IA 50011.
E-mail: wgallus@iastate.edu

It is well known that QPF verification performed for short time intervals (e.g., 6 h) will result in lower skill scores than verification performed over longer (e.g., 24 h) periods (e.g., see the Web site http://www.hpc.ncep.noaa.gov/html/hpcverif.shtml), because timing errors have a less adverse effect on skill scores evaluated over the longer time periods. What may be less well known is the impact of the *spatial* resolution of the verifying analysis on the skill scores obtained. Just as a longer time interval for verification can reduce the impact of temporal errors in simulated rainfall on skill scores, a coarser verifying grid can reduce the impact of small spatial errors.

Tustison et al. (2001) discuss in detail how the methods used to verify model QPF with gauge data can affect objective verification measures. The total error used to compute common skill measures consists not only of observation and model error but also of representativeness error, the error related to interpolation or averaging of either model gridpoint values or gauge observations. Tustison et al. show that when area-to-point techniques are used (model grid-box values are assumed to represent a point in the middle of the box, with interpolation to gauge sites), representativeness errors worsen for *coarser* resolutions. Thus improvements in skill scores that may be noted for refined horizontal resolution reflect at least in part the improvement in representativeness error; model performance itself may not necessarily be better. On the other hand, for point-to-area conversion, the technique commonly used at NCEP, representativeness error worsens as resolution is *refined.* Thus, there may be a tendency for refined resolution to result in worse threat scores, if verification is performed on the model's native grid. Complicating the issue further is the fact that coarser grids may inadequately resolve mesoscale processes important in generating a correct rainfall forecast, which should decrease threat scores.

As a first attempt at quantifying the impact that the verifying grid resolution may have on ETSs when point-to-area conversion is used, a subset of the cases and model variants used in Gallus and Segal (2001) was reevaluated. A subset of 11 cases was chosen, representing thirty-one 6-h periods of active observed convective system rainfall. ETSs were computed and averaged over all 31 events using the same methods as in Gallus and Segal (2001). In addition, ETSs were recomputed using the 10-km model output averaged onto a coarser 30-km grid. Last, the 11 events also were simulated using a 30-km version of the model, whose gridpoint locations matched those of the 10-km results averaged onto a 30-km grid. Further details about the data used and methodology are provided in section 2. ETSs computed using the different verification techniques and averaged over the sample of cases, along with the impacts of the initialization adjustment techniques and differing moist physics, are discussed in section 3. The final section provides a summary and short discussion.

2. Data and methodology

Eleven cases were selected from the warm seasons (May–August) of 1998, 1999, and 2000 in which mesoscale boundaries, usually convectively induced, were present at 0000 or 1200 UTC, and significant mesoscale convective system (MCS) precipitation followed within the next 12–18 h over the upper Midwest. The cases were chosen using surface data, radar data from the National Climatic Data Center (NCDC) online archive, 4-km horizontal grid spacing NCEP stage-IV precipitation observations (Baldwin and Mitchell 1997), and reports from *Storm Data,* published by NCDC. These 11 events are a subset of the 20 events examined in Gallus and Segal (2001), and they were chosen to ensure roughly equal contributions of 6-h periods from cases initialized at 0000 and at 1200 UTC. The cases had originally been simulated using 10-km grid spacing in a workstation version of the NCEP Eta Model, run for 24 h over a small domain covering roughly 1000 km × 1000 km. Standard initial and boundary condition data were provided by 40-km NCEP Eta output. In addition to the 10-km simulation, the Eta Model was also run for the same cases using a 30-km grid version.

The Eta Model version used was similar to that used operationally at NCEP in early 2000, and included the same physical parameterizations present in the operational model [see Janjić (1994) and Rogers et al. (1998) for more details]. In addition to the Betts–Miller–Janjić (BMJ) convective scheme (Betts 1986; Betts and Miller 1986; Janjić 1994) used operationally, simulations were repeated using a version of the Kain–Fritsch (KF) scheme (Kain and Fritsch 1993) adapted for use in the Eta Model (J. Kain 2000, personal communication).

For each of the cases, a total of 12 different variants of the model were run [a subset of the 14 used in Gallus and Segal (2001)], consisting of modifications to improve the initialization of mesoscale features, and changes in the moist physics. The initialization modifications included (i) the cold pool initialization scheme (CP hereinafter) discussed in Stensrud et al. (1999), (ii) vertical assimilation of mesoscale surface observations (MO hereinafter) using the model's own vertical diffusion following the concepts of Ruggiero et al. (1996), and (iii) setting a minimum relative humidity threshold for all initial data (RH hereinafter) of 80% in the lower and middle troposphere (all levels warmer than $-10°C$) at all grid points where an organized radar echo was present. The RH adjustment generally resulted in rapid activation of the convective scheme at those points [in some ways like the technique proposed by Rogers et al. (2000)]. [More details about the techniques can be found in Gallus and Segal (2001).] All of these adjustments were made in runs using both the BMJ and the KF convective schemes. In addition to these variants, runs were performed using both MO and RH together. One additional run was performed using no convective scheme (denoted NONE), and another used both the

Table 1. ETSs for BMJ and KF control runs, plus runs alternating between both schemes, and using no convective scheme, for five precipitation thresholds, averaged over thirty-one 6-h periods, for 10-km output verfied on the 10-km grid (−10), 10-km output averaged to and verified on a 30-km grid (−10ave30), and 30-km output verfied on the 30-km grid (−30). Changes that are statistically significant at the 90% and 95% confidence levels using a Wilcoxon rank test are indicated with italics and boldface, respectively.

	Precipitation threshold (mm)				
Run	0.254	2.54	6.35	12.7	25.4
BMJ-10	0.246	0.175	0.114	0.054	0.006
BMJ-10ave30	**0.264**	**0.183**	**0.122**	0.051	*0.007*
BMJ-30	**0.273**	**0.207**	*0.135*	0.064	**0.023**
KF-10	0.216	0.154	0.090	0.036	0.012
KF-10ave30	**0.241**	0.159	0.094	0.034	0.009
KF-30	0.220	0.169	0.106	0.039	0.013
ALT-10	0.266	0.200	0.123	0.053	0.011
ALT-10ave30	**0.280**	0.205	0.129	0.051	0.011
ALT-30	0.283	0.216	0.137	*0.067*	**0.032**
NONE-10	0.113	0.077	0.052	0.020	0.006
NONE-10ave30	**0.119**	**0.082**	0.054	0.021	0.007
NONE-30	**0.087**	**0.055**	0.037	0.016	0.007

BMJ and KF schemes alternating within one run (denoted ALT).

For verification of the model runs, stage-IV precipitation observations were used. When available, the multisensor stage-IV data were used; otherwise, the stage-IV gauge measurements were used. For the cases simulated in the present study, these precipitation data (multisensor and gauge) did not differ substantially. These data were examined and compared with radar and surface reports to disregard small-scale spurious features that occasionally occurred directly over radar sites (manifested as isolated heavy rainfall amounts not part of a larger region of precipitation). Following procedures typically used at NCEP (M. Baldwin 2000, personal communication) the 4-km stage-IV observations were areally averaged onto the model's own grid for computation of skill scores. Thus the 10-km model results were verified on a 10-km grid, and the 30-km results on a 30-km grid. An additional test was performed in which the 10-km model results were averaged onto a 30-km grid, and then verified on that 30-km grid (which was collocated with the 30-km grid Eta version). The averaging of 10-km output to a 30-km grid was performed by equally weighting the nine grid points present in every 30 km × 30 km square in the 10-km grid structure, where the squares were chosen to not overlap (matching the grid used in the 30-km version of the model).

To evaluate the forecasts, traditional objective measures such as the ETS (Schaefer 1990) and bias were computed for all cases for a range of precipitation thresholds reflecting rainfall exceeding 0.254, 2.54, 6.35, 12.7 and 25.4 mm (0.01, 0.1, 0.25, 0.5 and 1.0 in.). The ETS is defined as

$$\mathrm{ETS} = \frac{(\mathrm{CFA} - \mathrm{CHA})}{(F + O - \mathrm{CFA} - \mathrm{CHA})}, \tag{1}$$

where CFA is the number of grid points where rainfall was correctly forecast to exceed the specified threshold (a ''hit''), F is the total number of grid points where rainfall was forecast to exceed the threshold, O is the number of observed grid points where rainfall exceeded the threshold, and CHA is a measure of the number of grid points where a correct forecast would occur by chance, where CHA is

$$\mathrm{CHA} = O\frac{F}{V} \tag{2}$$

and V is the total number of grid points evaluated. The bias (BIA) is the ratio of all grid points forecast to have rainfall to the number of grid points where rainfall was observed,

$$\mathrm{BIA} = \frac{F}{O}. \tag{3}$$

It is acknowledged that each individual measure of skill provides only its own unique information about the performance of a mesoscale forecast or the benefit of changes to a model. The best evaluation of model performance is one that uses multiple measures (e.g., Murphy 1991). For instance, as discussed in Mason (1989), high ETSs are often accompanied by overly high BIA scores. In Gallus and Segal (2001), a bias-adjustment technique to ETS (Hamill 1999) was used to allow formal hypothesis testing of the impacts of changes in the initialization to better represent mesoscale features. In the present note, unadjusted ETSs will be discussed but possible impacts of variations in BIA on the ETSs will be mentioned. A Wilcoxon signed-rank test (Wilks 1995) was used to determine the statistical significance of changes in the ETSs occurring when the verification grid-box size was adjusted. Hamill (1999) has cautioned that the Wilcoxon test is sensitive to small changes in the population of contingency table elements for high thresholds where sample size is smaller. Because of this and the fact that little skill exists for rainfall of 12.7 mm or more in 6 h (e.g., Stensrud et al. 2000; Gallus and Segal 2001), discussion in this note will emphasize lighter rainfall thresholds (6.35 mm or less).

3. Results

In the Eta runs examined in this study, the impact of the size of the verifying grid box on the ETSs averaged over thirty-one 6-h periods of active convective system rainfall was large. Table 1 shows the impacts for simulations using the BMJ scheme, the KF scheme, alternation of both schemes within the same model run, and fully explicit moist physics with no convective scheme. When the 10-km Eta Model results were averaged onto 30-km grid boxes and then compared with the observations averaged onto the same 30-km boxes (Table 1 rows marked with the −10ave30 suffix), the ETSs in-

Table 2. Biases for BMJ and KF control runs, plus runs alternating between both schemes, and using no convective scheme, for five precipitation thresholds, averaged over thirty-one 6-h periods, for 10-km output verified on the 10-km grid (−10), 10-km output averaged to and verified on a 30-km grid (−10ave30), and 30-km output verified on the 30-km grid (−30).

Run	Precipitation threshold (mm)				
	0.254	2.54	6.35	12.7	25.4
BMJ-10	1.280	1.565	1.459	1.234	0.534
BMJ-10ave30	1.254	1.532	1.451	1.423	0.501
BMJ-30	1.005	1.467	1.463	1.508	1.639
KF-10	0.797	0.877	0.891	1.129	2.150
KF-10ave30	0.846	0.881	0.874	1.223	1.975
KF-30	0.750	0.980	0.920	1.008	1.932
ALT-10	1.265	1.496	1.285	0.916	0.366
ALT-10ave30	1.261	1.449	1.241	1.022	0.526
ALT-30	0.969	1.315	1.210	1.164	1.326
NONE-10	0.420	0.458	0.619	1.161	3.719
NONE-10ave30	0.416	0.474	0.651	1.411	3.917
NONE-30	0.283	0.304	0.439	0.976	3.858

creased for all model variants (as compared with the verification on a 10-km grid, marked with the −10 suffix) for rainfall amounts of 6.35 mm or less. The biggest improvements occurred for the lightest rainfall threshold, when the ETS of the BMJ run, for instance, increased by 0.018. The changes in the BMJ runs were statistically significant at the 95% confidence level for thresholds of 6.35 mm or less. For the KF runs, ETS values increased even more, by as much as 0.025, but statistical significance was limited to the lightest threshold. For all but the heavier thresholds (where sample size was limited and results must be interpreted with caution) in the BMJ and KF runs, BIA did not change substantially (Table 2) as the 10-km output was averaged to 30 km, so the improvements in ETS were not due to overprediction of precipitation areal coverage.

For the lighter rainfall amounts, these increases in ETSs were nearly comparable to the largest improvements that occurred when initial conditions were modified to better represent mesoscale features (Table 3). Table 3 shows the change in ETS that occurred when initialization was modified for a particular convective scheme evaluated with a particular verifying grid resolution. The increase in the KF run ETS of 0.025 at the 0.254-mm threshold that occurred when verification was performed on a grid of 30 km instead of 10 km compares to maximum improvements of 0.026 at the 12.7-mm threshold when MO was used with the BMJ scheme, and 0.024 at the 2.54-mm threshold when the RH adjustment was used with BMJ (for verification on a 10-km grid; rows in Table 3 marked with the −10 suffix).

Table 3. Change in ETSs from the appropriate control run (run having the same convective scheme but no initialization modification and verified with the same grid resolution) for simulations using initialization modifications for five precipitation thresholds, averaged over thirty-one 6-h periods, for 10-km output (−10), 10-km output averaged to and verified on a 30-km grid (−30ave), and 30-km output verified on the 30-km grid (−30). Notation cp represents cold pool adjustments, mo represents vertical assimilation of mesoscale observations, rh represents relative humidity adjustment based on radar, and morh uses both mo and rh.

Run	Precipitation threshold (mm)				
	0.254	2.54	6.35	12.7	25.4
Variants of BMJ					
BMJcp-10	−0.001	−0.006	−0.003	−0.002	−0.001
BMJcp-10ave30	−0.002	−0.008	−0.007	−0.005	−0.001
BMJcp-30	−0.004	−0.006	−0.010	−0.003	0.001
BMJmo-10	0.004	0.018	0.019	0.026	0.016
BMJmo-10ave30	0.000	0.016	0.017	0.023	0.012
BMJmo-30	0.002	0.010	0.014	0.017	0.014
BMJrh-10	0.018	0.024	0.020	0.022	0.013
BMJrh-10ave30	0.018	0.022	0.019	0.020	0.014
BMJrh-30	0.016	0.021	0.023	0.018	0.023
BMJmorh-10	0.017	0.022	0.024	0.033	0.018
BMJmorh-10ave30	0.015	0.017	0.023	0.033	0.016
BMJmorh-30	0.016	0.020	0.030	0.034	0.029
Variants of KF					
KFcp-10	−0.003	−0.003	0.000	0.000	0.000
KFcp-10ave30	−0.002	−0.002	−0.001	−0.001	0.003
KFcp-30	−0.002	−0.002	−0.005	0.004	0.005
KFmo-10	0.005	0.012	0.021	0.017	0.013
KFmo-10ave30	0.002	0.016	0.022	0.018	0.016
KFmo-30	−0.001	−0.002	0.003	0.008	−0.001
KFrh-10	0.025	0.022	0.020	0.011	0.007
KFrh-10ave30	0.016	0.016	0.012	0.002	0.012
KFrh-30	0.008	0.008	0.011	0.014	−0.002
KFmorh-10	0.019	0.027	0.030	0.021	0.014
KFmorh-10ave30	0.016	0.034	0.032	0.018	0.019
KFmorh-30	0.016	0.013	0.020	0.013	0.002

For the KF scheme, the best improvement for MO was 0.021 at the 6.35-mm threshold, and 0.030 at the same threshold with RH. These results suggest that if ETS is compared between models having differing horizontal grid spacings, there will be a tendency for the ETSs to be lower for the model with finest resolution, if the verification is performed on each model's native grid. As suggested by Tustison et al. (2001), this implies the larger representativeness error at higher resolutions is having a stronger negative impact than any true forecast improvement in the model.

Results are even more interesting when a 30-km version of the Eta Model is run for the same cases. As Table 1 shows, the best ETSs in the BMJ runs at all thresholds occurred in the 30-km run (rows marked with the −30 suffix). These improvements in ETSs over those in the 10-km verification were statistically significant at either the 90% or 95% confidence level for all thresholds of 6.35 mm or less. (Note that increased variability among cases can result in less statistical significance despite having larger increases in the average ETSs between some runs at some thresholds.) For heavier thresholds (where the smaller sample size requires caution in interpretation), the substantial improvements in the 30-km ETSs when compared with the 10-km ETSs remained, unlike in the 10-km results averaged onto a 30-km grid. The improvements were as large as 0.031 for one BMJ variant (not shown) at the heaviest (25.4 mm) threshold. Table 2 shows that BIA actually decreased noticeably for the 30-km runs for the lighter two thresholds. The ETS improvement here was thus not related to overprediction of rainfall area. For the heaviest two thresholds, the BIA in the 30-km runs increased substantially, possibly explaining the increases showing up in ETS (Mason 1989) for those amounts.

In contrast, when the KF scheme was used, the changes in ETSs between the 10- and 30-km model runs were very different (Table 1). The 30-km model ETS exhibited almost no change from the 10-km ETS for the lightest threshold and was substantially less than that obtained when 10-km results were averaged to a 30-km grid. At other thresholds, only slight improvements in ETS occurred in comparison with the 10-km output and 10-km output averaged to 30 km. BIA scores (Table 2) generally did not change as much as in the BMJ runs and, at all but the lightest threshold, were closer to 1.0 than with the other verifying grids.

The difference in the impacts of the size of the verifying grid box between the different convective schemes is consistent with the findings of Gallus (1999). That study found that the BMJ scheme tended to aggressively dry the atmosphere after activation, minimizing grid-scale contributions of rainfall. Thus, peak rainfall amounts did not vary as horizontal resolution was refined nearly as much as what occurred when the KF scheme was used. The results of the current study suggest that the KF scheme likely benefits more than the BMJ scheme does from the improved resolution of mesoscale circulations when horizontal grid resolution is refined. This result is consistent also with Stensrud et al. (1999) who found that the cold pool scheme seemed to have a bigger impact when the KF scheme was used than when the BMJ scheme was used, because the KF scheme includes a parameterized downdraft that is able to sustain a cold pool. The KF scheme likely is better able to generate small-scale circulations that, on the one hand, may be helpful in some cases allowing that scheme to generate a realistic precipitation forecast, but on the other hand, are more negatively affected when horizontal grid spacing is insufficiently fine.

Table 1 also shows impacts on ETSs for runs where both convective schemes were alternated within the same run. As discussed in Gallus and Segal (2001), this configuration often generated higher ETSs than either the BMJ or KF scheme alone for lighter thresholds. The ETSs here show a pattern that seems to be an average of the individual BMJ and KF results. The 30-km Eta output generally verified with the highest ETSs but without the statistical significance of the BMJ runs alone. Trends in BIA score (Table 2) between the 30- and 10-km runs followed the trends that occurred in the BMJ runs.

Although in general the above results suggest that higher ETSs are favored in 30-km simulations than in 10-km simulations, it should be noted that one model variant received much lower ETSs at 30 km than at 10 km. As expected, the run using no convective scheme (fully explicit run), which tended to have much lower ETSs than other variants no matter what verification grid-box size was used, had significantly worse forecasts at 30 km. The ETSs improved in a statistically significant way for thresholds of 2.54 mm or less when the 10-km results were averaged and verified on a 30-km grid but were statistically significantly worse in the 30-km runs. BIA scores (Table 2) declined in the 30-km results relative to the 10-km results for all but the heaviest threshold.

The sensitivity to grid-box size shown above raises questions about the generality of results that compare different models or model variants. For instance, Gallus and Segal (2001) found that impacts of improved initialization of mesoscale features on ETSs in a 10-km grid spacing model were seldom large and varied greatly among cases, so that averaged impacts were small. When the BMJ scheme was used, (i) CP generally had no effect; (ii) MO did improve ETSs, particularly for heavier rainfall amounts; and (iii) RH resulted in the largest, most consistent improvements, which were statistically significant for rainfall amounts of 6.35 mm or less. The same pattern is evident in the present study, which uses a subset of the 54 Gallus and Segal cases (rows in Table 3 identified with the −10 suffix can be directly compared with those of Gallus and Segal). Table 3 reveals that despite the sensitivity of ETS to the verifying grid resolution, the use of different verifying grid-box sizes did not markedly change the findings of Gallus and Segal (2001) regarding the role of these mesoscale initialization adjustments. For most variants, changes in the ETSs in comparison with appropriate control runs were comparable between runs verified on the model's 10-km grid (−10 suffix), runs in which the 10-km output was averaged onto a 30-km grid for verification (−10ave30 suffix), and 30-km-grid-spacing Eta runs verified on the same 30-km grid (−30 suffix). The most noticeable exceptions were for runs using the KF scheme with the MO or RH variants. For these runs, improvements in ETSs from the control run tended to be somewhat smaller in the 30-km version of the model than in the 10-km version of the model.

4. Discussion and conclusions

The impact of verification grid-box size on ETSs was evaluated by studying Eta Model output from 11 warm-season convective system events. The events composed a subset of those studied by Gallus and Segal (2001). Likewise, a subset of 12 variants of the Eta Model was also chosen from that study for analysis in the present note. Verification of rainfall amounts was performed on simulations that used 10-km grid spacing, simulations that used 30-km grid spacing, and on the 10-km output averaged onto a 30-km grid. In all cases, the observations were taken from NCEP's stage-IV precipitation analysis, which has a 4-km grid spacing. The observations were areally averaged onto the model's own grid, either 10 or 30 km.

An evaluation of model results showed that ETSs increased substantially for light rainfall thresholds when the 10-km output was averaged onto a 30-km grid for verification instead of having the verification take place on the model's own 10-km grid. The increase was nearly as large as any improvement seen in ETSs when initial conditions were modified to better represent mesoscale features. When a 30-km version of the model was compared with the 10-km version, the BMJ runs experienced substantial increases in ETSs for all rainfall thresholds, generally more so than what happened when 10-km results were averaged onto 30-km grids. The improved ETSs for lighter thresholds were also associated with less of a high bias. When the KF scheme was used, some improvement in ETS was noted in the 30-km output when compared with the 10-km output, but the changes were less than with the BMJ scheme, especially for the lightest rainfall threshold.

These results suggest that it may be difficult to show improvements in ETSs for models with increasingly fine horizontal resolution if the verification is done on the model's own grid using point-to-area conversion. Fine-resolution models in effect are being penalized more for small spatial errors (disagreements between location of observed and forecast rain regions) than coarser models. Back-of-the-envelope calculations with a simple skill measure such as mean absolute error (MAE) show that MAE will always be better (lower) when it is computed using averaged values instead of the nonaveraged ones, unless the sign of the error is the same at all grid points.

The surprisingly higher skill scores found for 30-km output in comparison with both the 10-km output and the 10-km results averaged onto a 30-km grid also raise questions about possible negative impacts in the general forecast from model depiction of spurious small-scale features. Szunyogh and Toth (2002) have found that a truncation of a global model (i.e., coarsening of the grid spacing) after 3 days can improve the longer-range forecasts from a single control model, because it eliminates small-scale features that are unpredictable at those time ranges and that can adversely affect the general forecast. The results in the present note suggest that similar behavior may be present in short-range convective system rainfall forecasts simulated at high spatial and temporal resolution. It is possible that spurious very small scale features that have poor predictability adversely affect skill scores of a 10-km forecast, even when averaged onto a 30-km grid. Such spurious features may in effect be filtered from the 30-km version of the model. Traditional skill measures may be higher for those grid spacings that are coarse enough to effectively ''filter'' out these phenomena, which are as yet not consistently predictable using current model configurations.

It is also possible, as suggested by Tustison et al. (2001), that the worsening in representativeness error as resolution is refined outweighs any improvements in the actual forecast, so that the ETSs are worse for finer-resolution verification. In addition, the results may reflect problems with the use of common convective parameterizations at grid spacings on the order of 10 km, which are finer than those for which the schemes were originally designed and where the parameterizations are physically justifiable (Molinari and Dudek 1992). Last, as suggested by Mass et al. (2002), it is possible that the higher ETSs for the 30-km model in comparison with those for the 10-km model reflect inadequacies in the use of ETS as a verification approach in high-resolution models. That study stated that the realism of small-scale features was generally better in the finer-resolution models despite objective accuracy measures changing little.

The large impact of the resolution of the verifying analysis on objective accuracy measures suggests that caution be used when comparing results from different model configurations and among different studies. Verification (using point-to-area conversion) performed on a 10-km grid will likely yield lower scores than that performed on a coarser grid (to a certain point; eventually the coarseness of the grid will harm the ability of the model to resolve mesoscale forcing mechanisms important in the generation of rainfall). The results presented in this note suggest that ETSs will improve if 10-km output is averaged onto a 30-km grid. It is unclear if such trends would remain consistent when going from 30-km grid spacing, for instance, to 80 km (e.g., Wandishin et al. 2001).

Despite the changes in ETSs that occur when the resolution of the verification data is changed, the impacts of the initialization modifications discussed in Gallus and Segal (2001) remained unchanged.

Acknowledgments. This work was sponsored by the National Science Foundation/National Oceanic and Atmospheric Administration Grant ATM-9908932 in the USWRP program.

REFERENCES

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

Betts, A. K., 1986: A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quart. J. Roy. Meteor. Soc.,* **112,** 677–692.

——, and M. J. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, and arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.,* **112,** 693–709.

Black, T. M., 1994: The new NMC mesoscale Eta Model: Description and forecast examples. *Wea. Forecasting,* **9,** 265–278.

Brooks, H. E., C. A. Doswell, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting,* **7,** 120–132.

Gallus, W. A., Jr., 1999: Eta simulations of three extreme precipitation events: Impact of resolution and choice of convective parameterization. *Wea. Forecasting,* **14,** 405–426.

——, and M. Segal, 2001: Impact of improved initialization of mesoscale features on convective system rainfall in 10-km Eta simulations. *Wea. Forecasting,* **16,** 680–696.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.,* **122,** 927–945.

Kain, J. S., and J. M. Fritsch, 1992: The role of the convective ‘‘trigger function’’ in numerical forecasts of mesoscale convective systems. *Meteor. Atmos. Phys.,* **49,** 93–106.

——, and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: The Kain–Fritsch scheme. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.,* No. 46, Amer. Meteor. Soc., 165–170.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Meteor. Mag.,* **37,** 75–81.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.,* **83,** 407–430.

McDonald, B. E., and J. D. Horel, 1998: Evaluation of precipitation forecasts from the NCEP’s 10 km mesoscale Eta Model. Preprints, *12th Conf. on Numerical Weather Prediction,* Phoenix, AZ, Amer. Meteor. Soc., J27–J30.

Molinari, J., and M. Dudek, 1992: Parameterization of convective precipitation in mesoscale numerical models: A critical review. *Mon. Wea. Rev.,* **120,** 326–344.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.,* **119,** 1590–1601.

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting,* **10,** 498–511.

Rogers, E., and Coauthors, 1998: Changes to the NCEP operational ‘‘early’’ eta analysis/forecast system. NWS Tech. Procedures Bull. 447, National Oceanic and Atmospheric Administration/National Weather Service, 14 pp. [Available from Office of Meteorology, National Weather Service, 1325 East–West Highway, Silver Spring, MD 20910.]

Rogers, R. F., J. M. Fritsch, and W. C. Lambert, 2000: A simple technique for using radar data in the dynamic initialization of a mesoscale model. *Mon. Wea. Rev.,* **128,** 2560–2574.

Ruggiero, F. H., K. D. Sashegyi, R. V. Madala, and S. Raman, 1996: The use of surface observations in four-dimensional data assimilation using a mesoscale model. *Mon. Wea. Rev.,* **124,** 1018–1033.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Stensrud, D. J., and J. M. Fritsch, 1994: Mesoscale convective systems in weakly forced large-scale environments. Part III: Numerical simulations and implications for operational forecasting. *Mon. Wea. Rev.,* **122,** 2084–2104.

——, G. S. Manikin, E. Rogers, and K. E. Mitchell, 1999: Importance of cold pools to NCEP mesoscale Eta Model forecasts. *Wea. Forecasting,* **14,** 650–670.

——, J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.,* **128,** 2077–2107.

Szunyogh, I., and Z. Toth, 2002: The effect of increased horizontal resolution on the NCEP global ensemble mean forecasts. *Mon. Wea. Rev.,* **130,** 1125–1143.

Tustison, B., D. Harris, and E. Foufoula-Georgiou, 2001: Scale issues in verification of precipitation forecasts. *J. Geophys. Res.,* **106,** 11 775–11 784.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.,* **129,** 729–747.

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