NOTES AND CORRESPONDENCE

Evaluation of Probabilistic Precipitation Forecasts Determined from Eta and AVN Forecasted Amounts

WILLIAM A. GALLUS JR.

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

MICHAEL E. BALDWIN* AND KIMBERLY L. ELMORE

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

(Manuscript received 6 October 2005, in final form 19 June 2006)

ABSTRACT

This note examines the connection between the probability of precipitation and forecasted amounts from the NCEP Eta (now known as the North American Mesoscale model) and Aviation (AVN; now known as the Global Forecast System) models run over a 2-yr period on a contiguous U.S. domain. Specifically, the quantitative precipitation forecast (QPF)–probability relationship found recently by Gallus and Segal in 10-km grid spacing model runs for 20 warm season mesoscale convective systems is tested over this much larger temporal and spatial dataset. A 1-yr period was used to investigate the QPF–probability relationship, and the predictive capability of this relationship was then tested on an independent 1-yr sample of data. The same relationship of a substantial increase in the likelihood of observed rainfall exceeding a specified threshold in areas where model runs forecasted higher rainfall amounts is found to hold over all seasons. Rainfall is less likely to occur in those areas where the models indicate none than it is elsewhere in the domain; it is more likely to occur in those regions where rainfall is predicted, especially where the predicted rainfall amounts are largest. The probability of rainfall forecasts based on this relationship are found to possess skill as measured by relative operating characteristic curves, reliability diagrams, and Brier skill scores. Skillful forecasts from the technique exist throughout the 48-h periods for which Eta and AVN output were available. The results suggest that this forecasting tool might assist forecasters throughout the year in a wide variety of weather events and not only in areas of difficult-to-forecast convective systems.

1. Introduction

Hamill and Colucci (1997) showed that for ensemble simulations of precipitation, the probability of occurrence of precipitation increases with increased forecasted probability. These more successful forecasts of precipitation occurrence can be attributed to the fact that the ensemble variability in the initialization and/or physical formulation has not affected the prediction of precipitation by multiple ensemble members at a given grid point. Because they provide probabilistic forecasts that may be of more value to users than a deterministic forecast, ensemble forecasts are increasingly being used by operational forecasters. However, such forecasts require multiple simulations to be performed, such that computational costs may restrict the creation of ensemble forecasts to operational centers or a few research institutions.

Nearly two decades ago, Wilks (1990) explored the relationship between quantitative precipitation amounts and probability forecasts. Specifically, he determined that heavier precipitation amounts were more likely to occur when the subjectively forecasted probability of precipitation was high than when the forecasted probability was low. Gallus and Segal (2004) addressed the reverse situation—the relationship between

* Current affiliation: Department of Earth and Atmospheric Sciences, Purdue University, West Lafayette, Indiana.

Corresponding author address: William A. Gallus Jr., Iowa State University, 3025 Agronomy, Ames, IA 50011.
E-mail: wgallus@iastate.edu

DOI: 10.1175/WAF976.1

the probability of rainfall occurring and the quantitative precipitation amount forecasted by a model. Examining 10-km grid spacing forecasts of 20 convective events, they showed that in subdomains consisting of model grid points at which large amounts of precipitation are predicted, the probability of experiencing a lighter rain amount was higher than that valid for the entire simulation domain. In addition, they suggested that skillful probabilistic forecasts over the entire domain could be issued based on a quantitative precipitation forecast (QPF) amount. They argued that this relationship might assist in the operational forecasting of precipitation, particularly for warm season events for which objective skill measures are generally very low.

The present study extends the conclusions of Gallus and Segal (2004) to a much larger dataset having coarser grid spacing. Specifically, the study will (i) investigate the relationship between the likelihood of occurrence of precipitation and the forecasted precipitation amount, and (ii) investigate the predictive capability of this relationship, as an approach for creating probabilistic forecasts of precipitation occurrence based on the output of a single model. Simulated 3-hourly accumulated precipitation interpolated to a 40-km grid from the National Centers for Environmental Prediction (NCEP) Eta [now referred to as the North American Mesoscale (NAM)] model (Mesinger et al. 1988; Janjic 1994; Rogers et al. 2001) and Aviation [AVN, now referred to as the Global Forecast System (GFS)] models (Global Climate and Weather Modeling Branch 2003) for a 1-yr period running from 1 September 2002 to 31 August 2003 is examined to determine the relationship between QPF amount and the probability of precipitation. These relationships are then applied to an independent dataset for a 1-yr period from 1 September 2003 to 31 August 2004. The discrimination ability, reliability, and accuracy of these probability forecasts are verified using relative operating characteristic (ROC) curves, reliability diagrams, and Brier skill scores. As in Gallus and Segal (2004), the two models have different bias characteristics, and they use different cumulus parameterization schemes: the Betts–Miller–Janjić (Betts and Miller 1986; Janjić 1994) scheme in the Eta and a simplified Arakawa–Schubert scheme (Pan and Wu 1995; Grell 1993; Arakawa and Schubert 1974) in the AVN. It should also be noted that the present study evaluates 3-hourly accumulated precipitation, which is more difficult to forecast than the 6-hourly accumulations examined by Gallus and Segal.

2. Data and methodology

To achieve the goals outlined above, conditional probabilities of precipitation must first be estimated, followed by the verification of forecasts based on these probabilities. To determine conditional probabilities, Eta and AVN simulations run operationally at NCEP initialized at both 0000 and 1200 UTC during the period 1 September 2002–31 August 2003 were used. Forecasts from both models were archived through 48 h, and the evaluation examined separately accumulated precipitation in 3-h periods within the first (hereafter called day 1) and second (day 2) 24 h of the forecast. The domain of the archived model output covered the contiguous United States.

Conditional probabilities of rainfall were determined by comparing the model predictions for 3-h periods with 4-km horizontal resolution NCEP stage IV precipitation observations (Baldwin and Mitchell 1997). Multisensor stage IV output, which includes both radar and gauge observations, was used. The observations were areally averaged onto the 40-km grid for which model output was available using procedures similar to those used at NCEP.

To evaluate the predictive capability of using the conditional probabilities, 0000 and 1200 UTC model runs from 1 September 2003 through 31 August 2004 were used and the resulting forecasts compared with stage IV observations for this period. It is important to note that both the Eta and AVN models were undergoing minor changes during both time periods. Ideally, the relationship between forecasted rainfall amounts and conditional probabilities should be determined from static models and applied to the same models. Changes in the models may affect the performance of the QPF–probability relationship.

3. Results

a. Analysis approach

To determine if the probability of precipitation (PoP) varies directly with the amount of precipitation predicted, we compute the conditional probability of a specified observed precipitation event, given a forecast of precipitation within a predetermined range of values (QPF bin). The observed precipitation events were defined as 3-h accumulated precipitation exceeding three threshold amounts: 0.01, 0.10, and 0.25 in. (0.01 in. = 0.254 mm). QPF bins were chosen to generally match standard operational verification thresholds, including <0.01 (no rain), 0.01–0.05, 0.05–0.10, 0.10–0.25, 0.25–0.50, and ≥ 0.50 in. $(3\ \text{h})^{-1}$. Using the contingency table for a given observed event and QPF bin (Table 1), the PoP is defined by $a/(a + b)$ where $a + b$ is the total number of grid points at which precipitation is forecasted to fall within the specified QPF bin, and a represents the number of "hits"—those grid points at

TABLE 1. Contingency table for a given event.

		Obs		
		Yes	No	Tot
Forecast	Yes	a	b	$a + b$
	No	c	d	$c + d$
	Tot	$a + c$	$b + d$	$a + b + c + d$

which a specified observed precipitation event also occurred.

Once the QPF–PoP relationship is established, the probabilities can be verified using other quantities computed from the traditional 2×2 contingency table (Table 1) for dichotomous forecasts. To verify probabilities, a "yes" forecast is given at each point where the PoP exceeds a given threshold value. A yes observed event is given whenever the observed precipitation exceeds a specified threshold value. The probability of detection (POD) is given by $a/(a + c)$, where $a + c$ is the total number of grid points where the observed precipitation event occurred, and a is the number of correct yes forecasts. The probability of false detection (POFD), defined as $b/(b + d)$, indicates the ratio of the area where an event was predicted to occur but was not observed (b), to the area where the event was not observed ($b + d$). Using the PoP values corresponding to the QPF bins, ROC curves were computed where POD is plotted as a function of POFD for yes–no forecasts made based on forecast probability thresholds that vary from 0% to 100%. ROC curves indicate the ability of a forecast to distinguish between observed events and nonevents, based on various decision thresholds. Using a bootstrap methodology (see the appendix for details), mean and 95% confidence intervals of ROC areas were calculated for each probabilistic forecast.

b. Relationship between PoP and QPF

Estimated PoPs for observed precipitation events exceeding the thresholds 0.01, 0.10, and 0.25 in. during 3-hourly periods taken from the first 24 h of a forecast, for the Eta, AVN, and a simple average of the two (AVG) predicted precipitation amounts within specified bins, are shown in Table 2. In both models, PoPs rise with increasing QPF amount. In both models, the estimated probability of any precipitation being observed when the forecast is for less than 0.01 in. is less than 5%. The probability for greater than 0.25 in. is less than 0.5%. The probabilities rise steadily as the forecasted amounts increase toward 0.5 in. or greater. For QPF amounts exceeding 0.5 in., the probability of any measurable precipitation is roughly 80% or greater in both models. The probability of greater than 0.25 in. exceeds 30% in both models.

Table 2 also shows the sample climatology for the three observed precipitation events. For all model configurations and thresholds, the sample climatology lies between the PoP associated with zero QPF and the PoP associated with a QPF of 0.01 in. or more. Thus, precipitation is less likely to occur in those areas where the models indicate no precipitation than it is elsewhere in the domain; it is more likely to occur in those regions where precipitation is predicted, especially where the predicted precipitation amounts are largest.

During the day 2 forecast period (Table 3), the estimated PoPs generally show the same trends as during day 1, although the PoPs associated with zero QPF increase slightly as compared with day 1. The strength of the association between QPF and PoP has also decreased, because the probability of observing rain when heavier rain is forecasted is not as high as it is in the day 1 period. Both trends are consistent with decreasing forecast skill for longer-range forecasts. The peak probability of measurable precipitation is around 70%, when the QPF is greater than 0.50 in.

c. Verification of PoP forecasts

Probabilistic forecasts are typically evaluated using reliability and ROC diagrams, and measures of accuracy such as the Brier score. Figure 1 shows reliability diagrams for observed events based upon rainfall thresholds of 0.01, 0.10, and 0.25 in. for the day 1 forecasts from the Eta, AVN, and AVG, valid during the 1 September 2003–31 August 2004 period using the PoP values shown in Table 2. Reliability diagrams show the relative frequency a given event is observed as a function of forecast probability. A perfectly reliable forecast will be observed with the same frequency as is predicted, and fall along the main diagonal of the reliability diagram. This figure shows that all of the probability forecasts obtained from the QPF–PoP relationship are almost perfectly reliable; that is, the observed relative frequency of each event in the September 2003–August 2004 period is almost identical to that found 1 yr earlier (and used as the basis for the QPF–PoP association).

A commonly used measure of accuracy for probability forecasts is the Brier score (Brier 1950), which is basically the mean-squared error of the probability forecasts (here given in %):

$$\mathrm{BS} = \frac{1}{n}\sum_{k=1}^{n}(p_k - o_k)^2, \tag{1}$$

where, for a given case k of n total cases, p_k is the forecast probability and o_k is the observed probability ($o_k = 100\%$ if the event occurs, $o_k = 0\%$ if the event

TABLE 2. Estimated PoP (%) exceeding thresholds of 0.01, 0.10, and 0.25 in. for 3-hourly predicted rainfall amounts in the specified ranges during the day 1 period, 1 Sep 2002–31 Aug 2003. The sample climatology (observed frequency) is given in the first column for each threshold. Results are presented for the Eta, AVN, and AVG (average of Eta and AVN QPFs).

Obs rainfall threshold (in)	Sample climatology (%)	PoP (%) for given predicted rainfall amount (in.)					
		<0.01	0.01–0.05	0.05–0.10	0.10–0.25	0.25–0.50	>0.50
				Eta			
0.01	9.6	4.7	27.1	43.5	58.4	74.7	85.3
0.10	2.5	0.8	5.6	11.7	22.9	42.5	59.9
0.25	0.9	0.3	1.9	4.1	8.7	19.6	34.3
				AVN			
0.01	9.6	3.8	23.2	42.4	58.4	75.5	78.5
0.10	2.5	0.6	4.3	10.9	22.6	43.7	53.2
0.25	0.9	0.2	1.5	3.6	8.1	19.9	31.2
				AVG			
0.01	9.6	3.3	23.9	45.5	65.2	81.5	83.1
0.10	2.5	0.5	4.4	12.1	27.3	51.6	60.9
0.25	0.9	0.2	1.5	4.1	10.3	24.8	38.9

does not occur). As reviewed by Wilks (1995, p. 259), a skill score, known as the Brier skill score (BSS) in the form of

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \tag{2}$$

is often computed in order to provide information on the accuracy of the forecasts relative to some standard. Usually, $\mathrm{BS}_{\mathrm{ref}}$ is computed using Eq. (1) with p_k equal to the climatological event frequency. Murphy (1973) showed how Eq. (1) could be partitioned into three components, measuring the degree of reliability, resolution, and uncertainty in the forecasts and observations. Here, the verification dataset is assumed to contain a discrete number I of probability forecast values, where N_i is the number of cases in the ith forecast category. For each forecast category, the average relative frequency of the observed events is computed:

$$\bar{o}_i = \frac{1}{N_i} \sum_{k \in N_i} o_k. \tag{3}$$

Also, the overall sample climatology of the observed event is computed:

$$\bar{o} = \frac{1}{n} \sum_{k=1}^{n} o_k. \tag{4}$$

Given these, Eq. (1) can be rewritten as

$$\mathrm{BS} = \frac{1}{n} \sum_{i=1}^{I} N_i (p_i - \bar{o}_i)^2 - \frac{1}{n} \sum_{i=1}^{I} N_i (\bar{o}_i - \bar{o})^2 + \bar{o}(1 - \bar{o}), \tag{5}$$

TABLE 3. As in Table 2 but for 3-hourly simulated rainfall amounts in the specified ranges during the day 2 period.

Obs rainfall threshold (in.)	Sample climatology (%)	PoP (%) for a given predicted rainfall amount (in.)					
		<0.01	0.01–0.05	0.05–0.10	0.10–0.25	0.25–0.50	>0.50
				Eta			
0.01	9.6	5.5	25.5	38.8	50.5	63.3	71.1
0.10	2.5	1.1	5.8	10.7	18.7	31.1	41.2
0.25	0.9	0.4	2.0	3.9	7.3	13.3	19.6
				AVN			
0.01	9.6	4.6	21.8	37.7	50.5	64.0	66.9
0.10	2.5	0.9	4.7	10.4	18.9	32.3	39.1
0.25	0.9	0.3	1.6	3.7	7.1	14.1	20.2
				AVG			
0.01	9.6	4.0	22.9	41.0	56.6	70.8	71.2
0.10	2.5	0.7	4.9	11.8	22.9	38.7	45.1
0.25	0.9	0.2	1.7	4.3	9.0	17.6	24.2

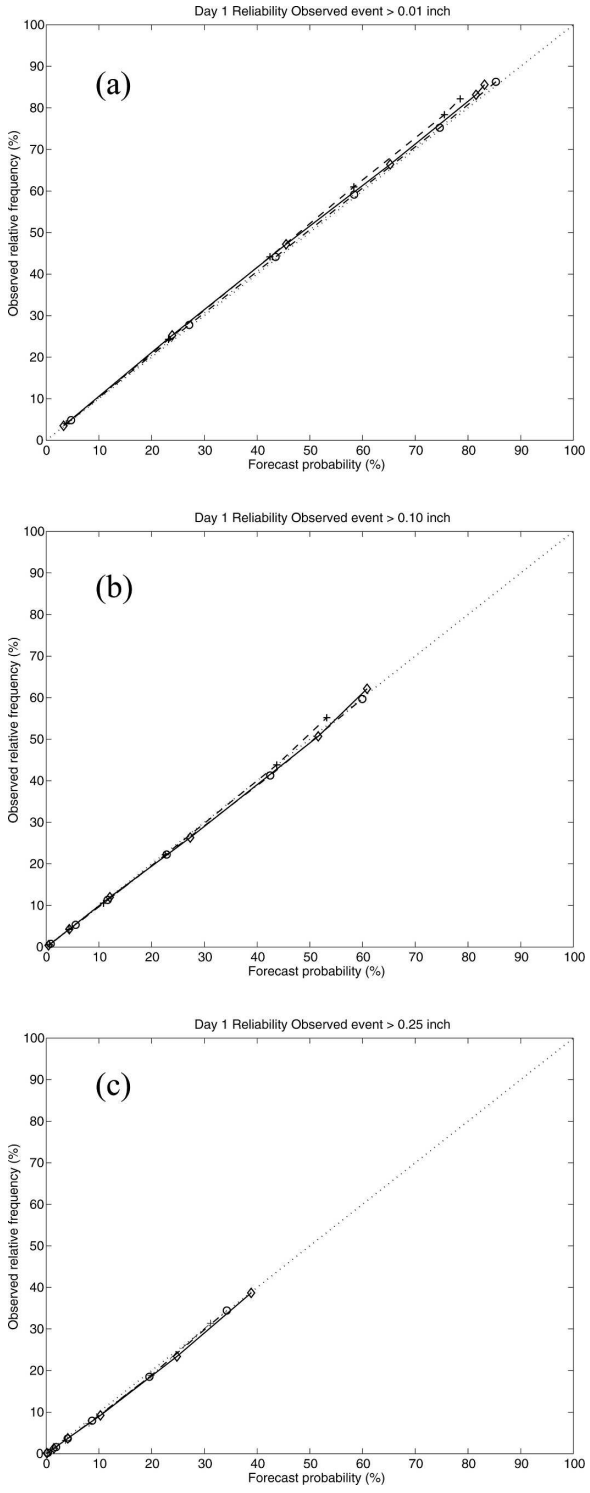

FIG. 1. Reliability diagrams for the day 1 forecast period for the AVN (dashed line with plus signs), Eta (dash–dot line with circles), AVG (solid line with diamonds), and perfect reliability curve (dotted line) for (a) >0.01, (b) >0.10, and (c) >0.25 in. observed events.

TABLE 4. Accuracy of day 1 PoP (%) forecasts as measured by the Brier score and BSS. The uncertainty, reliability, and resolution components of the Brier score, as decomposed by Murphy (1973), are also given.

	Obs rainfall threshold (in.)	AVN	Eta	AVG
Brier score	0.01	6.66	6.95	6.43
	0.10	1.87	1.93	1.83
	0.25	0.70	0.71	0.69
BSS	0.01	23.3	20.0	26.0
	0.10	14.8	12.4	16.8
	0.25	7.3	6.1	8.4
Uncertainty	0.01	8.69	8.69	8.69
	0.10	2.20	2.20	2.20
	0.25	0.76	0.76	0.76
Reliability	0.01	5.1×10^{-3}	0.8×10^{-3}	3.9×10^{-3}
	0.10	0.2×10^{-3}	0.3×10^{-3}	0.3×10^{-3}
	0.25	0.4×10^{-3}	0.4×10^{-3}	0.5×10^{-3}
Resolution	0.01	2.03	1.73	2.26
	0.10	0.33	0.27	0.37
	0.25	0.06	0.05	0.06

where the first summation term is the reliability term, the second summation term is the resolution, and the last term is the uncertainty. The reliability term quantifies the information provided in the reliability diagram (Fig. 1). This is the weighted average of the squared differences between the forecast probabilities and the relative frequencies of the observed event, across I forecast categories. The resolution term provides information on the forecast system's ability to sort events into subsamples with different relative frequencies. The uncertainty term is a function of the observed sample climatology alone and quantifies the variability of the observed events. The uncertainty term is equal to BS when Eq. (1) is computed using the climatological event frequency as p_k.

Table 4 summarizes the accuracy of the probability forecasts for the day 1 period using the Brier score partitioning. The probability forecasts from the Eta Model display the least amount of accuracy for all observed events, and the forecasts using the average of the Eta and AVN QPFs are shown to be the most accurate (lowest Brier score). The BSS values show over 20% improvement in accuracy over using a sample climatology for the >0.01 in. observed event, with the AVG probability forecasts showing a BSS of 26%. BSS values decrease as the threshold for the observed event increases, demonstrating a decrease in accuracy relative to the climatology for these probabilistic forecasts for heavier rain events. The nearly perfect reliability of

TABLE 5. As in Table 4 but for day 2 PoP forecasts.

	Obs rainfall threshold (in.)	AVN	Eta	AVG
Brier score	0.01	7.27	7.48	7.03
	0.10	2.00	2.03	1.96
	0.25	0.73	0.73	0.72
BSS	0.01	16.4	13.9	19.1
	0.10	9.0	7.6	10.8
	0.25	4.0	3.4	4.9
Uncertainty	0.01	8.69	8.69	8.69
	0.10	2.20	2.20	2.20
	0.25	0.76	0.76	0.76
Reliability	0.01	1.2×10^{-3}	0.3×10^{-3}	1.2×10^{-3}
	0.10	0.6×10^{-3}	0.4×10^{-3}	0.5×10^{-3}
	0.25	0.5×10^{-3}	0.4×10^{-3}	0.6×10^{-3}
Resolution	0.01	1.42	1.21	1.66
	0.10	0.20	0.17	0.24
	0.25	0.03	0.03	0.04

these probability forecasts is quantified in the reliability terms for each of the forecast systems, with values on the order of 10^{-3}. The verification information shows that the QPF–PoP relationship is well calibrated. The improvement of these forecast systems over climatology is primarily found in the systems' ability to resolve situations where the likelihood of an observed event is more (or less) than the overall sample climatology. The resolution term decreases substantially as the precipitation threshold for observed events increases, indicating the increasing difficulty in predicting heavier rainfall events. Table 5 displays similar accuracy information for the day 2 forecast period. The Brier scores are higher (and BSS lower) than the day 1 period, indicating a decrease in accuracy with a longer forecast range. Again, the forecasts are nearly perfectly reliable. BSS values are largest for the AVG probability forecasts for the >0.01 in. event, showing nearly 20% improvement over the sample climatology. BSS values drop to near 10% for the >0.10 in. event, and below 5% for the >0.25 in. event in the day 2 period.

Figure 2 shows ROC diagrams for the three different observed events (observed precipitation >0.01, 0.10, and 0.25 in.) for the day 1 forecasts for the Eta, AVN, and AVG. ROC diagrams summarize the ability of a forecast system to discriminate between observed events and nonevents. More discrimination ability is found for ROC curves closest to the upper left-hand corner of the plot, where POD = 1 and POFD = 0 indicates a perfect forecast. It can be seen in the figure that all three curves lie above the diagonal no-skill line (where false alarms are as likely as hits) for all three

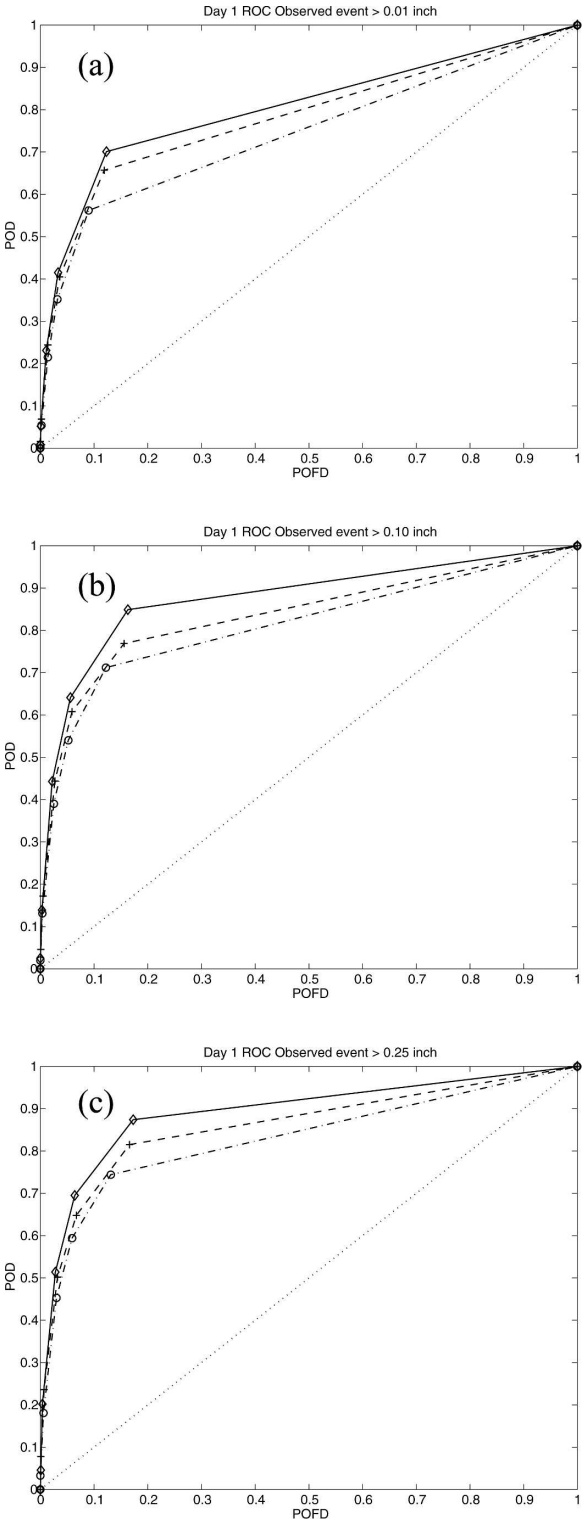

FIG. 2. ROC diagrams for the day 1 forecast period for (a) >0.01, (b) >0.10, and (c) >0.25 in. observed events. Lines and symbols are as in Fig. 1, except the dotted line indicates no skill curve.

TABLE 6. Areas under the ROC curves for the three forecasts for the day 1 period for the given 3-h precipitation thresholds.

	Obs event		
	>0.01 in.	>0.10 in.	>0.25 in.
AVG	0.80	0.87	0.89
AVN	0.78	0.85	0.86
Eta	0.74	0.81	0.83

TABLE 7. Same as Table 6 but for the day 2 forecast period.

	Obs event		
	>0.01 in.	>0.10 in.	>0.25 in.
AVG	0.76	0.83	0.84
AVN	0.74	0.79	0.80
Eta	0.70	0.75	0.77

observed events. Thus, in all cases, the area under the ROC curve calculated using the trapezoidal method, shown in Table 6, exceeds 0.5, implying the potential for a useful forecast (Buizza et al. 1999). The areas for the forecasts obtained by averaging the Eta and AVN QPF are noticeably higher than those for either model individually for all three thresholds shown. For many parameters, the ensemble mean has been shown to provide a more accurate forecast than a single deterministic forecast (Leith 1974). Table 7 also shows the magnitude of the area under the ROC curve for forecasts verifying in the day 2 period (24–48 h). The 95% confidence intervals determined using the bootstrap method (not shown) are very small; all differences between the means at a specified threshold for a given day are statistically significant with p values less than 0.001. While the differences are statistically significant, likely due to the large sample size, in practical terms the ROC areas and BSSs show the AVG forecast quality to be only slightly greater than the AVN, which is only slightly greater than the Eta.

For the day 1 period, the area under the ROC curve for the QPF–PoP relationship applied to the AVN output exceeds 0.8 and approaches 0.9 for the two heavier thresholds. Values from the relationship applied to the Eta output are more noticeably lower (~6%). The ROC areas based on the average of the QPFs in both models are higher than either model individually, but only by around 2% compared with the AVN forecasts. Ebert (2001) points out that a simple ensemble mean applied to the QPF leads to a large bias in rain area for light amounts and an underestimate of maximum rainfall. Despite these problems, the QPF–probability relationship worked well for the average of the Eta and AVN forecasts, suggesting that this technique may work well when applied to ensembles. In addition, Ebert (2001) has suggested that better methods to determine an ensemble mean for precipitation may exist, and it is possible that even more skill would be present if these methods were applied.

For the day 2 period the areas under the ROC curves for forecasts made using the QPF–probability technique decrease substantially, by roughly 0.05, for each threshold in both models. The technique applied to the average QPF evidences less of a decrease and becomes relatively more skillful compared with its application using individual models. For these data, the QPF-based technique performs significantly better when applied to AVN output than when applied to Eta output. As with day 1 forecasts, the technique shows the ability to discriminate more for heavier thresholds than for lighter ones.

The relationship shown in this study therefore appears to be robust and applicable throughout large regions at any time during the year. It could be used by forecasters in their standard issuance of subjectively determined probabilistic precipitation forecasts.

4. Conclusions

It was determined that the QPF amount–PoP relationship found to exist for warm season convective system rainfall in the Upper Midwest (Gallus and Segal 2004) is also present when output from the NCEP Eta and AVN models for 2 yr over the contiguous United States is evaluated. The estimated PoP exceeding a specified threshold increases substantially as the Eta and the AVN models predict increasingly heavier precipitation amounts. The estimated probabilities were determined from model QPF output for a 1-yr period, and then these PoPs were used as forecasts on an independent 1-yr set of Eta and AVN output. These probability forecasts were determined to be both reliable and skillful. Forecasters can be more confident of at least light amounts of precipitation occurring if either of these operational model runs produces heavy precipitation at a point. Additionally, at grid points where the model QPF amount is zero, precipitation is less likely to occur than the climatological PoP, computed as an average throughout the domain.

The skill of these PoP forecasts, shown in reliability and ROC diagrams as well as Brier scores, implies that both models are more likely to indicate the regions where atmospheric processes are most favorable for precipitation (where the models generate enhanced amounts) than they are able to accurately predict the actual amounts of observed precipitation. The QPF–

probability relationship evaluated in the present note can be used by forecasters as guidance for issuing probabilistic forecasts from a single deterministic forecast. In addition, forecasters can apply the technique to ensemble mean forecasts of rainfall. Future work should compare the skill of probabilistic forecasts based on this technique applied with ensemble mean precipitation with the skill from traditional ensemble methods that determine probabilities based upon the number of members indicating rainfall above a threshold. In addition, regional and seasonal analyses to determine if the applicability of the technique varies spatially or temporally would be beneficial to forecasters, and would ensure that the skill is not primarily related to variations in climatology across the large domain.

Acknowledgments. Software to perform some of the ROC computations was kindly provided by Matthew Wandishin. The paper was substantially improved by the helpful comments of three anonymous reviewers. This study was supported by NSF Grants ATM-0226059 and ATM-0537043.

APPENDIX

Bootstrap Methodology

For each day, each evaluated model has an associated ROC area, or area under the ROC curve, which is a measure of skill. Statistical significance testing of differences in the ROC areas associated with each model, and an average of the two models' QPF amounts, are performed using a permutation test (Efron and Tibshirani 1993). By definition, the difference between the areas over each day is not normally distributed because the values lie within the interval [0, 1], and the probability density function will be far more dense within the interval [0.5, 0]. A permutation test, particularly a matched-pairs permutation test, of the difference between means is an ideal instrument to determine whether these differences are statistically significant. The permutation test may be thought of as an analog to a *t* test of the difference between means. An advantage of the permutation test is that it is exact (in the limit of using all possible permutations) and is completely nonparametric.

As an example, let the contingency table elements required to generate the ROC area for each day of the Eta and AVN model day 1 forecasts be placed into vectors $\mathbf{e} = e_1, e_2, \ldots, e_n$ and $\mathbf{a} = a_1, a_2, \ldots, a_n$, respectively. Let $\mathbf{d} = \mathbf{e} - \mathbf{a}$, so $\mathbf{d} = d_1, d_2, \ldots, d_n$, where $d_1 = e_1 - a_1$, $d_2 = e_2 - a_2$, etc., and let the mean of $\mathbf{d}$ be $\bar{d}$. The permutation test uses, as the null hypothesis (H_0), that the data in **e** and **a** are drawn from the same distribution (or, at least, distributions that will provide the same mean). Thus, under the null hypothesis, the value of $\bar{d}$ is unaffected by a random reassignment of the membership associated with each element in **d**.

Let each permuted mean of **d** be denoted by $\bar{d}_i^*$, where i ranges from 1 to B, and B is the number of permutations taken (5000 for the data in the present study). The achieved significance level (ASL) is then $\text{ASL} = \Pr_{H_0}(\bar{d}_i^* \geq \bar{d})$, which is simply the number of $(\bar{d}_i^* \geq \bar{d})/B$.

REFERENCES

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large-scale environment, Part I. *J. Atmos. Sci.,* **31,** 674–701.

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

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.,* **78,** 1–3.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting,* **14,** 168–189.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.,* **129,** 2461–2480.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap*. Chapman and Hall, 436 pp.

Gallus, W. A., Jr., and M. Segal, 2004: Does increased predicted warm season rainfall indicate enhanced likelihood of rain occurrence? *Wea. Forecasting,* **19,** 1127–1135.

Global Climate and Weather Modeling Branch, 2003: The GFS Atmospheric Model. NOAA/NWS/NCEP Office Note 442, 14 pp. [Available online at http://www.emc.ncep.noaa.gov/officenotes/newernotes/on442.pdf.]

Grell, G. A., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.,* **121,** 764–787.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

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,** 928–945.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Mesinger, F., Z. I. Janjić, S. Nickovic, D. Gavrilov, and D. G. Deaven, 1988: The step-mountain coordinate: Model description and performance for cases of alpine lee cyclogenesis and for a case of an Appalachian redevelopment. *Mon. Wea. Rev.,* **116,** 1493–1518.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **10,** 155–156.

Pan, H.-L., and W.-S. Wu, 1995: Implementing a mass flux

convection parameterization package for the NMC Medium-Range Forecast Model. NMC Office Note 409, 40 pp. [Available from NCEP, 5200 Auth Rd., Washington, DC 20233.]

Rogers, E., T. Black, B. Ferrier, Y. Lin, D. Parrish, and G. DiMego, cited 2001: Changes to the NCEP Meso Eta Analysis and Forecast System: Increase in resolution, new cloud microphysics, modified precipitation assimilation, modified 3DVAR analysis. NWS Tech. Procedures Bull. [Available online at http://www.emc.ncep.noaa.gov/mmb/mmbpll/eta12tpb/.]

Wilks, D. S., 1990: Probabilistic quantitative precipitation forecasts derived from PoPs and conditional precipitation amount climatologies. *Mon. Wea. Rev.,* **118,** 874–882.

——, 1995: *Statistical Methods in the Atmospheric Sciences*. Cambridge University Press, 547 pp.