

Improved Probability Forecasting of Weather and Weather-Related Variables

Pfeffer, Niu and Gahrs Florida State University (FSU)

[[Download PDF version of this document.](#)]

Introduction

We have developed and tested a new method of probability forecasting that provides greater skill than standard methods in use today. The method involves a statistical combination of a suite of linear and nonlinear statistical techniques applied to a set of relevant predictors. The statistical techniques are used to select the predictors from any available source, such as observations, numerical analyses and predictions, as well as other potential predictors (e.g., subjective forecasts of meteorological variables, road conditions or other weather-related variables), and to determine the optimal weights to assign to each predictor. The combination of statistical techniques improves forecast skill by drawing from the strengths of different linear and nonlinear relationships between the predictor sets and the predictand. We have found this methodology to be more robust and less susceptible to overfitting than generalized additive modeling. Our statistical ensemble methodology can be tailored to the prediction of any and all weather-related variables, including precipitation type and amount, maximum and minimum temperatures, ceiling and visibility, fog, road-surface conditions, and sea-surface conditions, to mention a few.

Application to Probabilistic Quantitative Precipitation Forecasting (PQPF)
(development supported by NSF and NOAA under USWRP and CSTAR grants;
implementation supported by UCAR under a COMET grant)

After making extensive comparisons of statistical methods that might be applied to PQPF ([Appelquist et. al. 2002](#); [Gahrs et. al. 2003](#), we have, most recently, developed a statistical ensemble forecast system consisting of [linear regression](#), [logistic regression](#) and a [classification and regression tree \(CART\)](#), which we applied to the prediction of precipitation accumulations exceeding various thresholds up to .50" per 6 hours, out to 48 hours. We used as predictors analyses and forecasts made by three numerical prediction models, namely NOAA's Aviation (AVN), nested-grid (NGM) and eta models. Our forecast system was tested at 398 stations in [7 regions](#) over the U.S.east of the Rockies using 3 years of model analyses and predictions (1999-2001). The first three quarters of the data record was used for training and the last quarter for verification. Linear and logistic regression were each applied to the data from each of the NOAA models to create 6 members of what we call the [statistical/dynamical ensemble](#). The seventh member was created by applying a classification and regression tree to predictors consisting of averages of model precipitation values over all three numerical prediction models. [details](#)

We used the [Brier Skill Score \(BSS\)](#), which measures the improvement of skill over that of a climatological forecast, to compare the skill of the statistical/dynamical ensemble forecast system with that of the individual statistical methods applied to each of the numerical prediction models. The scores of the statistical/dynamical ensemble were found to be significantly higher at the 99% confidence level in all regions than those for any other statistical method applied to any of the numerical models. Moreover, [attributes diagrams](#) reveal that the statistical/dynamical ensemble forecast system exhibits greater reliability and higher resolution than any of the other methods. In particular, the ensemble made more forecasts at the extremes (e.g., > 45% chance and < 5% chance) than the other methods, and the observed frequencies of precipitation were closer to the probabilities predicted by the ensemble in almost all ranges than by any other method. [details](#)

The [mean probability predicted](#) by the statistical/dynamical ensemble when the event occurred was significantly higher, and the mean probability predicted by the ensemble when the event did not occur was significantly lower, than that predicted by the other methods. [details](#).

The [relative operating characteristics \(ROC\)](#) diagrams, which plot the hit rate vs the false alarm rate, were generally

better for the ensemble than for the other methods. The ROC curves can be used to convert statistical forecasts into deterministic forecasts where needed by decision makers. [details](#)

Conclusions

The results of our work reveal that, based on the Brier skill score, the attributes diagram, the mean probability predicted and the relative operating characteristics, the statistical ensemble methodology, and, in particular, the statistical/dynamical ensemble forecast system for probabilistic quantitative precipitation forecasting, exhibits greater skill than do any of the individual methods. The statistical ensemble methodology can be applied to any set of predictors and any forecast problem. By drawing from the strengths of different linear and nonlinear relationships between the predictor sets and the predictand, this methodology holds promise for more accurate forecasts in all applications, and for reducing annual costs for preparing for meteorological events.

Acknowledgments

The development of the statistical/dynamical forecast system for probabilistic quantitative precipitation forecasting was made possible by NSF and NOAA under USWRP Grant ATM 9714414 and NOAA CStar Grant NA17WA1010. Implementation of this forecast system was made possible by UCAR COMET Partners Award S04-44699. The Florida State University Geophysical Fluid Dynamics Institute also provided generous financial support and facilities without which this work would not have been possible.

References

Applequist, S., G. E. Gahrs, R. L. Pfeffer, and X.-F. Niu, 2002: Comparison of Methodologies for Probabilistic Quantitative Precipitation Forecasting. *Wea. Forecasting*, **17**, 783-799.

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

Gahrs, G. E., S. Applequist, R. L. Pfeffer, and X.-F. Niu, 2003: Improved Results for Probabilistic Quantitative Precipitation Forecasting. *Wea. Forecasting*, **18**, 879-89

[Top](#)

Statistical Methods

Linear Regression involves fitting a linear function of independent variables (predictors) onto dependent variables (predictands) by minimizing the least square difference between the predicted and observed values of the predictands.

Logistic regression, also known as a special class of generalized additive modeling (GAM), can be considered a degenerate case of a neural network with no hidden layer. It relates the predictand y to the predictors x_k by the formula

$$y = \{1 + \exp[-\alpha_0 - \sum \alpha_k x_k]\}^{-1}$$

where the a_k are the coefficients to be determined from the training data. An optimal set of a_k is found that minimizes the squared difference between the predicted and observed values of the predictand.

Classification and Regression Tree (CART) is a method which finds an optimal relationship between a predictor and a predictand by splitting the predictor values in the training data into distinguishable ranges and determining the homogeneity of each range based on the values of predictand in that range. For predictions of a parameter, such as precipitation or maximum temperature exceeding a specific threshold, CART finds an optimum value of the predictor which splits the data into occurrences and non occurrences of the event. The predicted probability of occurrence for any value of the predictor is the ratio of the number of occurrences to the total number of cases within the range in which that predictor value falls.

Map of 7 Regions



[Top](#)

Statistical/Dynamical Ensemble Forecast System

The statistical/dynamical ensemble that we applied successfully to probabilistic quantitative precipitation forecasting (PQPF) consists of the following 7 members:

- probability forecasts based on the application of logistic regression to NGM variables (analyzed and predicted). The predictors are selected from a large pool of potential predictors by stepwise selection using logistic regression in the selection phase, as well as in the prediction phase.
- probability forecasts based on the application of logistic regression to eta model variables. While the stepwise selection procedures are the same, the predictors chosen from the eta model and the coefficients applied to these predictors, are not necessarily the same as those for the NGM.
- probability forecasts based on the application of logistic regression to Aviation model variables. Once again the predictors, the coefficients and the probabilities derived from this model will typically differ from those derived from the other two models.
- probability forecasts based on the application of linear regression to NGM variables (analyzed and predicted). The predictors are selected from a large pool of potential predictors by stepwise selection using linear regression in the selection phase, as well as in the prediction phase. The predictors from the NGM selected by linear regression, as well as the coefficients and the probabilities, are characteristically different from those derived by logistic regression applied to the same numerical model.
- probability forecasts based on the application of linear regression to eta model variables.
- probability forecasts based on the application of logistic regression to Aviation model variables.
- probability forecasts based on the application of a classification and regression tree (CART) to model-predicted precipitation averaged over all three numerical models.

The seven probability forecasts for each meteorological station are then combined using another layer of statistics to

determine the final probability forecast for that station. While we have thus far used numerical model analyses and forecasts, other types of information (such as observations, subjective forecasts, road conditions, etc.) can also be incorporated into the statistical component of the ensemble.

[Back](#) [Top](#)

Brier Skill Score

We use the Brier Skill Score (BSS; Brier 1950)

$$BSS = 1 - BS/CS$$

to assess the skill of each forecast methodology. Here BS is the Brier score, which is defined as the mean squared difference between the predicted probabilities and the observations. The observations are taken as unity when the precipitation accumulation equals or exceeds a given threshold and zero when it does not. The parameter CS is the Brier score of the climatological forecast. The Brier skill score, which measures the degree of improvement achieved by each methodology over climatological forecasts ranges from negative infinity to unity. A value of unity represents a perfect score. A value of zero indicates that the method has the same skill as a climatological forecast. Negative values represent scores worse than a climatological forecast.

Tables 1a,b,c and d compare the Brier skill scores for each threshold, method, and forecast interval over all 7 regions during the cool season. In each table we have bold-faced the highest score for each forecast interval. With only three exceptions out of 32 comparisons, the statistical/dynamical ensemble exhibits the highest BSS for every forecast interval. The T-scores in the tables below the tables of scores indicate the level of significance of the differences over the 8 forecast time intervals between the scores for the statistical/dynamical ensemble and each of the individual methods that make up the ensemble. To the right of this lower table is a key to the significance levels for different T-scores. These tables show that most of the differences are significant at the 99% level or higher.

[Back](#) [Top](#)

Table 1a .01" Cool Season - average over 7 regions

cool season .01"	0-6hr	6-12hr	12-18hr	18-24hr	24-30hr	30-36hr	36-42hr	42-48hr
statistical/dynamical ensemble	.568	.486	.498	.519	.449	.438	.420	.397
classification and regression tree	.512	.472	.445	.465	.432	.374	.357	.258
Ngm logistic regression	.535	.436	.457	.469	.387	.383	.355	.340
eta logistic regression	.547	.458	.474	.497	.432	.420	.396	.380
Avn logistic regression	.535	.439	.455	.473	.394	.387	.376	.358
Ngm linear regression	.490	.405	.427	.444	.367	.369	.339	.321
eta linear regression	.508	.427	.453	.477	.410	.408	.386	.360
Avn linear regression	.495	.410	.436	.454	.377	.380	.364	.343

cool season .01" T-scores - statistical/dynamical ensemble vs.							Significance Level	T-Scores
classification and regression tree	NGM logistic regression	eta logistic regression	AVN logistic regression	NGM linear regression	eta linear regression	Avn linear regression	99%	3.499

							95%	2.365
4.2435	13.729	15.401	18.561	45.067	11.091	21.761	90%	1.895

Table 1b .10" Cool Season - average over 7 regions

cool season .10"	0-6hr	6-12hr	12-18hr	18-24hr	24-30hr	30-36hr	36-42hr	42-48hr
statistical/dynamical ensemble	.568	.541	.490	.486	.415	.426	.388	.336
classification and regression tree	.519	.529	.432	.424	.401	.379	.321	.170
Ngm logistic regression	.534	.487	.437	.430	.339	.353	.305	.277
eta logistic regression	.528	.508	.460	.458	.374	.398	.360	.319
Avn logistic regression	.529	.449	.436	.430	.332	.350	.328	.299
Ngm linear regression	.501	.447	.405	.404	.333	.326	.287	.255
eta linear regression	.527	.495	.447	.455	.385	.407	.365	.295
Avn linear regression	.531	.468	.436	.429	.355	.364	.328	.279

cool season .10" T-scores - statistical/dynamical ensemble vs.							Significance Level	T-Scores
classification and regression tree	NGM logistic regression	eta logistic regression	AVN logistic regression	NGM linear regression	eta linear regression	Avn linear regression	99%	3.499
							95%	2.365
3.513	10.994	11.348	8.794	21.553	9.733	16.194	90%	1.895

Table 1c .25" Cool Season - average over 7 regions

cool season .25"	0-6hr	6-12hr	12-18hr	18-24hr	24-30hr	30-36hr	36-42hr	42-48hr
statistical/dynamical ensemble	.512	.483	.416	.417	.335	.373	.312	.272
classification and regression tree	.438	.487	.346	.334	.309	.299	.239	.109
Ngm logistic regression	.478	.431	.348	.342	.257	.286	.236	.232
eta logistic regression	.470	.428	.367	.396	.298	.350	.279	.254
Avn logistic regression	.460	.390	.350	.348	.269	.302	.238	.245
Ngm linear regression	.451	.417	.348	.331	.249	.272	.235	.209
eta linear regression	.481	.457	.394	.386	.312	.361	.310	.244
Avn linear regression	.482	.429	.366	.379	.295	.318	.264	.233

cool season .25" T-scores - statistical/dynamical ensemble vs.							Significance Level	T-Scores
classification and	NGM	eta logistic	AVN	NGM linear	eta linear	Avn linear		

regression tree	regression	regression	regression	regression	regression	regression		
							95%	2.365
4.0993	9.310	7.264	9.630	15.288	6.123	14.178	90%	1.895

Table 1d .50" Cool Season - average over 7 regions

cool season .50"	0-6hr	6-12hr	12-18hr	18-24hr	24-30hr	30-36hr	36-42hr	42-48hr
statistical/dynamical ensemble	.445	.343	.303	.215	.251	.292	.167	.119
classification and regression tree	.348	.310	.215	.182	.201	.151	.150	.049
Ngm logistic regression	.403	.286	.268	.244	.140	.196	.083	.073
eta logistic regression	.410	.295	.262	.280	.202	.269	.109	.087
Avn logistic regression	.371	.302	.236	.263	.216	.222	.109	.099
Ngm linear regression	.382	.277	.267	.249	.148	.190	.105	.071
eta linear regression	.413	.319	.307	.288	.218	.262	.157	.094
Avn linear regression	.405	.313	.267	.276	.223	.239	.132	.087

cool season .50" T-scores - statistical/dynamical ensemble vs.							Significance Level	T-Scores
classification and regression tree	NGM logistic regression	eta logistic regression	AVN logistic regression	NGM linear regression	eta linear regression	Avn linear regression	99%	3.499
							95%	2.365
4.532	3.576	2.003	2.792	3.649	.763	1.935	90%	1.895

Attributes Diagrams

Attributes diagrams provide a measure of the "reliability" and "resolution" of different forecast methodologies.

Reliability is a measure of the closeness of the observed frequencies to the predicted probabilities for the different categories of probabilities predicted. A forecast method is said to exhibit perfect reliability when, for each predicted probability, the observed frequency is identical to the predicted probability. Perfect reliability does not, however, ensure a high Brier skill score, unless the resolution is high for a large number of forecasts. For example, forecasts of probabilities close to the climatological frequency of occurrence will have a very low Brier skill score, even when they exhibit perfect reliability.

The resolution measures the degree to which the forecast system succeeds in identifying subsample forecast periods in which the frequencies of occurrence are much below or much above the climatological frequency.

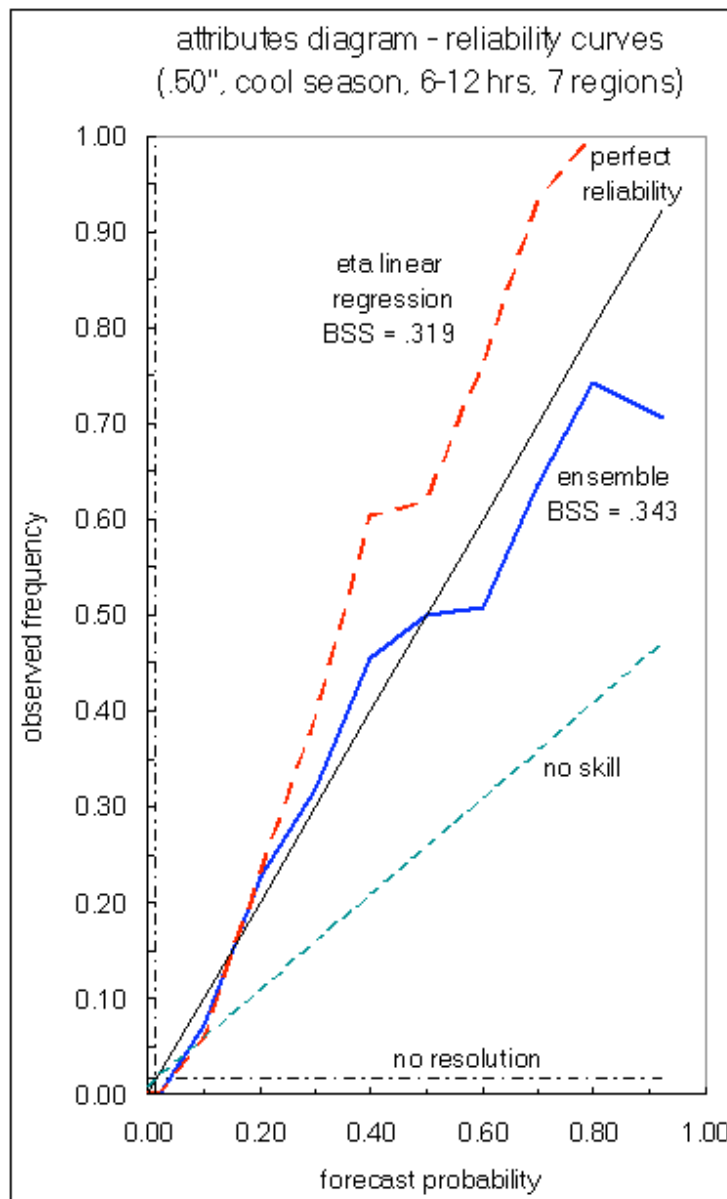
The upper graph in **Fig. 2** compares the reliability of the ensemble method (heavy blue solid curve) with that of linear regression applied to the eta model (heavy red dashed curve). The diagonal line in the graph represents perfect reliability. The thin dashed line is the no-skill line, and the dot-dashed horizontal and vertical lines represent the climatological frequency (no resolution). These results are based on averages over 7 regions (398 stations) of the U.S. east of the Rockies for cool season precipitation exceeding .50" during the 6-12 hr forecast interval. Linear regression applied to the eta model was selected for this comparison because it exhibits the highest average Brier skill score

among the linear regression forecasts. The bar graph below the reliability graph compares the resolution of the two forecast methodologies.

The tables below each of the graphs in this figure give the numerical values plotted in the graphs. The negative differences in the first table indicate that the ensemble forecasts are closer to perfect reliability than are the linear regression forecasts. Positive differences in the second table indicate the ranges in which ensemble makes more forecasts than does linear regression. The most significant feature here is that the statistical/dynamical ensemble method gives many more predictions of 0%, and above 45% probability, where the climatological frequency is .0149. In particular, there are 2,088 more forecasts of probabilities of 0% by the ensemble than by linear regression. In the range 45-100%, the ensemble gives 199 more forecasts than linear regression. The attributes curves and the corresponding table, showing the departures from perfect reliability for both models, reveal that the forecasts made with the statistical/dynamical ensemble have a high degree of reliability in comparison with those made by linear regression (negative differences). The greater reliability, together with the greater resolution of very high and very low probabilities of the event, result in significantly higher Brier skill scores for the statistical/dynamical ensemble at the 99% confidence level.

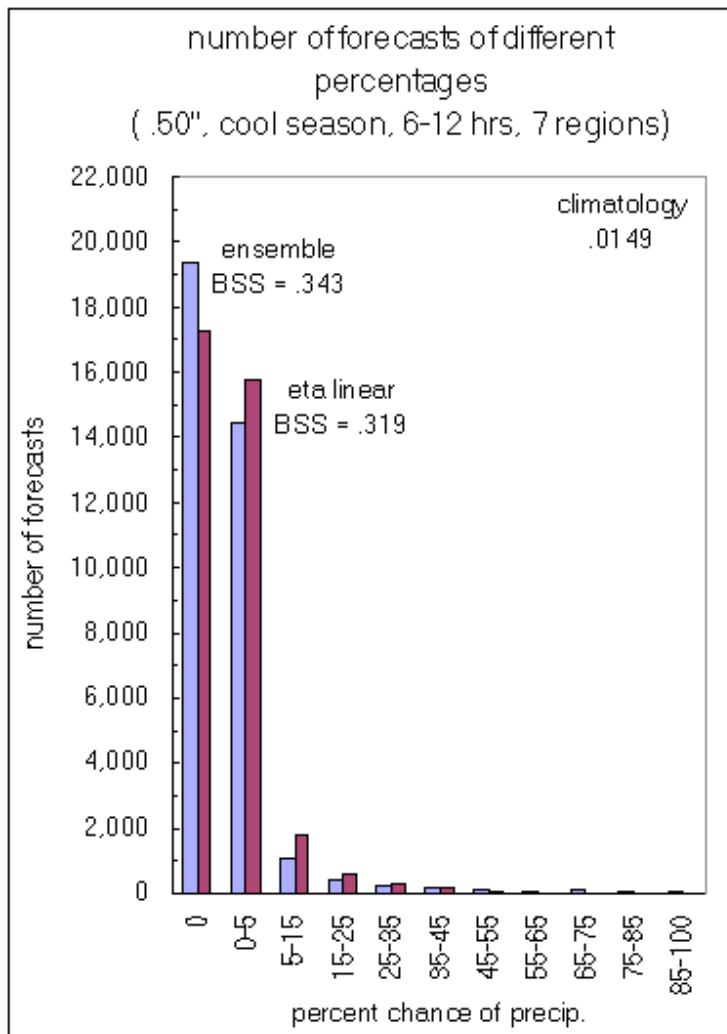
[Back](#) [Top](#)

Figure 2. Attributes Diagram and Bar Graph



7 regions climatology = .0149

.50" cool season	magnitude of departure from perfect reliability		6-12 hr forecasts
range (%)	ensemble BSS=.343	eta linear regression BSS=.319	difference
0	.000	.000	.000
0-5	.022	.023	-.001
5-15	.028	.042	-.015
15-25	.027	.031	-.004
25-35	.019	.090	-.072
35-45	.056	.203	-.147
45-55	.000	.118	-.118
55-65	.092	.162	-.070
65-75	.064	.233	-.170
75-85	.056	.200	-.144
85-100	.220	.075	.145



7 regions climatology = .0149

.50" cool season	number of cases predicted	6-12 hr forecasts
------------------	---------------------------	-------------------

range (%)	ensemble BSS=.343	BSS=.319	difference
0	19355	17267	2,088
0-5	14429	15742	-1,313
5-15	1050	1798	-748
15-25	383	576	-193
25-35	226	264	-38
35-45	136	131	5
45-55	98	76	22
55-65	63	21	42
65-75	77	15	62
75-85	39	5	34
85-100	44	5	39

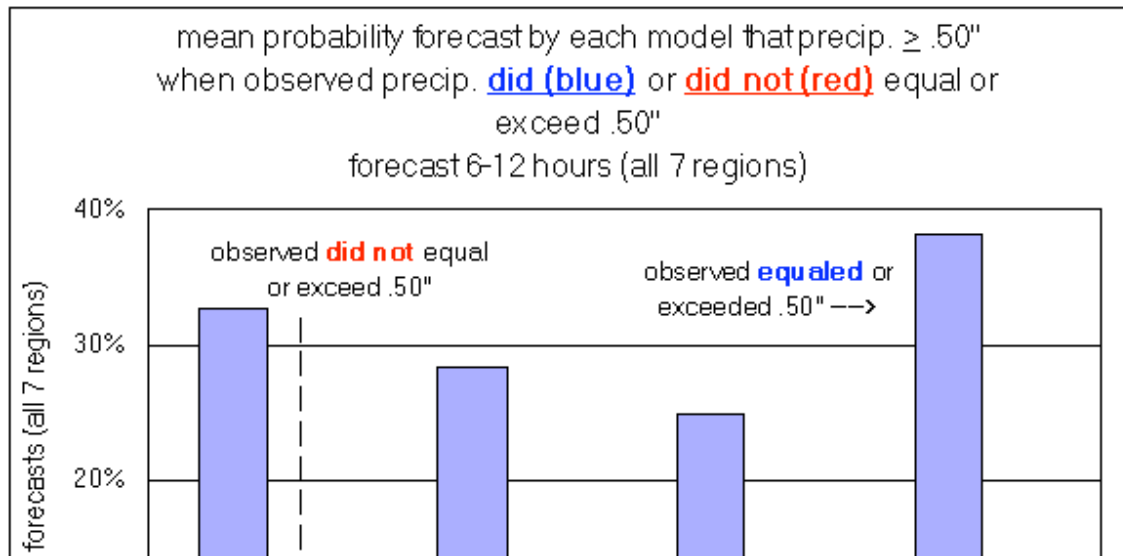
Mean Probability Predicted

Fig. 3 and Table 2 compare the mean probabilities predicted during the cool season for the forecast period 6-12 hrs by the statistical/dynamical ensemble with those predicted by linear regression applied to each of the numerical models (standard MOS approach). The blue bars correspond to the days in which precipitation was observed to equal or exceed .50" during the forecast period, and the red bars correspond to the days in which it was not. It is significant that the statistical/dynamical ensemble gives higher probability forecasts than linear regression applied to any of the numerical models on days for which precipitation $\geq .50$ " and lower probability forecasts on days for which precipitation $< .50$ ". The ratio of the two probability forecasts (for the cases of observed to not observed) is more than 70% greater for the statistical/dynamical ensemble than for the AVN linear regression model, which has the highest ratio among the linear regression forecasts.

Fig. 4 displays these results with error bars corresponding to the standard error of the mean. Fig. 4a, in particular, compares the probability forecasts for the precipitation $\geq .50$ " cases and fig. 4b compares them for the precipitation $< .50$ " cases. In both cases the error bars for statistical/dynamical forecasts exclude by wide margins those for linear regression applied to each of the numerical models.

[Back](#) [Top](#)

Fig. 3 **Forecast Statistics for each model (6-12 hour forecast over all 7 regions)**. Comparison of mean predicted probability for each model when the observed equaled or exceeded .50" vs. when it did not.



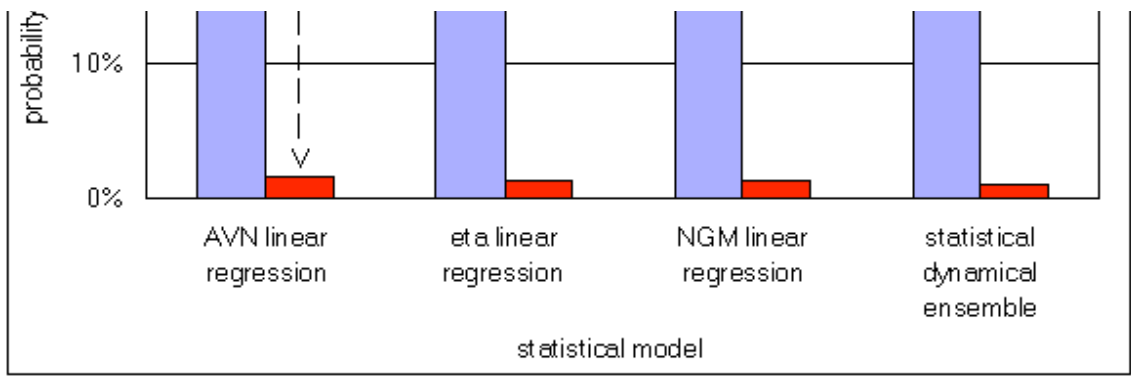
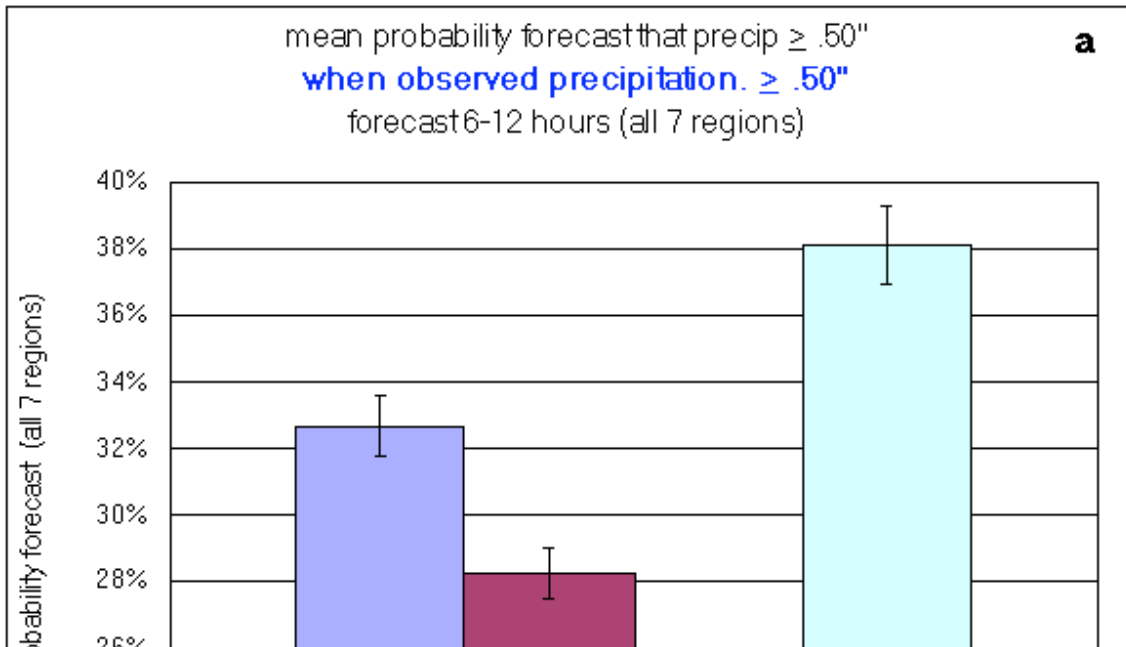


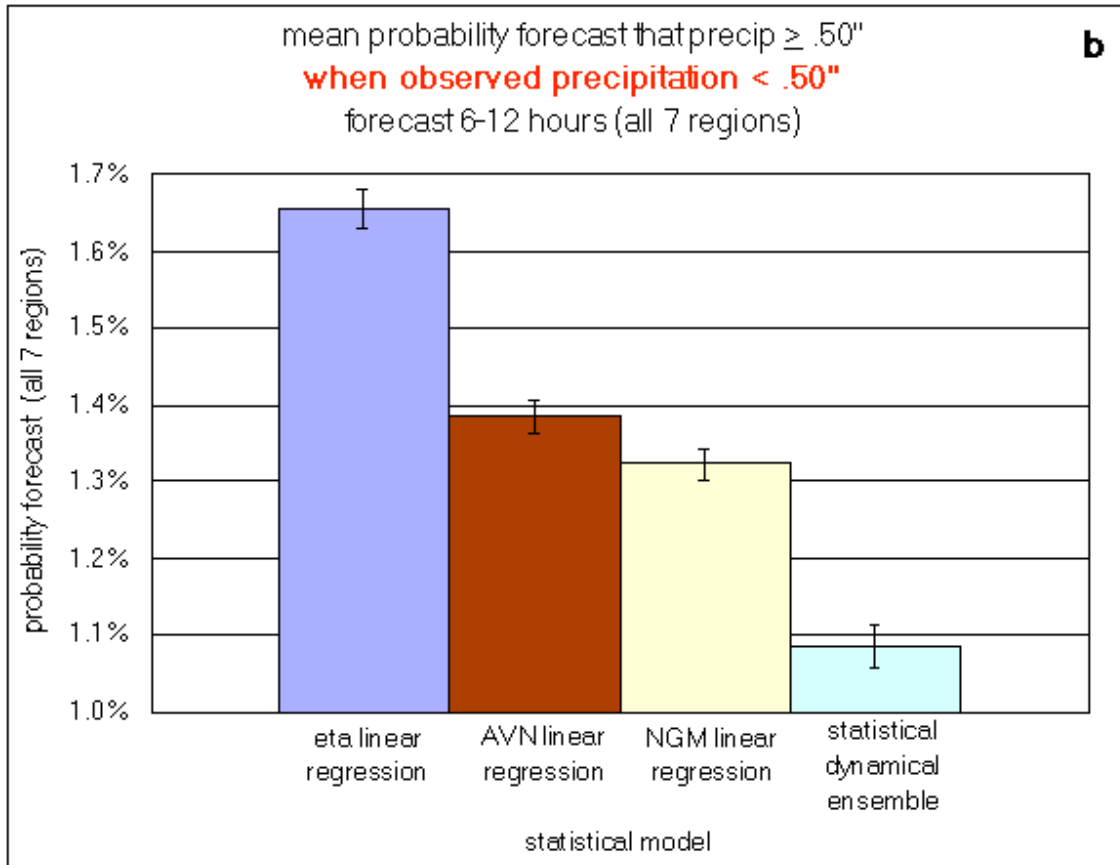
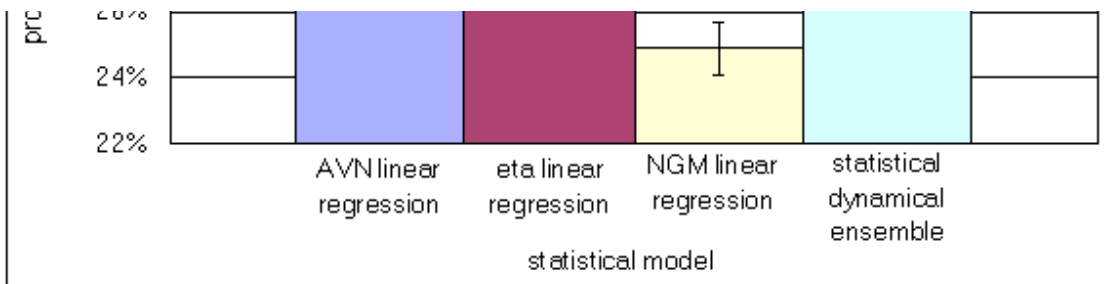
Table 2 Mean predicted probability of precipitation equal to or exceeding .50" for each model and ratio of forecast probabilities on days in which observed equaled or exceeded .50".

all 7 regions forecast 6-12 hours probability of precip $\geq .50"$	eta linear regression	AVN linear regression	NGM linear regression	Statistical/Dynamical ensemble
observed - yes	32.7%	28.2%	24.9%	38.1%
observed - no	1.7%	1.4%	1.3%	1.1%
ratio -> yes / no	19.74	20.40	18.81	35.09

Fig 4 Standard error of the mean for each model - 6-12 hour forecast (7 regions)

- Mean predicted probability and standard error of the mean for each model when observed precip. equaled or exceeded .50".
- Mean predicted probability and standard error of the mean for each model when observed precip. was less than .50".





Relative Operating Characteristics (ROC)

Decision-makers must determine when to prepare for a meteorological event, such as fog, icing of roads, flooding, heavy snow, tornadoes, minimum or maximum temperature exceeding a critical value, visibility (or ceiling) below safe levels for aircraft landings, etc.. Should the transportation department close roads or send out sanding vehicles and/or snow plows? Should the city or county evacuate people from the area? Should schools, government offices and businesses close for the day? Should citrus growers activate smudge pots or the power company divert extra power to an affected area? Should airport authorities divert flights to another airport? Affirmative answers to any of these questions add costs, whether or not the forecast is correct.

The easiest course for a decision-maker to follow is to prepare for an event if a deterministic forecast (say, from a numerical model, a consensus of numerical models or a forecast service) says that the phenomenon will exceed a predetermined threshold. **This might not, however, be the most cost-effective method of approach.** Depending on the cost of preparing for an event and the cost incurred if the event happens when no preparation is made in advance, a probability forecast can offer a more cost-effective way of making the decision. To understand this we need only recognize that all forecast methods have uncertainty.

A given deterministic forecast method has a certain rate of hits (successful predictions) vs. false alarms (predictions of the event when it does not happen). Statistical methods, on the other hand, have a continuum of hit vs. false-alarm rates, corresponding to probability forecasts ranging from 0% to 100%. Depending on the cost scenario for a particular

phenomenon, the decision-maker can use the point on a relative operating characteristics (ROC) curve corresponding to the threshold probability forecast that minimizes the costs over all events. The ROC curve is a plot of hit rates vs. false alarm rates for a particular forecast methodology. Each deterministic forecast method is represented by a single point on a ROC diagram, whereas each probability forecast method is represented by a continuous curve. We illustrate this in Fig. 5 using the data for the 6-12 hr. cool-season forecasts for the .50" threshold averaged over all 7 regions.

In the figure, the heavy solid curve is for the statistical/dynamical forecast. The area under the curve, which for this forecast methodology is .972, is a measure of forecast skill. A perfect forecast method would have an area of 1.0, representing all hits and no false alarms. A method with no skill (diagonal line in the figure) would have an area of 0.50, representing forecasts for which there are as many false alarms as there are hits. A forecast method giving more false alarms than hits is considered to have negative skill. The ROC curves for the individual statistical methods applied to the individual numerical models are not too different from the one shown in Fig. 5, but the area under the curve corresponding to the statistical/dynamical ensemble, averaged over all regions and time periods, is greater by a small, but statistically significant, amount at a significance level of 99% in all but one case and 90% in the remaining case.

We shall address the minimization of costs by considering the following cost scenario:

- the cost for preparing for an event when it does not happen is 10 units,
- the total cost if prepared when the event happens is 500 units and
- the cost if not prepared when the event happens is 990 units.

The open square (which correspond to a forecast of 4% chance of the event) represents the threshold that minimizes the costs for the statistical/dynamical forecast methodology. That is, any forecast with a probability equal to or greater than 4% should be taken as a signal to prepare for the event in order to minimize overall costs for the cost scenario given above. While small, 4% is greater than the 1.49% climatological probability over the 7 regions of cool season precipitation exceeding .50". If the decision-maker uses the statistical/dynamical ensemble 4% threshold, the costs would be a total of 8,440 units during the period for which forecasts were made on the independent data set. For other cost scenarios, there will be different thresholds that minimize costs.

The diamond, pyramid and circle in the figure represent the AVN, eta and NGM model forecasts, respectively. The cost for preparing for the event when each of these models is employed to make the determination is shown next to the point for that model. It is notable that all three costs are significantly greater than the cost of preparing when the statistical/dynamical forecast gives a 4% chance or higher of the event. The costs of 11,430, 12,010 and 12,900 for the AVN, eta and NGM models are 35%, 42% and 53% greater than that for the statistical/dynamical model. It is worth noting also that the hit rate corresponding to the optimum point on the ROC curve for the ensemble is .934, whereas the hit rates for the AVN, eta and NGM model forecasts are .461, .378 and .255, respectively.

One might suspect that, with a cost of only 10 units for preparing for the event, a decision-maker might do well by ignoring the forecasts altogether and spending his or her money to be prepared at all times for the possibility of the event happening. The fact is, however, that for the time period of our independent data set, the cost would have been 17,290 units, which compares unfavorably with the costs incurred by employing any of the forecast methodologies discussed above as a basis for preparing for the event.

At the opposite extreme, it might be thought that, with an event so rare as to occur 1.49% of the time (the climatological frequency), a decision-maker might do well by ignoring the forecasts and never preparing for the event, thereby incurring the occasional cost of 990 units. Based on the number of times the event happened in the independent data set, however, the cost for following this course would have been 14,730 units, which again compares unfavorably with the costs incurred by employing any of the forecast methodologies discussed above as a basis for preparing for the event.

[Back](#) [Top](#)

Fig. 5 Roc Diagram for the .50" threshold for the f6-12 cool season forecast period.

Fig. 5 ROC Diagram and costs

.50" cool season forecasts 6-12 hrs

