Ceiling and Visibility Forecasts via Neural Networks

CAREN MARZBAN

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma, and Department of Statistics, University of Washington, Seattle, Washington

STEPHEN LEYTON

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

BRAD COLMAN

National Weather Service, Seattle, Washington

(Manuscript received 26 February 2006, in final form 22 August 2006)

ABSTRACT

Statistical postprocessing of numerical model output can improve forecast quality, especially when model output is combined with surface observations. In this article, the development of nonlinear postprocessors for the prediction of ceiling and visibility is discussed. The forecast period is approximately 2001–05, involving data from hourly surface observations, and from the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model. The statistical model for mapping these data to ceiling and visibility is a neural network. A total of 39 such neural networks are developed for each of 39 terminal aerodrome forecast stations in the northwest United States. These postprocessors are compared with a number of alternatives, including logistic regression, and model output statistics (MOS) derived from the Aviation Model/Global Forecast System. It is found that the performance of the neural networks is generally superior to logistic regression and MOS. Depending on the comparison, different measures of performance are examined, including the Heidke skill statistic, cross-entropy, relative operating characteristic curves, discrimination plots, and attributes diagrams. The extent of the improvement brought about by the neural network depends on the measure of performance, and the specific station.

1. Introduction

Coastal locations around the United States are susceptible to marine stratus and low ceilings. Occasionally, the stratus extends to the surface, resulting in dense fog conditions and a significant reduction in horizontal visibility. When these adverse conditions occur, they often disrupt daily activities, such as aviation and shipping interests, sometimes to the point of jeopardizing human safety. It is imperative, then, to provide forecasters with the most accurate guidance for these high-impact weather phenomena.

Marine stratus is especially problematic along the Pacific coast of the United States during the summer months. This is primarily because of a semipermanent area of surface high pressure over the northern Pacific Ocean during the summer months, the continuous flow of relatively cool ocean waters southward along the Pacific Coast via the California Current, and the warm daytime temperatures that occur over inland locations. When the warmer, less dense air over the land rises, it is replaced by cooler, more dense air offshore that moves inland and ultimately creates a large-scale circulation, similar to a sea breeze. The inland extent of the marine air depends on the differential heating of the land and water and the resulting strength of the circulation. This cycle of marine stratus development is explained in more detail by Hilliker and Fritsch (1999). In addition, Hilliker and Fritsch (1999) displayed the frequency of occurrence of these marine stratus events through an annual number of instrument flight rules (IFR) events at various airports across the United States. It is shown that San Francisco International Air-

Corresponding author address: Caren Marzban, Dept. of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322.
E-mail: marzban@caps.ou.edu

DOI: 10.1175/WAF994.1

© 2007 American Meteorological Society

port experiences nearly 200 IFR events per year while Seattle–Tacoma International Airport experiences about 125 IFR events per year.

In addition to the very common inland intrusions of marine fog and stratus are episodes of inland basin (radiational) fog. During spells of prolonged ridging over the western United States in the cool season, shallow layers of persistent fog form in the inner basins, including the Columbia River basin, and the Sacramento River valley. These events are hazardous to all kinds of transportation and commerce, but especially to aviation. Additionally, recurring winter storms bring restricted flying conditions to all terminals across the Pacific Northwest. These events include strong wind, rain, fog, snow, and freezing precipitation.

Traditionally, the primary source of guidance for forecasting surface weather conditions has been statistically postprocessed numerical model output. In particular, model output statistics (MOS) derived from the Nested Grid Model (NGM) and Aviation Model/Global Forecast System (AVN/GFS) provide forecasts of weather parameters at 3-h intervals out to 48 or 60 h, depending on the model.

This type of forecasting guidance has at least two limitations: 1) the models are run only a few times daily, allowing forecasts to become several hours old before an updated product is made available, and 2) the MOS equations are linear. To alleviate the first limitation, studies have been undertaken to investigate the use of an observations-based forecasting system (Vislocky and Fritsch 1997). In this system, a network of surface observations is used as predictors in a multiple regression technique. It was demonstrated that this approach could improve the accuracy of ceiling and visibility forecasts for the hours between the times that the output from the numerical models is released. Moreover, Leyton and Fritsch (2003, 2004) showed that the introduction of high-density observation networks and high-frequency observations lead to further improvement.

The linearity of traditional MOS equations may be a limitation if the underlying relations are nonlinear. To allow for more general relations, nonlinear generalizations of multiple regression have been utilized to model any nonlinearity and interactions of the underlying processes. Of course, the nonlinearity of the statistical model does not prevent it from capturing linear relations as well. For example, temperature forecasts from the Advanced Regional Prediction System have been postprocessed via neural networks, displaying a reduction in bias and error variance of the forecasts (Marzban 2003). There, it is found that the optimal neural network is indeed nonlinear. As such, the nonlinear statistical postprocessing yields temperature forecasts that are more accurate than the model forecasts as well as MOS forecasts. Some recent applications of neural networks include nowcasting of visibility from surface observations (Pasini et al. 2001), prediction of snowfall (Roebber et al. 2003), and fog forecasts at an airport in Brazil (Costa et al. 2006).

Here, these two solutions have been combined to produce superior forecasts. In other words, a nonlinear postprocessor has been developed that takes as input (i.e., predictors) not only model data but also surface observations, and it produces forecasts of ceiling and visibility. In fact, 39 postprocessors are developed for 39 terminal aerodrome forecast (TAF) stations. TAF forecasters rely heavily on statistical postprocessors, because not all of the parameters included in TAFs are directly output by numerical weather prediction models. Specifically, explicit information about visibility, cloud, and fog layers is generally absent. Forecasters thus rely on experience, climatology, and statistical tools. The numerical model for which postprocessors are developed is the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5). The MM5 real-time mesoscale weather prediction system run by the Northwest Modeling Consortium and the University of Washington is one of the longest-running and mature regional modeling efforts in existence (Grimit and Mass 2002). The statistical model underlying these postprocessors is a neural network (NN). In what follows, the development of these NNs is outlined in detail, and it is shown that their performance is generally superior to that of logistic regression and MOS derived from the GFS, with the extent of the improvement dependent on the measure of performance and the specific station. As shown below, the exceptions include stations where insufficient data exist for a proper development of a nonlinear postprocessor.

2. Data

Three datasets are employed. The first dataset is an archive of standard hourly surface [Automated Surface Observing System (ASOS) and Automated Weather Observing System (AWOS)] data, also acquired from the National Center for Atmospheric Research (NCAR). This dataset includes reports of temperature, dewpoint, wind speed and direction, cloud cover, visibility, and precipitation. The surface observations employed for the analysis span the dates 7 November 2001–31 January 2005. Although the data are available on an hourly basis, only the 0000 UTC values are used, except for ceiling and visibility, which are used at 0600 and 1200 UTC—the valid times.

The second dataset is an archive of MM5 data. The

TABLE 1. The categories for ceiling and visibility adopted here.

Class (category)	Ceiling [ft (km)]	Visibility [mi (km)]
1	≤500 (0.15)	≤1 (1.61)
2	500–1000 (0.15–0.30)	1–3 (1.61–4.83)
3	1000–3000 (0.30–0.91 km)	≥3 (4.83)
4	≥3000 (0.91)	

MM5 has been run at the University of Washington on a twice-daily basis since late 1997. The standard configuration includes a triple nest of 36-, 12-, and 4-km grids. The current configuration consists of a 12-km grid that covers a broad area of the Pacific Northwest, extending from western Montana to the coast and southward to northern Nevada and California. An archive of 12-km runs provides the model data used in the study. The model data were archived for the period 7 November 2001–31 March 2005, and were used every 6 h. Here, the 0000 UTC runs of the MM5 were employed.

The last dataset consists of MOS forecasts derived from the GFS model, spanning the dates 1 November 2001–28 February 2005.[1] MOS forecasts are issued at 0000 UTC, are categorical, and are available every 6 h. This dataset is utilized as a benchmark for comparing the performance of the neural network.

The categories for ceiling and visibility are given in Table 1. The categories defined in MOS are finer, and so, in order to compare the neural nets developed here with MOS, the categories in the latter are combined into larger categories that coincide with those adopted here.

The list of the TAF stations (and their station identifiers) for which specific neural nets are developed is given in Table 2. They are scattered across Washington, Oregon, Idaho, Montana, Nevada, and California; see Fig. 1.

3. Methodology

Recently, there has been a surge of interest in the use of NNs among meteorologists (Hsieh and Tang 1998). Among the numerous applications, postprocessing of model output has been beneficial (Casaioli et al. 2003; Marzban 2003). Given that NNs belong to the class of nonlinear statistical models, the postprocessing activity can be considered a nonlinear version of MOS. Although the choice of a nonlinear method is not unique, NNs are generally useful because in addition to being able to approximate a large class of functions, compared with many other models, they are less prone to overfit data. Overfitting can occur when a statistical model has sufficient flexibility to fit features in the data that are purely statistical fluctuations. As such, a statistical model that has overfit some dataset is likely to perform poorly on data not represented in that dataset. The number of parameters in a statistical model is one measure of model flexibility, and so a model with too many parameters is more likely to overfit a dataset than another model with less parameters. For example, the number of parameters in polynomial regression grows exponentially with the number of independent variables (i.e., predictors); by contrast, for NNs the growth is only linear.

The prediction of ceiling and visibility can be cast into a classification problem. The NN is designed to produce a probability of belonging to one of the classes/categories shown in Table 1, given the atmospheric conditions at the time; the latter are quantified by the predictors, namely the model output variables and surface observations.

The development of the NNs requires substantial preprocessing of the data. Seasonal cycles in the data are removed by subtracting the monthly average of each variable from the data on that variable. More sophisticated techniques (e.g., seasonal differencing) for filtering out periodicity were examined, but showed no significant difference. Outliers are removed by simply excluding any data beyond 5 standard deviations of the mode. For variables with highly skewed distributions, outliers are identified visually. Square-root and cube-root transformations are applied to the inputs to render their distribution more normal. Pairwise correlations between the predictors are employed to exclude one member of highly correlated pairs. Principal components of the data were considered as inputs to the NN, but no significant improvement was detected.

After this preprocessing, 20 predictors remain, and the surviving sample size is 878. The predictors are normalized in the manner of z scores (i.e., the mean of each variable is subtracted from each case, and the result is divided by the standard deviation of that variable). Jitter (or Gaussian noise) with a mean of 0 and standard deviation of 0.2 is added to each variable. To inflate the sample size, this is repeated 10 times, increasing the sample size from 878 to 8780.[2]

[1] The data archive for MOS_NGM is highly incomplete, and for that reason, the NNs will be compared with the MOS_GFS forecasts for which the data archive is quite complete.

[2] This manner of increasing sample size may seem like trickery; however, it is a relatively standard technique in nonlinear modeling, with the sole purpose of restraining overfitting. The choices of 0.2 for the standard deviation of the jitter, and 10 for the multiplication factor, are based on trial and error. The final results are generally insensitive to the specific values of these quantities. See, for example Koistinen and Holmstrom (1992).

TABLE 2. List of TAF stations.

Arcata, CA	KACV	Moses Lake, WA	KMWH
Walla Walla, WA	KALW	Whidbey, WA	KNUW
Astoria, OR	KAST	Olympia, WA	KOLM
Boeing Field (Seattle), WA	KBFI	Paine Field (Everett), WA	KPAE
Baker City, CA	KBKE	Pendleton, OR	KPDT
Bellingham, WA	KBLI	Portland International Airport, OR	KPDX
Burns, OR	KBNO	Pasco, WA	KPSC
Boise, ID	KBOI	Redding, CA	KRDD
Crescent City, CA	KCEC	Redmond, OR	KRDM
Port Angeles, WA	KCLM	Seattle–Tacoma International Airport, WA	KSEA
Coeur d'Alene, ID	KCOE	Felts Field (Spokane), WA	KSFF
The Dalles, OR	KDLS	Fairchild Air Force Base, WA	KSKA
Wenatchee, WA	KEAT	Salem, OR	KSLE
Elko, NV	KEKO	Sun Valley, ID	KSUN
Eugene, OR	KEUG	McChord, WA	KTCM
Geiger Field (Spokane), WA	KGEG	Troutdale, OR	KTTD
Hoquium, WA	KHQM	Twin Falls, ID	KTWF
Klamath Falls, OR	KLMT	Winnemucca, NV	KWMC
Medford, OR	KMFR	Yakima, WA	KYKM
Missoula, MT	KMSO		

Some NN-specific matters are as follows. The number of input nodes is 20, and the number of output nodes is four or three, respectively, for ceiling or visibility. The number of hidden nodes is determined via bootstrapping. In its simplest version, bootstrapping calls for repeated partitioning of the data into a training set and a validation set. The number of hidden nodes is taken to be that which minimizes the average of the validation errors over the bootstrap trials. The number of bootstrap trials is four. Given the small number of bootstrap trials, the statistical significance of some of the results found here cannot be established. The results from the four bootstrap trials are often displayed separately, and should be interpreted as a rough estimate of the sampling variation.

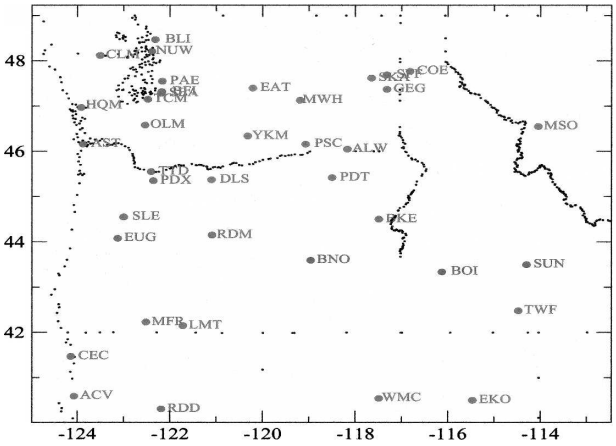

FIG. 1. A map of the 39 TAF stations superimposed over the northwest region of the United States.

One way of displaying many of these quantities is through a "TV diagram" (Marzban 2000), which is basically a scatterplot of training and validation performance numbers for different hidden nodes and different bootstrap trials (e.g., Fig. 6).

The activation functions for all the nodes is the logistic function, and the error function being minimized is cross-entropy (a generalization of mean-squared error, but appropriate for categorical data). This setup assures that the outputs of the NN represent the posterior probability of belonging to a certain category of ceiling or visibility, given the inputs. In other words, the outputs of the NN are forecast probabilities.

TABLE 3. List of the predictors after variable selection.

x_1	Sea level pressure (PSEALVLC)
x_2	Temperature at 2 m (T2)
x_3	Mixing ratio at 2 m (Q2)
x_4	Wind u component at 10 m (U10)
x_5	Wind v component at 10 m (V10)
x_6	Tot precipitation since initialization (RAINTOT)
x_7	Geopotential height at 850 mb (H)
x_8	Mixing ratio at 850 mb (Q)
x_9	Temperature at 850 mb (T)
x_{10}	Wind u component at 850 mb (U)
x_{11}	Wind v component at 850 mb (V)
x_{12}	Sine of Julian day (sin)
x_{13}	Cosine of Julian day (cos)
x_{14}	Sky
x_{15}	Precipitation occurrence (Precip)
x_{16}	Wind direction (Wind_D)
x_{17}	Wind speed (Wind_S)
x_{18}	Dewpoint
x_{19}	Dewpoint depression
x_{20}	Relative humidity

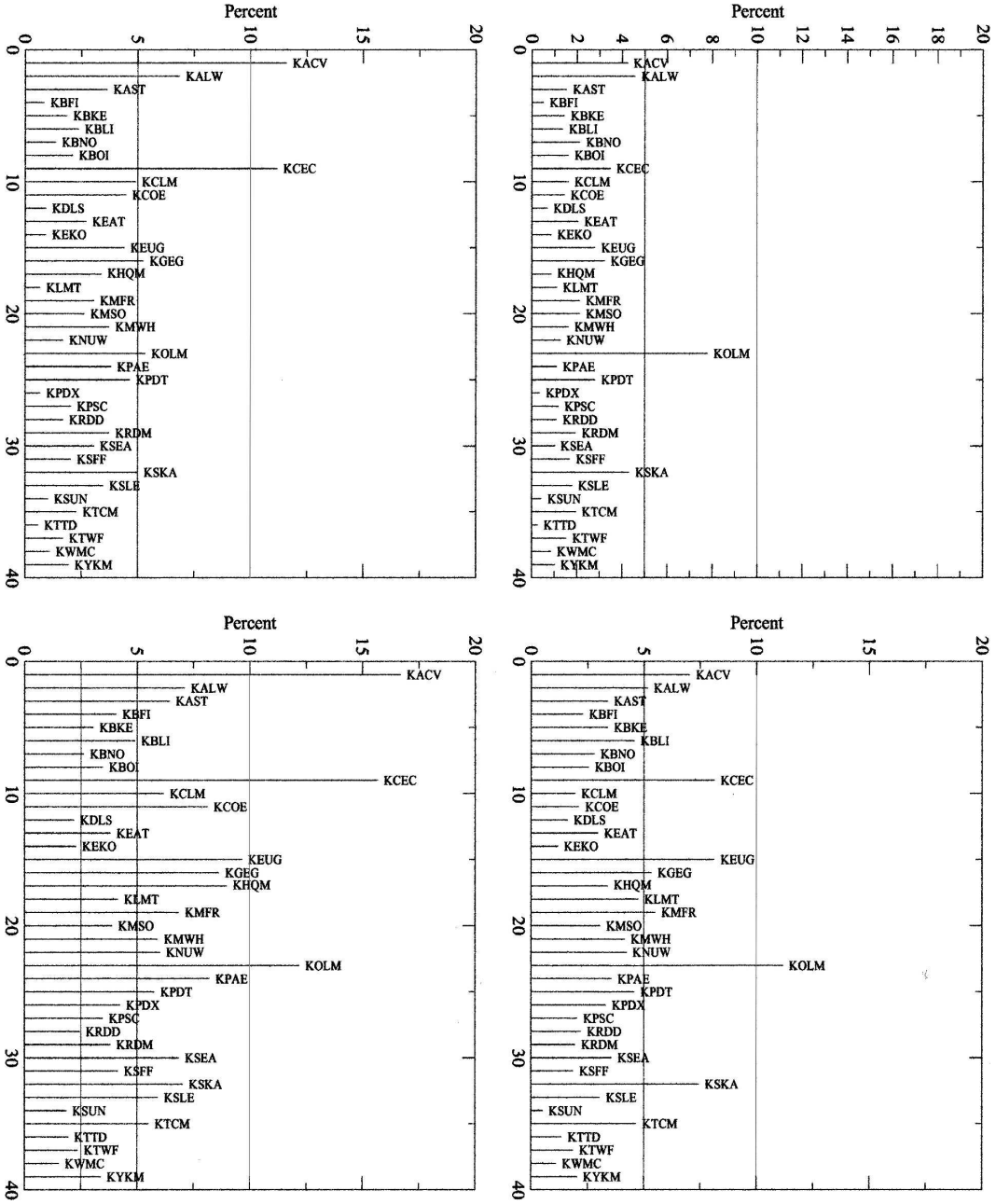

FIG. 2. The climatological frequency (%) of the lowest category of (left) ceiling and (right) visibility, at (top) 0600 and (bottom) 1200 UTC. The vertical lines at 5% and 10% are drawn for visual aid.

As mentioned previously, for some stations the occurrence of the lowest ceiling and visibility classes is extremely rare. So rare, in fact, that no satisfactory NN can be developed. For these stations, categories of ceiling (and visibility) are combined to render it binary; the highest category is considered as a distinct class, and the remaining categories are combined into a second class. The NNs based on this revised dataset are referred to as "two class," and the NNs based on the noncombined classes are labeled as "four class" or "three class," depending on whether the predictand is ceiling or visibility, respectively.

The performance of the NN is compared with that of logistic regression and MOS in terms of several measures of performance. For the two-class NNs, in addition to relative operating characteristic (ROC) diagrams, performance is gauged within the framework of probabilistic forecasting developed by Murphy and

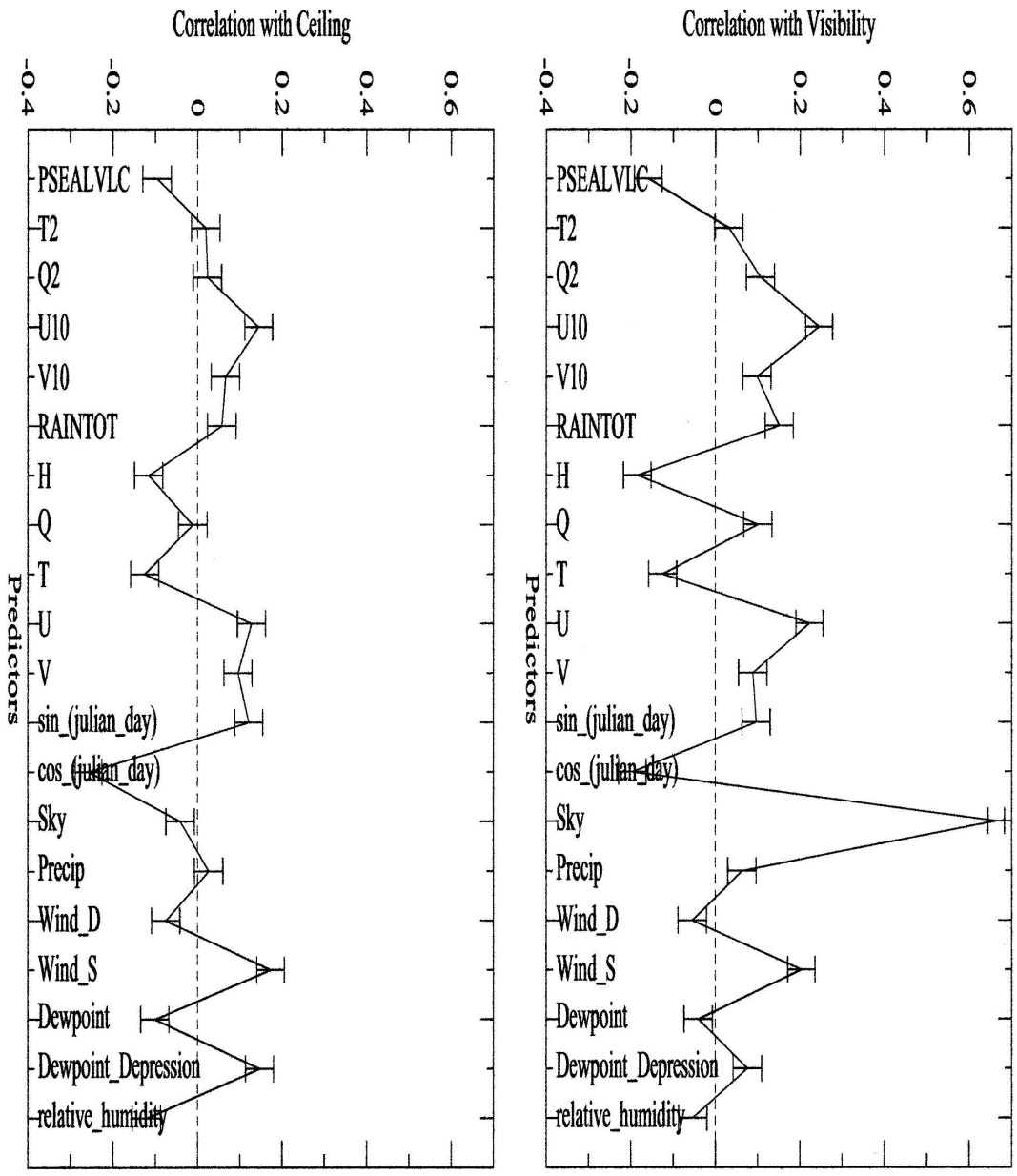

FIG. 3. The correlation coefficient between the predictors (Table 3) at 0000 UTC and (left) ceiling and (right) visibility at 1200 UTC for KOLM.

Winkler (1987, 1992)—specifically, reliability, discrimination, refinement, resolution, and attributes diagrams. Some scalar measures representing the quality of the probabilistic forecasts are also computed, including cross-entropy and ranked probability score. Because MOS forecasts are categorical, some scalar (nonprobabilistic) measures of performance are also computed, for example, the Heidke skill scores (HSS); this measure has been advocated as a relatively "healthy" measure (Marzban 1998) and has a natural generalization to nonbinary classes and, so, is utilized to asses the performance of the three-class and four-class NNs.

Two sets of NNs are developed for two different valid times: 0600 and 1200 UTC. The former utilize the model predictors at 0600 UTC, and the surface observations at 0000 UTC. These NNs produce forecasts with a lead time of 6 h. The NNs producing forecasts with a 12-h lead time are developed on model predictors at 1200 UTC, and surface observations at 0000

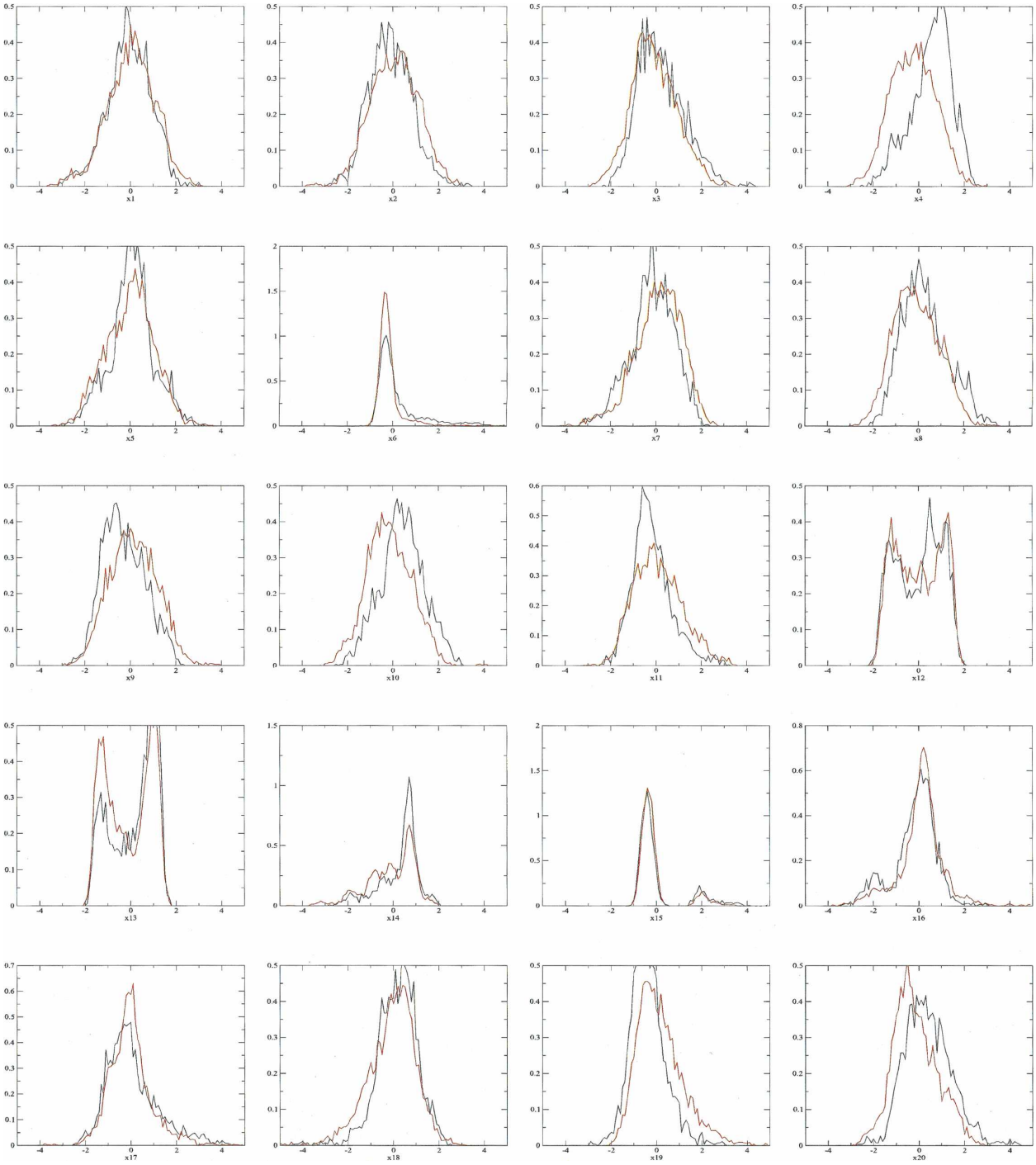

FIG. 4. The conditional distribution of the predictors (Table 3) at 0000 UTC, conditioned on low (black) and high (red) ceiling at 1200 UTC.

UTC. These NNs produce forecasts of ceiling and visibility at 1200 UTC.[3]

The number of NNs developed here is unwieldy. The 39 stations, binary and nonbinary classes, ceiling and visibility, five different hidden nodes (0, 2, 4, 6, and 8), four seeds, and 0600 and 1200 UTC forecasts altogether yield 6240 NNs. The verification task is even more complex because each of these NNs can be assessed in terms of a multitude of verification

[3] In principle, it is possible to arrange for model and surface observation variables to be at 0000 UTC, while the predictand is at the valid time. That possibility has been examined in Marzban et al. (2006).

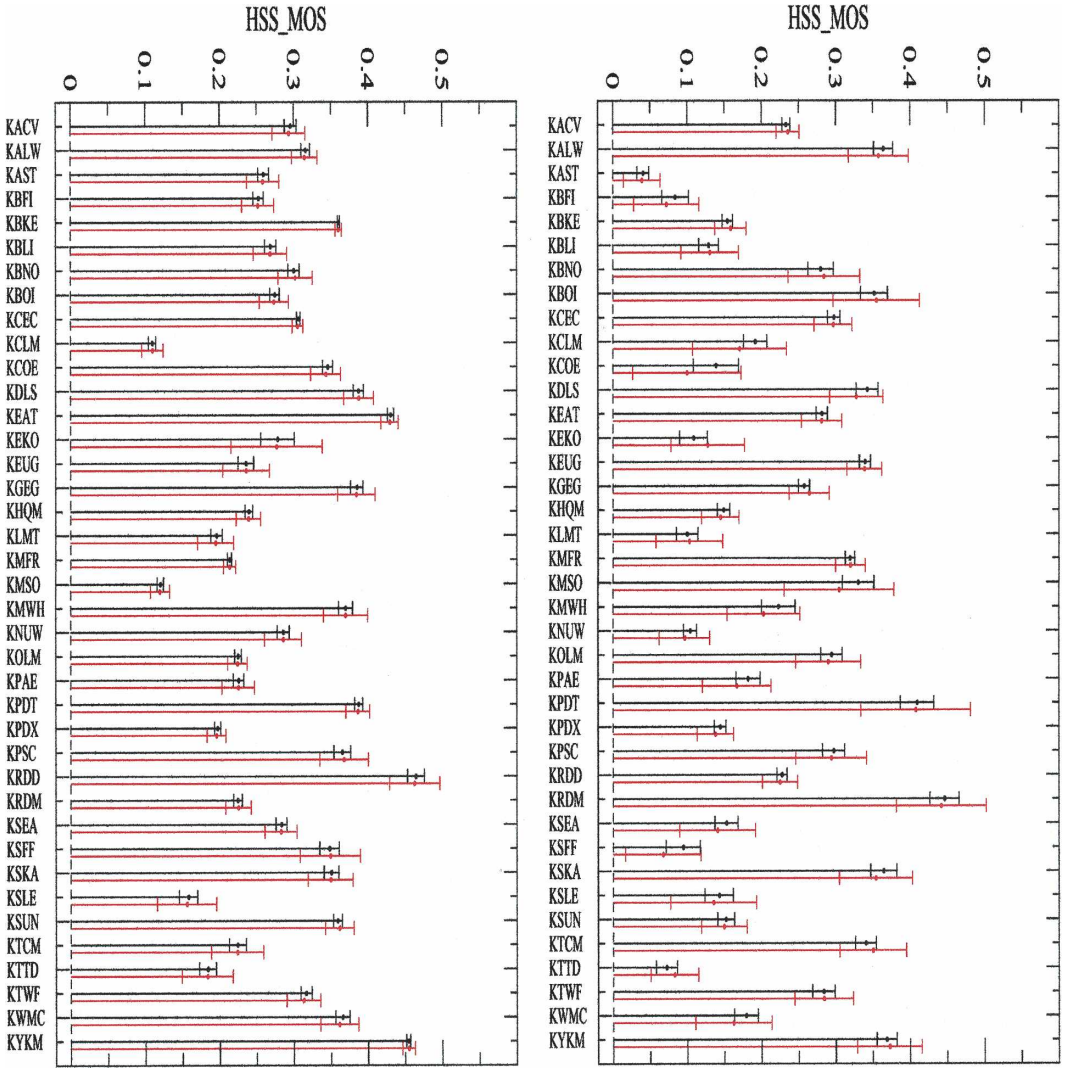

FIG. 5. HSS values according to MOS forecasts of (right) ceiling and (left) visibility, for training (black) and validation (red) sets, separately. The error bars are derived from bootstrapping.

measures. Here, only a sample of all the results is presented.

4. Results from preprocessing

The initial variables available for analysis numbered in the hundreds. However, after the preprocessing, only 20 remain—11 from MM5 and 8 from surface observations. The final list of variables is shown in Table 3. Of course, in addition to these predictors, the data include measurements of ceiling and visibility at every station and at 6-h intervals. The variables in the table are self-explanatory; the only ones that call for further clarification are variables 14–16. The variable "sky" refers to sky coverage, and takes the following values: clear, few, scattered, broken, overcast, and obscured. Variable 15 is binary and labels the occurrence or nonoccurrence of precipitation. Wind direction is recorded as NE, E, SE, S, SW, W, NW, and N. As a test of colinearity, the correlation coefficients between all pairs of variables are computed; no statistically significant correlations are found.

As mentioned previously, for some stations the data contain an exceedingly small number of low ceilings or low visibilities. Figure 2 shows the climatological frequency (in percent) of the lowest categories at 0600 and at 1200 UTC.[4] Low ceiling and visibility occur most

[4] The stations with low climatological frequency of low ceiling and visibility are those farther away from a body of water.

frequently at KACV, KCEC, KEUG, KOLM, and KSKA (see Table 2 for a list of station identifiers), with some variability between 0600 and 1200 UTC. The stations at which these conditions occur least frequently depend more strongly on the time of day: at 0600 UTC, they are KBFI, KDLS, KEKO, KPDX, KSUN, KTTD, KWMC, and KYKM; at 1200 UTC, they are KBKE, KBNO, KBOI, KDLS, KEKO, KPSC, KRDD, KSUN, KTTD, KWMC, and KYKM. The stations with the lowest occurrence of ceiling and visibility are expected to have poorly performing NNs.

As discussed in Marzban et al. (1999), the notion of a "best predictor" is ill-defined, at least outside of the framework of a statistical model. Nevertheless, this is a natural place in the analysis to explore the relationships among the various variables in more detail. Figure 3 shows the linear correlation coefficient between each of the predictors at 0000 UTC and the predictands—ceiling and visibility—at 1200 UTC for KOLM. Evidently, all of the predictors of ceiling have generally equal strength (within the error bars). The exceptions are T2, Q2, and Q, which have near-zero correlation with ceiling. However, this does not imply that they are poor predictors as far as the NN is concerned, because, first, their correlations may be nonlinear, and second, their interaction with the other variables may be important for predicting ceiling. This is why they are not excluded as inputs to the NNs. The best predictors of visibility again have comparable strength, with the exception of sky (sky coverage), which is highly correlated with visibility. The weakest predictors are T2 and dewpoint, but they are not excluded from the analysis for the aforementioned reasons. In short, for these reasons, further expounded upon in Marzban et al. (1999), no attempt is made to perform any variable selection prior to NN development.

A more complete representation of predictive strength is given in terms of conditional distributions. Figure 4 shows the conditional distribution of the predictors for low ceiling and high ceiling, separately.[5] It can be seen, for example, that x_1 (sea level pressure) is a poor predictor, while x_4 (u component wind at 10 m) is a relatively good predictor. Note that a binary variable such as x_{15} (precipitation occurrence) appears as a continuous quantity, because of the jitter introduced in the data (see section 3). Such plots provide a more meaningful assessment of the predictive strength of the predictors, but they are still univariate in nature and, so, are not utilized in determining the inputs to the NN.

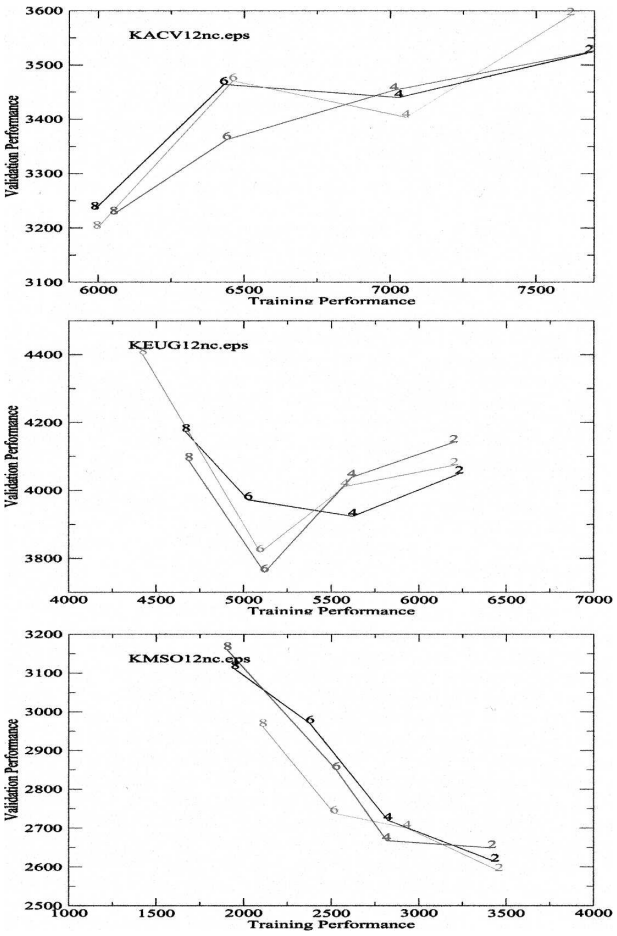

FIG. 6. The TV diagrams, with cross-entropy as the performance measure, for three stations with different levels of nonlinearity in the data: (top) extreme, (middle) midrange, and (bottom) no nonlinearity.

5. Results

To compare the NNs with MOS, the HSS scores for MOS are computed first. Figure 5 shows the HSS scores for 1200 UTC, 12-h MOS forecasts. The HSS is computed from 4×4 contingency tables, that is, for the classes adopted here (Table 1). Each station has two bars: one for the training set and one for the validation set. The error bars on each bar are computed from bootstrapping. The left panel is for ceiling and the right panel is for visibility. Note that MOS performs differently for ceiling and visibility. For example, the best ceiling forecasts occur at KEAT and KRDD, while the best visibility forecasts take place at KPDT and KRDM. Moreover, the uncertainty in the HSS values (i.e., error bars) is not consistent across stations. For

[5] It is possible to plot the conditional distributions for all four classes of ceiling, but the results are visually unappealing.

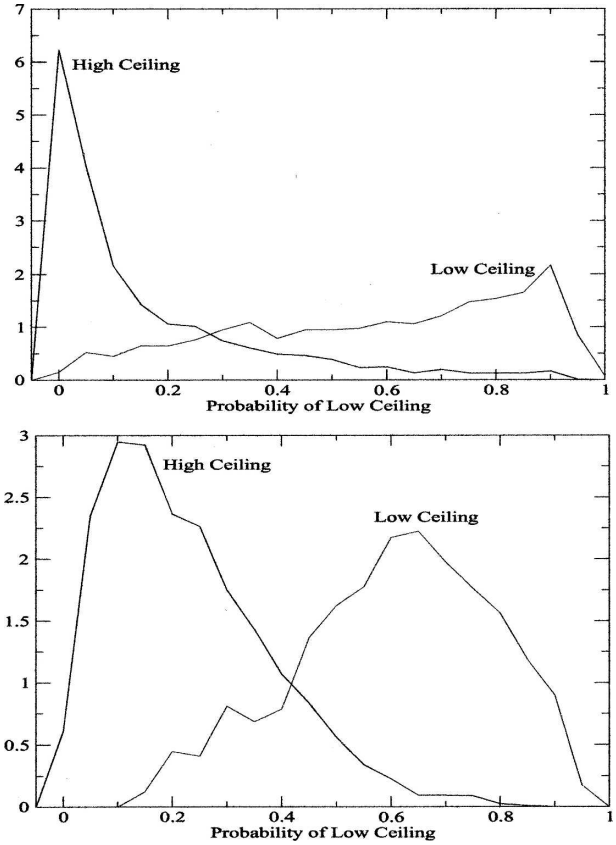

FIG. 7. Discrimination diagrams for (top) NN and (bottom) logistic regression forecasts for KNUW. The forecasts are for 1200 UTC ceiling, and the (optimal) number of hidden nodes is six.

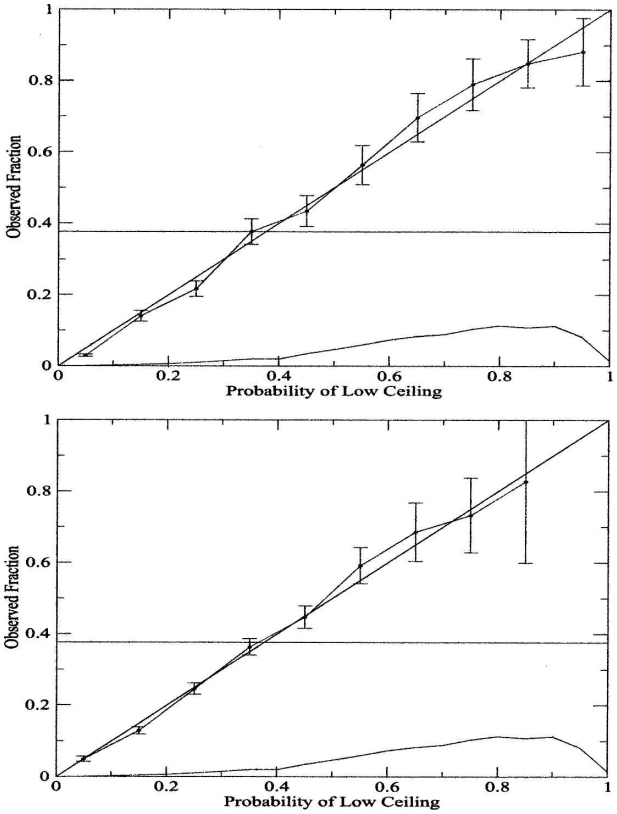

FIG. 8. Attributes diagrams for (top) NN and (bottom) logistic regression forecasts for KNUW. The forecasts are for 1200 UTC ceiling, and the (optimal) number of hidden nodes is six.

example, for ceiling, whereas stations like KBKE, KCEC, and KYKM are consistently in the high-HSS range, KEKO and KRDD display a large variability in performance.

As mentioned previously, the optimal number of hidden nodes is determined via bootstrapping. The data from different stations have different amounts of nonlinearity, leading to different number of hidden nodes. The TV diagrams, with cross-entropy as the performance measure, for three stations and three bootstrap trials are shown in Fig. 6. KACV, for example, has lower training and validation errors for more hidden nodes. This suggests a highly nonlinear relationship between the predictors and the predictand (here, ceiling).[6] KEUG, however, shows a strong preference for an NN with six hidden nodes; larger NNs lead to lower training but higher validation errors for each of three bootstrap trials. By contrast, any amount of nonlinearity introduced into the NN for KMSO causes overfitting.[7]

The discrimination diagram for the NN and logistic regression are shown in Fig. 7. These figures are for KNUW; the forecasts are 12-h forecasts of ceiling for 1200 UTC, and the NN has six hidden nodes (the optimal for that station). It can be seen that the NN makes more discriminatory forecasts; that is, the conditional distribution of the forecasts is concentrated around small values for cases with high ceiling, and the distribution for low ceiling is concentrated around high values (~0.9). By contrast, logistic regression has broader distributions for low and high ceilings, leading to a greater overlap between the two.

The attributes diagram for the NN and logistic regression are shown in Fig. 8. The station and the NN are the same as those described for Fig. 7. Both the NN and logistic regression produce highly reliable forecasts, as

[6] As a final step, in order to avoid overfitting, the largest allowed number of hidden nodes is eight.

[7] An NN with zero hidden nodes is equivalent to logistic regression; for clarity, the performance measures for logistic regression are shown in later figures.

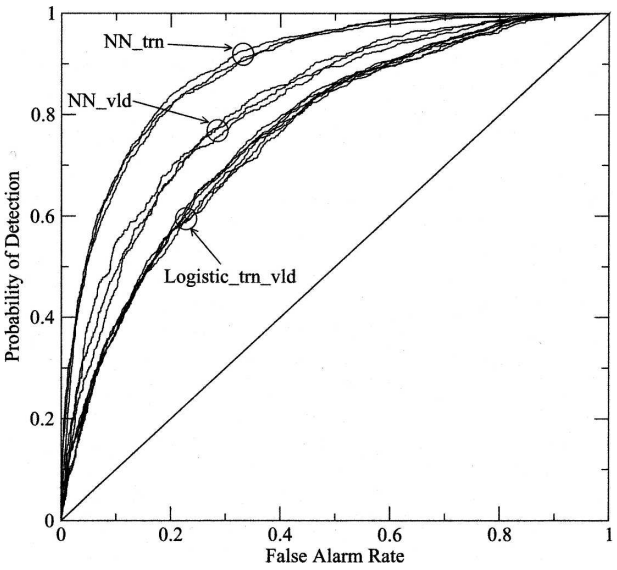

FIG. 9. ROC curves for NN and logistic regression, separately for training and validation data, and for three different seeds (see text).

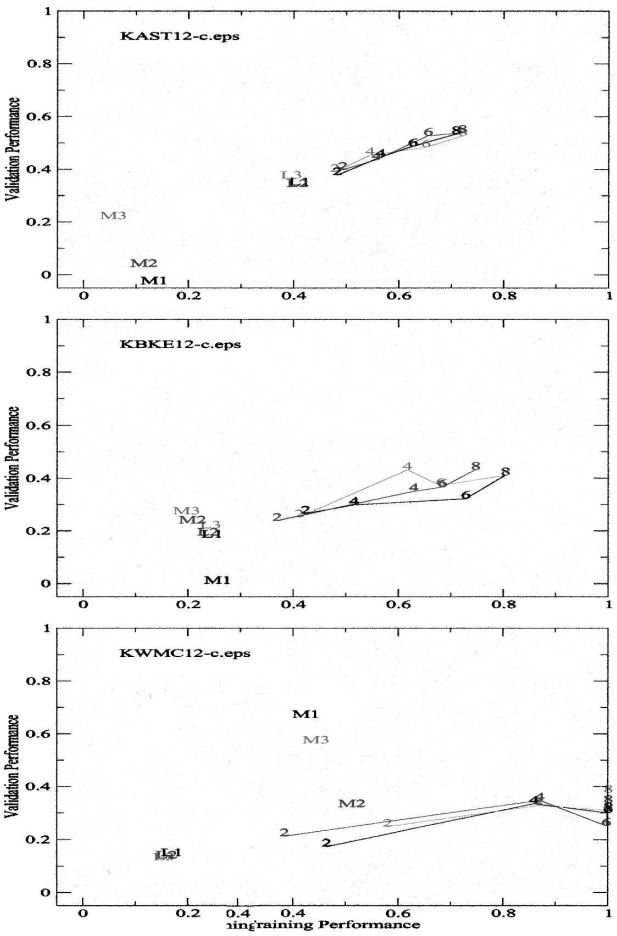

FIG. 10. The TV diagrams displaying HSS values of NNs with two, four, six, and eight hidden nodes, logistic regression (L), and MOS (M), for three bootstrap trials (labeled 1, 2, and 3). The underlying forecasts are two class.

evident from the overlap of the reliability curve and the diagonal line. One difference is that whereas the NN produces reliable forecasts (at least within the error bars) even when the probability of low ceiling is in the 0.9 range, logistic regression produces no forecasts in that range at all. Moreover, a comparison of the error bars between the two figures suggests that the NN's forecasts are more certain than those of logistic regression.

The bell-shaped curve in Fig. 7 is often called the refinement diagram and is simply the distribution of the forecasts without regard to low or high ceilings. Its skewed nature is simply a reflection of the rare nature of low-ceiling conditions.

Figure 9 shows the ROC curves for the NN and logistic regression. Here, the training and validation curves are plotted separately. The most concave curves correspond to three different training sets (i.e., seeds), and the intermediate curves are the associated validation sets. The least concave curves are for logistic regression; the training and validation curves are not as distinguishable as the NN curves. Again, it is evident that the NN outperforms logistic regression.

To compare NNs, logistic regression, and MOS, simultaneously, ceiling and visibility must be binary, because logistic regression models only two-class predictands. Furthermore, because MOS does not produce probabilistic forecasts, a categorical verification measure, such as HSS, must be employed. The TV diagrams showing HSS for three stations are shown in Fig. 10. The results from the bootstrap trials are displayed separately in order to convey some sense of sampling variation. The top panel in Fig. 10 illustrates that for three bootstrap trials, the NN outperforms logistic regression, which in turn outperforms MOS. This conclusion follows because the training and validation HSS values for the three NNs (regardless of the number of hidden nodes) are all superior to those of logistic regression; moreover, more hidden nodes lead to better performance on the training and validation sets. The middle panel in Fig. 10 illustrates a situation where MOS and logistic regression are comparable in their performance, but the NNs still outperform both. An ambiguous situation that does arise is exemplified in the bottom panel of Fig. 10. Here, MOS displays validation HSS values that are superior to both logistic regression and NN, in spite of the latter's wide spread in the training HSS values. As such, it is not possible to assess the relative performance of the three models.

TABLE 4. A coarse classification of the stations according to how significantly the NN outperforms MOS. The forecasts are 12-h forecasts for 1200 UTC, and ceiling is binary (i.e., takes two classes).

Significantly	Moderately	Questionably
KACV	KALW	KBNO
KAST	KBKE	KBOI
KBFI	KCEC	KEKO
KBLI	KCOE	KEUG
KCLM	KGEG	KMSO
KDLS	KMWH	KRDM
KEAT	KOLM	KSUN
KHQM	KPDT	KTCM
KLMT	KPDX	KWMC
KMFR	KPSC	
KNUW	KSKA	
KPAE	KSLE	
KRDD		
KSEA		
KSFF		
KTTD		
KTWF		
KYKM		

Analogous figures for all 30 stations have been produced, but are not shown here.[8] To distill that information, a coarse categorization of the results is in order. Given that logistic regression is an NN with zero hidden nodes, an analysis of such figures for all 39 stations allows for a coarse comparison of NN with MOS. Labeling the above-mentioned three situations as "significantly," "moderately," and "questionably," Table 4 lists the stations falling in each category. These categories are based on a visual inspection of the figures, and assess the relative separation between the performance values for the three models, but also with reference to the amount of spread among the three bootstrap trials.

HSS allows a comparison of the various models when ceiling and visibility are not binary. The results for the above-mentioned three stations are shown in Fig. 11, but an analysis of the figures for all 39 stations suggests that the difference in performance between the NN and MOS is generally smaller than in the binary case. Nevertheless, the NN does generally outperform MOS for many of the stations. For some stations, however, MOS outperforms NN. For instance, binary MOS forecasts of 0600 UTC ceiling for KEUG are superior to the NN forecasts. All of these comparisons are shown in Table 5. The letter N implies that the NN outperforms MOS,

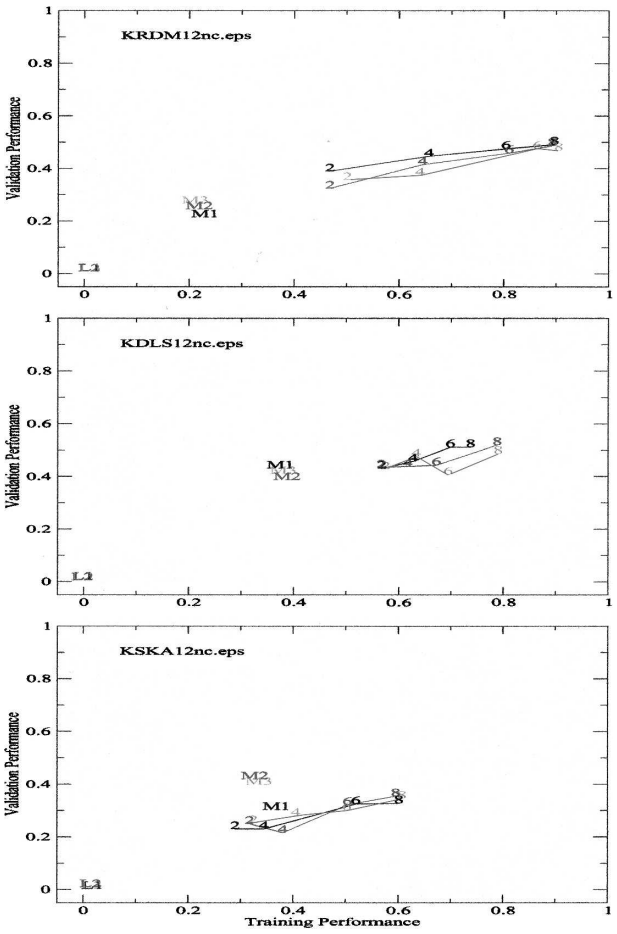

FIG. 11. Same as in Fig. 10 but for a ceiling having four classes.

and the letter O implies the opposite. Meanwhile, there are many situations where it is not at all clear which forecasts are best, and these are labeled with a dash; the cases where the comparison is ambiguous are generally when the amount of data is exceedingly small, and the occurrence of low ceiling and visibility is exceedingly rare.

6. Conclusions and discussion

MM5 model output and surface observations at 39 TAF stations are utilized to develop NNs for predicting ceiling and visibility with 6- and 12-h lead times. A number of verification measures are employed to compare the performance of these NNs with logistic regression and with traditional MOS. It is shown that for stations for which sufficient data exist for developing an NN, the NN outperforms logistic regression and MOS, with the degree of improvement depending on a number of factors, including lead time and the choice of the

[8] All the figures can be found in the appendix of http://www.stat.washington.edu/www/research/reports/2005/tr490.pdf.

TABLE 5. A coarse comparison of NN (marked with an N) and MOS (marked with an O) at all stations, for forecasts of ceiling and visibility at 0600 and 1200 UTC.

	Ceiling				Visibility			
	1200 UTC		0600 UTC		1200 UTC		0600 UTC	
	Two class	Four class	Two class	Four class	Two class	Three class	Two class	Three class
KACV	N	N	—	N	N	N	—	—
KALW	N	N	—	—	—	—	—	—
KAST	N	N	N	N	N	N	N	—
KBFI	N	N	—	N	N	N	N	—
KBKE	N	—	—	—	N	N	—	—
KBLI	N	N	N	N	N	N	—	—
KBNO	—	—	—	—	—	—	—	—
KBOI	N	N	—	—	—	—	—	—
KCEC	N	N	O	—	N	—	—	—
KCLM	N	N	—	—	N	N	—	—
KCOE	N	—	—	—	N	N	N	—
KDLS	N	—	—	—	N	—	—	—
KEAT	N	—	N	—	N	—	—	—
KEKO	N	N	N	—	N	N	N	—
KEUG	—	N	O	—	—	—	—	—
KGEG	N	—	—	O	N	—	—	—
KHQM	N	N	N	N	N	N	N	—
KLMT	N	N	—	—	N	N	—	—
KMFR	N	N	—	N	—	—	—	—
KMSO	N	N	—	N	N	—	—	—
KMWH	N	N	—	—	N	N	—	—
KNUW	N	N	—	—	N	N	N	N
KOLM	N	N	—	—	N	—	—	—
KPAE	N	N	—	—	N	N	—	—
KPDT	N	N	—	—	N	—	—	—
KPDX	N	N	O	—	N	N	—	—
KPSC	N	N	—	—	N	N	—	—
KRDD	N	N	—	—	N	N	—	—
KRDM	—	N	—	—	—	—	—	—
KSEA	N	N	N	—	N	N	—	—
KSFF	N	N	N	—	N	N	N	—
KSKA	N	—	—	O	—	—	—	O
KSLE	N	N	O	—	N	N	—	—
KSUN	—	N	N	—	N	N	N	N
KTCM	N	N	O	—	N	—	—	—
KTTD	N	N	—	—	N	N	—	—
KTWF	N	N	—	—	N	N	—	—
KWMC	—	—	—	—	N	—	N	—
KYKM	N	N	N	—	N	—	—	—

station itself. This suggests that the NNs should be employed in real time, especially because they are developed to produce highly reliable and discriminatory probabilistic forecasts (in contrast to MOS's categorical forecasts).

The comparison of statistical models is contingent on the data at hand, and therefore, the above conclusions are valid only for the data employed herein. In fact, the manner in which the NNs are developed is apt to render their performance positively biased. Specifically, although a number of steps are taken to avoid overfitting, it is still possible that some degree of overfitting has occurred. This is mostly because the training and validation bootstrap data are not entirely independent of one another; although seasonal effects are removed, some autocorrelation is apt to remain, thereby rendering a given training set and validation set dependent. This may appear to unfairly favor NN over MOS; however, it must be pointed out that the regression models underlying the MOS forecasts employed here are likely developed on the same data used for comparing MOS with NN. In other words, the MOS forecasts are also positively biased. A better comparison of MOS and NN would involve truly independent data; these data are

TABLE 6. A typical conditional (on ceiling) probability distribution between ceiling and visibility.

	Visibility (%)			
Ceiling	35	32	33	100
	5	18	77	100
	1	7	92	100
	0.04	0.33	99.63	100

being currently archived, and the comparison will be reported upon at a later time.

Changes to the MM5 can be accommodated in several ways: 1) by running an older postprocessor on new model output, 2) by developing a new postprocessor developed from only new data produced from new runs of the new model, and 3) on new runs of the new model on archived data. A priori it is not clear which approach would produce better forecasts. These options will be tested in the future.

Other future work involves developing separate models for warm and cool seasons, more sophisticated preprocessing (e.g., removal of autocorrelations), and variable selection. An attempt to employ principal components of the data as inputs to the NNs yielded no noticeable improvement.

A more promising direction is the development of NNs that utilize the relationship between ceiling and visibility to improve the forecasts. All of the NNs described here were designed to predict either ceiling or visibility. There is, however, some evidence that an NN that predicts multiple predictands (e.g., ceiling and visibility) simultaneously may outperform separate NNs designed to predict each predictand, separately. The advantage of the former is most evident when the predictands are correlated. A typical conditional (on ceiling) probability distribution between ceiling and visibility is shown in Table 6. The structure of this table implies that when the ceiling is high, the visibility is high as well. By contrast, a low ceiling does not imply low visibility. The complex nature of the association between ceiling and visibility is ideal for developing an NN that predicts both ceiling and visibility, simultaneously. This idea will be examined in the future.

Acknowledgments. The authors are grateful to Patrick Tewson of the Applied Physics Laboratory, University of Washington, for providing the MM5 data. Partial funding for this work was provided under a cooperative agreement between the National Oceanic and Atmospheric Administration (NOAA) and the University Corporation for Atmospheric Research (UCAR). The views expressed herein are those of the authors and do not reflect the views of NOAA, its subagencies, or UCAR.

REFERENCES

Casaioli, M., R. Mantovani, F. P. Scorzoni, S. Puca, A. Speranza, and B. Tirozzi, 2003: Linear and nonlinear post-processing of numerically forecasted surface temperature. *Nonlinear Processes Geophys.,* **10,** 373–383.

Costa, S. B., F. Carvalho, R. Amorim, A. Campos, C. Ribeiro, V. Carvalho, and D. dos Santos, 2006: Fog forecast for the international airport of Maceio, Brazil, using artificial neural networks. *Proc. Eighth Int. Conf. on Southern Hemisphere Meteorology and Oceanography,* Foz do Iguacu, Brazil, INPE, 1741–1750.

Grimit, E. P., and C. F. Mass, 2002: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecasting,* **17,** 192–205.

Hilliker, J. L., and J. M. Fritsch, 1999: An observations-based statistical system for warm-season hourly probabilistic forecasts of low ceiling at the San Francisco International Airport. *J. Appl. Meteor.,* **38,** 1692–1705.

Hsieh, W. W., and B. Tang, 1998: Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bull. Amer. Meteor. Soc.,* **79,** 1855–1870.

Koistinen, P., and L. Holmstrom, 1992: Kernel regression and backpropagation training with noise. *Neural Info. Process. Syst.,* **4,** 1033–1039.

Leyton, S. M., and J. M. Fritsch, 2003: Short-term probabilistic forecasts of ceiling and visibility utilizing high-density surface weather observations. *Wea. Forecasting,* **18,** 891–902.

——, and ——, 2004: The impact of high-frequency surface weather observations on short-term probabilistic forecasts of ceiling and visibility. *J. Appl. Meteor.,* **43,** 145–156.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting,* **13,** 753–763.

——, 2000: A neural network for tornado diagnosis. *Neural Comput. Appl.,* **9,** 133–141.

——, 2003: A neural network for post-processing model output: ARPS. *Mon. Wea. Rev.,* **131,** 1103–1111.

——, E. D. Mitchell, and G. Stumpf, 1999: The notion of "best predictors": An application to tornado prediction. *Wea. Forecasting,* **14,** 1007–1016.

——, S. Sandgathe, and E. Kalnay, 2006: MOS, perfect prog, and reanalysis. *Mon. Wea. Rev.,* **134,** 657–663.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

——, and ——, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting,* **7,** 435–455.

Pasini, A., V. Pelino, and S. Potesta, 2001: A neural network for visibility nowcasting from surface observations: Results and sensitivity to physical input variables. *J. Geophys. Res.,* **106,** 14 951–14 959.

Roebber, P., S. Bruening, D. Schultz, and J. V. Cortinas Jr., 2003: Improving snowfall forecasting by diagnosing snow density. *Wea. Forecasting,* **18,** 264–287.

Vislocky, R. L., and J. M. Fritsch, 1997: An automated, observations-based system for short-term prediction of ceiling and visibility. *Wea. Forecasting,* **12,** 31–43.