A Bayesian Neural Network for Severe-Hail Size Prediction

CAREN MARZBAN

National Severe Storms Laboratory, and Cooperative Institute for Mesoscale and Meteorological Studies and Department of Physics, University of Oklahoma, Norman, Oklahoma

ARTHUR WITT

National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 23 May 2000, in final form 26 February 2000)

ABSTRACT

The National Severe Storms Laboratory has developed algorithms that compute a number of Doppler radar and environmental attributes known to be relevant for the detection/prediction of severe hail. Based on these attributes, two neural networks have been developed for the estimation of severe-hail size: one for predicting the severe-hail size in a physical dimension, and another for assigning a probability of belonging to one of three hail size classes. Performance is assessed in terms of multidimensional (i.e., nonscalar) measures. It is shown that the network designed to predict severe-hail size outperforms the existing method for predicting severe-hail size. Although the network designed for classifying severe-hail size produces highly reliable and discriminatory probabilities for two of the three hail-size classes (the smallest and the largest), forecasts of midsize hail, though highly reliable, are mostly nondiscriminatory.

1. Introduction

The National Severe Storms Laboratory (NSSL) has developed numerous algorithms for the detection of atmospheric phenomena. These algorithms reside collectively within an ‘‘umbrella’’ program called the Severe Storm Analysis Package (SSAP). Two such algorithms designed for the detection of tornadoes and/or damaging wind have recently been supplemented with neural networks (NNs) (Marzban and Stumpf 1996/ 1998) and their performance has been compared with conventional statistical methods in Marzban et al. (1997). NSSL has also developed a Hail Detection Algorithm (HDA) that computes a number of attributes believed to be relevant for the detection/prediction of severe hail (Witt et al. 1998a). The improvement brought about by the tornado and damaging wind NNs renders it natural to develop a similar NN to complement the HDA; in this scheme, the HDA and other algorithms in SSAP compute the attributes that are employed as the NN's inputs, while the output of the NN is based on ground truth (i.e., reported severe-hail size).

The statistical theory of NNs and the methodology of employing them is now well developed (Bishop 1996). It has become common practice in many statistical model-building tasks to examine NNs, as a candidate, alongside traditional methods. There exist regression models developed for the same purpose—a linear regression model for predicting size, and a logistic regression model for classifying different size classes (Billet et al. 1997). Both models are specific instances of NNs, and so no attempt is made to compare NNs with other statistical models. There are some differences between the models developed here and those of Billet et al. (e.g., choice of the predictors). Therefore, prior to ‘‘fielding’’ a hail size prediction algorithm, a performance comparison should be made.

The question of whether NNs are superior to traditional methods has been extensively examined (see ftp://ftp.sas.com/pub/neural/FAQ.html). The general consensus (Bishop 1996) appears to be that most commonly employed (if not all) statistical methods are in fact equivalent to some NN.[1] As such, NNs are not necessarily superior to other methods and represent a convenient all-in-one ‘‘packaging.’’ At the same time, the manifestations of that convenience can often be quite dramatic.

An NN's ability to represent a wide range of models is usually accompanied by one disadvantage, namely an NN's tendency to overfit data. Overfitting occurs when

Corresponding author address: Dr. Caren Marzban, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: marzban@nssl.noaa.gov

[1] For example, an NN with a linear activation function and a mean square error cost function is equivalent to linear regression. Similarly, an NN with a logistic activation function, minimizing cross entropy (see below), is equivalent to logistic regression.

a model captures the statistical noise in the data rather than the underlying signal. The problem of overfitting is not peculiar to NNs; indeed, all regression models can suffer from it to various degrees. A simple linear regression model is prone to overfitting if the independent variables have been transformed in some highly nonlinear fashion. Polynomial regression, although capable of representing a wide range of functions (i.e., all polynomials), can easily overfit data because of the exponential growth of the number of parameters as a function of the number of predictors. In that respect, NNs occupy a special state in that the growth of the number of parameters is only linear with the number of predictors [see ''curse of dimensionality'' in Bishop (1996)]. At the same time, this restrained growth does not adversely affect an NN's ability to be a universal approximator (Hornik et al. 1989).

In spite of being the least susceptible, NNs are capable of and do overfit. A number of popular methods for identifying a model that overfits utilize several independent datasets; these methods fall under the general classes of split-sample and resampling methods. Examples are cross-validation, jackknifing, and bootstrap (Bishop 1996). In these methods, one or more data subsets are employed for estimating the parameters of the NN (i.e., training), and the remaining (validation) subsets are employed for determining the optimal complexity (nonlinearity) so as to prevent overfitting. These methods are employed to *identify* the onset of overfitting. Of course, knowledge of the overfitting model allows for the identification of the optimal model (that does not overfit).

Additionally, there exist means of *restraining* the overfitting problem (Sarle 1995). For example, the introduction of a weight-decay term into the error function can restrict the overfitting problem. Indeed, a weight-decay term can be arranged to preclude overfitting altogether but only at the cost of rendering the NN linear. Using methods of Bayesian inference, it is possible to arrive at a weight-decay term that is optimal in the sense that the NN's ability to overfit can be limited but without compromising its nonlinearity. The details of this Bayesian approach are beyond the scope of this article, but further details and an application to tornado prediction can be found in MacKay (1996), Neal (1996), Wolpert (1993), and Marzban (1998). A non-Bayesian NN for severe-hail size prediction developed recently (Marzban and Witt 2000) has displayed some symptoms of overfitting, and so, with the aim of limiting the overfitting problem, the NN developed herein is Bayesian (as defined here). Meanwhile, bootstrapping will be employed to identify the onset of overfitting.

The development of an NN model for severe-hail size prediction can be divided into two subtasks: one of developing a model that predicts the occurrence of severe hail, and another that predicts the size of severe hail, given that severe hail has occurred or is expected to occur. Only the latter will be considered in this article, as the former model is currently under construction.

TABLE 1. The nine predictors.

No.	Description
1	Cell-based vertically integrated liquid
2	Severe-hail index
3	Storm-top divergence (delta-V in m s^{-1})
4	Midaltitude rotational velocity (m s^{-1})
5	Height of the wet-bulb zero [km above sea level (MSL)]
6	Height of the melting level (km MSL)
7	Vertically integrated wet-bulb temp
8	Wind speed at the equilibrium level (m s^{-1})
9	Storm-relative flow at −20°C level (m s^{-1})

An NN for the prediction of size alone can be developed in two independent ways: one can develop an NN that predicts the size of hail in some physical unit (e.g., in. or mm), or one can assess the probability of belonging to some size range. The data suggest that severe hail reports fall naturally into three different classes, corresponding to coin size, golfball size, and baseball size, in an ordinal fashion. As such, it is possible to assess the probability of a severe hail report belonging to each of these three classes. The former approach models a continuous quantity, and so, falls in the domain of regression, while the latter is an example of a classification problem. Both NNs will be considered herein, and will be referred to as the regression NN and the classification NN, respectively.

In what follows, the data and the methodology are further described, and multidimensional (e.g., distribution based) measures of performance are set forth in terms of which the performance of the NNs is gauged. Finally, a discussion section offers some final thoughts on the matter of an NN for hail size prediction.

2. Data

The input variables provided to the NN include a mix of Doppler radar–derived parameters along with several parameters representing the near-storm environment. The radar parameters include two based on reflectivity data, cell-based vertically integrated liquid (Johnson et al. 1998) and the severe hail index (Witt et al. 1998a), as well as two based on velocity data, storm-top divergence (Witt and Nelson 1991), and midaltitude rotational velocity (Witt 1998). The near-storm environment parameters include three based on thermodynamic data and two based on kinematic data (Table 1). The vertically integrated wet-bulb temperature parameter is computed by integrating the wet-bulb temperature profile from the surface to the height of the wet-bulb zero. These near-storm environment parameters are either calculated within the SSAP using numerical model data, or they can be calculated from sounding data and manually entered into an adaptable parameter file. For this study, all the near-storm environment parameters were calculated from sounding data. For each individual

TABLE 2. Summary of the 81 storm cases analyzed.

Region	No. of days	No. of hailstorms
Western United States	13	21
High plains	7	27
Southern plains	22	115
Midwest	15	85
SE United States	19	94
NE United States	6	44
Total	81	386

''storm event'' analyzed,[2] a single sounding was used, with the most representative sounding being chosen from among the available candidates. Factors affecting the choice were proximity to the midpoint, in time, of the storm event, being in the ''inflow sector'' of the event, and being reasonably close (within 400 km) to the event (Rasmussen and Blanchard 1998). Incomplete soundings that did not allow for the calculation of all the environmental parameters, and soundings that appeared to be contaminated by convection, were disqualified.

The verification data on hail size comes from *Storm Data*. Because *Storm Data* is a collection of severe weather reports, the minimum hail size in this study is 19 mm (0.75 in.). There are numerous problems associated with using *Storm Data* for verifying radar-based algorithm predictions (Witt et al. 1998a,b). For prediction of maximum hail size, the primary concern involves the possibility that any given hail report is not representative of the maximum size being produced by the storm (at the time of the report). To minimize the impact of this possibility, the analysis was restricted to the maximum size observed per hailstorm (Witt 1998). For each severe hailstorm, a 20-min time window (Witt et al. 1998a) was used to relate the predictor variables to the maximum reported hail size. For the radar-based parameters, the maximum value within the time window was used, whereas for the near-storm environment parameters, an average value was used. Since a single sounding was used for each severe event, the only environmental parameter that actually required averaging across the 20-min time window was the storm-relative flow (due to changes in the storm motion vector).

In situations where there are multiple reports of the same maximum hail size for a storm, the report corresponding to the time period when the storm appeared to be weaker was used, since these radar characteristics were indicative of the minimum strength necessary to produce the observed maximum hail size. For example, suppose one is using vertically integrated liquid (VIL) to predict maximum hail size. Then, if there are two reports of golfball-size hail with a storm, and the VIL is 50 for one report and 60 for the other, it is reasonable to assume that a VIL of 50 is the representative value associated with golfball-size hail for this storm. However, this condition (using the weaker period) only applies to multiple reports where all the predictor variables have been measured within the time window; that is, the storm-top divergence or midaltitude rotational velocity parameters are not ''missing'' due to range folding. In cases where the choice was between using a stronger period with a full set of data or a weaker period that includes missing data, then the stronger period was used.

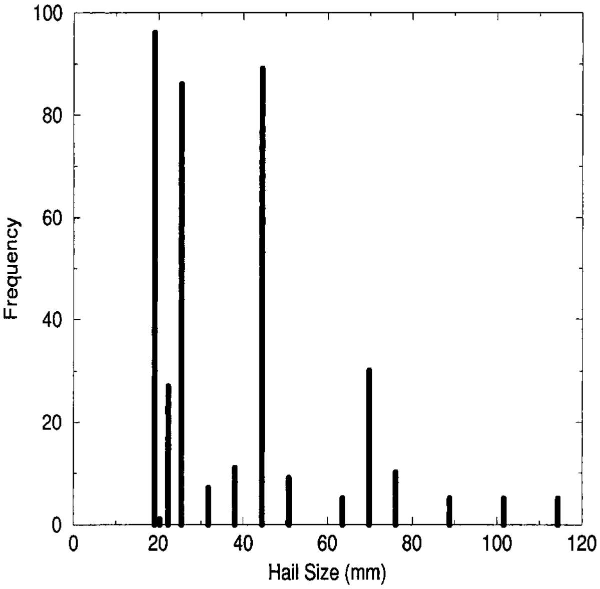

FIG. 1. The distribution of reported hail size.

The HDA and other algorithms within SSAP were employed to compute the NN input values for 81 storm days from across the United States (Table 2), on which 386 severe hailstorms were observed. The distribution of maximum reported hail sizes for these 386 reports is shown in Fig. 1. The common practice of reporting hail size using familiar circular or spherical objects (e.g., various coins or balls) is clearly evident, as reports tend to be clustered along discrete sizes. The highest frequency corresponds to dime-, nickel-, and quarter- (coin) size hail (19–25 mm), golfball-size hail (44 mm), and baseball-size hail (70 mm). It would appear that few hail reports are actually measured to obtain a precise reading of their size and that most hail sizes are estimated. Hence, one must assume that a certain amount of ''rounding off'' error exists in the observations, and this error appears to increase as hail size increases. Because of all this, the prediction of maximum hail size was approached from two distinct directions: regression and classification.

3. Bayesian neural network

In NNs, nonlinearity is determined by two quantities, the number and the magnitude of the parameters

[2] A storm event is defined as a continuous period of time (up to 24 h long) when convective activity is occurring within 230 km of a WSR-88D site.

(weights). The former is determined by the number of hidden nodes (if the inputs are not collinear), and the latter is determined by the variance of the weights. If the number of independent weights greatly exceeds the sample size, and the magnitude of the weights greatly exceeds unity (if the inputs are normalized in some fashion), then an NN can overfit data. If, however, either of these conditions is not met, then overfitting is unlikely. For instance, if the magnitude of the weights is restricted to be much less than one, then most activation functions are linear, and so the NN simply becomes linear regression, regardless of the number of hidden nodes.

As such, it is possible to restrain overfitting, even with a large NN (i.e., with many hidden nodes), if the weights are prevented from becoming too large. The magnitude of the weights can be constrained by introducing a weight-cay term into the error function that is to be minimized (Bishop 1996). The optimal value of the coefficient of this term can be inferred by Bayesian techniques. One of the advocates of this method is MacKay (1996), and the methodology followed in this article is also that of MacKay. The NN program employed in this research is a variation on that developed by MacKay (which is available on the Web at http://wol.ra.phy.cam.ac.uk/mackay).

Both the split-sample and resampling methods rely on splitting the data into three sets: a training set, employed to estimate the parameters of the model; a validation set for optimizing the complexity of the model; and a test set for estimating the unbiased performance of the model. So, for the purpose of identifying the optimal complexity (i.e., number of hidden nodes, or the magnitude of the weights) it is sufficient to split the data into only two sets: a training and a validation set. For small sample sizes, it is preferable to employ a resampling method instead of a split-sample method such as cross-validation (Bishop 1996). An example of a resampling method for model selection is bootstrapping (Efron and Tibshirani 1993). In its simplest form, one repeatedly trains with subsamples of the data, and the optimal NN is selected to be the one with the lowest average error over the unused subsamples. From the variance (over the subsamples) of the validation errors one can construct a confidence interval for the performance of the NN.

In addition to the variance due to resampling, in NNs there exists a variance due to local minima. In other words, there is no assurance that all the NNs trained on the different subsamples rest in the same minimum of the error function. Of course, in practice, only the deepest, practically reachable local minimum is of interest. Specifically, the distribution (over the different subsamples) of the values of the error function at the local minima can aid in assessing the likelihood of obtaining a deeper local minimum than a given one. Thus, a local minimum that is highly unlikely to be won over by a deeper one is considered to be the deepest local minimum. This issue has been thoroughly discussed in Marzban (2000) and Marzban and Witt (2000); suffice it to say that the ‘‘global’’ minimum referred to in the current article is the deepest local minimum as outlined above.

4. Methodology

As mentioned previously, the problem of predicting maximum hail size can be approached either from the point of view of regression or classification. This can be illustrated by examining the distribution of the hail size dataset (Fig. 1). The number of distinct values (i.e., 14) is sufficiently large to warrant a regression analysis wherein the predictand is hail size. On the other hand, the existence of clusters/peaks in that distribution suggests three distinct classes for hail size, corresponding to coin size, golfball size, and baseball size. As such, a classification approach is also feasible. Both approaches are fruitful in that the former provides estimates for hail size in some physical unit, and the latter can assess the probability of belonging to one of the three classes of hail size.

In the regression approach, an appropriate measure of error is the mean square error (MSE), defined as

$$\mathrm{MSE} = \frac{1}{N}\sum [t - y(\mathbf{x}, \boldsymbol{\omega})]^2, \qquad (1)$$

where $\mathbf{x}$ is the vector of inputs (attributes), $\boldsymbol{\omega}$ is the vector of the weights, and t is the target value that is to be estimated by the output $y(\mathbf{x}, \boldsymbol{\omega})$. The summation is over the number of cases, N, in the relevant dataset (training, validation, etc.). The activation function for the hidden nodes is taken to be the logistic function, $f(x) = 1/[1 + \exp(-x)]$, while that of the single output node is the linear function. The former introduces the necessary nonlinearity into the NN, and the latter allows for the output node to take the full range of values taken by the target.

In the classification approach, the appropriate error function is cross-entropy, defined as

$$S = -\frac{1}{N}\sum \{t \log y(\mathbf{x}, \boldsymbol{\omega}) + (1 - t)\log[1 - y(\mathbf{x}, \boldsymbol{\omega})]\}. \qquad (2)$$

In the c-class case, with c output nodes representing class membership, and with an appropriate choice of activation functions [i.e., the softmax function (Bishop 1996), a generalization of the logistic function to multiple output nodes], the minimization of S yields outputs that can be interpreted as the posterior probability of class membership, given the inputs (Richard and Lippmann 1991). Such a probability is precisely what is required for forecasting purposes. An additional weight-decay term of the form

$$S_\omega = \alpha\frac{1}{2}\sum \boldsymbol{\omega}^2$$

restrains the size of the weights (and therefore, overfitting) without affecting the probabilistic interpretation of the output nodes. The total error function is, therefore, $S + S_{\omega}$.

Another assumption of Richard and Lippmann (1991) is that 1-of-*c* coding be employed for coding the classes among the output nodes. This means that three output nodes are required to represent three classes, with the largest output node designating the corresponding class. Accordingly, the NNs described herein have three output nodes. This is in contrast to the NNs discussed in Marzban and Witt (2000) wherein the problem of predicting the three classes was decomposed into three two-class problems. In other words, three NNs were developed: one for discriminating between coin-size hail and otherwise, a second for discriminating between golfball-size hail and otherwise, and a third NN for discriminating between baseball-size hail and otherwise. As such, each network required only one output node, and a simple application of Bayes' theorem showed that the output nodes represented posterior probabilities for the corresponding classes. The reason for this was to reduce the chance of overfitting by reducing the number of parameters between the hidden and the output layer. In the current analysis, however, overfitting is restrained through the weight-decay term in the error function, and so it is safe to allow for three output nodes.

In both the regression and the classification schemes the number of input nodes is nine (or seven, if there are missing data; see the last paragraph in this section), namely, the total number of hail attributes. Experimentation with smaller numbers of input nodes suggests that nothing is gained by employing subsets of the nine attributes as inputs (see the discussion section).[3] Each attribute is centralized by subtracting the mean and dividing by the respective standard deviation. This can stabilize and even expedite convergence to the minimum of the error function.

As mentioned previously, the number of hidden nodes (on one hidden layer) was determined via bootstrapping, and it was found unnecessary to go beyond one hidden layer. As expected, what is gained by employing the Bayesian procedure for inferring the strength of the weight-decay term (α) is that no loss of performance is found even for a larger number of hidden nodes. In other words, the Bayesian method yields a *range* of numbers of hidden nodes within which the NN's performance is insensitive to the precise number of hidden nodes.

There is also the matter of missing data. One common approach is to simply replace any missing data with the average of the nonmissing data for each predictor. For a skewed distribution, however, this approach can be unsatisfactory. In the present application, the fortunate situation arises in which the missing data always appear in only two of the predictors (storm-top divergence and midaltitude rotational velocity). As such, it is possible to develop two NNs, one based on all the data and all nine inputs (when all nine inputs are available), and another trained on only seven inputs for which there are no missing data. This was done for both the regression and the classification NN.

For both the classification and regression NNs, the $N = 386$ cases were divided into a training set (250) and a validation set (136), and this partitioning was repeated four times for the purpose of bootstrapping. The size of the training set is (approximately) $(1 - e^{-1})N \sim (2/3)N$ as suggested in a bootstrapping scheme with replacement (Efron and Tibshirani 1993).

5. Performance measures

Although single, scalar (i.e., one-dimensional) measures such as MSE and cross-entropy are minimized during the training phase, it is important to assess the performance of the NN in a multidimensional sense such as with scatterplots, distribution plots, or attributes diagrams (Wilks 1995; Murphy and Winkler 1992).

The performance of the regression network can be assessed in terms of the scatterplot of the actual versus predicted hail size. Such a plot can display regions of forecasts that may be problematic as well as regions where the NN performs superbly. Of course, it is possible to distill this two-dimensional measure into a single, scalar measure such as R^2, which reflects the amount of variance explained by the model, but such a reduction is apt to lead to loss of information. Another multidimensional view of performance is offered by a residual plot (Draper and Smith 1981). This is a plot of the residues (i.e., actual minus predicted) as a function of predicted size.

For the classification network, there are many more ways of expressing performance. Classification performance can be assessed not only in terms of scalar and multidimensional measures, but also through categorical and probabilistic measures. This gives rise to four possibilities: an example of a scalar, categorical measure is the Heidke skill Score (Wilks 1995), and an example of a scalar, probabilistic measure is the ranked probability score (a multiclass generalization of the Brier score; Wilks 1995). As for multidimensional measures (typically, diagrams), an example of a categorical measure is the relative operating characteristic (ROC) diagram, and an example of a probabilistic measure is the attributes diagram (more below).

The starting point for computing categorical measures is the contingency table. The elements of a contingency table, C_{ij} are the number of observations in the ith class that are forecast as belonging to the jth class. One can

[3] One may also employ some variable selection method (see ftp://ftp.sas.com/pub/neural/importance.html), but it is generally believed that in NNs the best such method is the brute-force one, i.e., examining all possible combinations of the inputs. Given its impracticality, it is common practice to explore only a subset of all the combinations to assure that a ''reasonable set'' (if not the best set) has been found. Further, collinearity of the inputs is not of concern if an NN is employed as a ''black box'' for making predictions, without examining the individual parameters of the NN.

construct numerous categorical measures, each of which captures a different facet of performance. A nonexhaustive examination of such measures based on 2 × 2 contingency tables is given in Marzban (1998). While many such measures are easily generalized to the 3 × 3 case, others are constructed from more basic quantities that are defined only for a two-class problem. That problem is exacerbated when forecasts are probabilistic. For instance, reliability and attributes diagrams do not generalize to multiple classes.[4]

Typically, one reduces the three-class problem to three two-class problems. In other words, the problem of classifying three objects—labeled 1, 2, and 3—is broken down to three two-class problems: classification of class 1 or otherwise, class 2 or otherwise, and class 3 or otherwise. Although this treatment of a three-class problem can lead to some loss of information (Wilks 1995), the loss is partially compensated by the multidimensionality brought about via reliability or attributes diagrams. The three-class problem in the present application is reduced in this fashion.

In a two-class problem (i.e., with yes/no observations and forecasts), with 1 (2) representing no (yes), two basic quantities are the probability of detection (POD) and the false alarm ratio (FAR), defined as

$$\mathrm{POD} = \frac{C_{22}}{C_{21} + C_{22}}, \qquad \mathrm{FAR} = \frac{C_{12}}{C_{12} + C_{22}}.$$

Another basic quantity is the false alarm rate (FAT), defined as

$$\mathrm{FAT} = \frac{C_{12}}{C_{11} + C_{12}}.$$

One multidimensional measure of performance is the ROC diagram. A ROC diagram (Masters 1993) is simply a parametric plot of POD versus FAT, as a probability threshold is varied from 0 to 1. It is easy to show that a classifier with no ability to discriminate between two classes yields a diagonal line of slope one and intercept zero; otherwise, the ROC curve lies in the region above the diagonal line. The area under the curve is often taken as a scalar measure of the classifier's performance, and so, a perfect classifier would have an area of 1 under its ROC curve, while a random classifier would have an area of 0.5 (i.e., the area under the diagonal line). In addition to its multidimensionality, another virtue of a ROC diagram is its ability to express performance without a specific reference to a unique probability threshold. A specific choice of the threshold calls for knowledge of the costs of misclassification, which are user dependent. In this way, a ROC diagram offers a user-independent assessment of performance. Because the ROC diagram is based on inherently two-dimensional quantities (POD and FAT), it has no natural extension to the three-class case. Therefore, three ROC diagrams must be considered, one for each of the three classes.

Since the network produces probabilistic forecasts, it is possible to assess the quality of the forecasts without categorizing the forecasts. One commonly employed probabilistic measure is the Brier skill score (BSS) which is based on the MSE in Eq. (1), and is defined as (Wilks 1995)

$$\mathrm{BSS} = 1 - \frac{\mathrm{MSE}}{p(1 - p)}, \tag{3}$$

where p is the climatological probability of hail of a given size. Due to its scalar nature, BSS fails to capture all the facets of probabilistic forecasts. Consequently, the quality of the probabilities will be assessed within a probabilistic scheme (Murphy and Winkler 1987, 1992; Wilks 1995). According to Murphy and Winkler (1992), the various aspects of the quality of probabilistic forecasts can be captured by three diagrams: reliability (or calibration), refinement, and discrimination diagrams. In terms of the conditional distribution of forecasts, f, and observations, x, a reliability diagram is a plot of $p(x = 2 \mid f)$ as a function of f.[5] A refinement plot is the plot of the marginal distribution $p(f)$ as a function of f, and a discrimination plot is a plot of $p(f \mid x = 1), p(f \mid x = 2), \ldots, p(f \mid x = c)$ as a function of f, where c is the total number of classes in the problem (here 3). Reliability diagrams can be supplemented with the contribution of the forecasts to BSS and their resolution, with the resulting diagram referred to as an attributes diagram (Murphy and Winkler 1992). Resolution is the variance of the difference between the unconditional observations and conditional (on forecast) observations (Murphy and Daan 1985).

The reliability portion of an attributes diagram displays the extent to which the frequency of an event at a given forecast probability matches the actual forecast probability. A refinement diagram displays the sharpness of the forecasts, that is, the extent to which very high or very low probabilities are issued, and a discrimination plot exhibits the extent to which the forecasts discriminate between the classes. All of these together present a fairly complete representation of performance quality.

6. Results

As mentioned previously, two NNs have been developed: a regression NN designed to estimate hail size,

[4] One generalization of reliability diagrams has been introduced by Hamill (1987). However, it appears that the proposed diagrams assess not reliability but some other facet of performance quality. The proposed diagrams have been computed for the current application; however, they suggest that the facet being assessed is totally "perfect." In other words, deviations from perfect quality are all statistically insignificant. As such, the proposed diagrams do not offer a useful assessment of performance, and are not presented herein.

[5] Here, $x = 1$ and $x = 2$ refer to the nonexistence and existence of an event, respectively.

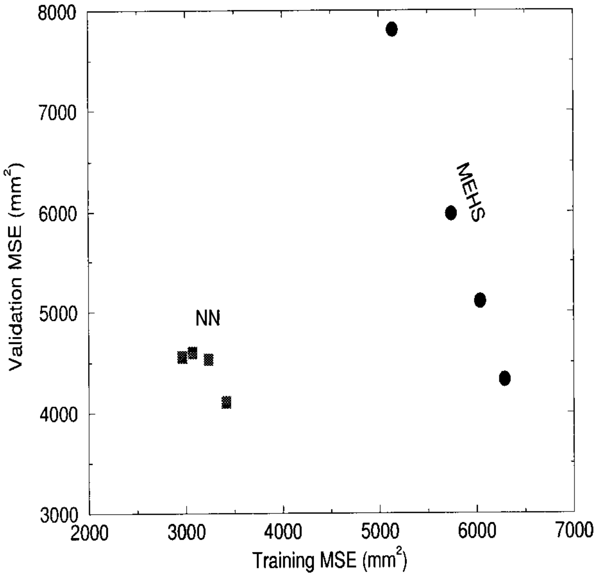

FIG. 2. The training and validation mean square errors of the NN (square) and MEHS (circle) for four bootstrap trials.

and a classification NN for modeling the probability of the three hail size classes. The optimal number of hidden nodes for both networks was found via bootstrapping, as described in the method section. However, as expected within the Bayesian approach, the final results are quite insensitive to the precise number of hidden nodes. In particular, for the regression NN it was found that any number of hidden nodes between 0 and 4 yields results that are statistically equivalent. For the classification NN the range of the number of hidden nodes leading to comparable performance is 2–6. As a result, the number of hidden nodes was set at 2 for the regression NN, and at 4 for the classification NN.

According to the bootstrap approach, in order to assure that the results are not jeopardized due to sampling effects, one must repeat the entire training/validation process with different training and validation data. For the current analysis, four such trials are made and the results are presented either separately, or averaged over the trials.

Having identified the optimal NNs, one can turn to their performance. Again, only multidimensional measures (i.e., diagrams) will be employed in order to avoid any loss of information brought about by the use of scalar measures.

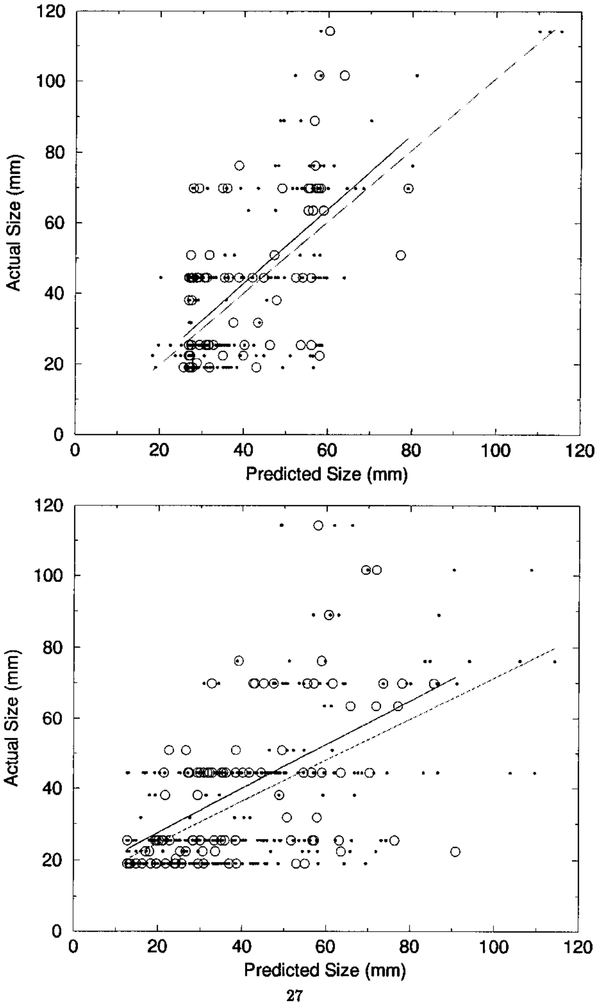

FIG. 3. Scatterplot of the actual vs predicted hail size of the (left) NN and (right) MEHS for the training (dark dots) and validation (circles) data. Also shown are the corresponding regression fits.

We begin with the regression NN. In this case, it is possible to compare the performance of the NN with an existing model that is currently in operation in the Weather Surveillance Radar-1988 Doppler (WSR-88D) HDA. The existing model is an empirically derived equation for the maximum expected hail size. The model, hereafter referred to as MEHS, is a nonlinear model based only on the severe hail index (Witt et al. 1998a). Figure 2 shows the training and the validation errors for the four bootstrap trials of the NN and the MEHS models. It can be seen that not only does the NN have training and validation mean square errors that are lower than those of MEHS, they are also more clustered. In other words, the NN consistently outperforms MEHS in terms of the mean square error of the forecasts.

Figure 3 shows the scatterplot of the NN and MEHS for one of the bootstrap trials. The dark dots are the training data and the circles are the validation data. From the general pattern of these figures, it is evident that the NN outperforms MEHS. Regressing to a scalar measure, like r^2 of the fits, provides a more quantitative measure of the goodness of fit. Values of r^2 approaching 1 imply a better fit. The r^2 of the training and validation data for the NN are 0.51 and 0.40, respectively. By comparison, the same quantities for MEHS are 0.34 and 0.29. Clearly the NN provides a far better fit to the data than MEHS does. Also shown, are the regression fits to the corresponding plots. It can be seen that whereas the NN's fit to both the training and validation datasets pro-

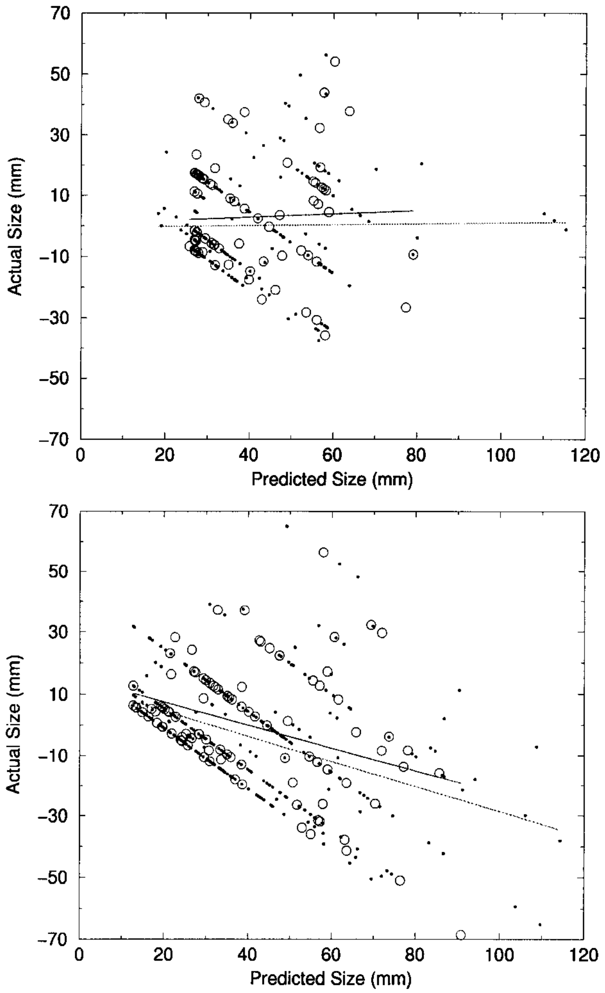

Fig. 4. Residual plot of the (left) NN and (right) MEHS for the training (dark dots) and validation (circles) data. Also shown are the corresponding regression fits.

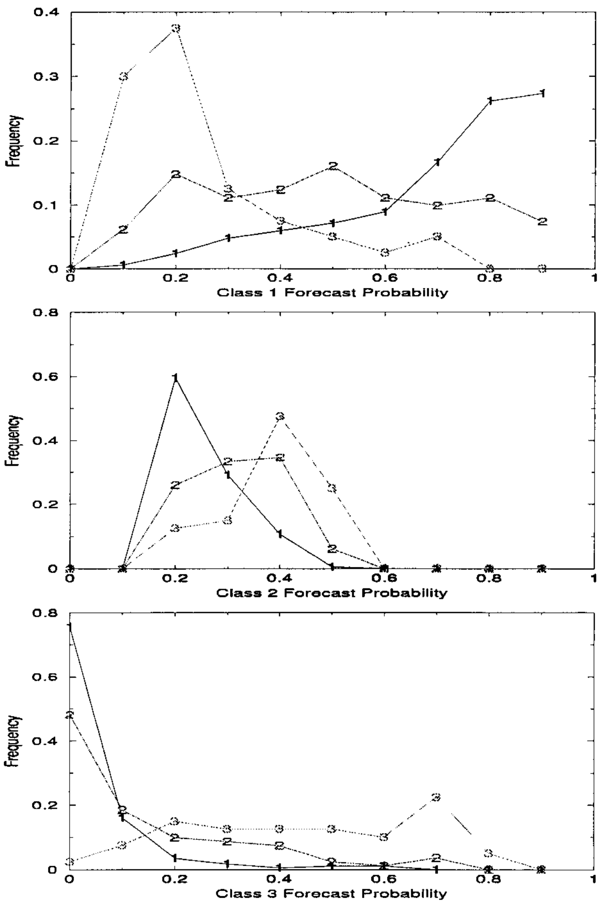

Fig. 5. Discrimination diagrams for (top) class 1, (middle) class 2, and (bottom) class 3 forecasts.

duces diagonal lines of slope 1, the MEHS slopes fall short of that ideal value. This means that if MSE is the measure of error (or agreement), then on the average there is near-perfect agreement between NN-predicted size and the actual size; by contrast, the MEHS-predicted size is typically higher than the actual size (i.e., it has an overforecasting bias).

An examination of the residual plots (Fig. 4) is also informative. First, the performance of both the NN and MEHS deteriorates with increasing hail size.[6] That, of course, is partially a consequence of the skewed nature of the distribution of hail size (Fig. 1). Furthermore, the NN displays far less scatter about the horizontal line than MEHS does. As such, the NN's predictions are more accurate than those of MEHS. Again, a scalar but quantitative reflection of these observations is offered by the r^2 of the residuals. Here a small r^2 is desirable, for that implies more accurate predictions. For the NN, the training and the validation r^2 are 0.0002, and 0.0017, respectively, while the same two quantities for MEHS are 0.21, and 0.13. Also, the regression fits to the NN data yield nearly horizontal lines for both the training and the validation sets. This is desirable because it implies that the NN has correctly captured the underlying relation between the predictors and hail size. By contrast, the same fit for the MEHS data yields lines that have significant negative slopes. In other words, MEHS's predictions not only deteriorate with the magnitude of the prediction (as do also the NN's), but they are also heavily biased toward larger estimates.

Continuing with performance issues, that of the classification networks is assessed in terms of ROC, discrimination, refinement, and attributes diagrams. The former requires the introduction of a probability threshold and, so, will be treated last. The discrimination diagrams for the three classes are displayed in Fig. 5. It can be seen that the class 1 forecasts clearly discriminate between the three classes. The distribution of class 1 observations is peaked to the right, while those of the

[6] This heteroscedasticity is detrimental only if the NN parameters are subjected to hypothesis testing; otherwise, this instability in the variance of the residuals is not of great concern.

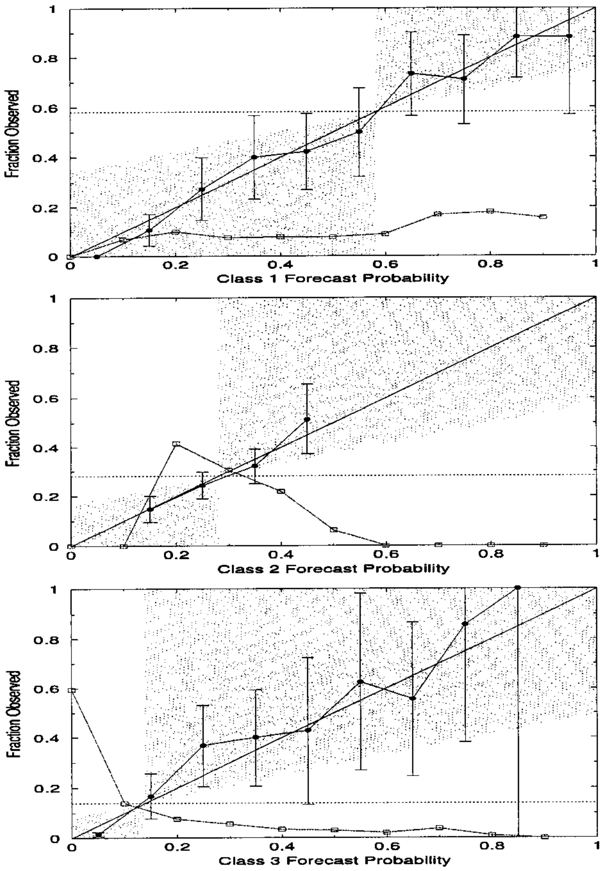

FIG. 6. The attributes and refinement diagrams for the (top) class 1, (middle) class 2, and (bottom) class 3 forecasts. Also shown are the 95% confidence intervals due to sampling. The curve consisting of the squares is the refinement diagram, the hashed region corresponds to forecasts that contribute to BSS, and the horizontal line defines forecasts with no resolution.

other two classes are either flat or peaked to the left. This is a desirable result, although ideally one would want one curve peaked to the right and two curves peaked to the left. By contrast, class 2 forecasts display little or no discriminatory capability. Although, the class 1 curve is peaked to the left (as desired), the curve that is peaked to the right is the class 3 curve, and not the class 2 curve. Finally, class 3 forecasts are quite discriminatory, but with an interesting twist. They derive their discriminatory capability not only from the identification of class 3 observations, but also (in fact, mostly) from the identification of observations that do *not* belong to class 3.

Several facets of the quality of the forecasts can be assessed through attributes diagrams. Figure 6 shows these diagrams for forecasts belonging to each of the three classes. It can be seen that the reliability of nearly all the forecasts is within statistical limits of perfect forecasts (i.e., the diagonal line). The error bars are 95% confidence intervals due to sampling. The horizontal line corresponds to forecasts that have no resolution, and the bisector of the angle formed by it and the diagonal marks the locus of forecasts with no skill (i.e., BSS = 0). The shaded area defines forecasts that contribute positively to skill. Therefore, it can be seen that in addition to being highly reliable, all the forecasts also contribute positively to skill. This is true even for the class 2 forecasts for which almost no discriminatory capability exists (Fig. 5, middle). The lack of discriminatory class 2 forecasts and their positive contribution to BSS may seem paradoxical; however, it must be noted that the vicinity of the reliability curve (dark circles) in Fig. 6b to the region of no skill, in conjunction with the size of the error bars, suggests that the contribution of the forecasts to skill may not be statistically significant.

Although the probabilities are all highly reliable, the range of the forecasts is quite varied. Class 1 forecasts span a relatively wide range of 10%–90%, whereas class 2 forecasts are restricted to the range 20%–50%. This reflects the difficulty in predicting class 2 hail with a high degree of confidence. It is interesting that class 3 forecasts can reach probabilities nearing 100%, albeit rarely.

It is convenient to superimpose the refinement diagram upon the attributes diagram, the former labeled with squares in Fig. 6. Evidently, the three forecast classes have distinct levels of refinement. The class 1 forecasts display a mild degree of the desired U-shaped pattern. Class 2 forecasts have an uncommon and undesirable bell-shaped pattern, indicating that most of the forecasts are in the vicinity of 20%. The highly left-peaked forecasts for class 3 suggest that the most common forecasts are at 0%. This is partially a consequence of the rarity of class 3 observations in the data.

Finally, the introduction of a probability threshold can dichotomize the forecasts and allow for the computation of ROC diagrams (Fig. 7). These diagrams support the previous findings that class 3 forecasts appear to have the highest performance, followed by class 1, and class 2 forecasts, respectively. In fact, from the validation data (Fig. 7b), one might suspect that class 2 forecasts are nearly random because the diagonal line goes through the corresponding error bars.

7. Summary and discussion

Two neural networks have been developed for maximum hail size prediction, one for estimating the size in some physical dimension, and a second for assessing the probability of belonging to one of three hail size classes. The parameters of the networks have been inferred via Bayesian methodology in order to alleviate the problem of overfitting. Four Doppler radar, and five environmental variables, have been employed as inputs to the networks. Performance is assessed in terms of multidimensional measures. It is shown that the network designed to predict maximum hail size outperforms the existing method in the WSR-88D Hail Detection Al-

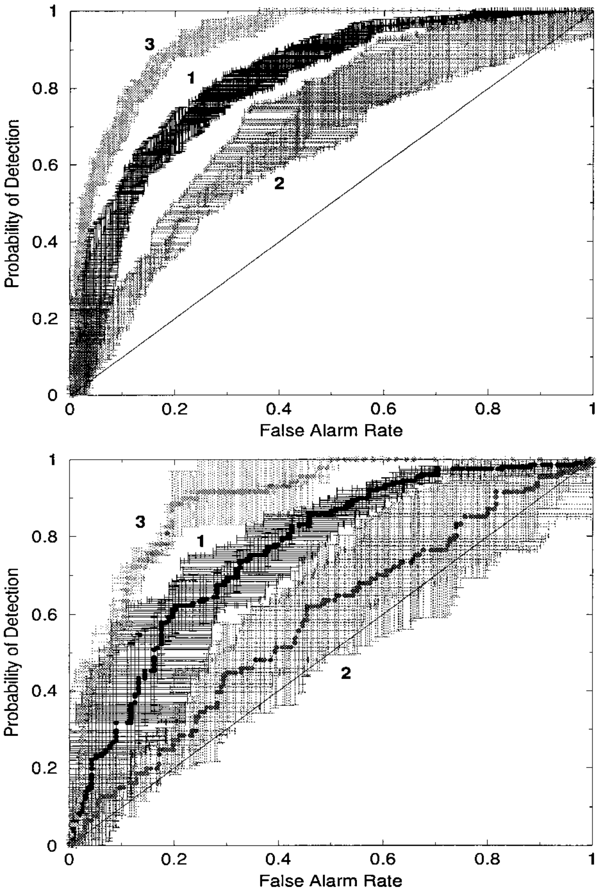

FIG. 7. ROC diagrams for class 1, 2, and 3 forecasts, based on (top) the training set and (bottom) the validation set. The error bars in the *x* and *y* directions are the one standard deviation intervals based on bootstrapping.

gorithm. Since the existing method does not produce probabilities, it is not possible to compare the performance of the second network; however, it is found that the network-produced probabilities constitute high quality forecasts, with quality gauged in terms of reliability, refinement, and discrimination diagrams. Attributes diagrams suggest that the forecasts contribute positively to skill. The one possible exception is the forecasts for midsize hail (~ 40mm), which display no statistically significant skill. This is easy to understand, for discriminating between extreme events is a relatively simple task. In contrast, disambiguating intermediate events from the extreme ones is more difficult. It is possible that all three classes can be reasonably discriminated as more data become available.

To better understand the relation between the individual attributes (predictors) and hail size, it is useful to compute the corresponding linear correlation coefficients, *r*. The *r* between hail size and the four ‘‘radar attributes’’ (attributes 1–4 in Table 1) all have approximately equal correlation with hail size, namely 0.5 (standard error = 0.04). The ‘‘environmental attributes’’ (attributes 5–9 in Table 1) have linear correlations that vary from 0.01 to 0.1, but are practically zero within the standard errors (0.05). Of course, as evidenced by the nonzero number of hidden nodes in the NNs, the true underlying relationship is nonlinear. In fact, the exclusion from the NN of the variables with a small correlation with hail size leads to inferior performance. This implies that in spite of their weak linear correlation with hail size, the environmental attributes do play a significant role in a nonlinear and interactive model (e.g., NN) wherein the prediction equation for severe-hail size includes interaction terms (e.g., products of the predictors).

Furthermore, the *r* between the various attributes themselves is important in ascertaining the collinearity among the inputs. Identifying collinear inputs (i.e., a pair of inputs with a large *r*) and the exclusion of one member of the pair as an input to the NN can reduce the likelihood of overfitting the data. The most collinear pair of attributes is (5, 6) with $r = 0.83$, followed by the pair (6, 7) with $r = 0.70$. Neither of these *r*’s is sufficiently large to justify the exclusion of either member of either pair. As a result, no attributes were excluded as input nodes.

Since the data used to develop the neural networks consist only of severe-hail reports, the hail size predictions are *conditional* in nature. In order words, once it is determined that severe hail is occurring, or is expected to occur, then the predictions made by the neural networks can be used to estimate the maximum hail size. The process would be similar to other situations where conditional forecasts are made, for example, the conditional probability of severe thunderstorms given that thunderstorms actually occur. Once a sufficiently large dataset is collected of hail reports of all sizes (not just 19 mm and larger), then new neural networks can be developed to make nonconditional predictions of maximum hail size.

As to the operational utilization of these new techniques, some users, particularly those from regions that are not well represented in the dataset (e.g., the northeast and western United States), might be concerned about the robustness of the neural networks. We plan to address this issue by analyzing more data specifically from these regions, and to retrain the neural networks on the larger dataset before they are added to any operational system. Another potential concern of users might be that of computation of the near-storm environment parameters. Although sounding data were used in the development of the neural networks, it is expected that real-time numerical model data will be utilized by the SSAP in an operational setting to determine these parameters (Lee et al. 1998), thus relieving users of the burden of this task.

What remains to complete the augmentation of the Hail Detection Algorithm with neural networks is the development of the latter for assessing the probability of severe hail, regardless of size. Such a network would issue a probability for the existence of severe hail (within some time window) after which the networks outlined in this article would provide information regarding the size of the forecast hail. The former network is currently in development and will be discussed in a separate article.

Acknowledgments. The authors are grateful to Harold Brooks, V. Lakshmanan, and Conrad Ziegler for valuable discussions. Mark DeLisi and an anonymous reviewer are acknowledged for insightful reviews. This study was partially supported by the WSR-88D Operational Support Facility.

REFERENCES

Billet, J., M. DeLisi, and B. G. Smith, 1997: Use of regression techniques to predict hail size and the probability of large hail. *Wea. Forecasting,* **12,** 154–164.

Bishop, C. M., 1996: *Neural Networks for Pattern Recognition.* Clarendon Press, 482 pp.

Draper, N. R., and H. Smith, 1981: *Applied Regression Analysis.* John Wiley and Sons, 709 pp.

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

Hamill, T. M., 1997: Reliability diagrams for multicategory probability forecasts. *Wea. Forecasting,* **12,** 736–741.

Hornik, K., M. Stinchcombe, and H. White, 1989: Multilayer feedforward networks are universal approximators. *Neural Networks,* **4,** 251–257.

Johnson, J. T., P. L. MacKeen, A. Witt, E. D. Mitchell, G. J. Stumpf, M. D. Eilts, and K. W. Thomas, 1998: The Storm Cell Identification and Tracking algorithm: An enhanced WSR-88D algorithm. *Wea. Forecasting,* **13,** 263–276.

Lee, R. R., G. J. Stumpf, and P. L. Spencer, 1998: Should geographic region or near-storm environment dictate WSR-88D algorithm adaptable parameter settings? Preprints, *19th Conf. on Severe Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 784–787.

MacKay, D. J. C., 1996: *Models of Neural Networks III.* Springer-Verlag, 311 pp.

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.

——, and G. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.,* **35,** 617–626.

——, and ——, 1998: A neural network for damaging wind prediction. *Wea. Forecasting,* **13,** 151–163.

——, and A. Witt, 2000: A neural network for hail size prediction. Preprints, *Second Conf. on Artificial Intelligence,* Long Beach, CA, Amer. Meteor. Soc., 38–44.

——, H. Paik, and G. Stumpf, 1997: Neural networks vs. Gaussian discriminant analysis. *AI Appl.,* **10,** 49–58.

Masters, T., 1993: *Practical Neural Network Recipes in C++.* Academic Press, 493 pp.

Murphy, A. H., and H. Daan, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 379–437.

——, and ——, 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.

Neal, R. M., 1996: *Bayesian Learning for Neural Networks.* Cambridge University Press, 183 pp.

Rasmussen, E. N., and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting,* **13,** 1148–1164.

Richard, M. D., and R. P. Lippmann, 1991: Neural network classifiers estimate Bayesian a-posteriori probabilities. *Neural Comput.,* **3,** 461–483.

Sarle, W. S., 1995: Stopped training and other remedies for overfitting. *Proc. 27th Symp. on the Interface of Computing Science and Statistics,* Cary, NC, SAS, 352–360,

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

Witt, A., 1998: The relationship between WSR-88D measured midaltitude rotation and maximum hail size. Preprints, *19th Conf. on Severe Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 740–743.

——, and S. P. Nelson, 1991: The use of single-Doppler radar for estimating maximum hailstone size. *J. Appl. Meteor.,* **30,** 425–431.

——, M. D. Eilts, G. J. Stumpf, J. T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998a: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting,* **13,** 286–303.

——, ——, ——, E. D. Mitchell, J. T. Johnson, and K. W. Thomas, 1998b: Evaluating the performance of WSR–88D severe storm detection algorithms. *Wea. Forecasting,* **13,** 513–518.

Wolpert, D. H., 1993: On the use of evidence in neural networks. *Advances in Neural Information Processing Systems 5,* C. L. Giles, S. J. Hanson, and J. D. Gowan, Eds., Morgan Kaufmann, 539–546.