OSF MOU Task 4.1, Deliverable D4.1.2: Bayesian Neural Networks for Severe Hail Prediction

Caren Marzban and Arthur Witt

November 27, 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 several neural networks are developed for the prediction of the occurrence and the size of severe hail. Performance is assessed in terms of multi-dimensional (i.e., nonscalar) and scalar measures. It is shown that the respective networks outperform their existing non-network counterparts.

1 Introduction

Neural Networks' (NN) ability to represent a wide range of models is usually accompanied by one disadvantage, namely their tendency to overfit data. Overfitting occurs when a model captures the statistical noise in the data rather than the underlying signal. A number of popular methods for identifying a model that overfits utilize several independent data sets; these methods fall under the general classes of split-sample and resampling methods. Examples are cross-validation, jack-knifing, and bootstrapping (Bishop 1996). In these methods, one or more of 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 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. 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 NNs ability to overfit can be limited but without compromising its nonlinearity. The details of this Bayesian approach are presented in Marzban and Witt (2000a,b), and Marzban (1998).

The development of a NN model for severe hail prediction can be divided into two sub-tasks: One of developing a model that predicts the probability of severe hail, and another that predicts the size of severe hail, given that severe hail has occurred or is expected to occur. Further, a NN for the prediction of size alone can be developed in two independent ways: One can develop a NN that predicts the size of hail in some physical unit (e.g., inches or millimeters), or one can assess the probability of belonging to some size range. The data suggests 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. For size prediction, the former approach falls in the domain of regression, while the latter is an example of a classification problem. Thus, there will be three NNs considered herein, and will be referred to as the Probability of Severe Hail (POSH) NN, the regression NN, and the classification NN, respectively.

In what follows, the data and the methodology are further outlined, and scalar and multidimensional (e.g., distribution-based) measures of performance are set forth to gauge the performance of the NNs.

2 Data

The input variables provided to the NNs include a mix of Doppler-radar derived predictors along with several predictors representing the near-storm environment. The predictors available for the POSH NN are listed in Table 1, and those for the regression and classification NN are given in Table 2. The radar predictors consist of four based on reflectivity data, including 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 base reflectivity and base height predictors correspond to the lowest-altitude 2D component of each storm cell detected by the Storm Cell Identification and Tracking (SCIT) algorithm (Johnson et al. 1998). The near-storm environment predictors include four based on thermodynamic data and two based on kinematic data. The vertically-integrated wet-bulb temperature predictor is computed by integrating the wet-bulb temperature profile from the surface to the height of the wet-bulb zero. For this study, all the near-storm environment predictors were calculated from sounding data. For each individual "storm event" analyzed[1], 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. Incomplete soundings that did not allow for the calculation of all the environmental predictors, and soundings that appeared to be contaminated by convection, were disqualified.

The verification data 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 inch). There are numerous problems associated with using *Storm Data* for verifying radar-based algorithm predictions (Witt et al. 1998a,b). For predicting the occurrence of severe hail, the primary concern is the need to infer "no" observations from the lack of a hail report in *Storm Data.* Because this inference is dubious in rural areas, a population density filter is used. 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).

Different methods were used for relating severe hail reports to algorithm output (i.e., the predictor variables) for development and testing of the different NNs. For the POSH NN, algorithm output was generated for a number of volume scans for each storm event that was analyzed. Hail-truth files were produced for each storm event, and a scoring

[1] A "storm event" is defined as a continuous period of time (up to 24 hours long) when convective activity is occurring within 230 km of a radar site.

Table 1: The 14 inputs of the POSH NN, and two additional predictors (†) not employed as NN inputs.

No.	Description
1. †	Storm cell azimuth
2.	Storm cell range (km)
3.	Maximum reflectivity (dBZ)
4.	Base reflectivity (dBZ)
5.	Base height (km)
6.	Cell-based Vertically Integrated Liquid (VIL; kg(m)$^{-2}$)
7.	Severe Hail Index (SHI)
8.	Midaltitude rotational velocity (m/s)
9.	Storm-top divergence (delta-V in m/s)
10. †	Volume Coverage Pattern (VCP)
11.	Height of the wet-bulb zero (km mean sea level [MSL])
12.	Height of the melting level (km MSL)
13.	Height of the -20 C level (km MSL)
14.	Vertically-integrated wet-bulb temp
15.	Wind speed at the Equilibrium level (m/s)
16.	Storm-relative flow at the -20 C level (m/s)

Table 2: The 9 predictors of the regression and classification NNs, and 6 additional predictors (†) not employed as NN inputs.

No.	Description
1 †	Date
2 †	Time (UTC)
3 †	Hail size (mm)
4	Cell-based VIL (kg(m)$^{-2}$)
5	SHI
6	Storm-top divergence (delta-V in m/s)
7	Midaltitude rotational velocity (m/s)
8 †	Storm cell range (km)
9 †	VCP
10 †	Geographic region (see Figure 1)
11	Height of the wet-bulb zero (km MSL)
12	Height of the melting level (km MSL)
13	Vertically-integrated wet-bulb temp
14	Wind speed at the equilibrium level (m/s)
15	Storm-relative flow at $-20°$ C level (m/s)

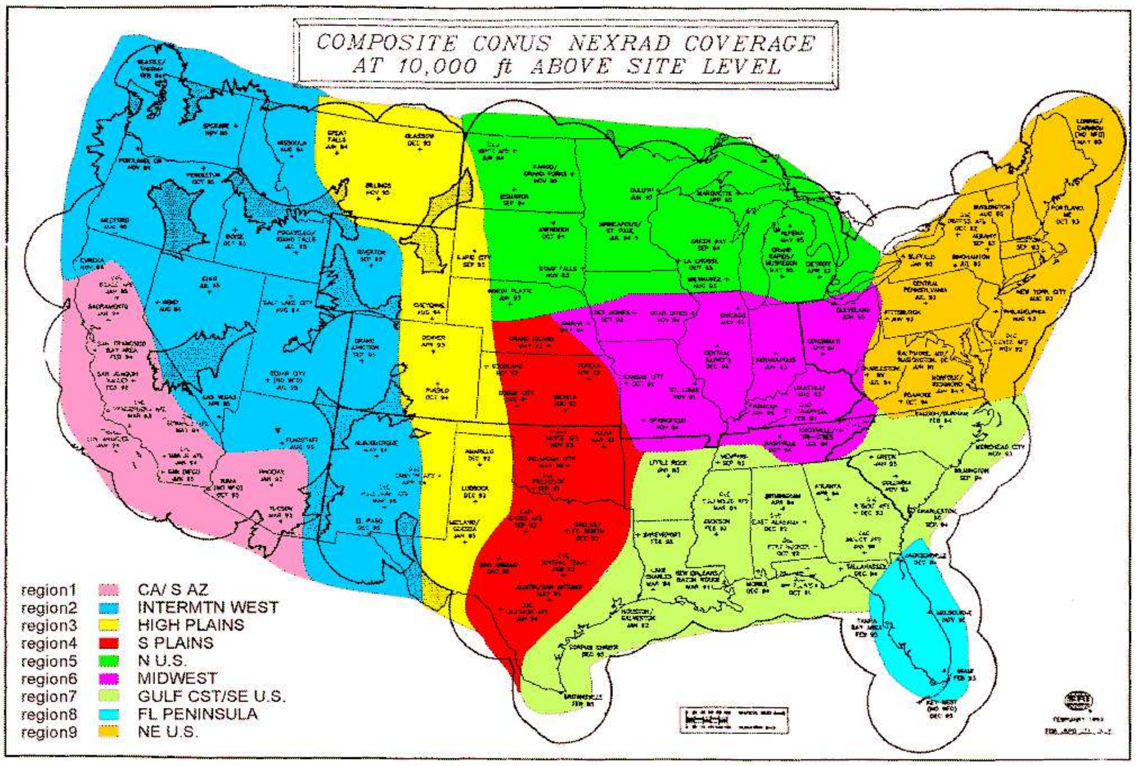

Figure 1: Map showing the different geographical regions listed in Table 3.

program was run to relate severe hail reports to algorithm output using a 20 minute time window (see Witt et al. 1998 for additional details). For algorithm output not associated with severe hail reports (i.e., the inferred "no" observations), a population density filter was applied at a threshold of 100 per 4 km^2. For the POSH NN, individual storm cells where identified via the SCIT algorithm.

For development and testing of the NNs that predict maximum hail size, a 20 minute time window was used to relate the predictor variables to the maximum reported hail size (per storm[2]). For the radar-based predictors, the maximum value within the time window was used, whereas for the near-storm environment predictors, an average value was used. Since a single sounding was used for each severe event, the only environmental predictors that actually required averaging across the 20-min time window was the storm-relative flow (due to changes in the storm motion vector). If the largest reported hail size for a

[2] For the hail size NNs, storm cells where identified manually. Because the sample size used to develop these NNs is much smaller than the sample size used for the POSH NN, manual analysis was done to minimize errors.

Table 3: Summary of the 130 storm events analyzed. A POSH case corresponds to an individual SCIT detection (on a volume-scan by volume-scan basis). See Figure 1 for a definition of each region.

Region	No. events	No. hailstorms	No. reports	No. POSH cases
Western U.S.	19	31	45	1048
High Plains and S Plains	35	179	465	3356
N U.S. and Midwest	22	93	249	7624
SE U.S. and FL	36	152	355	14638
NE U.S.	18	113	232	10891
Total	130	568	1346	37557

storm occurred when the storm was in the radar's cone-of-silence[3] or at ranges $> 230km$, or when no radar data was collected, then that storm was not included in the development and testing of the hail-size NNs.

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.[4] For example, suppose one is using 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. This condition (using the weaker period) only applies to multiple reports where all the predictors have been measured within the time window, i.e., the storm-top divergence or midaltitude rotational velocity predictors are not "missing" due to range folding. In cases where one (or more) of the periods had missing data, the period(s) without missing data was used.

The Hail Detection Algorithm (HDA) and other algorithms[5] in the Severe Storm Analysis Package[6] were employed to compute the NN input values for 130 storm events from across the U.S. (Table 3), on which 568 severe hailstorms were observed. The 130

[3] For the development of the hail-size NNs, a storm is considered to be in the radar's cone-of-silence if its reflectivity at the highest elevation angle is $\geq 50(dBZ)$.

[4] Storm strength was determined by a composite of the four radar-based predictors in Table 2.

[5] Specifically, the SCIT algorithm, the Mesocyclone Detection Algorithm (Stumpf et al. 1998) and the Upper-level Divergence Algorithm.

[6] The Severe Storm Analysis Package is the name given to the entire group of National Severe Storm Laboratory (NSSL) severe-storm-analysis algorithms.

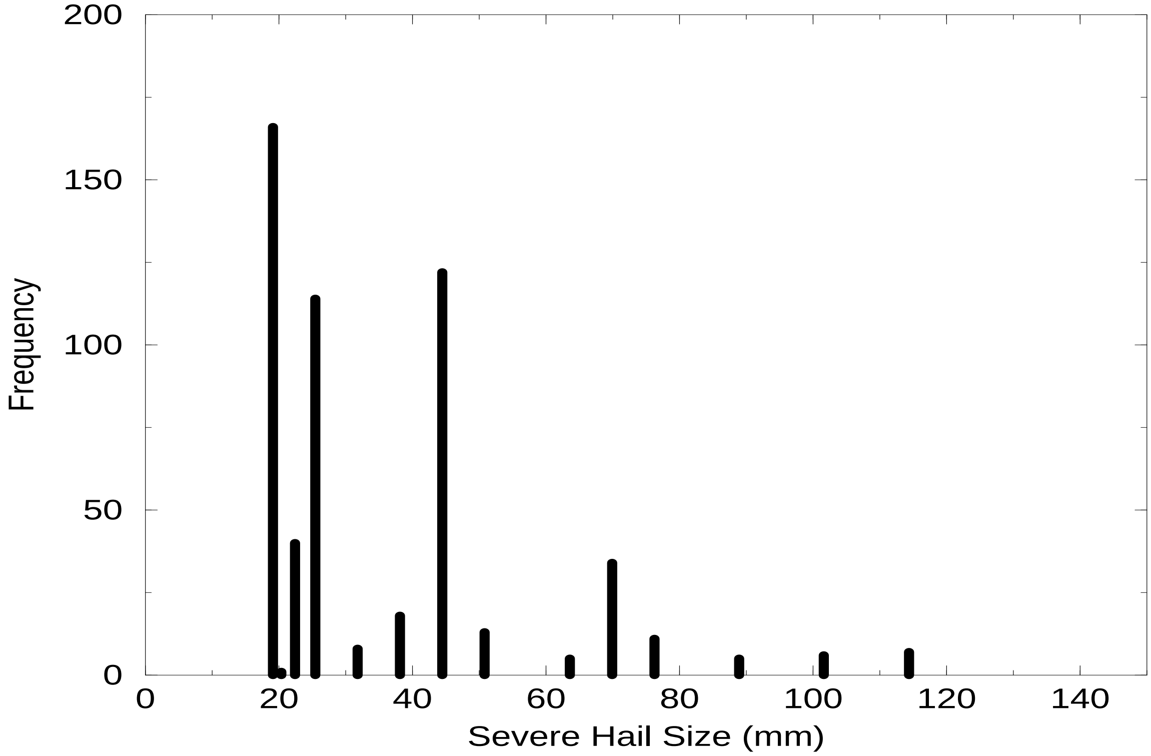

Figure 2: The distribution of reported hail size.

storm events were selected to produce a data set with broad geographic (Figure 1) and seasonal diversity. The distribution of maximum reported hail sizes (from 550 hailstorms) is shown in Figure 2. 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.[7] 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.

[7] This statement is based on the assumption that the true hail-size distribution is continuous in nature.

3 Bayesian Neural Network

All nonlinear regression and classification models can overfit data; overfitting occurs when the nonlinearity of a model allows it to fit a data set or a decision boundary to such high accuracy that the fit is driven by the statistical fluctuations in the data. Consequently, such a model has no predictive capability. The true cause of this phenomenon is the finiteness of the sample size, and it occurs mostly when the model is allowed to be highly nonlinear.

It is possible to restrain overfitting if the NN weights are prevented from becoming too large. The question is then "How large is too large?" The magnitude of the weights can be constrained by introducing a weight-decay term into the error function that is to be minimized. The question then becomes one of determining the optimal value of the coefficient of this term. It is for this purpose that techniques of Bayesian inference can be employed. Additionally, bootstrapping will be employed to identify (and thereby avoid) the onset of overfitting. 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.

4 Methodology

Three quantities are to be produced by the three NNs: POSH, severe hail size, and the probability of belonging to one of three classes of hail size. Since the first and the last deal with a classification problem, they require the minimization of cross entropy

$$S = -\frac{1}{N}\sum[t\log y(x,\omega) + (1-t)\log(1-y(x,\omega))] + \alpha\frac{1}{2}\sum\omega^2,$$

while the appropriate error function for the regression NN is the mean square error

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

where x is the vector of inputs (predictors), ω is the vector of the weights, t is the target value that is to be estimated by the output $y(x,\omega)$, N is the sample size of the relevant data set (training, validation, etc.), and α is the weight-decay coefficient. The minimization of these error functions assures that the output nodes have the correct interpretation, be it size or probability. The number of output nodes for the POSH NN is 2 (for event and nonevent), 1 (size) for the regression NN , and 3 (one for each class of hail size) for the classification NN.

The number of predictors available for the development of the various NNs is 14 for the POSH NN (Table 1), and 9 for the regression and classification NNs (Table 2). For the development of the NNs, it was deemed unnecessary to include predictors designated with a †.

The potential for the storm-top divergence and midaltitude rotational velocity predictors to be "missing" due to range folding actually calls for the development of two POSH NNs; one NN to make forecasts when all 14 inputs are available (NN1 in Figure 3), and another NN to make forecasts when only 12 predictors are available (NN2 in Figure 3). As such, the general prediction of POSH would involve the conjunction of NN1 and NN2 (NN1+NN2). The regression and classification NNs deal with missing data in the same fashion (i.e., they both actually consist of two NNs). However, unlike the regression and classification NNs, which perform best in this dual-NN mode, it was found that the POSH NN with only 12 inputs (NN2), based on all the cases in the data set, outperforms the combination of NN1 and NN2. This would seem to indicate that storm-top divergence and midaltitude rotational velocity do not provide any useful information in assessing POSH. This was confirmed by comparing NN1 with NN3, which is a 12-input NN based on the same number of cases employed in the development of NN1. Further experimentation with fewer input nodes suggests that nothing is gained by employing subsets of the predictors as inputs.

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

For the POSH NN, the number of available cases is 37,557; however, as described below, the behavior of base reflectivity suggests a simple preprocessing rule that in turn

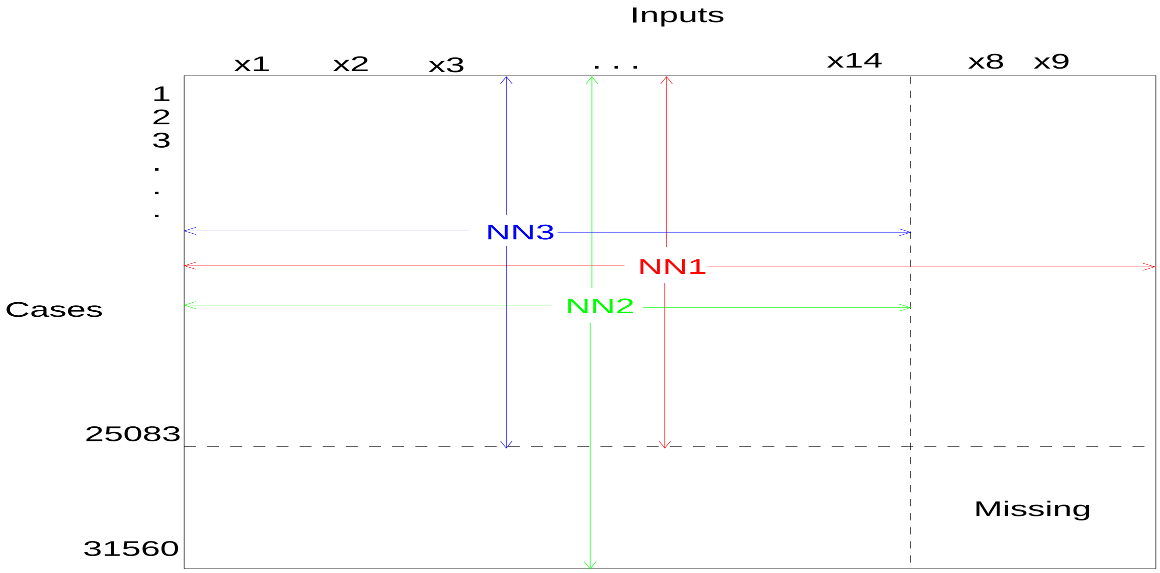

Figure 3: The affect on sample size of excluding predictors from the POSH NN.

reduces the number of cases to 31,560. The sample size for the regression and classification NNs is 550. In all the NNs the respective data sets were randomly partitioned into training and validation sets according to the 2/3-rule in bootstrapping with replacement, i.e. 2/3 of the cases are used in training, and the remaining 1/3 are employed for validation.

Finally, the linear correlation coefficient, r, between the various predictors 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 predictors in Table 1 are (3,4) and (12,13) both with $r = 0.93$. Figure 4 shows the scatterplot for the first pair. In addition to partially illuminating the reason for the collinearity, namely that one predictor is the upper-bound for the other, this figure also suggests a natural and simple pre-processing rule: in particular, all cases with base reflectivity $\leq 40(dBZ)$ are excluded from the training process. For evaluation purposes, such as observation is assigned the prior probability for such an event (i.e., 9/5939), because only 9 out of 5939 cases are associated with severe hail. The exclusion of either member of correlated pairs was found to lead to inferior performance, and so, no collinear predictors are excluded as inputs to the NNs. The finding that the inclusion of both members of a correlated pair improves performance is reflective of the imperfect correlation

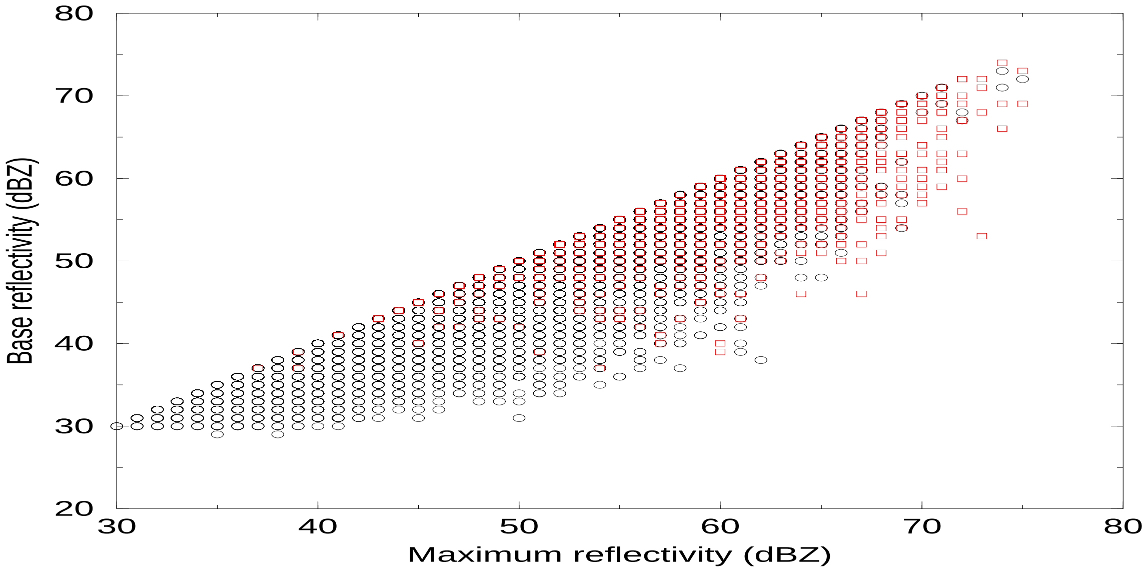

Figure 4: The scatterplot between two highly correlated predictors in Table 1. The black circles are associated with non-occurrence of severe hail, while the red circles are associated with the occurrence of severe hail.

between the members (i.e., $r \neq 1$).

As for the predictors in Table 2, the pairs (11,12) with $r = 0.83$, followed by the pair (12,13) with $r = 0.70$ are the most correlated predictors. However, neither of these r's is sufficiently large to justify the exclusion of either member of either pair. As a result, no predictors were excluded as input nodes.

5 Performance Measures

Both scalar and multi-dimensional measures are employed to assess performance. For the regression NN, scatterplots and residual plots, and scalar measures based thereupon will be employed. For the POSH and the classification NNs, performance is gauged with ROC diagrams, distribution plots, attributes diagrams (including reliability, and refinement), and scalar measures based thereupon. For the definitions of these measures consult Marzban and Witt (2000a,b). The four possibilities are as follows:

	Scalar	Multi-dimensional
Categorical	Heidke Skill Score	ROC Diagram
Probabilistic	Ranked Probability Score	Attributes Diagram

6 Results

To better understand the relation between the individual predictors (predictors) and the predictand (e.g., hail size and the occurrence of severe hail) it is useful to compute the corresponding linear correlation coefficients, r (Figure 5). As compared to the environmental predictors, it can be seen that the radar-based predictors are generally better correlated with the respective predictand. Within the error-bars, predictors 1 and 10 from Table 1, and predictors 1, 2, 8, 9, and 11-13 from Table 2, appear to be completely uncorrelated (linearly) with the respective predictands. However, as discussed below, it was found that all of the predictors are nonlinearly correlated with the respective predictands.

A few experiments were performed to identify the best set of predictors. In one experiment, only the radar-based predictors were used, whereas in another experiment, only the environmental predictors were employed. It was found that the NN with the lowest validation error is the one trained on both sets of predictors. Therefore, even though the environmental predictors have a negligible linear correlation with hail size, they do contribute significantly as inputs to the (nonlinear) NN. In other words, although several predictors have a low linear correlation coefficient with the predictand, an NN with those predictors outperforms an NN without them.

Figure 6 shows the training and the validation errors for NN1 (one of the POSH NNs; Figure 3) for a range of number of hidden nodes; in this case, it can be seen that 12 is the optimal number of hidden nodes, since at that point the validation error begins to rise. Similar plots for NN2 suggest 20 as the optimal number of hidden nodes. In the same fashion the regression and classification NNs are found to be optimal with only 2 hidden nodes.

6.1 The POSH NN

Figure 7 displays numerous multidimensional measures of performance for the POSH NN. The top two figures are discrimination diagrams; ideally one would expect the two curves to have distinct peaks with minimum overlap between them. It can be seen that the NN (left) has peaks that are more distinct than the existing algorithm, i.e., the Weather Surveillance Radar-1988 Doppler (WSR-88D) HDA. In fact, the latter shows no peak in

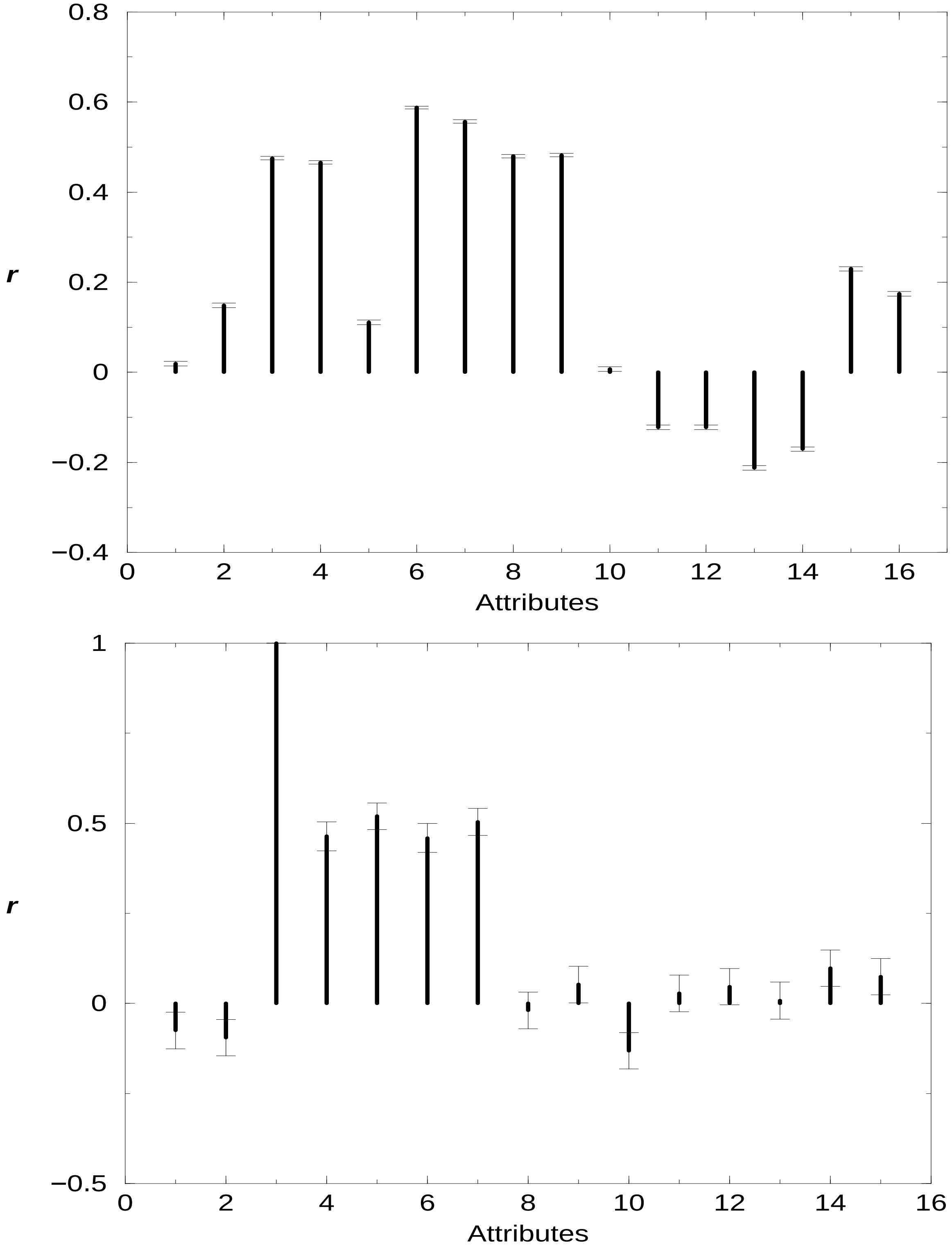

Figure 5: The linear correlation coefficients between the various predictors in Table 1 and the occurrence of severe hail (top), and the predictors in Table 2 and the size of severe hail (bottom). The standard errors are shown as error bars.

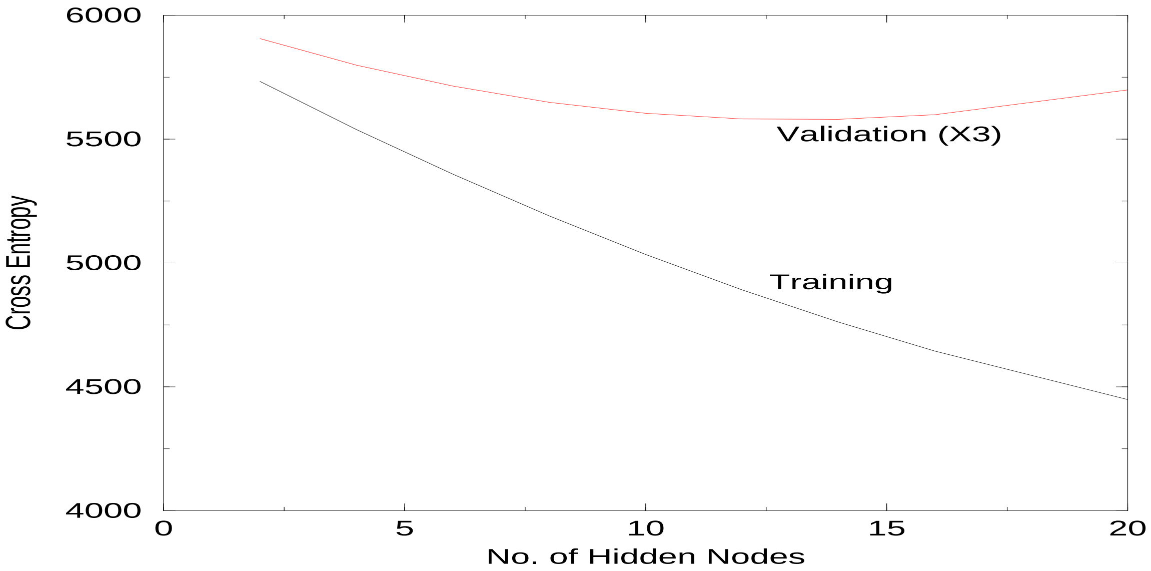

Figure 6: Training and validation errors for NN1 for a range of number of hidden nodes. Clearly 12 is the number of hidden nodes beyond which the NN overfits data.

the (red) curve corresponding to the occurrence of severe hail. As such, the NN better discriminates between the occurrence and non-occurrence of severe hail. The middle figure shows the ROC curve for the two algorithms, again for the validation set; clearly, the NN curve (black) is always above the WSR-88D HDA curve (red). Recall that an ROC curve is a parametric curve of POD versus the false alarm rate (not ratio) as the probability threshold varies from 0 to 1. Therefore the NN outperforms WSR-88D HDA for all range of probability thresholds placed on POSH. The bottom diagrams are the attribute diagrams with the refinement curves superimposed thereupon. Again, clearly the NN outperforms the WSR-88D HDA in terms of reliability and refinement.

These multidimensional diagrams can be distilled to scalar measures. Table 4 compares the performance of the two algorithms in terms of a number of scalar measures. Cross entropy is the appropriate error function for a classification algorithm, because the estimated parameters then coincide with the maximum likelihood estimates. It is a probabilistic, scalar measure because it is computed from the estimated probabilities. Lower values of cross entropy imply higher performance. The Critical Success Index (CSI) and the Heidke Skill Score (HSS) are scalar measures that are nonprobabilistic, for they are computed from a contingency table. POD and FAR are the Probability of Detection and the False Alarm Ratio, respectively. ROC refers to the area under the ROC curve (Figure

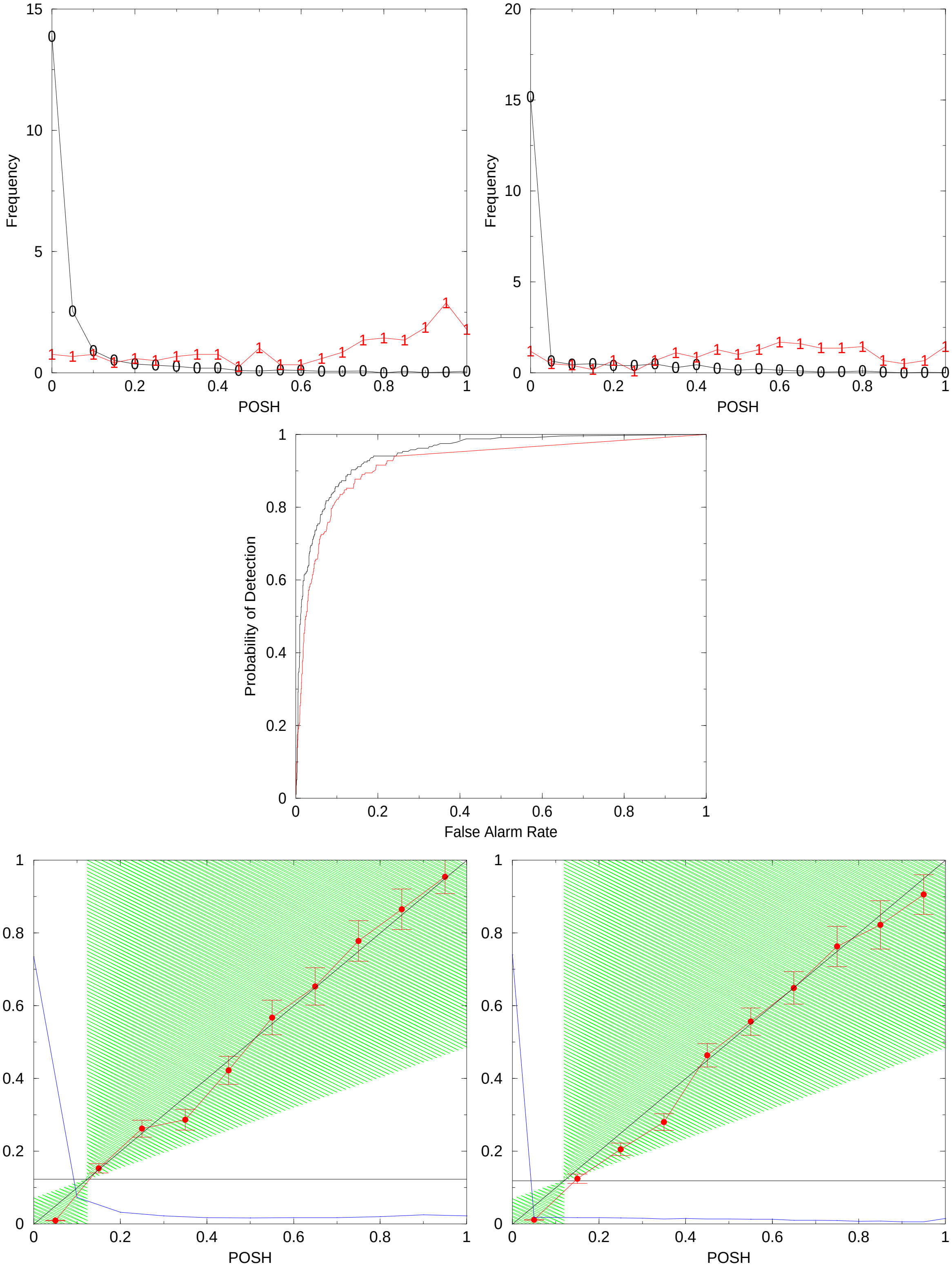

Figure 7: Top: The discrimination diagrams for the POSH NN (left) and WSR-88D HDA (right). Middle: The ROC curve for NN (black) and WSR-88D HDA (red). Bottom: The attributes diagrams for the POSH NN (left) and the WSR-88D HDA (right), with the refinement plots superimposed thereupon. The error bars are the 95% confidence intervals.

Measure	NN	WSR-88D HDA
Cross entropy	27	44
POD	68	62
FAR	22	26
CSI	57	51
HSS	70	64
ROC	97	94
CON	5.4	6.7
REL	0.14	0.37

Table 4: Scalar measures of performance (in %) for the POSH NN and WSR-88D HDA for one validation set.

7, middle); for a random classifier ROC=0.5, and for a perfect classifier ROC=1.0. CON refers to convolution and is the area of the overlap of the distribution plots (Figure 7, top); for a random classifier it is 1, and is zero for a perfect classifier. REL is the mean square error of the reliability curve (Figure 7, bottom) about the diagonal line; a perfect classifier would have REL=0, and there is no upper-bound to it. Evidently, the NN outperforms the WSR-88D HDA in terms of all the measures of performance.

6.2 The Regression NN

Table 5 shows the training and validation errors for the four bootstrap trials of the NN, NSSL's Warning Decision Support System (WDSS; Eilts et al. 1996) HDA, and the WSR-88D HDA. It can be seen that not only does the NN have training and validation errors that are lower than those of the WDSS and WSR-88D HDA, they are also more clustered. In other words, the NN consistently outperforms the WDSS and WSR-88D HDA in terms of both mean square and mean absolute error of the forecasts. The average improvement in the mean square errors is 20% and 33% for the training error (compared to the WDSS and WSR-88D HDA, respectively), and 13% and 27% for the validation error. The average improvement in the mean absolute errors is 8% and 17% for the training error, and 5% and 14% for the validation error.

Figure 8 (left) shows the scatterplot of the NN, the WDSS HDA and the WSR-88D HDA for one of the bootstrap trials. From the general pattern of this figure, it is evident that the NN outperforms both the WDSS HDA and the WSR-88D HDA. Clearly the NN provides a better fit to the data than either of the existing HDA algorithms. Also shown

	Training			Validation		
Trial	NN	WDSS HDA	WSR-88D HDA	NN	WDSS HDA	WSR-88D HDA
	Mean Square Error ($10^{-4}in^2$)					
1	3616	4703	5568	4093	4188	5190
2	3711	4528	5494	3868	4713	5412
3	3883	4750	5741	3321	4044	4666
4	3385	4316	5090	4643	5353	6628
	Mean Absolute Error ($10^{-2}in$)					
1	44.4	49.4	55.3	48.7	49.7	53.8
2	45.9	49.4	54.9	48.4	49.6	55.0
3	47.6	50.6	56.2	42.5	46.2	51.2
4	44.2	48.5	53.3	48.8	52.4	59.8

Table 5: The mean square and mean absolute training and validation errors of the regression NN, WDSS HDA, and the WSR-88D HDA for 4 bootstrap trials.

are the regression fits to the corresponding plots. It can be seen that whereas the NN's fit produces a diagonal line of slope 1, the WSR-88D HDA slope falls 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 (i.e., there is no overall bias); by contrast, the WDSS HDA and the WSR-88D HDA-predicted sizes are typically higher than the actual size.

An examination of the residual-plots (Figure 8, right) is also informative. Evidently, the NN displays far less scatter about the horizontal line than either the WDSS HDA or the WSR-88D HDA does. As such, the NN's predictions are more accurate than those of the WDSS HDA or the WSR-88D HDA.

6.3 The Classification NN

The classification NN 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 Figure 9. It can be seen that the class 1 (coin-size hail) forecasts clearly discriminate between the three classes; the distribution of class 1 observations is peaked to the right, while those of the 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 (golfball-size hail) forecasts display less

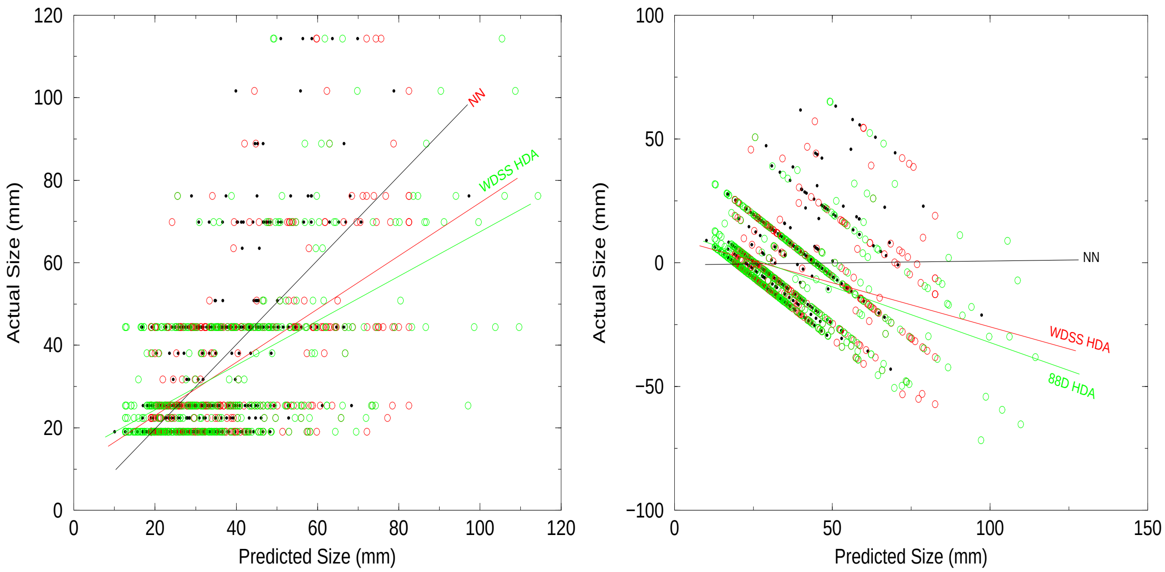

Figure 8: The scatterplot (left) and residual plot (right) of the NN (black), the WDSS HDA (red), and the WSR-88D HDA (green). For visual purposes the regression fits to the scatterplots are also shown. Based on the slope of the regression fits, it is evident that the better algorithm (i.e., with a slope ∼ 1 for the scatterplot and slope ∼ 0 for the residual plot) is the NN, followed by the WDSS HDA and the WSR-88D HDA.

discriminatory capability. These forecasts discriminate between class 1 and class 2 events, and between class 1 and class 3 (baseball-size hail) events, but not between class 2 and class 3 events. 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 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 10 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., Brier Skill Score=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.

Although the probabilities are all highly reliable, the range of the forecasts is quite

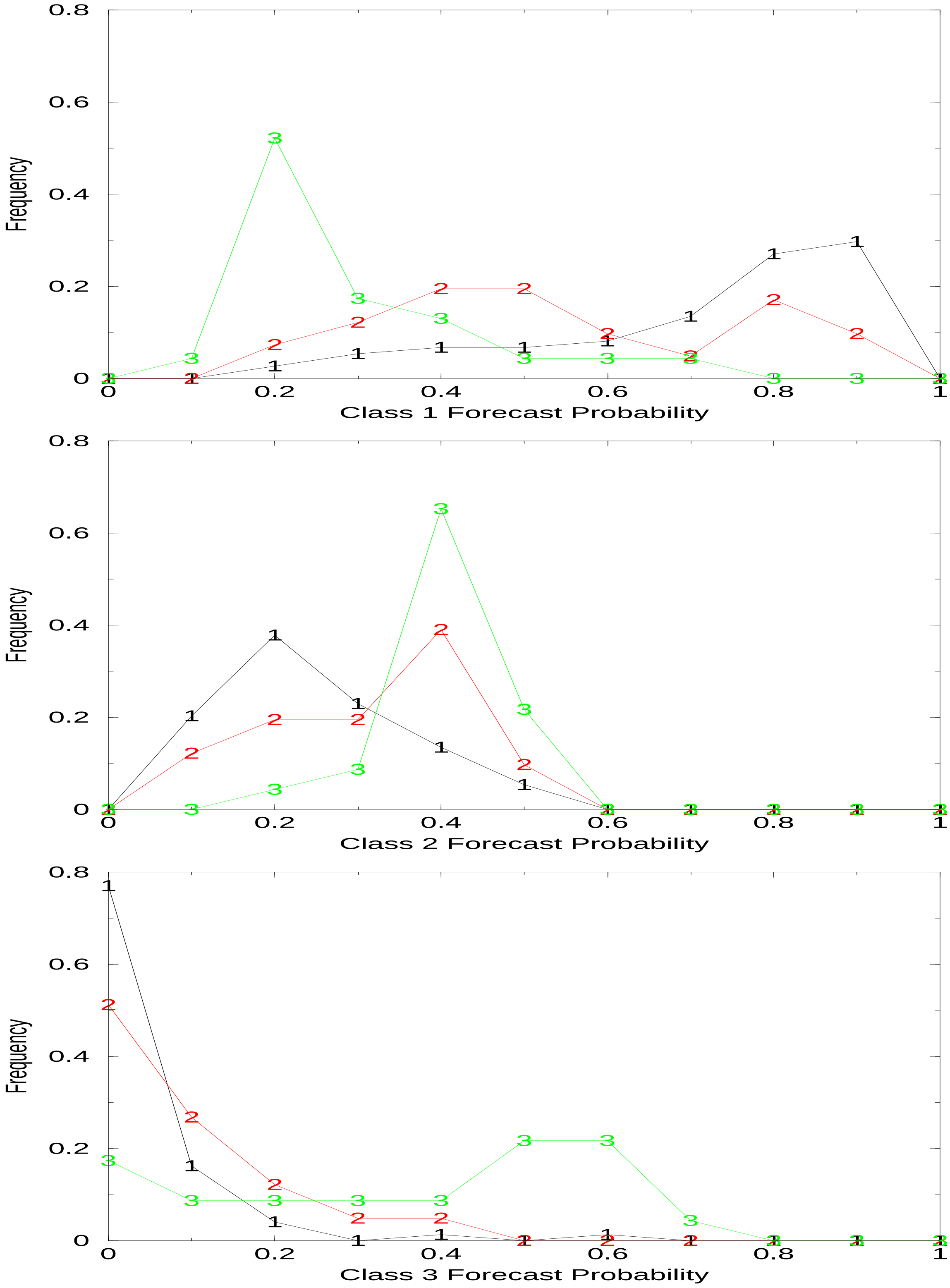

Figure 9: Discrimination diagrams for class 1 (top), class 2 (middle), and class 3 (bottom) forecasts for one validation set.

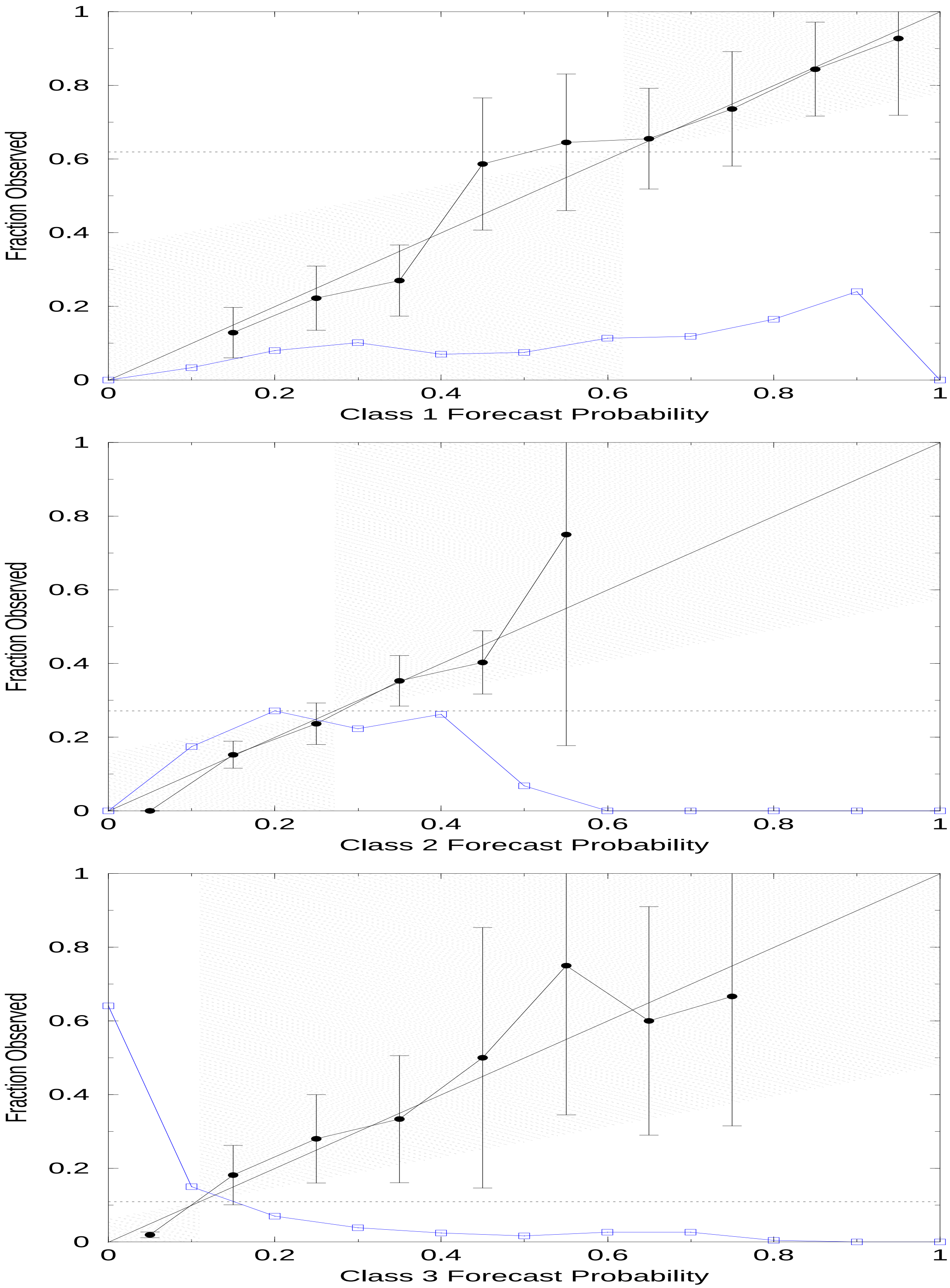

Figure 10: The attributes and refinement diagrams for the class 1 (top), class 2 (middle), and the class 3 (bottom) forecasts. Also shown are the 95% confidence intervals due to sampling. The curve consisting of the squares is the refinement diagram, the hashed region corresponds to forecasts that contribute to Brier Skill Score, and the horizontal line defines forecasts with no resolution.

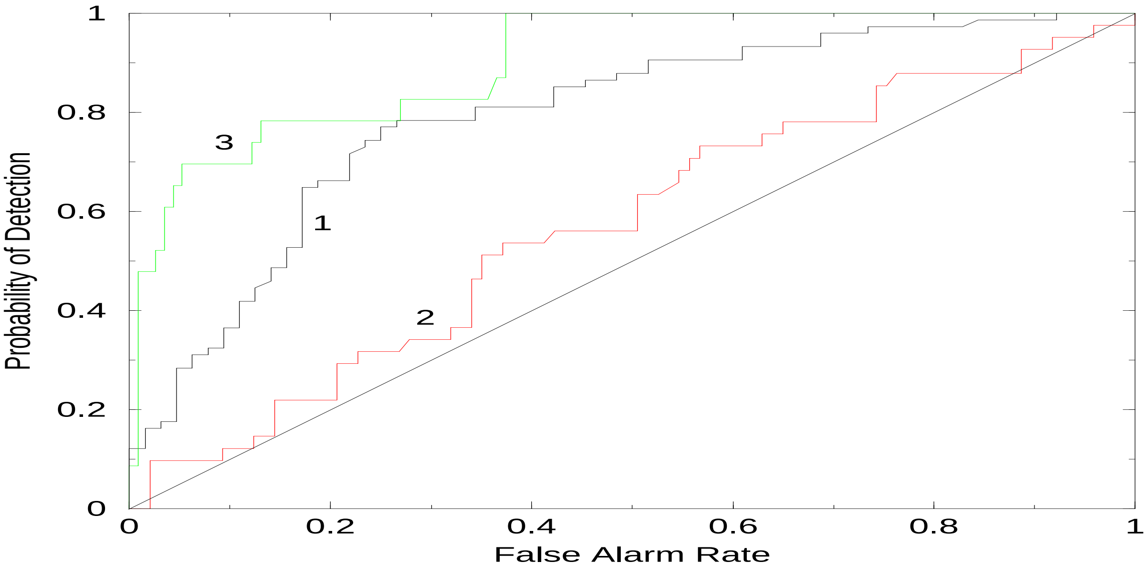

Figure 11: ROC diagrams for class 1, 2, and 3 forecasts, based on the validation set.

varied. Class 1 forecasts span a relatively wide range of 10% to 90%, whereas class 2 forecasts are restricted to the range 10%-50%. This reflects the difficulty in predicting golfball-size hail with a high degree of confidence. Class 3 forecasts reach probabilities of about 70%.

It is convenient to superimpose the refinement diagram upon the attributes diagram, the former labeled with blue squares in Figure 10. Evidently, the 3 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 (Figure 11). 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.

7 Summary

Numerous neural networks have been developed for the prediction of the occurrence and the maximum size of severe hail. The parameters of the networks have been inferred via Bayesian methodology in order to alleviate the problem of overfitting. Performance is assessed in terms of multidimensional and scalar measures. It is shown that the NNs outperform the existing counterparts. Attributes diagrams suggest that the forecasts contribute positively to skill. Even forecasts of mid-size hail ($\sim 40mm$) which initially (based on a smaller data set) displayed no statistically significant skill now display skill, because the current (larger) data set allows for the NNs to identify the underlying nonlinear relations.

8 References

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

Eilts, M. D., J. T. Johnson, E. D. Mitchell, S. Sanger, G. Stumpf, A. Witt, K. W. Thomas, K. D. Hondl, D. Rhue, and M. Jain, 1996: Severe weather warning decision support system. Preprints, *18th Conf. on Severe Local Storms*, San Francisco, CA, Amer. Meteor. Soc., 536-540.

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.

Marzban, C., 1998: Bayesian inference in neural networks. Preprints, *78th meeting of the American Meteorological Society*, Probability and Statistics Session, Phoenix Arizona, January.

Marzban, C., and A. Witt, 2000a: A Neural Network for Hail Size Prediction. Preprints, *2nd Conf. on Artificial Intelligence*, Long Beach, CA, Amer. Meteor. Soc., 38-44.

Marzban, C., and A. Witt, 2000b: A Bayesian Neural Network for Severe-Hail Size Prediction. Conditionally accepted by *Wea. & Forecasting.*

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory Mesocyclone Detection Algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 304-326.

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.

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

Witt, A., 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.

Witt, A., M. D. Eilts, G. J. Stumpf, 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.