Improving Tropical Cyclogenesis Statistical Model Forecasts through the Application of a Neural Network Classifier

CHRISTOPHER C. HENNON
Department of Atmospheric Sciences, University of North Carolina—Asheville, Asheville, North Carolina

CAREN MARZBAN
Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma, and Department of Statistics and Applied Physics Laboratory, University of Washington, Seattle, Washington

JAY S. HOBGOOD
Department of Geography, The Ohio State University, Columbus, Ohio

(Manuscript received 23 March 2004, in final form 22 May 2005)

ABSTRACT

A binary neural network classifier is evaluated against linear discriminant analysis within the framework of a statistical model for forecasting tropical cyclogenesis (TCG). A dataset consisting of potential developing cloud clusters that formed during the 1998–2001 Atlantic hurricane seasons is used in conjunction with eight large-scale predictors of TCG. Each predictor value is calculated at analysis time. The model yields 6–48-h probability forecasts for genesis at 6-h intervals. Results consistently show that the neural network classifier performs comparably to or better than linear discriminant analysis on all performance measures examined, including probability of detection, Heidke skill score, and forecast reliability. Two case studies are presented to investigate model performance and the feasibility of adapting the model to operational forecast use.

1. Introduction

Skillful forecasting of a rare event is a difficult challenge. A classic example is the forecasting of tropical cyclogenesis (TCG). On average, more than 80% of all Atlantic Basin tropical cyclone “seedlings” (tropical waves or other organized convection) fail to develop into a tropical depression despite the favorable thermodynamic environment frequently in place during the development season of summer and autumn (Avila et al. 2000; Hennon and Hobgood 2003). There is an increasing body of literature (e.g., Emanuel 1989; Bister and Emanuel 1997; Montgomery and Enaganio 1998) that suggests that smaller-scale features (submodel grid size) within cloud clusters are the important discriminating mechanisms between development and nondevelopment. Operational dynamical models constrained by insufficient resolution to capture these smaller-scale interactions have historically performed poorly in forecasting TCG, although significant improvement had been noticed during the 2001–03 Atlantic seasons (Pasch et al. 2002; J. G. Jiing 2003, personal communication).

Hennon and Hobgood (2003, hereafter HH) adopted a statistical approach to TCG forecasting by deriving large-scale predictors from the National Centers for Environmental Prediction—National Center for Atmospheric Research (NCEP–NCAR) reanalysis (Kalnay et al. 1996). Even though the data were coarse (2.5° horizontal resolution), HH showed that skillful predictions were possible with a linear discriminant analysis (LDA) classifier. A brief summary of this work will be presented in section 2. Although the specific physical mechanisms of TCG are complicated and uncertain, it is generally accepted that the process is nonlinear (e.g., Ritchie and Holland 1997; Simpson et al. 1997), at least at subsynoptic scales. This is not in contradiction with
the assumptions of HH (where linearity was implied by using LDA) because the coarse resolution of the reanalysis dataset precluded any consideration of nonlinear, mesoscale effects. But it is still possible that there are nonlinear relationships in the data itself that by definition cannot be recognized by the LDA. To determine whether improvements in forecasts may result from a more robust classifier, a neural network (NN) has been applied to the data from HH. These NNs are able to detect linear and nonlinear patterns in data. Therefore, they can be a very powerful tool for forecasting applications if they are designed and used properly. Although they are a more recent innovation than traditional statistical techniques, NNs have already been used with success in several meteorological applications, including cloud classification (Bankert 1994), tornado prediction (Marzban and Stumpf 1996), and postprocessing of model output (Marzban 2003).

The purpose of this paper is to develop and evaluate the performance of an NN in the forecasting of tropical cyclogenesis events. While performance will be measured by skill scores, a comparison will also be made to the LDA forecasts originally performed in HH. The following section will briefly review the data and methodology. Section 3 follows with descriptions of the LDA and the NN classifiers. This is followed by a brief overview of the measures of performance in section 4. Section 5 will present the results of each performance measure. Two diverse case studies are shown in section 6 that highlight model performance on individual systems. Finally, conclusions and future work are discussed in section 7.

2. Data and methodology

In HH, infrared Geostationary Orbiting Environmental Satellite-East (GOES-East) and Meteosat-7 imagery were used to identify and track 291 cloud clusters, defined as organized areas of convection with the potential to develop into tropical depressions, during the 1998–2000 Atlantic hurricane seasons. To assess the probability of a cloud cluster developing into a tropical depression (TD), eight large-scale predictors of TCG from the NCEP–NCAR reanalysis were selected a priori and calculated for each analysis time. Those predictors were latitude, daily genesis potential (DGP; McBride and Zehr 1981), maximum potential intensity (MPI; Holland 1997), low-level moisture divergence, 24-h pressure tendency, precipitable water, and 6-h surface and 700-mb relative vorticity tendency.

After the cloud clusters were identified, they were stratified into developing (DV) and nondeveloping (ND) classes. A DV case was one in which the cloud cluster developed into a TD within 48 h. An ND case was defined as a cloud cluster that formed into a TD beyond 48 h into the future, or not at all. Using the predictors described above, LDA was then used to obtain a probability for development given the atmospheric and oceanic conditions at the analysis time. Forecasts were made out to 48 h in 6-h increments. Significant forecast skill was demonstrated, but it was concluded that there were several areas where significant improvement in the model could potentially be realized. Hennon and Hobgood (2003) suggested several factors that may have limited the skill of the forecast model. Among those was the linearity of the classifier. To test this hypothesis, we use the same dataset, predictors, and methodology as in HH with the exception of the inclusion of an additional season of cloud clusters (2001).

3. Classifiers

One of the simplest statistical classifiers is discriminant analysis (McLachlan 1992). It is based on the assumption that the predictors in each class have a normal distribution. As such, the parameters of the classifier that must be inferred from data are the elements of the covariance matrix in each class. To reduce that number, one often assumes that the covariance matrices are the same across the different classes. It can be shown that this assumption of homoscedasticity leads to a classifier capable of fitting only linear decision boundaries, and for that reason it is called linear discriminant analysis (LDA). It is a robust model in that in spite of its somewhat severe assumptions it has been extremely successful in modeling a wide range of problems.

If, however, nonlinearities are expected in the data, then it is appropriate to explore a nonlinear classifier. As pointed out earlier, recent research has suggested that tropical cyclogenesis is inherently a nonlinear process in which events on a range of spatial and temporal scales interact favorably to form a self-sustaining vortex (e.g., Simpson et al. 1997). Although discriminant analysis without the assumption of homoscedasticity is nonlinear, it is capable of handling at most quadratic decision boundaries. By contrast, NNs are capable of fitting any decision boundary (Bishop 1995). Neural networks are a generalization of regression models in the sense that they consist of a number of “weights”—

1 In HH the two classes were assumed to have equal climatological probability, whereas in this study the prior probabilities are estimated from the climatology of the data. This does not fundamentally change the performance of LDA, only the magnitude of the forecast probabilities.
anals of regression coefficients—that must be estimated from data. The nonlinearity of NNs can be attributed to the existence of a parameter referred to as the number of hidden nodes, $H$. Large values of $H$ can lead to a highly nonlinear NN capable of overfitting the data, thereby rendering the classifier ineffective for classifying future cases. A small $H$ can lead to underfitting of data. A major task in NN modeling is the determination of an optimal number of hidden nodes.

A class of methods for estimating the optimal value of $H$ ($H_o$) is based on the idea of cross validation (Bishop 1995). There, one employs a sample of the data (the training set) to estimate the weights of the NN, and the remainder of the data (validation set) to estimate $H_o$. The procedure is then repeated for different samples. It can be shown that the number of hidden nodes that optimizes the performance of the NN on the average of the unused data is an unbiased estimate of the optimal value of $H$. In the present work, a series of 10 random partitions were performed for each forecast hour dataset (thus eight independent networks were trained), with 2/3 of the data sent (~1500 cases) into the training set and 1/3 (~750 cases) into the validation set. For each partition, the network was trained 20 times, each with a different random initialization of the weights (200 trials in total). This was performed by varying $H$ from 0 to 7, resulting in 1600 trials for all forecast hours (200 trials × 8 forecast hours). Minimized network cross-entropy errors from each trial were used to determine $H_o$. Thus, $H_o = 6$ was found for all forecast hours except the 12-h period ($H_o = 7$).

The NN used in this study was designed and coded within a Matlab environment (Kolenda et al. 2000). It is a three-layer, feed-forward backpropagation network with eight input nodes (layer 1), six or seven hidden nodes ($H$, described above), and one output node (layer 3). It is designed to produce outputs that are posterior probabilities of the data belonging to the “developing” group. This is achieved by the choice of the logistic activation function and adopting cross entropy as the error function. For more details on the network functions used here, see Bishop (1995). Specifically, the posterior probability of TCG is the conditional probability of TCG, given the values of the corresponding predictors (Richard and Lippmann 1991).

4. Measures of performance

Performance is a multifaceted quantity. To achieve a complete evaluation of the performance of a system, one must calculate an exhaustive suite of performance measures. It is not uncommon to find that system 1 outperforms system 2 on one skill score, but system 2 comes out ahead on a different score. Murphy and Winkler (1987) argue that a contingency table, which represents the joint probability of observations and forecasts, is a convenient way of encapsulating all components of performance. In this case, the contingency table is a $2 \times 2$ table of observations (0 for ND; 1 for DV) versus forecasts (0 and 1 as well). A number of scalar measures of performance can be calculated based on the values in the contingency table. However, many of these are not reliable for rare (low prior probability) events such as TCG and it has been argued that any attempt to optimize any single measure produces forecast bias (Marzban 1998). We can only conclude that the results presented here are valid for these performance measures only, although it is reasonable to expect they would be similar for other performance measures.

The contingency table is a convenient method of representing the joint probability of observations and forecasts. In the $2 \times 2$ case, let $a$ be the number of 0s correctly classified as 0, and $b$ be the number of 0s misclassified as 1. Also let $c$ and $d$ label the number of 1s classified as 0 and 1, respectively. A number of scalar skill scores can be derived from the contingency table. Two of the more common scores are the probability of detection (POD) and the false alarm rate (FAR), defined as

$$\text{POD} = \frac{d}{(c + d)} \quad \text{and} \quad \text{FAR} = \frac{b}{(a + b)},$$

Of course, one wishes to maximize POD and minimize FAR. The Heidke skill score is one skill score that provides a suitable measure of performance in a rare-event situation (Marzban 1998). It is defined as

$$\text{HSS} = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}.$$

Random forecasts yield $\text{HSS} = 0$, and perfect forecasts yield $\text{HSS} = 1$. Since the forecasts produced by the classifiers are probabilistic, a (decision) threshold is required to reduce them to binary forecasts. A range of thresholds is examined by incrementing them from 0 to 1, calculating scalar measures at each interval. The final result is a plot of each scalar measure as a function of the threshold.

The quality of the probabilistic forecasts is best expressed in terms of measures that do not require the

---

2 The magnitude of the weights can also affect nonlinearity.
introduction of a threshold. To adhere to the multidimensionality of forecasts, their quality is best represented in terms of diagrams. One such diagram is the reliability diagram, defined as a plot of the observed relative frequency of an event as a function of the forecast probability. Perfectly reliable forecasts will reveal themselves as a diagonal line in the diagram. Points falling above (below) the diagonal correspond to the under- (over-) forecasting an event. See Wilks (1995) for more information on interpreting reliability diagrams.

5. Results

a. HSS, POD, and FAR

The NN performance results were derived from the same 10 partitions described in section 3 but with only one random initialization. For the LDA results, 10 separate random partitions of the data were performed using identical proportions of training (2/3) and validation (1/3) data as the NN trials. Figures 1–4 show the HSS and POD for the NN and LDA classifications for the 12-, 24-, 36-, and 48-h forecast periods, varying with decision boundaries of 0 through 1. This provides information on how well each classifier performs at a different choice of decision threshold. The FARs for both classifiers were very low and similar in magnitude (~3%–5% at an optimized decision boundary). Therefore, they are not presented in graphical form here. In an operational environment, one may choose the maximum in the HSS curve to determine the optimal threshold value. The error bars in the plots represent the 90% confidence intervals if one assumes that the error values are normally distributed.

For the 12-h forecast period (Fig. 1a), there is little difference in performance between the NN and the LDA near the optimal decision boundary (~0.20). However, the NN is noticeably more robust at classifications when the decision threshold is higher. This means that the NN is less sensitive to the selection of a threshold than is LDA. This point becomes more obvious in the 24- and 36-h forecasts (Figs. 2a and 3a). Note that as the forecast lead time increases, the LDA HSS values become zero at smaller and smaller thresholds while the NN HSS values are positive across more of the spectrum. This is an indication that the quality of the NN probabilistic forecasts is higher—they are more refined, in that they span a wider range of possible values (Murphy and Winkler 1987; Murphy 1992). The 48-h forecast results (Fig. 4a) suggest a performance advantage for the NN at the optimal decision threshold in addition to the higher forecast refinement. Maximum HSS values for the NN (LDA) are approximately 0.26 (0.20) for that time, falling within the confidence intervals. The corresponding POD plots (panel b in Figs. 1–4) tell a similar story. For low thresholds (<0.10), the NN and LDA have very similar POD values for all forecast periods. However, as the threshold increases, the NN detects a higher percentage of incipient systems than does LDA for every forecast period.

In summary, although it appears that the use of a nonlinear NN has provided only a slight edge over LDA in terms of optimal classification performance, a closer look at the results across the range of thresholds clearly shows that the NN is less sensitive to the choice of a threshold and is thus a more robust classifier. The NN forecasts are also more refined than those of LDA.

---

3 Generally, an attributes diagram provides more information to the forecaster, including forecast resolution and the identification of forecasts that contribute to forecast skill. In this case, however, the forecast resolution is so small (~0.03) that nearly all forecasts contribute to forecast skill. Thus, the cleaner reliability diagrams are shown.

4 The performance of the NN on any single validation set is optimistically biased, because the validation set is employed in inferring $H_o$. However, since the NN is being compared to LDA within the framework of cross validation, this bias is not of concern.
b. Reliability diagrams

Figure 5 shows the reliability diagrams for the NN training set, validation set, and LDA for the 12-, 24-, 36-, and 48-h forecast periods. First, as expected, the forecast reliability generally decreases for both classifiers as the forecast lead time increases. Also note that the reliability curves are much more irregular for higher forecast probabilities. For rare event forecasting, this behavior is normal. Typically, the higher probability bins only contain between 0 and 10 events, whereas the lower probability bins contain hundreds of cases. An examination of the reliability curves in the region of lower forecast probabilities shows that the forecasts there are much more reliable (and with higher confidence) than those with higher probabilities.

For the 12- and 24-h forecasts, the NN performs with similar skill as the LDA, although one could argue that the NN is more reliable over a greater number of forecast probabilities (especially at 12 h). As the forecast lead times become greater, the NN increasingly outperforms the LDA. For the 36- and 48-h forecasts, there is little question that the NN has become a much more reliable classifier than the LDA. The LDA is unable to detect any developing events with a probability > 0.7 in that forecast range.

6. Forecasting application: Case studies

We now present two case studies with the objective of exploring the relationship between the predictor values and the probability output of the classifiers. A secondary purpose is to evaluate the performance that may be experienced in an operational forecast setting. In its current configuration, the model is not suitable for operational use because it was developed using a reanalysis dataset. Work is under way to assess the performance of the model with Global Forecast System (GFS) analyses serving as the training data. By examining a few case studies, we cannot only determine the feasibility of transferring this model to an operational one, but also extract some insight into model tendencies and possible systematic issues that may assist forecasters as they use it.

The first case is a developing system that was handled well by both the LDA and NN versions of the model. This case is followed by one where both classifiers “missed” a developing case. A brief discussion on the values of the predictors will follow for each case. Readers who desire a more detailed discussion on the model performance in the case studies as well as some insight into the relative significance of each predictor are encouraged to read HH. That paper included an LDA
analysis of the following case plus a more ambiguous, nondeveloping case in which the model-produced probabilities were more difficult to interpret.


The system that would eventually develop into Hurricane Keith originated as an easterly wave that moved off the African coast around 16 September 2000 (Fig. 6). Convection remained strong and relatively persistent for several days as the wave moved in a west-northwestward fashion—but development did not occur. Around 50°W the convection appeared to separate and move northward from the main wave axis. Around this time the cloud cluster could not be tracked for a couple of days as the deep convection disappeared. About 2 days prior to genesis, strong convection associated with the wave regenerated in the Caribbean Sea. The system was declared a TD by the National Hurricane Center (NHC) at 1800 UTC 28 September. For a more detailed description of Keith, see Franklin et al. (2001).

Probabilistic forecasts for genesis were issued for each 6-h period when a cloud cluster was observed. Table 1 contains the mean predictor values for all ND cases, DV cases 24 and 48 h prior to genesis, and Keith’s predictors for the same time period. For brevity, other forecast hours are omitted. The data suggest a very favorable environment for development. At 24 h prior to genesis, seven out of eight predictors (shown in boldface) have values that are more similar to all DV cases, including all thermodynamic (MPI, moisture convergence, precipitable water) predictors. At the 48-h period, they remain favorable but to a lesser degree. It is not surprising that both classifiers assigned high probabilities of development to this case.

A time series of the 24- and 48-h forecasts for the LDA and NN classifiers is shown in Fig. 7. The genesis time is shown as a filled circle and the approximate decision threshold (as discussed in section 5) is shown as a yes–no line at right. Both versions of the model performed well in both the 24- and 48-h forecast periods for the nondeveloping period from 18 September through 24 September. Probabilities of development were generally near zero for the entire period. As the convection regenerated around 26 September, both models indicate an increased chance of genesis near 28 September, with development probabilities of 85% (62%) for the NN 24-h (48 h) forecasts. The LDA forecasts are comparable, but show consistently lower probabilities for development than the NN. It should be noted that the climatological probability (from the dataset) that any cloud cluster will develop into a depression is approximately 3% (1%) at 24 (48) h.

b. Danielle (1998)

An easterly wave also provided the preexisting disturbance that would eventually form Hurricane Danielle (Fig. 8). Convection associated with the wave was initially widespread and rather disorganized but began to consolidate during the day of 22 August. About 42 h prior to genesis (1200 UTC 22 August), significant convection developed near a poorly defined circulation center. By 0600 UTC 24 August, intensity classifications from the cloud pattern (Dvorak technique) provided sufficient evidence to declare the system a tropical depression. A more thorough synoptic discussion of Danielle can be found in Pasch et al. (2001).

Both versions of the statistical model (LDA and NN) missed the genesis of Danielle. Probability forecasts for 24 and 48 h (Fig. 9) were between <1% and 5% for the 2 days preceding TCG. Table 2 shows the predictor values for Danielle 24 and 48 h prior to genesis. Although the data show a strengthening of the system at the surface (both pressure tendency and surface vorticity tendency predictors are favorable), other predictors suggest that the environment was not especially favorable for genesis. Danielle was at relatively low latitude,
and the DGP predictor suggests that she was embedded in a less than optimal dynamical environment. That leads to the question of how Danielle intensified into a tropical depression without strong large-scale support. Although a definite answer is not certain without additional investigation, there are a couple of possibilities that may explain the development. First, it is not uncommon for mesoscale features, such as convective bursts (Rodgers et al. 2000), to rapidly alter the structure and intensity of a tropical system in such a way to lead to intensification. Second, the available moisture may have become more favorable as the system emerged from a dry Saharan air layer (SAL; Dunion and Velden 2004). The SAL is a layer of much drier air.

Fig. 5. Reliability diagrams of the NN (dark) and discriminant analysis (light) for the (a) 12-, (b) 24-, (c) 36-, and (d) 48-h forecast periods.
originating from the African continent that frequently is blown out over the Atlantic during the hurricane season. It is poorly realized in the reanalysis data as well as operational analyses such as the GFS.

Each of these factors is unrecognized by the current model configuration. Since the model only considers conditions at analysis time near the cloud cluster, it is unable to predict rapid changes in the environment or evaluate conditions just ahead of the system’s track. It is interesting to note that Danielle continued to rapidly intensify after genesis, achieving hurricane status just 30 h later at 1200 UTC 25 August.

### 7. Summary, conclusions, and future work

The performance of two fundamentally different classifiers is presented in the context of a rare meteorological event, tropical cyclogenesis. Developing and nondeveloping cloud clusters are identified for the 1998–2001 Atlantic hurricane seasons. Eight large-scale predictors of TCG are selected, as in Hennon and Hobgood (2003). The forecast performance of each classifier is presented in the form of probability of detection (POD), Heidke skill scores (HSSs), and reliability diagrams.

Across nearly all possible decision boundaries, the NN outperforms LDA in terms of POD and HSS. The difference in skill becomes increasingly apparent as the forecast lead time increases. The FAR for both classifiers is very low—although the NN is slightly lower at

#### Table 1. Predictor values for all nondeveloping cases (ND mean), developing cases 24 h prior to genesis (24-h DV mean), Keith 24 h prior to genesis (Keith 24 h), developing cases 48 h prior to their genesis (48-h DV mean), and Keith 48 h prior to genesis (Keith 48 h).

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Units</th>
<th>ND mean</th>
<th>24-h DV mean</th>
<th>Keith 24 h</th>
<th>48-h DV mean</th>
<th>Keith 48 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latitude</td>
<td>10^(-5) s^-1</td>
<td>3.43</td>
<td>4.53</td>
<td>4.28</td>
<td>4.07</td>
<td>3.74</td>
</tr>
<tr>
<td>DGP</td>
<td>10^(-5) s^-1</td>
<td>0.40</td>
<td>1.23</td>
<td>1.16</td>
<td>1.05</td>
<td>0.76</td>
</tr>
<tr>
<td>MPI</td>
<td>mb</td>
<td>906.7</td>
<td>903.0</td>
<td>868.1</td>
<td>904.4</td>
<td>868.2</td>
</tr>
<tr>
<td>Moisture divergence</td>
<td>10^-7 kg s^-1</td>
<td>0.01</td>
<td>-0.03</td>
<td>0.00</td>
<td>-0.04</td>
<td>-0.08</td>
</tr>
<tr>
<td>Pressure tendency</td>
<td>mb (24 h)^-1</td>
<td>0.02</td>
<td>-0.21</td>
<td>-0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precipitable water</td>
<td>kg m^-2</td>
<td>40.6</td>
<td>41.5</td>
<td>43.1</td>
<td>41.3</td>
<td>42.6</td>
</tr>
<tr>
<td>Vorticity tendency at the surface</td>
<td>10^-5 s^-1 (6 h)^-1</td>
<td>-0.00</td>
<td>0.08</td>
<td>0.05</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Vorticity tendency at 700 hPa</td>
<td>10^-5 s^-1 (6 h)^-1</td>
<td>-0.01</td>
<td>0.05</td>
<td>-0.07</td>
<td>-0.02</td>
<td>-0.17</td>
</tr>
<tr>
<td>Development climatology</td>
<td>%</td>
<td>3%</td>
<td>1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P (LDA)</td>
<td>%</td>
<td>20%</td>
<td>5%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P (NN)</td>
<td>%</td>
<td>85%</td>
<td>62%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
all forecast times (not shown). The reliability diagrams indicate that the NN is conclusively more reliable than LDA, especially at longer forecast lead times. In general, reliability is very high (low) for low (high) forecast probabilities, an expected symptom of rare event forecasts.

We believe that the degree of robustness of these results is damped by weaknesses in the dataset, including the representation of moisture (vitaly important for the TCG process) and the difficulties of finding skillful TCG predictors from larger-scale data. We speculate that improvements in those areas would not change the fundamental conclusion of this paper that the NN can produce skillful forecasts of TCG of at least comparable, if not better, skill than an LDA classifier. However, in terms of operational implementation, the biggest weakness with using a nonlinear model is its inability to expose the underlying relationships of the data. As such, nothing can be said about the predictive strength of the predictors. This type of information is usually a critical consideration in an operational environment, and is generally more accessible within a linear framework. Hence, as future development of this model turns toward operational implementation, it may be wise to use LDA or a quadratic form of discriminant analysis as the classification method—especially since the LDA performance is not grossly inferior to the NN.

There are several areas where further work in this area should yield beneficial returns.

1) **Using higher-resolution data.** Although the NCEP–NCAR reanalysis came with the benefit of having a uniform analysis system across all years of the study, we speculate that its coarse resolution dampened the signal from developing systems, especially in the moisture and vorticity predictors. The use of a higher-resolution operational model, such as the GFS, could potentially amplify these signals at the expense of a nonuniform analysis procedure across hurricane seasons. This is currently being investigated. A preliminary test from the 2001 Atlantic hurricane season using LDA has shown that the GFS model skill scores are of similar magnitude to the research model trained with the reanalysis data.

2) **Test for better predictors.** In an effort to keep the first generation of this model simple, only eight predictors were chosen a priori. It would probably be beneficial to choose many more predictors at first, and then keep only the significant contributors by running a preprocessing routine such as principal component analysis.

3) **Add more cases.** The addition of more hurricane seasons would increase the number of developing cases in the dataset. It follows that the DV signal would be less likely to be lost in the noise of the ND cases, which far outnumber the developers. This would increase forecast skill. Of course more ND
cases would also be added, but since they are already plentiful the impact would be less.

4) **Apply model to other basins.** This model was developed solely from Atlantic Basin systems for Atlantic Basin forecasting. It is reasonable to assume that although the fundamental basis for genesis is similar in other basins, there would be small but significant differences that would have to be accounted for in order to produce a skillful model. For example, nearly half of all Atlantic tropical storms form within easterly waves. This number is smaller in the Pacific Basin.

We have shown that a statistical model with an NN classifier produces skillful forecasts of TCG. We believe that this model can add value to hurricane forecast operations as objective guidance for TCG, especially since the performance of dynamical forecast models in this area has been inconsistent. If dynamical models begin to show more improvement in forecasting TCG, this model could be adopted to serve as a baseline performance measure for them. In any event, results presented here as well as in other meteorological applications have shown that NNs can be a valuable resource for improving forecasts.

**Acknowledgments.** The authors thank the NOAA–CIRES Climate Diagnostics Center for online access to the NCEP–NCAR reanalysis data. Archived GOES imagery was provided by the Space Science Engineering Center (SSEC) at the University of Wisconsin. EUMETSAT provided the archived Meteosat-7 imagery. Sigurdur Sigurdsdsson of the Technical University of Denmark kindly answered several questions regarding the implementation and structure of their neural network.

**REFERENCES**


