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

Christopher C. Hennon[1]

Caren Marzban[2,3]

Jay S. Hobgood[4]

[1] UCAR Visiting Scientist Program, NOAA Tropical Prediction Center/National Hurricane Center, Miami FL

[2] Center for Analysis and Prediction of Storms
University of Oklahoma, Norman OK

[3] Department of Statistics, and the Applied Physics Lab.
University of Washington, Seattle WA

[4] Department of Geography, The Ohio State University, Columbus OH

Submitted March 11, 2004

Corresponding Author:

Christopher C. Hennon
NOAA TPC/National Hurricane Center
11691 SW 17th Street
Miami, FL 33165

Email: Christopher.Hennon@noaa.gov

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 which 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 a probability forecast for genesis at 6 hour intervals out to 48 hours prior to the event. Results consistently show that the neural network classifier outperforms linear discriminant analysis on all performance measures examined, including probability of detection, false alarm rate, Heidke Skill Score, and forecast reliability**.**

1. Introduction

Skillful forecasting of a rare event is a difficult challenge. A classic example is the forecasting of tropical cyclogenesis (TCG). Approximately 90% of all Atlantic Basin tropical cyclone 'seedlings' fail to develop into a tropical depression despite the favorable thermodynamic environment almost always in place during the development season of summer and autumn. There is an increasing body of literature (e.g. Emanuel 1989, Bister and Emanuel 1997, Montgomery and Enaganio 1998) which suggests that smaller scale features (sub-model grid size) within cloud clusters are the important discriminating mechanisms between development and non-development. Operational dynamical models constrained by insufficient resolution to capture these smaller scale interactions have historically performed poorly in forecasting TCG, although significant improvement has been noticed during the 2001 - 2003 Atlantic seasons (Pasch et al. 2002, J. G. Jiing 2003, personal communication).

Hennon and Hobgood (2003, hereafter HH) took a different tack in developing a probabilistic statistical model of TCG derived from large-scale predictors in the National Centers for Environmental Prediction (NCEP) / National Center for Atmospheric Research (NCAR) Reanalysis. Even though the data were large-scale (2.5˚ horizontal resolution), HH showed that useful 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 sub-synoptic 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 affects. However a non-linear classifier could potentially improve forecasts since it has the ability to detect more complex relationships in the data. A logical step, therefore, would be to apply a nonlinear classifier to the same dataset, which theoretically would improve the forecasts. A neural network (NN) was chosen as the vehicle for this purpose. Neural networks are able to detect nonlinear patterns in data and 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 post processing of model output (Marzban 2003).

The purpose of this paper is to evaluate the performance of a NN compared to LDA in forecasting TCG. 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. Finally, conclusions and future work are discussed in section 6.

2. Data and Methodology

In HH, infrared Geostationary Orbiting Environmental Satellite (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, 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 (McBride and Zehr 1984), maximum potential intensity (Holland 1997), low-level moisture divergence, 24-hour pressure tendency, precipitable water, and 6-hour 850 mb and 700 mb relative vorticity tendency.

After the cloud clusters were identified, they were stratified into developing (DV) and non-developing (ND) classes. A DV case was one in which the cloud cluster developed into a TD within 48 hours. A ND case was defined as a cloud cluster which formed into a TD beyond 48 hours 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 hours in 6 hour increments. Results were promising, but it was concluded that there were several areas where significant improvement in the model could be realized.

HH suggested several factors which may have limited the skill of the forecast model. Among those was the linearity of the classifier. To test this hypothesis, we used the same dataset, predictors, and methodology as in HH with the exception of the inclusion of an additional season of cloud clusters (2001).[1]

[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.

3. Classifiers

One of the simplest statistical classifiers is Discriminant Analysis (McLaughlan 1992). It is based on the assumption that the predictors in each class have a normal distribution. As such, the parameters of the classifier which 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 heteroelasticity 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. Although discriminant analysis without the assumption of heteroelasticity is nonlinear, it is capable of handling at most quadratic decision boundaries. By contrast, Neural Networks (NN) are capable of fitting any decision boundary (Bishop 1996). NNs are a generalization of regression models in the sense that they consist of a number of "weights" - analogs of regression coefficients - which 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.[2] Large values of H can lead to a highly nonlinear NN capable of overfitting the data, thereby rendering the classifier useless in 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.

[2] The magnitude of the weights can also affect nonlinearity.

A class of methods for estimating the optimal value of H (H_o) is based on the idea of cross validation (Bishop 1996)**.** 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 which 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, with 2/3 of the data sent into the training set and 1/3 into the validation set. For each partition, the network was trained 10 times, each with a different random initialization of the weights. This was performed by varying H from 0-7. Network errors from each trial were used to determine H_o. $H_o = 6$ was found for all forecast hours except the 12-hour period ($H_o = 7$). The NN used in this study was designed and coded within a Matlab environment (Kolenda et al. 2002). It is a three-layer, feed forward backpropagation network, structured in such a way that outputs of the network can be interpreted as posterior probabilities.

4. Measures of Performance

Performance is a multifaceted quantity. In order 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 2x2 table of observations (‘0 for ND, ‘1’ for DV) vs.

forecasts (0 and 1 as well). A number of scalar measures of performance, such as the Heidke Skill Score (HSS), can be calculated based on the values in the contingency table. However, many of these are not reliable in a rare-event situation (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 2X2 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

$$POD = \frac{d}{(c+d)}$$

$$FAR = \frac{b}{(a+b)}$$

Of course, one wishes to maximize POD and minimize FAR. The Heidke Skill Score is one skill score that is relatively healthy in the rare-event situation (Marzban 1998). It is defined as:

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

Random forecasts yield HSS=0, and perfect forecasts yield HSS=1. Since the forecasts produced by the classifiers are probabilistic, a (decision) threshold is required to reduce them to binary forecasts. The value of the threshold is incremented from 0 to 1, and the scalar measures are computed at each step. 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 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. The diagonal line represents perfectly reliable forecasts. Points falling above (below) the diagonal correspond to the under (over) forecasting an event. See Wilks (1995) for more information on interpreting reliability diagrams.[3]

[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.

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.[4] 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. Figure 1 shows the HSS, POD, and FAR for the NN and LDA classifications for the 6-24 forecast periods. Figure 2 is the same figure but for the 30-48 hour forecast periods. This provides information on how well each classifier performs at a different choice of the decision threshold. In an operational environment one may choose the maximum in the HSS curve to correspond to the optimal value of the threshold. Error bars are not shown due to aesthetic reasons. Typical standard errors for the NN HSS curves range from **+-** .04 to .08**,** with lower errors near the decision threshold extremes.

For the 6-hour forecast period, it appears that the NN generally performs better than the LDA near their respective optimal decision thresholds (~0.15 for LDA and 0.35 for NN). Average HSSs exceed 0.5 for the NN and 0.45 for LDA. At the 12-hour forecast period, the NN and LDA performance is nearly equal, although the broader maximum for the NN suggests that it is less sensitive to the selection of a threshold than

[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.

LDA. This point is even more evident in the 18-42 hour forecasts. 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. The 48-hour forecast results (Figure d) suggest, like the 6-hour forecast period, a performance advantage for the NN in addition to the higher forecast refinement. Maximum HSS values for the NN (LDA) are approximately 0.26 (0.20) for that time.

The corresponding POD and FAR plots, shown to the right of the HSS figures, tell a similar story. FARs for both the NN and LDA are generally low (~2%) near the optimal threshold for each. For low thresholds (< 0.10), the NN and LDA have very similar POD values for all forecast periods. But as the threshold increases, the NN detects a higher percentage of incipient systems than LDA for every forecast period except 6 hours (where LDA may have a slight edge at thresholds above 0.8).

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's forecasts are also more refined than those of LDA.

b. Reliability Diagrams

Figure 3 shows the reliability diagrams for the NN training set, validation set, and LDA for all forecast hours. The error bars shown are standard error, which can be interpreted as 90% confidence intervals assuming the errors are normally distributed.

First, as expected, the forecast reliability generally decreases for all 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. In essence, when the model issues a low (high) probability of genesis, the forecaster should have high (low) confidence in that forecast.

For the 6-hour forecast, the NN slightly outperforms the LDA, which is over forecasting events at higher forecast probabilities (indicated by the LDA curve below the diagonal line). As the forecast lead times increase, the NN increasingly outperforms the LDA. For the 30-48 hour forecasts, there is no 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 the 30-48 hour forecast range.

6. 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 non-developing 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), false alarm rate (FAR), Heidke Skill Scores (HSS), and reliability diagrams.

Across nearly all possible decision boundaries, the NN outperforms LDA in terms of POD and HSS and FAR. 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 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 the robustness of these results is capped by weaknesses in the dataset, including the representation of moisture (vitally important for the TCG process) and the difficulties of finding skillful TCG predictors from large-scale data. However, we do not believe that improvements in those areas would change the fundamental conclusion - that the NN is a more valuable classifier of TCG than LDA. Future development of this model should focus on improving the probabilistic forecasts with the NN classifier in place. There are several areas where further work in this area should yield beneficial returns: 1) Higher *Resolution Data.* Although the NNR 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 Global Forecast System (GFS), could potentially amplify these signals at the possible expense of a non-uniform analysis procedure across hurricane seasons. 2) *Better Predictors.* In an effort to keep the first generation of this model simple, only 8 predictors were chosen *a priori*. It would be beneficial to choose many more predictors at first, and then keep only the significant contributors by running a pre-processing 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. 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 which 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 much smaller in the Pacific Basin.

We have shown that a statistical model with a NN classifier performs better than a linear counterpart. If dynamical models continue to improve in forecasting TCG, this model may better 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 are a valuable resource for improving forecasts. As dynamical models continue to improve, the ultimate use of NNs in this context may more usefully be applied to post-model output processing.

Acknowledgments. The authors would like to thank the NOAA-CIRES Climate Diagnostics Center for on-line access of the NCEP/NCAR Reanalysis data. Archived GOES imagery was provided by the Space Science Engineering Center (SSEC) at the University of Wisconsin. EUMETSAT kindly provided archived Meteosat-7 imagery.

REFERENCES

Bankert, R.L., 1994: Cloud classification of AVHRR imagery in maritime regions using a probabilistic neural network. *J. Appl. Meteor.*, **33**, 909-918.

Bishop, C. M., 1996: Neural networks for pattern recognition. Clarendon Press, Oxford, 482 pp.

Bister, M., and K.A. Emanuel, 1997: The genesis of Hurricane Guillermo: TEXMEX analyses and a modeling study. *Mon. Wea. Rev.*, **125**, 2662-2682.

Emanuel, K.A., 1989: The finite-amplitude nature of tropical cyclogenesis. *J. Atmos. Sci.*, **46**, 2599-2620.

Hennon, C.C., and J.S. Hobgood, 2003: Forecasting tropical cyclogenesis over the Atlantic Basin using large-scale data. *Mon. Wea. Rev.*, **131,** 2927-2940.

Holland, G.J., 1997: The maximum potential intensity of tropical cyclones. *J. Atmos. Sci.*, **54**, 2519-2540.

Kolenda, T., S. Sigurdsson, O. Winther, L.K. Hansen, and J. Larsen, 2002: DTU: Toolbox. ISP Group at Institute of Informatics and Mathematical Modelling at the Technical University of Denmark. Internet. http://mole.imm.dtu.dk/toolbox/

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

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

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

McBride, J.L., and R. Zehr, 1981: Observational analysis of tropical cyclone formation. Part II: Comparison of non-developing vs. developing systems. *J. Atmos. Sci.*, **38**, 1132-1151.

McLachlan, G. J., 1992: Discriminant Analysis and Statistical Pattern Recognition, John Wiley and Sons, Inc., New York. 526 pp.

Montgomery, M.T. and J. Enaganio, 1998: Tropical cyclogenesis via convectively forced vortex Rossby waves in a three-dimensional quasigeostrophic model. *J. Atmos. Sci.*, **55**, 3176-3207.

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

Pasch, R.J., J-G. Jiing, F.M. Horsfall, H-L Pan, and N. Surgi, 2002: Forecasting tropical cyclogenesis in the NCEP global model. *Preprints, 25th Conf. Hurr. Trop. Meteor.*, San Diego, Amer. Meteor. Soc., 178-179.

Ritchie, E.A. and G.J. Holland, 1997: Scale interactions during the formation of Typhoon Irving. *Mon. Wea. Rev.*, **125**, 1377-1396.

Simpson, J., E.A. Ritchie, G.J. Holland, J. Halverson and S.R. Stewart, 1997: Mesoscale interactions in tropical cyclone genesis. *Mon. Wea. Rev.*, **125**, 2643-2661.

Wilks, D.S., 1995: Statistical methods in the Atmospheric Sciences. Academic Press, San Diego, 467 pp.

Figure 1. Heidke skill scores (left) and POD/FAR scores (right) for the NN (dark) and LDA (light). Forecast hours are a) 6, b) 12, c) 18, and d) 24. At right, the solid lines are POD and the dotted lines are the FAR.

Figure 2. As in Figure 1 except for forecast hours a) 30, b) 36, c) 42, and d) 48.

Figure 3. Reliability diagrams for the NN (dark) and LDA (light) validation datasets. The perfect reliability line is shown as the diagonal. Error bars represent standard error. Forecast times are a) 6, b) 12, c) 18, d) 24, e) 30, f) 36, g) 42, and h) 48 hours.

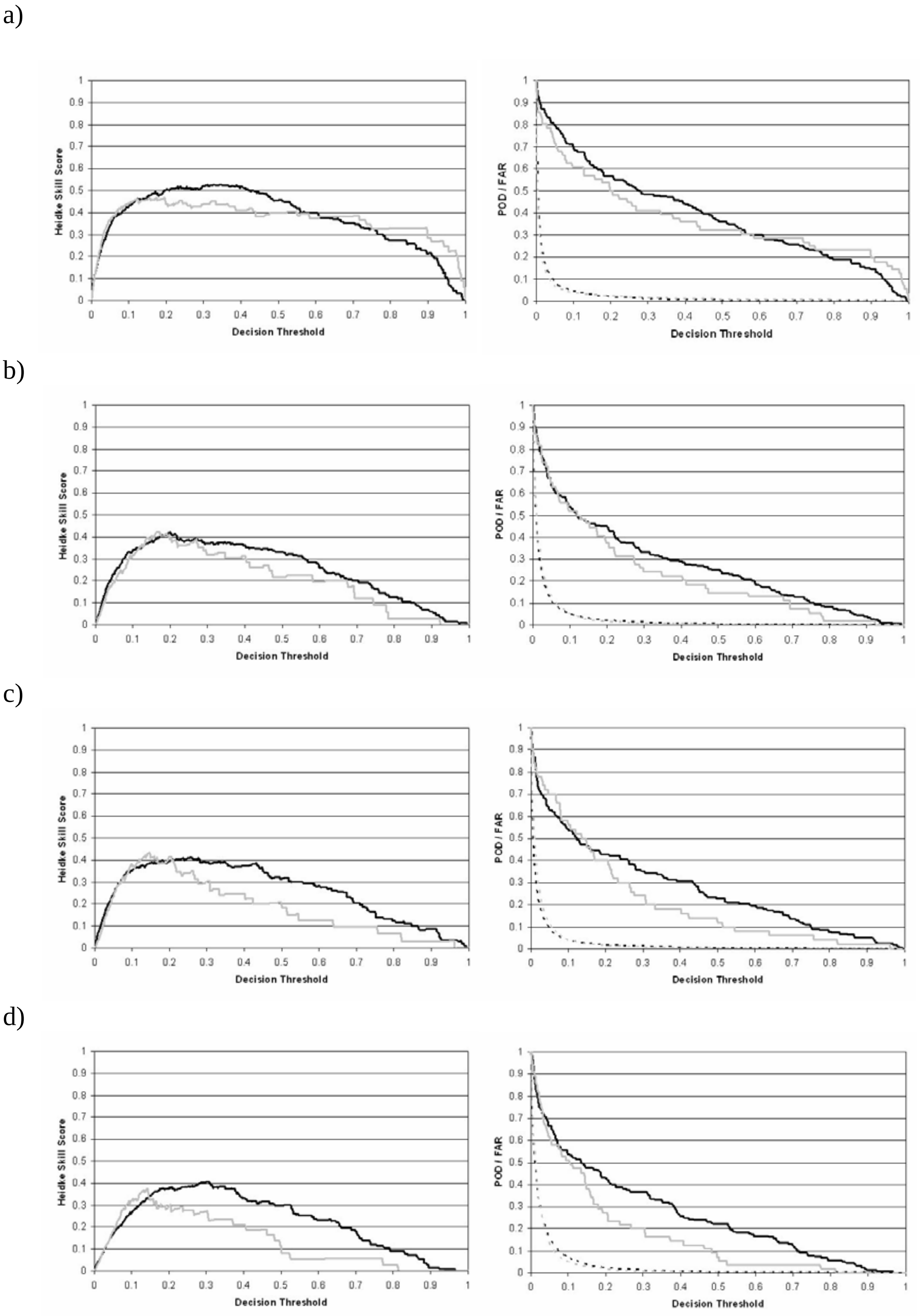

Figure 1. Heidke skill scores (left) and POD/FAR scores (right) for the NN (dark) and LDA (light). Forecast hours are a) 6, b) 12, c) 18, and d) 24. At right, the solid lines are POD and the dotted lines are the FAR.

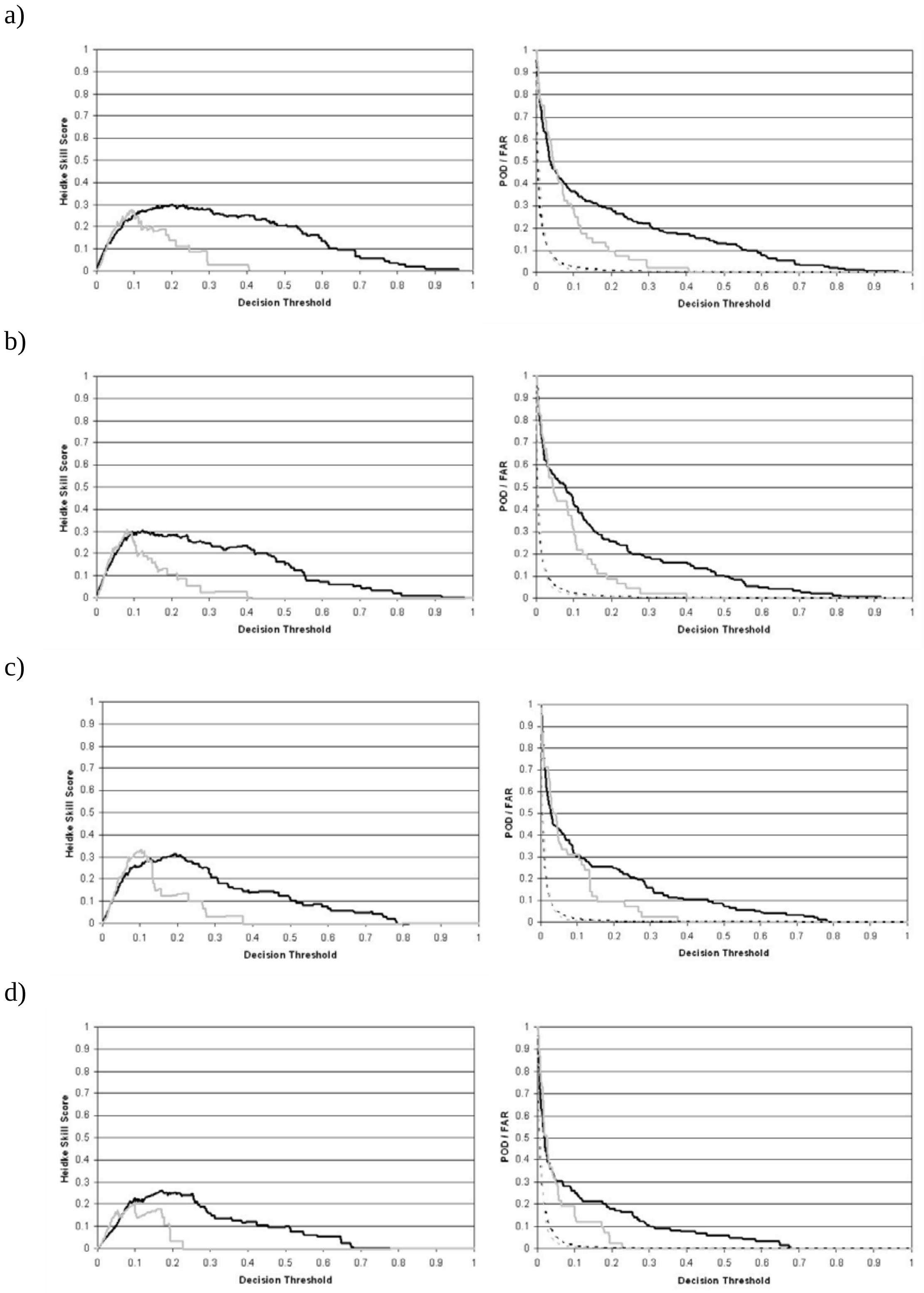

Figure 2. As in Figure 1 except for forecast hours a) 30, b) 36, c) 42, and d) 48.

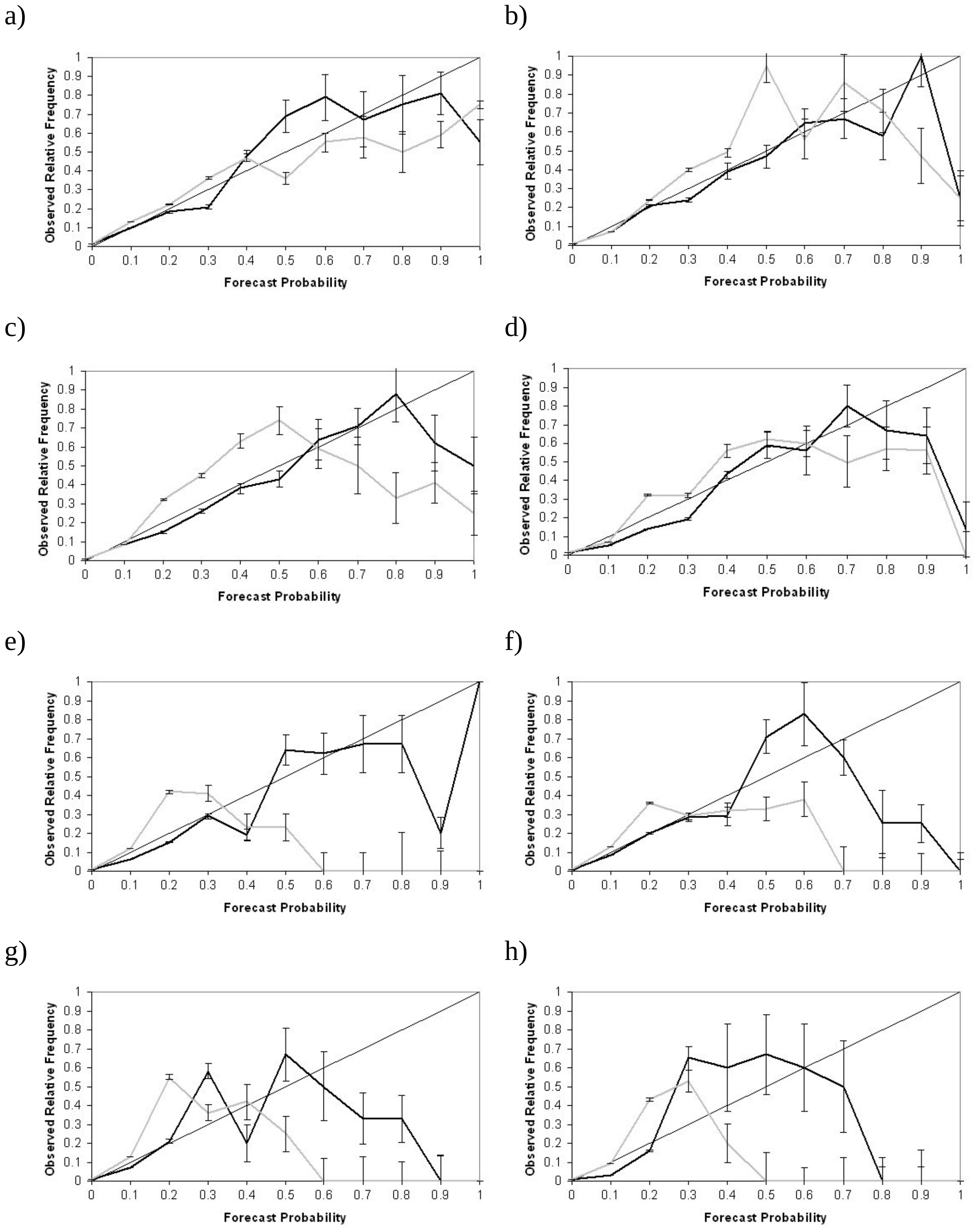

Figure 3. Reliability diagrams for the NN (dark) and LDA (light) validation datasets. The perfect reliability line is shown as the diagonal. Error bars represent standard error. Forecast times are a) 6, b) 12, c) 18, d) 24, e) 30, f) 36, g) 42, and h) 48 hours.