Improvements in Hurricane Intensity Forecasts from a Multimodel Superensemble Utilizing a Generalized Neural Network Technique

T. GHOSH AND T. N. KRISHNAMURTI[a]

Department of Earth, Ocean and Atmospheric Science, Florida State University, Tallahassee, Florida

(Manuscript received 23 January 2017, in final form 9 January 2018)

ABSTRACT

Forecasting tropical storm intensities is a very challenging issue. In recent years, dynamical models have improved considerably. However, for intensity forecasts more improvement is necessary. Dynamical models have different kinds of biases. Considering a multimodel consensus could eliminate some of the biases resulting in improved intensity forecasts as compared to the individual models. Apart from the ensemble mean, the construction of multimodel consensuses has always contributed to somewhat improved forecasts. The Florida State University (FSU) multimodel superensemble is one that, over the years, has systematically provided improved forecasts for hurricanes, numerical weather prediction, and seasonal climate forecasts. The present study considers an artificial neural network (ANN), based on biological principles, for the construction of a multimodel ensemble. ANN has been used for constructing multimodel consensus forecasts for tropical cyclone intensities. This study uses the generalized regression neural network (GRNN) method for the construction of consensus intensity forecasts for the Atlantic basin. Hurricane seasons 2012–16 are considered. Results show that with only five input models improved guidance for tropical storm intensities may be obtained. The consensus using GRNN mostly outperforms all the models included in the study and the ensemble mean. Forecast errors at the longer forecast leads are considerably less for this multimodel superensemble based on the generalized regression neural network. The skill and correlations of different models along with the developed consensus are provided in our analysis. Results suggest that this consensus forecast may be used for operational guidance and for planning and emergency evacuation management. Possibilities for future improvements of the consensus based on new advances in statistical algorithms are also indicated.

1. Introduction

Consensus forecasts for meteorological events were operationally used in the pioneering studies of Toth and Kalnay (1993 1997), Molteni et al. (1996), Houtekamer et al. (1996), and Goerss (2000). Krishnamurti et al. (1999) introduced the notion of a multimodel superensemble (MMSE) to combine multimodel forecast datasets using a linear multiple regression approach that utilized the mean-square error reduction principle. Studies reported on the efficiency of this consensus approach for the forecasting of tropical cyclones, including Krishnamurti et al. (1999, 2000), Williford et al. (2003), and Kumar et al. (2003). Cane and Milelli (2006) and Sanders (1973) had shown that a simple average of a set of forecasts produces better forecasts, which is often superior to the best model. This may be attributed to the notion that the arithmetic mean is mostly an unbiased estimator of the population mean of a statistical population. Here, one might assume that the model forecasts are the sample observations and the average of these forecasts is the sample mean. Another way of looking at this is to regard every model forecast as an estimate of the event to be forecasted, and the simple mean of the results is a standard combination of those estimates. This notion was also expressed in the works of Leslie and Fraedrich (1990), Mundell and Rupp (1995), and Goerss (2000), who examined tropical cyclone track forecasts. These studies show that a consensus, on the average, produces better results than the individual member model forecasts. The ensemble averages showed considerable improvement compared to the member models in the studies by Goerss et al. (2004) and Sampson et al. (2005) for forecasts of typhoons over

[a] We are extremely sorry to announce that Dr. Krishnamurti passed away during the editorial process of this manuscript. He contributed to this work immensely.

Corresponding author: T. Ghosh, tirthankarghosh2@gmail.com

DOI: 10.1175/WAF-D-17-0006.1

the western North Pacific and the Southern Hemisphere. The National Hurricane Center (NHC) introduced a consensus model called GUNA (Franklin 2006; Goerss 2000) in 2001. (A list of key consensus-naming acronyms and their expansions is provided in Table 1.) GUNA was based on the average of GFDL (Kurihara et al. 1993, 1995, 1998), Met Office Model (UKMO; Cullen 1993; Heming et al. 1995), NOGAPS (Hogan and Rosmond 1991; Goerss and Jeffries 1994), and the GFS model. GUNA has presently been discontinued by NHC. There are some ensembles where member models can vary. One is CONU (Goerss 2000). This was derived by taking a simple average of at least two of the five models: GFS, GFDL, the U.S. Navy's version of the GFDL model, NOGAPS, and the UKMO model. This type of consensus was started in 2000. NHC has been using several consensus aids constructed through varying and nonvarying member models. Presently, NHC uses consensus models, namely, ICON and IVCN for forecast guidance. ICON is a simple average of the following models: DSHP, HWFI, LGEM, and GHMI. Therefore, member models do not vary in ICON. IVCN is an average of at least two of the DSHP, GHMI, AVNI, and LGEM models. In the case of IVCN, member models vary. Therefore, depending on availability, sometimes it is the average of two models, sometimes it is the average of three models, and sometimes it is the average of all four models. The COAMPS-TC regional model has been included in IVCN since 2015. Unlike a simple consensus in superensemble methodology, variable weights are assigned to the models included for the construction of the consensus. This approach has been used extensively in hurricane, NWP, and climate forecasts. A summary of these works appears in Krishnamurti et al. (2016). Here, the unique aspect was in the number of weights used. Those weights vary in the three space dimensions, with time, with different variables, and with the number of models being considered. Goerss (2000) also reported that the construction of a consensus using simple averages of skilled models gives better forecasts for hurricane intensities. Construction of a weighted consensus for tropical intensity forecasts was also studied by Emanuel (2005) and Biswas et al. (2006). These studies revealed improvements in skill from the combination of forecasts, mostly outperforming the best-performing individual model.

The construction of ensemble forecasts for hurricane intensity is much needed. Intensity forecasts are still a challenge, and no dynamical model is currently performing reliably in a consistent manner. The construction of a weighted average of model forecasts was introduced by Krishnamurti et al. (1999) with the least squares principle in a linear regression. It requires some statistical assumptions on the relationships among the dependent and independent variables. However, assumptions on the causality and validity of such relationships may not always hold. In reality, individual model forecasts are available to be combined for making consensus forecasts. Whenever past forecasts and observed values are available, they can be used to construct a consensus model. Those historical values led to the construction of consensus forecasts using the principle of learning through experience. The concept of an artificial neural network (ANN) is used here to find an optimum consensus output (forecast) from the different model forecasts. This methodology works following the way the human brain makes a decision. For example, a child is shown a chair and is told the name of the object "chair" repeatedly. This is called the learning (training) phase. After some time, if the same object is shown to her, she can identify a chair correctly. She does this from experience gained during the learning phase. Information received through one or more neurons in the input layer is passed to different neurons in the next layers and finally to the brain. The brain makes an appropriate decision that may be treated as an outcome. It is worth noting that the brain makes the decision from its past experience so that the decision is supposed to be the best one for that kind of situation. ANN mimics this idea to get an optimum solution in a given situation when past observations are available. There are different kinds of ANN configurations and architectures. These are now widely used in computer science (image processing, speech recognition, etc.) and other fields as well. This method gives an optimal solution for a particular situation on the basis of past experience. The concept of ANN has also been used to derive better solutions for many atmospheric science problems. A very brief description of some of this previous work is given below.

Liu et al. (1997) have shown that neural network estimates of longwave net radiation at the sea surface are better than those found when using a regression approach. They used five input, one output, and two hidden layers. A longwave radiative transfer model was developed using ANN by Chevallier et al. (2000). Ali et al. (2004) have used ANN to estimate ocean subsurface thermal structure from surface parameters. Estimation of nonlinear interaction for wind-wave spectra using ANN was examined by Tolman et al. (2005). They found that their neural network–based interaction approximation provided reasonable results with a limitation of integration to models. Krasnoplsky et al. (2005) have developed a hybrid environmental numerical model by combining a deterministic model and an ANN model with improved results. Application of ANN in the estimation of ocean mixed layer depth was studied by Swain et al. (2006). Jain et al. (2007) have studied ocean sonic-layer depth estimation by applying ANN

TABLE 1. Description of models taken into consideration.

Name	ATCF ID	Type
Official NHC forecast	OFCL	
Previous cycle OFCL, adjusted	OFCI	Interpolated
National Weather Service (NWS) Global Ensemble Forecast System (GEFS)	AEMN	Consensus
Previous cycle AEMN, adjusted	AEMI	Consensus
NWS–GFDL model	GFDL	Multilayer regional dynamical
Previous cycle GFDL, adjusted	GFDI	Interpolated–dynamical
Previous cycle GFDL, adjusted using a variable intensity offset correction that is a function of forecast time; note that for track, GHMI and GFDI are identical	GHMI	Interpolated–dynamical
NWS Hurricane Weather Research and Forecasting Model (HWRF)	HWRF	Multilayer regional dynamical
Previous cycle HWRF, adjusted	HWFI	Interpolated–dynamical
Average of at least two of DSHP, LGEM, GHMI, HWFI, GFNI, and COAMPS-TC (since 2015)	IVCN	Consensus
Logistic Growth Equation Model	LGEM	Statistical–dynamical
Statistical Hurricane Intensity Prediction Scheme (SHIPS)	SHIP	Statistical–dynamical
SHIPS with inland decay	DSHP	Statistical–dynamical
Average of GHMI, EGRI, NGPI, and GFSI	GUNA	Consensus
Previous cycle GFS, adjusted	GFSI	Interpolated–dynamical
Previous cycle UKMO (EGRR), adjusted	EGRI	Interpolated–dynamical
Previous cycle NOGAPS (NGPS), adjusted	NGPI	Interpolated–dynamical

techniques. Surface parameters were taken as input in their study. Forecasts of ceiling and visibility using ANN from surface observations and model output were studied by Marzban et al. (2007). Sharma and Ali (2013) applied ANN to achieve high-resolution tropospheric temperature profiles using geostationary satellite observations. Sharma et al. (2013) have shown that the use of ANN for the prediction of cyclone intensity using the usual atmospheric parameters and ocean heat content could produce better predictions than other alternatives. Roebber (2015) has shown that ensemble forecasts considering evolutionary programming improve the temperature forecasts for Chicago, Illinois, as compared to the operational ensemble model output statistics. Roebber (2015) also demonstrated that the pooling of evolutionary programs and conventional ensembles produced improvements in the forecasts with respect to root-mean-square error (RMSE). In SHIPS (DeMaria and Kaplan 1994), the diagnostic synoptic parameters were used to develop a hurricane intensity forecast using the multiple linear regression technique. Sampson et al. (2008) have studied the construction of simple consensus models using simple averaging for intensity prediction. However, a study on tropical storm intensity forecasts using ANN has not been reported yet. In the present study, the effectiveness of ANN is examined; specifically, the role of the generalized regression neural network (GRNN) in constructing consensus forecasts using available single model forecasts on tropical storm intensities in the Atlantic basin is explored.

Section 2 describes the data and the study region. Sections 3 and 4 contain the methodology and results, respectively. A summary and discussion are provided in section 5.

2. Data

The Automated Tropical Cyclone Forecasting System (ATCF) data from the NHC deck were used. The data are available online (http://ftp.nhc.noaa.gov/atcf/archive/). These files carry the datasets of the various model forecasts at 6- or 12-hourly time intervals for all of the hurricanes and tropical storms. Best-track data are also available via the same link. Our study covers the data in the Atlantic basin for the years 2011–16. The geographic region on which this work is concentrated is shown in Fig. 1. Figure 1 also reveals storm paths, where different colors of the trajectories indicate the different intensity categories of the storms. Along with the observed data we have considered forecasts provided by the following models: AVNI, GHMI, HWFI, DSHP, LGEM, OFCI, and IVCN. OFCI and IVCN were considered only for performance comparison, not for developing the consensus forecast. Forecast models vary from year to year as a result of changes in the model physics and other empirical processes. The effort here is in increasing the number of cases within a single hurricane season. It is desirable to have more cases from the same year for real-time forecasting. However, for the initial storms of a season, there is no alternative but to

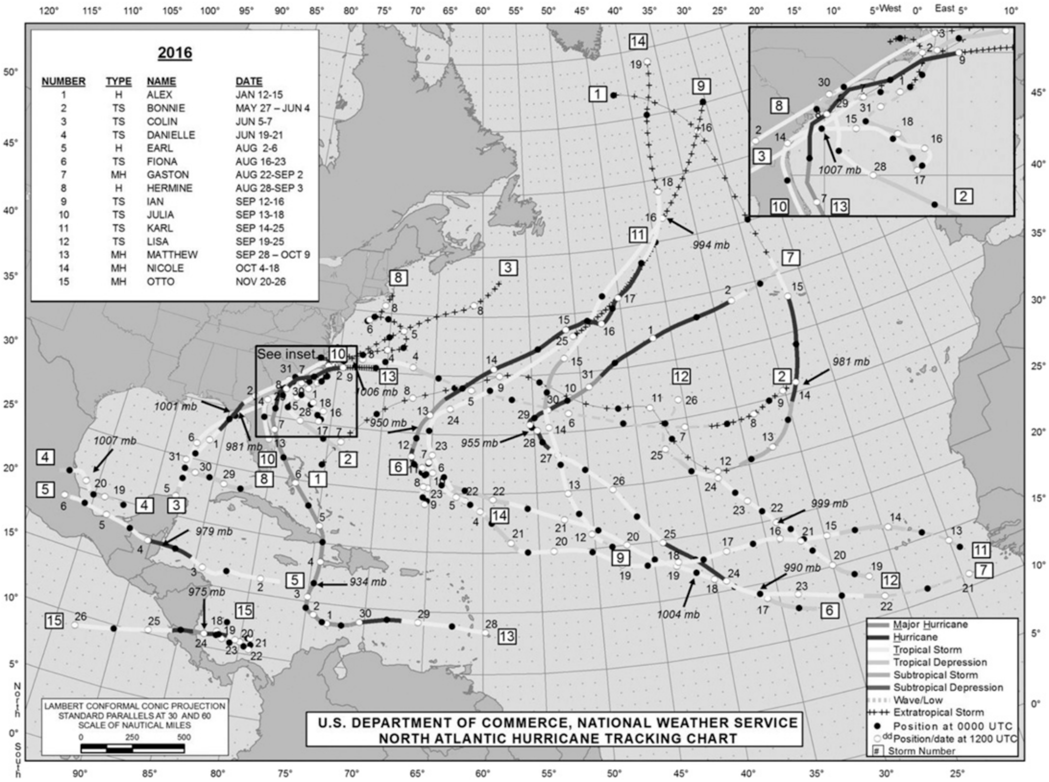


FIG. 1. The study area in the Atlantic basin. Storms during the year 2016 are shown. Different shades for a particular storm indicate the category of the storm according to the intensity scale. Image is courtesy of NHC, NOAA (http://www.nhc.noaa.gov/data/tracks/tracks-at-2016.png).

use the previous year's data. But as soon as a new season starts and storms occur, we should include the most-recent data so that the consensus model, which uses past data to generate forecasts, receives updated data with model changes to improve its performance.

3. Methodology

In making consensus forecasts using a multimodel, Krishnamurti et al. (1999) have taken a linear combination of the model forecasts. The coefficients of the linear combination were obtained by the least squares method from the previous cases. However, in the case of forecasting tropical storm intensities, it may be observed that the product moment correlation coefficient (which indicates the extent of a linear relationship) of model forecasts and observed values decreases with forecast leads. This happens, especially, for longer forecast leads. This indicates that in the longer forecast leads the relationship between the linear combination of model forecasts and observed intensities is nonlinear. The ANN methodology is well recognized for its better usefulness in the case of nonlinear situations. This methodology has been applied in optimization problems for many years. This is based on the principles of the human learning experience. It is the process of making an optimum decision by learning from past experiences. A large number of input–output examples are provided, on the basis of which a model is developed. That model gives the output corresponding to a new set of input(s). Generally, a neural network architecture comprises an input layer, one or more hidden layers, and an output layer. Information is received at the input layer and then processed at the hidden layers, and the final output is delivered at the output layer. Based on the architecture and the learning process, there are different types of neural networks. In this study the concentration is on GRNN (Specht 1991). It is a function approximation approach based on kernel regression or conditional expectation. Here, the network architecture is fixed, and there are two hidden layers between the input and output layers. The activation function is Gaussian with only one parameter, denoted by σ, called the bandwidth. Here, model forecasts are taken as input. A schematic diagram is provided in Fig. 2.

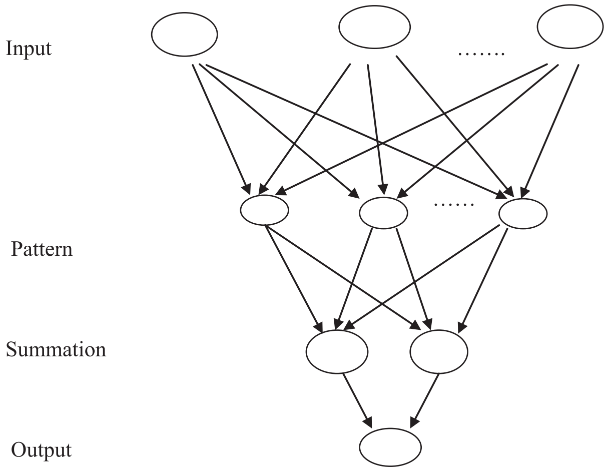


FIG. 2. Schematic diagram of the GRNN.

For forecasting a storm in a given year, cases from the previous year, as well as storms from earlier in the same year, including the observed intensities, were used. Irrespective of forecast lead hour each model forecast and observed intensity were used, respectively, as an input–output example. All such cases were taken together. This increases the number of recent cases with recent model changes. The datasets were prepared as below.

Intensity forecasts for all forecast leads (i.e., 12, 24, 36, 48, 60, 72, 84, 96, 108, and 120 h) were collected for the years 2011–16. The number of cases obtained are 1970 (in 2011), 2468 (in 2012), 836 (in 2013), 745 (in 2014), 798 (in 2015), and 1701 (in 2016). In forecasting the cases of a year, all cases of the previous year and the cases up to the current cases of that year were used together. This was done to avoid the heterogeneity of model forecasts over the years due to parametric changes in the models. In some of these cases, initially, not all model forecasts were available. For example, in a set for a single case, all included model forecasts and the corresponding observed data should be there. But, if any one or more than one of the model forecasts is not available, then those unavailable data or forecasts were treated as missing values. The numbers of missing values for years 2011, 2012, 2013, 2014, and 2015 are, respectively, 91, 75, 74, 22, and 40. If, in a set for a single case, at most two model forecasts were not available (i.e., missing), then the missing value(s) were replaced by the average of the available model forecasts of that particular case. This exercise helped to increase the number of cases, although such cases are very few with respect to the total number of cases. No standardization was carried out since all data are of the same type or unit. Then, GRNN was developed by the standard process of splitting the data into training and testing sets. For example, in the case of forecasting the 2013 storms, all 2468 observations or cases of 2012 were taken together with the available 2013 cases. Then, the dataset was split systematically into two sets, namely the training set and the validation set (sometimes called the verification set). The method developed by May et al. (2010) was used in this process. The training set contained 70% of the total cases, and the remaining cases were in the validation set. The data splitting can also be done randomly as well. Here, systematic splitting is preferable as it follows the principle of systematic sampling from a population having a systematic characteristic (i.e., linear trend). In tropical storm intensity, it may be found that the correlation coefficients of model forecasts and observed values gradually decrease from forecast leads of 12 to 120 h. Here, all forecast lead cases have been taken together. Therefore, it is expected that there would be a systematic pattern in the combined dataset. The data were split so that both the training set and the validating set contains cases from each forecast lead (e.g., 12 hourly, 24 hourly, . . . , 120 hourly) case. However, Sharma and Ali (2013) mentioned that the splitting of the data does not change the results significantly. Thefore, if data are split randomly, then similar results are also expected.

The training set is used to develop the ANN model, whereas the validation set is used for tuning the ANN model so that the model does not get overfitted and also gives better forecasts for the testing data. After the training set is used to train the model, the validating set is used to optimize the parameter bandwidth. Note that the validating set was outside the training sample so that those cases can be treated as new data. The bandwidth that gives the minimum mean square error (MSE) is taken for use in the forecast phase. This follows from May et al. (2010). This step may be skipped, and the trained GRNN model can be used to test or forecast directly without further optimizing the bandwidth parameter. During the process, cases to be forecasted were taken as the test set, and they were not part of developing the ANN model. For example, all cases of 12-, 24-, 36-hourly, etc. data of 2013 were put into the different testing sets. Those data were used to get the new forecasts employing the developed ANN model. So when 12-hourly values were forecasted, those data were not included in either the training or validating set. Therefore, the values to be forecasted were outside the training or validating sets and were not experienced by the model at all. The process was followed for every forecast lead. The effort was made to predict a set of data, using GRNN, outside of a given big dataset where all the predictors and predictands are located. The mean square errors, mean absolute errors (MAEs), and correlation coefficients are also calculated for the model by comparing the output of testing utilizing the target output.

In developing ANN, hidden layers are considered, mostly for the nonlinear relationships. Generally, the number of hidden layers is chosen by a trial and error method. The complexity of the network increases along with the increases in the hidden layers. GRNN does not suffer from this characteristic. The novelty in feeding the training period observed fields is that mostly only the values provided by the models and the observed values for the same forecast season were considered. It is assumed that there would not be changes to the models within a season.

The objective of this study was to examine how GRNN predicts a storm intensity given a large set of similar cases. Performances of different models are also not similar in each year. Therefore, it is preferable to include cases of each recent storm, as they reflect the models' latest characteristics, as soon as they are incorporated. At the same time, old cases may be truncated from the dataset as a result of old model configurations. For example, in this study for predicting cases of 2012, cases of 2011 and available cases of 2012 have been used. The training set for the 120-h forecast lead of 2012 contained 2958 cases, and 827 different cases were used to tune the bandwidth of the network.

4. Forecast results

In this section, the results for intensity forecasts for the hurricane seasons from 2012 through 2016 are included. The GRNN algorithm was run using datasets from the NHC (http://ftp.nhc.noaa.gov/atcf/archive/). Forecast errors for all of the named storms between 2012 and 2016 are included. The results for the seasonal summaries are shown as bar diagrams in Figs. 3a–e.

The summary of results for the 2012 hurricane season is shown in Fig. 3a. These results show a significant reduction in intensity forecast errors. The absolute errors between hours 24 and 120 were nearly constant around 7–9 kt (1 kt = 0.51 m s^{-1}). The 12-h absolute error of hurricane intensity, for most models, was around 6 kt, and the GRNN results show the least error compared to the other models. The total number of forecast cases varies between 177 (for 120 h) and 306 (for 12 h) for different forecast intervals. Figure 3a also indicates that the neural network–based ensemble produces at least 10% less error for the forecast leads of 48–120 h compared to models like IVCN, OFCI, and EM. It may be noted that such a number of cases carries rather robust results for the GRNN.

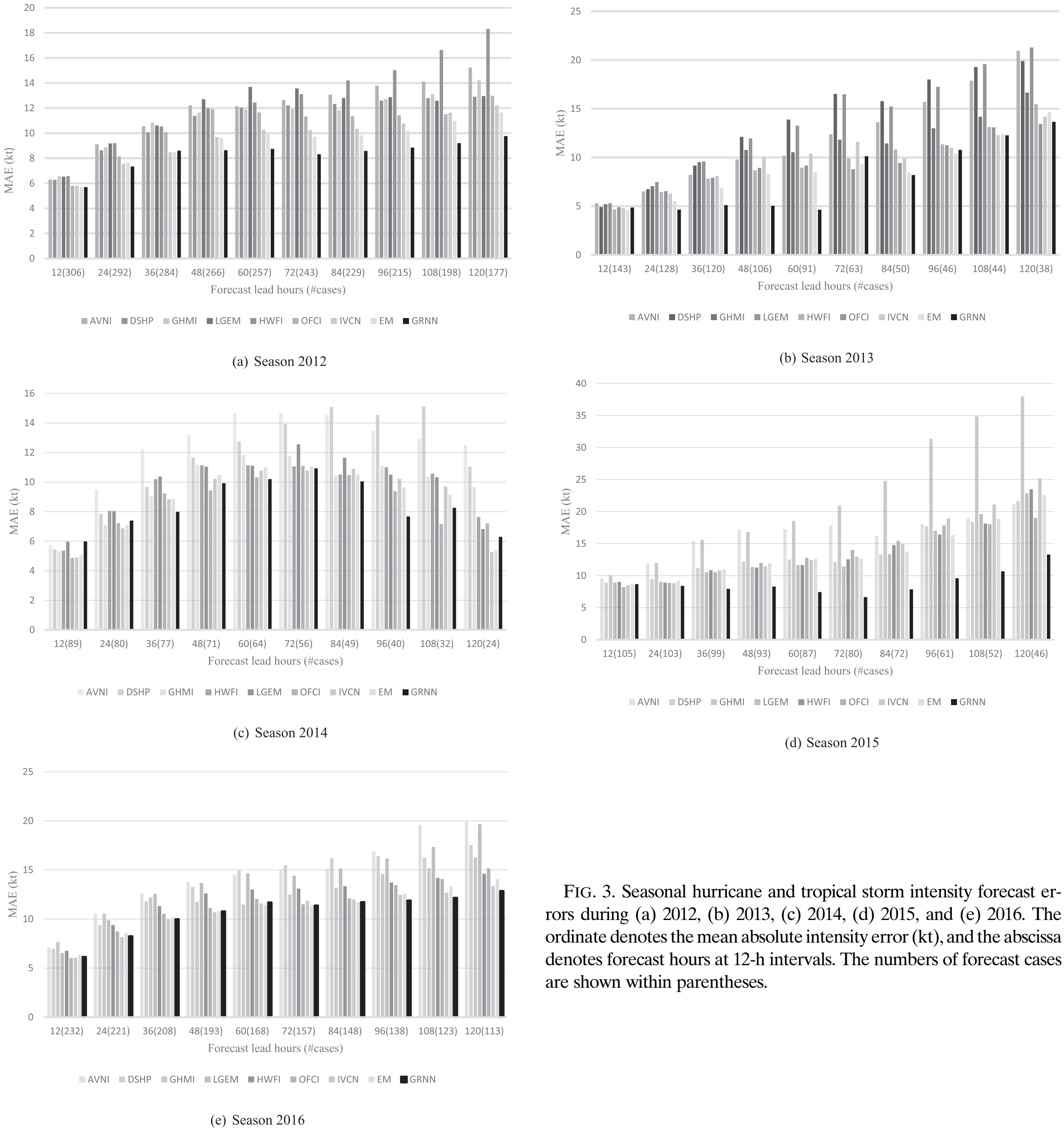
Seasonal forecast errors for the year 2013 can be seen in Fig. 3b. The numbers of cases vary between 38 (120 h) and 143 (12 h). Interestingly, here it is worth noting that GRNN gives very small errors, less than 5 kt, at the lower forecast leads. At the longer forecast leads GRNN produces almost the same error as for the ensemble mean (EM) except for 72 and 96 h. On these two occasions, GRNN has more errors than EM, which has the least error among all other models during this season.

The results for the 2014 season are shown in Fig. 3c. The minimum number of cases here is 24 at the 120-h forecast lead, and the maximum number of cases is 89 at the 12-h forecast lead. Since the number of cases is relatively low, marked improvements from the GRNN may not occur. Nevertheless, the forecasts from GRNN between hours 36 and 120 were comparable or better (as at hours 36 and 96) than those of the best-performing model. GRNN produces fewer forecast errors at the forecast leads of 36, 60, 84, 96, and 108 h than IVCN and the EM. OFCI produces higher MAEs than the GRNN at forecast leads of 60, 72, 84, and 96 h. Out of these, at 96 h the improvement is more than 10%.

Figure 3d depicts the forecast errors of different models for the 2015 season. Here, the number of hurricane forecast cases ranges between 46 (120 h) and 105 (12 h). Results show the best error reductions for the GRNN for 36 h onward. It may be noted that the individual member models show large forecast errors (more than 10 kt for leads of 36 h onward). However, the forecast errors of GRNN are less than 10 kt except for 108- and 120-h leads. GRNN produces fewer errors (10% or more) than all of the consensus models, EM, IVCN, and OFCI, especially for the forecast leads of 84–120 h.

The number of cases in 2016 is between 113 (120 h) and 232 (12 h). Forecast errors for this year show that GRNN almost consistently outperformed all member models beyond a forecast lead of 96 h.

In the 2016 season, the mean absolute errors were rather uniform around 12 kt between hours 48 and 120 for GRNN. Here, it may be seen that GRNN has the least forecast errors for the forecast leads of 72–120 h. GRNN produces 10% less error in comparison to OFCI for leads of 96–120 h. However, the improvement over IVCN varies from 2% for 108 h to 8% (72 h). At 96-h lead, the improvement of GRNN with respect to both EM and IVCN is about 2%. Overall, what is noteworthy here is that GRNN would be most valuable for real-time forecasting if the number of cases was more than 113. GRNN produces fewer forecast errors at longer forecast leads for the years 2012 (48 h onward) and 2016 (72 h onward) as well. In these years, the number of cases was relatively high. The



(a) Season 2012

(b) Season 2013

(c) Season 2014

(d) Season 2015

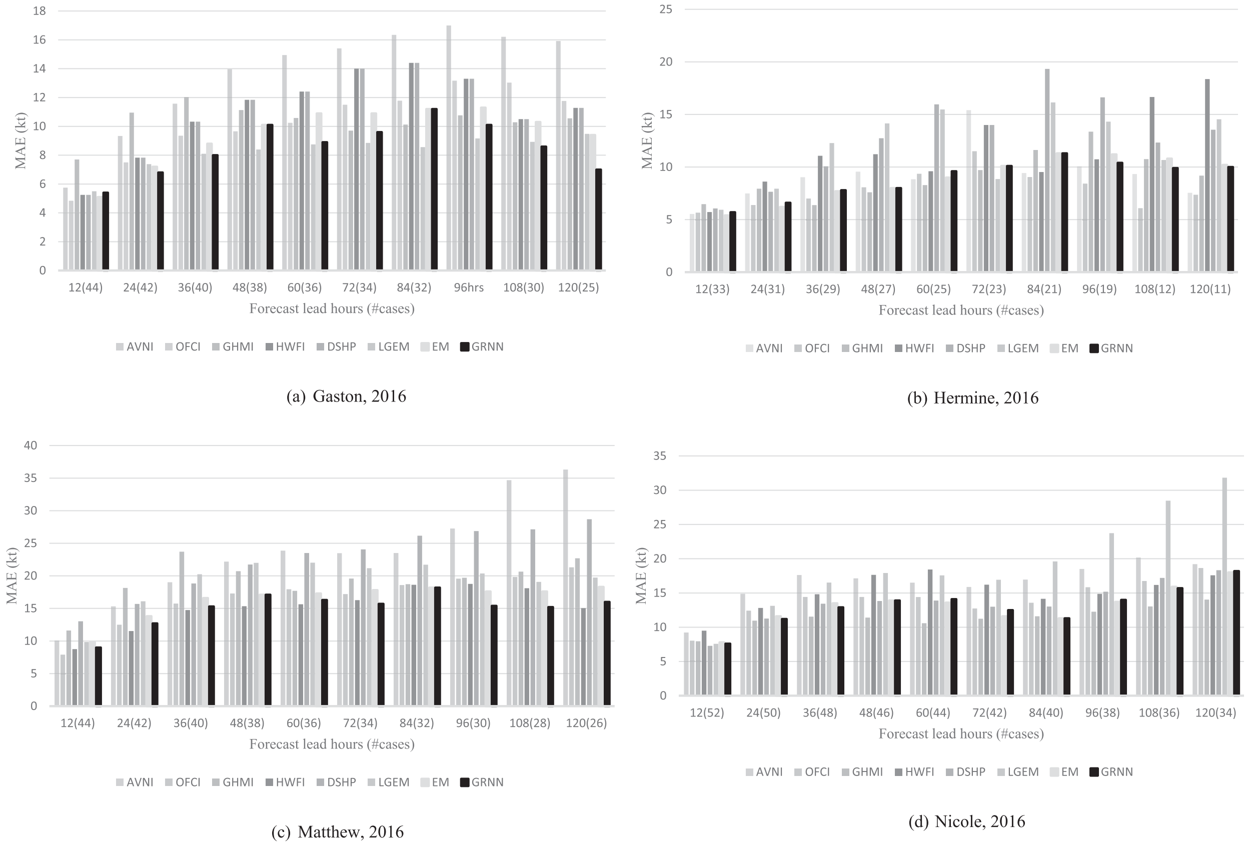
(e) Season 2016

FIG. 3. Seasonal hurricane and tropical storm intensity forecast errors during (a) 2012, (b) 2013, (c) 2014, (d) 2015, and (e) 2016. The ordinate denotes the mean absolute intensity error (kt), and the abscissa denotes forecast hours at 12-h intervals. The numbers of forecast cases are shown within parentheses.

higher the number of cases, the lower the forecast errors from GRNN. Therefore, it may be said that consistency in lower forecast errors for longer forecast leads of 36 h and onward is a notable characteristic of the GRNN forecasts. In the above seasonal summary, one important aspect is that for real-time forecasts as soon as a forecast is completed with ample cases, one can expect GRNN to provide the best hurricane intensity forecasts. Having such information for real-time cases could be very beneficial for forecast guidance.

a. *Performance with respect to some named storms*

Some individual storm forecasts for the 2016 season were also studied. The MAEs for the hurricane intensity forecasts for the "named storms" Gaston, Hermine, Matthew, and Nicole of the 2016 season are presented in Figs. 4a–d, respectively.



(a) Gaston, 2016

(b) Hermine, 2016

(c) Matthew, 2016

(d) Nicole, 2016

FIG. 4. Intensity forecast errors for (a) Gaston, (b) Hermine, (c) Matthew, and (d) Nicole during 2016. The ordinate denotes the mean absolute intensity error (kt), and the abscissa denotes forecast hours at 12-h intervals. The numbers of forecast cases are shown within parentheses.

The forecasts from GRNN for Gaston (Fig. 4a) carried the smallest mean absolute errors for intensity compared to nearly all of the member models at hours 24 and 120. The range of the number of member model forecast cases was between 25 (at hour 120) and 44 (at hour 12). The initial 12 h carries a large random error component related to the spinup of the model's hurricane intensity. This feature was noted in most model intensity forecasts. This calls for further research in the initial intensity specification for models. GRNN produces the second smallest error among the models at forecast leads of 36–96 h, while LGEM generates the smallest error at those hours. However, GHMI has fewer forecast errors than GRNN at 84 h. Gaston was a category 4 storm.

The summary of results for Hermine is shown in Fig. 4b. GRNN has not been adequately tested thus far for tropical depressions where there is a difficulty in tagging the intensity reference. Hermine was a category 1 storm. Results from this hurricane show that LGEM is one of the models that produces the largest error, even though it had the smallest error for Gaston. In the case of Hermine LGEM produced the largest errors for leads of 36 and 48 h and produced the second largest errors for leads of 24, 60, 84, 96, and 120 h. OFCI has the least error in forecasting the intensity of Hermine at the longer forecast leads. GHMI also produced much less error at the short- to medium-range forecast leads. The major drawback from the perspective of GRNN was that the number of cases ranged from 11 to 33 between 120 and 12 h. This number of cases was too low compared to Gaston, where the number of cases was between 25 and 44. However, the MAEs of EM and GRNN are almost the same for all forecast leads. The errors for GRNN range between 5 and 12 kt. This is an encouraging feature.

The forecast errors for major Hurricane Matthew are presented in Fig. 4c. This storm had a rapid intensification event. Consequently, the storm was also a difficult one to forecast with respect to intensity. Interestingly, here HWRF (HWFI) has made exceptionally good forecasts, but they were not better than those of GRNN. Beyond the forecast leads of 60 h, except at 120 h, GRNN produced fewer errors than HWRF (HWFI).

Therefore, in the case of rapid intensification as well, it may be expected that GRNN would have the smallest forecast errors for longer forecast leads. The mean absolute errors range approximately between 8 and 17 kt in the case of GRNN, which is much lesser than any models considered here including the OFCL (OFCI). It may be mentioned that EM produced the same error as GRNN at the 84-h forecast lead.

Figure 4d shows the results for Hurricane Nicole. It was also a category 4 storm. This was a storm where the GFDL's model, named GHMI, outperformed all other models after hour 12 of the forecast. GRNN was clearly the second-best model during the forecast history for Nicole. The excellent performance of HWRF/HWFI for Hermine and GFDL/GHMI for Nicole is something that may be expected because some models will outperform all others for a specific storm but without consistency. In its performance, GRNN is either the best or a close second for individual storms. The seasonal summary shows that it outperforms other models. This feature indicates the usefulness of the improved GRNN when using the neural network.

Therefore, from the intensity errors of some important storms in the 2016 season it may be observed that there is no single model that has uniformly minimum errors irrespective of storms and forecast leads. For example, forecasts of LGEM for Gaston were very impressive. But that is not seen for other storms like Hermine, Matthew, and Nicole, where LGEM forecasts carry significantly larger errors. OFCI forecasts have much smaller errors in the case of Hermine. However, this is not the case for Gaston, Matthew, and Nicole. HWFI had smaller errors for Matthew, but for Gaston, Hermine, and Nicole the HWFI forecast errors are quite large. GRNN shows consistently smaller forecast errors for the individual storms mentioned here, especially for the longer forecast leads. Matthew was a major hurricane with rapid intensification events. Therefore, it may be mentioned here that the relatively small forecast errors from GRNN in the case of Hurricane Matthew are quite encouraging. That implies, at least for longer forecast leads, that neural network–based consensus forecasts may be depended upon for better guidance. The utility of this lies in the fact that better forecasts for longer leads help with proper planning for evacuation, if necessary, as well as disaster management planning.

b. Forecast skills

Comparison of the forecast skills of different models, along with the ANN-based combined forecasts, to climatology–persistence (OCD5) was done. The computations of skills were made using the formula

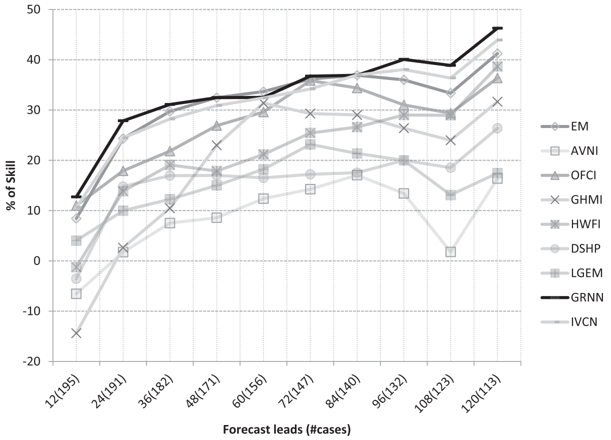


FIG. 5. Skills of different models with respect to climatology and persistence during the 2016 season. The ordinate denotes skill based on climatology and persistence (formula provided in the text), and the abscissa denotes forecast hours at 12-h intervals. The numbers of cases are shown in parentheses.

$$S_{k_f}(m)(\%) = 100 \times \left(e_{b_f} - e_{m_f}\right)/e_{b_f},$$

where e_{b_f} is the forecast error of the baseline model (CLIPER-Persistence or OCD5) and e_{m_f} is the forecast error of the model under consideration. Figure 5 relates to the 2016 cases only. It shows that GRNN has the highest skill among all models including the consensus models, IVCN, and the interpolated official forecasts (OFCI). At 72-h lead, OFCI and GRNN have the same skill levels. IVCN has equal skill to GRNN only at the forecast lead of 84 h. Another consensus model, EM, has slightly better skill at the forecast lead of 60 h. EM and GRNN have the same skill at the forecast leads of 48, 72, and 84 h. Therefore, it may be said that in the 2016 season GRNN is the most skillful model for intensity forecasting.

Forecast accuracy of the new GRNN methods for the 2016 season, in terms of other measures, is provided in Table 2. Table 2 contains the forecast accuracy of other consensus models (i.e., EM and IVCN as well) for comparison purposes. IVCN is NHC's operational consensus model whereas EM is the most commonly used consensus forecast.

Accuracy measures like bias (kt), correlation coefficient R between the observed and the forecasted intensity values, root-mean-square errors, and scatter index (SI; the ratio of RMSE to the data mean) are considered here. It may be noted from Table 2 that with respect to the considered accuracy measures the GRNN forecasts are better than the EM forecasts most of the time (i.e., the equal-weighted mean of the model forecasts). IVCN was included as it is an operational consensus model used by the NHC. It provides good forecasts. The biases of GRNN are always lower than

TABLE 2. Performance indicators of different consensus models during the 2016 season. The number in parentheses below each forecast lead time is the number of cases.

		Forecast lead (h)									
		12 (231)	24 (221)	36 (208)	48 (193)	60 (168)	72 (157)	84 (148)	96 (138)	108 (123)	120 (113)
Bias	GRNN	1.2	2.5	2.5	3.4	2.0	2.3	3.7	3.6	4.2	6.4
	EM	2.7	3.6	3.7	3.4	3.1	3.0	3.7	4.3	4.6	6.3
	IVCN	2.1	3.3	3.5	3.3	3.6	3.9	4.9	5.7	6.0	7.9
R	GRNN	0.96	0.92	0.88	0.85	0.86	0.87	0.85	0.84	0.81	0.78
	EM	0.95	0.91	0.87	0.85	0.85	0.85	0.85	0.82	0.79	0.76
	IVCN	0.95	0.91	0.88	0.85	0.85	0.85	0.84	0.83	0.82	0.80
RMSE	GRNN	6.5	9.8	11.9	13.9	11.4	10.1	11.1	10.5	11.9	13.1
	EM	7.6	10.8	12.9	13.9	12.2	11.0	11.1	11.4	12.2	13.1
	IVCN	7.1	10.7	12.9	13.9	12.6	11.4	11.9	11.6	11.8	12.3
SI	GRNN	1.1	1.2	1.2	1.3	0.97	0.88	0.95	0.88	0.98	1.0
	EM	1.2	1.3	1.3	1.3	1.1	0.96	0.95	0.90	0.92	0.94
	IVCN	1.2	1.3	1.3	1.3	1.1	0.96	0.99	0.93	0.94	0.92

those of IVCN except for at 48-h lead. The lowest bias for GRNN (IVCN) is 1.2 (2.1), while the greatest bias for GRNN (IVCN) is 6.4 (7.9). Correlation coefficients of GRNN are higher than or equal to those of IVCN except for leads of 108 and 120 h, while EM has smaller or equal correlation coefficients than GRNN for all forecast leads. Lower values of SI indicate less variance in the forecast errors. Therefore, a good model producing the lowest SI is preferable. GRNN has smaller RMSE than IVCN and EM, except for at 120-h forecast lead. Therefore, considering the above results, it may be claimed that generally GRNN provides better forecasts for hurricane intensities than IVCN and EM during the 2016 season. A similar analysis of other seasons has also been performed, but the results are not included here.

c. *Ensemble spread and forecast errors*

It is well known that, normally, the correlation between ensemble spread and forecast error is positive (Kalnay and Dalcher 1987; Murphy 1988; Buizza 1997) for shorter forecast leads. However, Barker (1991) showed that even in "perfect model" experiments the correlation coefficient between ensemble spread and error (sometimes called forecast skill) can be very low. An idealized model is one that is free from systematic biases. A detailed discussion of this topic may be found in Whitaker and Loughe (1998). In this study, the correlation coefficient between the ensemble spread and mean absolute errors has been computed. Computations were carried out for both the models, EM and GRNN, separately for each season. The results obtained do not show an encouraging relationship. All seasons showed very low correlation coefficients irrespective of forecast leads. Computed correlation coefficients are given in Table 3 for the years 2012, 2013, and 2016. It may be observed that for hurricane intensity forecasts both EM and GRNN show similar characteristics with respect to ensemble spread and forecast error correlation coefficients.

5. Summary and discussion

The construction of multimodel consensus forecasts based on an artificial neural network (ANN) was considered. The objective is to aid in the operational forecast process for the Atlantic basin. The GRNN method of ANN was deployed to examine whether consensus forecasts based on ANN are useful in forecasting tropical storm intensity. The models used for developing the GRNN consensus forecasts are AVNI, DSHP, GHMI, HWFI, and LGEM. Two important models, IVCN and OFCI, were considered for comparison purposes. OFCI may be regarded as one of the best

TABLE 3. Correlation coefficient between ensemble spread and mean absolute forecast errors for hurricane intensity forecasts in the Atlantic basin obtained for two ensembles: GRNN and EM.

Forecast	2012		2013		2016	
hours	GRNN	EM	GRNN	EM	GRNN	EM
12	−0.03	−0.002	0.16	0.17	0.28	0.28
24	−0.09	−0.1	0.10	0.23	0.11	0.13
36	0.02	−0.002	0.01	0.12	0.12	0.13
48	0.001	0.02	0.03	0.14	0.13	0.13
60	0.06	0.07	−0.09	−0.13	0.08	0.06
72	0.05	0.09	−0.2	−0.37	0.07	0.06
84	−0.01	0.08	−0.02	−0.3	0.16	0.16
96	0.04	0.02	−0.32	−0.2	0.09	0.09
108	0.06	0.09	−0.24	−0.13	0.08	0.11
120	0.09	0.1	−0.14	−0.07	0.21	0.19

consensus forecasts made by experienced meteorologists after considering all of the model forecast guidance. IVCN is an operational consensus model developed by NHC. The training sets were made so that the numbers of cases were maximized. This was done by considering the model and the observed intensities together, irrespective of forecast leads. This increases the number of cases in a hurricane season. The principle upon which the consensus model was based is input–output experience. This eliminates the annual variation of a model due to tuning or significant changes in the model parameters. It is evident from the various years' intensity forecast errors shown above that improved forecasts can be provided using neural network approaches. The improved ensemble gives minimum errors, especially for longer (48 h and beyond) forecast hours, which is very useful for planning and emergency management purposes. Most importantly, the ANN ensemble beats the arithmetic mean-based ensemble (EM). Storm-wise analyses of the year 2016 show the superiority of the new consensus forecasts. In this regard, storms Matthew and Nicole of the 2016 season may be mentioned. The skill of the new consensus is also noticeably better relative to climatology and persistence. The correlation coefficients of the observed and forecasted intensities of different models show that the new GRNN ensemble has the highest correlation for almost all forecast leads. In the seasonal results, as well as a storm-wise analysis, it may be noted that this neural network–based consensus is the best, or a very close second best. When it is the second best, the best model always varies. There is no single model that is uniformly best irrespective of forecast leads and seasons and storms. But the new consensus may be assumed to provide a consistent, superior level of performance for different seasons, storms, and forecast leads. Therefore, a new neural network–based consensus may be considered for operational use. Moreover, GRNN has the advantages of a defined network architecture and fixed hidden layers. GRNN does not strive to give a local optimal solution, as in multilayer perceptron neural networks. User choices are reduced in developing GRNN, which makes it easier to implement. Further improvement of this methodology may be achieved by including more member models as input. In this study, only five models have been used, since these models are considered to be the top intensity guidance models for the Atlantic basin. This machine learning–based ensemble may be improved further by using hybridization, a genetic algorithm, least absolute shrinkage, and a selection operator [least absolute shrinkage and selection operator (LASSO)] regression. These are the next areas that will be considered in forthcoming efforts on hurricane intensity forecasts. It is worth mentioning that the training phase largely uses the previous years' forecasts. When they are used for a current year, any model changes made during the forecast phase suffer from the use of the statistical weights of a previous year. This inconsistency can be avoided if the modeling groups provide retrospective forecasts for the previous year using the changes being made for a current year. Use of such updated data for training can further improve these multimodel ensemble forecasts.

Acknowledgments. This work was supported by a research grant from NOAA Award NA15OAR4320064. The authors benefited from discussion of this work with Dr. Mark DeMaria of the National Hurricane Center, Miami, Florida. The authors are thankful to the anonymous reviewers for their valuable comments that improved the paper.

REFERENCES

Ali, M. M., D. Swain, and R. A. Weller, 2004: Estimation of ocean subsurface thermal structure from surface parameters: A neural network approach. *Geophys. Res. Lett.*, **31**, L20308, https://doi.org/10.1029/2004GL021192.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate*, **4**, 733–742, https://doi.org/10.1175/1520-0442(1991)004<0733:TRBSAF>2.0.CO;2.

Biswas, M. K., B. P. Mackey, and T. N. Krishnamurti, 2006: Performance of the Florida State University Hurricane Superensemble during 2005. *60th Interdepartmental Hurricane Conf.*, Mobile, AL, Office of the Federal Coordinator for Meteorological Services and Supporting Research, https://www.ofcm.gov/meetings/TCORF/ihc06/Presentations/03%20session3%20Modeling%20and%20Prediction/s3-12biswas.pdf.

Buizza, R., 1997: Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **125**, 99–119, https://doi.org/10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2.

Cane, D., and M. Milelli, 2006: Weather forecasts obtained with a Multimodel SuperEnsemble technique in a complex orography region. *Meteor. Z.*, **15**, 207–214, https://doi.org/10.1127/0941-2948/2006/0108.

Chevallier, F., J. Morcrette, F. Che'ruy, and N. A. Scot, 2000: Use of neural network based longwave radiative transfer scheme in the ECMWF atmospheric model. *Quart. J. Roy. Meteor. Soc.*, **126**, 761–776, https://doi.org/10.1002/qj.49712656318.

Cullen, M. J. P., 1993: The Unified Forecast/Climate Model. *Meteor. Mag.*, **122**, 81–94.

DeMaria, M., and J. Kaplan, 1994: A Statistical Hurricane Intensity Prediction Scheme (SHIPS) for the Atlantic basin. *Wea. Forecasting*, **9**, 209–220, https://doi.org/10.1175/1520-0434(1994)009<0209:ASHIPS>2.0.CO;2.

Emanuel, K., 2005: Ensemble forecasting in hurricane intensity. *Proc. 59th Interdepartmental Hurricane Conf.,* Jacksonville, FL, Office of the Federal Coordinator for Meteorological Services and Supporting Research.

Franklin, J. L., 2006: 2005 National Hurricane Center forecast verification report. National Hurricane Center Rep., 52 pp., http://www.nhc.noaa.gov/verification/pdfs/Verification_2005.pdf.

Goerss, J. S., 2000: Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Wea. Rev.*, **128**, 1187–1193, https://doi.org/10.1175/1520-0493(2000)128<1187:TCTFUA>2.0.CO;2.

——, and R. A. Jeffries, 1994: Assimilation of synthetic tropical cyclone observations into the Navy Operational Global Atmospheric Prediction System. *Wea. Forecasting*, **9**, 557–576, https://doi.org/10.1175/1520-0434(1994)009<0557:AOSTCO>2.0.CO;2.

——, C. R. Sampson, and J. M. Gross, 2004: A history of western North Pacific tropical cyclone track forecast skill. *Wea. Forecasting*, **19**, 633–638, https://doi.org/10.1175/1520-0434(2004)019<0633:AHOWNP>2.0.CO;2.

Heming, J., J. Chan, and A. Radford, 1995: A new scheme for the initilisation of the tropical cyclones in the UK Meteorological Office global model. *Meteor. Appl.*, **2**, 171–184, https://doi.org/10.1002/met.5060020211.

Hogan, T. F., and T. E. Rosmond, 1991: The description of the Navy Operational Global Atmospheric Prediction System's Spectral Forecast Model. *Mon. Wea. Rev.*, **119**, 1786–1815, https://doi.org/10.1175/1520-0493(1991)119<1786:TDOTNO>2.0.CO;2.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242, https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2.

Jain, S., M. M. Ali, and P. N. Sen, 2007: Estimation of sonic layer depth from surface parameters. *Geophys. Res. Lett.*, **34**, L17602, https://doi.org/10.1029/2007GL030577.

Kalnay, E., and A. Dalcher, 1987: Forecasting forecast skill. *Mon. Wea. Rev.*, **115**, 349–356, https://doi.org/10.1175/1520-0493(1987)115<0349:FFS>2.0.CO;2.

Krasnopolsky, V. M., M. S. Fox-Rabinovitz, and D. V. Chalikov, 2005: New approach to calculation of atmospheric models physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Mon. Wea. Rev.*, **133**, 1370–1383, https://doi.org/10.1175/MWR2923.1.

Krishnamurti, T. N., C. M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved skills for weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, https://doi.org/10.1126/science.285.5433.1548.

——, ——, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel superensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216, https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

——, V. Kumar, A. Simon, A. Bhardwaj, T. Ghosh, and R. Ross, 2016: A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Rev. Geophys.*, **54**, 336–377, https://doi.org/10.1002/2015RG000513.

Kumar, T. S. V., T. N. Krishnamurti, M. Fiorino, and M. Nagata, 2003: Multimodel superensemble forecasting of tropical cyclones in the Pacific. *Mon. Wea. Rev.*, **131**, 574–583, https://doi.org/10.1175/1520-0493(2003)131<0574:MSFOTC>2.0.CO;2.

Kurihara, Y., M. Bender, and R. Ross, 1993: An initialization scheme of hurricane models by vortex specification. *Mon. Wea. Rev.*, **121**, 2030–2045, https://doi.org/10.1175/1520-0493(1993)121<2030:AISOHM>2.0.CO;2.

——, ——, R. Tuleya, and R. Ross, 1995: Improvements in the GFDL hurricane prediction system. *Mon. Wea. Rev.*, **123**, 2791–2801, https://doi.org/10.1175/1520-0493(1995)123<2791:IITGHP>2.0.CO;2.

——, R. Tuleya, and M. Bender, 1998: The GFDL hurricane prediction system and its performance in the 1995 hurricane season. *Mon. Wea. Rev.*, **126**, 1306–1322, https://doi.org/10.1175/1520-0493(1998)126<1306:TGHPSA>2.0.CO;2.

Leslie, L. M., and K. Fraedrich, 1990: Reduction of tropical cyclone position errors using an optimal combination of independent forecasts. *Wea. Forecasting*, **5**, 158–161, https://doi.org/10.1175/1520-0434(1990)005<0158:ROTCPE>2.0.CO;2.

Liu, Q., C. Simmer, and E. Ruprecht, 1997: Estimating longwave net radiation at sea surface from the Special Sensor Microwave/Imager (SSM/I). *J. Appl. Meteor.*, **36**, 919–930, https://doi.org/10.1175/1520-0450(1997)036<0919:ELNRAS>2.0.CO;2.

Marzban, C., S. Leyton, and B. Colman, 2007: Ceiling and visibility forecasts via neural networks. *Wea. Forecasting*, **22**, 466–479, https://doi.org/10.1175/WAF994.1.

May, R. J., H. R. Maier, and G. C. Dandy, 2010: Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks*, **23**, 283–294, https://doi.org/10.1016/j.neunet.2009.11.009.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

Mundell, D. B., and J. A. Rupp, 1995: Hybrid forecast aids at the Joint Typhoon Warning Center: Application and results. Preprints, *21st Conf. on Hurricanes and Tropical Meteorology*, Miami, FL, Amer. Meteor. Soc., 216–218.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.*, **114**, 463–493, https://doi.org/10.1002/qj.49711448010.

Roebber, P. J., 2015: Evolving ensembles. *Mon. Wea. Rev.*, **143**, 471–490, https://doi.org/10.1175/MWR-D-14-00058.1.

Sampson, C. R., J. S. Goerss, and A. J. Schrader, 2005: A consensus track forecast for Southern Hemisphere tropical cyclones. *Aust. Meteor. Mag.*, **54**, 115–119.

——, J. L. Franklin, J. A. Knaff, and M. DeMaria, 2008: Experiments with a simple tropical cyclone intensity consensus. *Wea. Forecasting*, **23**, 304–312, https://doi.org/10.1175/2007WAF2007028.1.

Sanders, F., 1973: Skill in forecasting daily temperature and precipitation: Some experimental results. *Bull. Amer. Meteor. Soc.*, **54**, 1171–1179, https://doi.org/10.1175/1520-0477(1973)054<1171:SIFDTA>2.0.CO;2.

Sharma, N., and M. M. Ali, 2013: A neural network approach to improve the vertical resolution of atmospheric temperature profiles from geostationary satellites. *IEEE Geosci. Remote Sens. Lett.*, **10**, 34–37, https://doi.org/10.1109/LGRS.2012.2191763.

——, ——, J. A. Knaff, and P. Chand, 2013: A soft-computing cyclone prediction scheme for the western North Pacific Ocean. *Atmos. Sci. Lett.*, **14**, 187–192, https://doi.org/10.1002/asl2.438.

Specht, D. F., 1991: A general regression neural network. *IEEE Trans. Neural Networks*, **2**, 568–576, https://doi.org/10.1109/72.97934.

Swain, D., M. M. Ali, and R. A. Weller, 2006: Estimation of mixed-layer depth from surface parameters. *J. Mar. Res.*, **64**, 745–758, https://doi.org/10.1357/002224006779367285.

Tolman, H. L., V. M. Krasnopolsky, and D. V. Chalikov, 2005: Neural network approximations for nonlinear interactions in wind wave spectra: Direct mapping for wind seas in deep water. *Ocean Modell.*, **8**, 253–278, https://doi.org/10.1016/j.ocemod.2003.12.008.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.*, **126**, 3292–3302, https://doi.org/10.1175/1520-0493(1998)126<3292:TRBESA>2.0.CO;2.

Williford, C. E., T. N. Krishnamurti, R. C. Torres, S. Cocke, Z. Christidis, and T. S. Vijaya Kumar, 2003: Real-time multimodel superensemble forecasts of Atlantic tropical systems of 1999. *Mon. Wea. Rev.*, **131**, 1878–1894, https://doi.org/10.1175//2571.1.