An Integrated Approach for Assessing Tropical Cyclone Track and Intensity Forecasts

WENQING ZHANG

Physical Oceanography Laboratory, Ocean University of China, Qingdao, Shandong, China, and Department of Marine, Earth, and Atmospheric Sciences, North Carolina State University, Raleigh, North Carolina

LIAN XIE

Department of Marine, Earth, and Atmospheric Sciences, North Carolina State University, Raleigh, North Carolina

BIN LIU

NOAA/NCEP/EMC/I. M. Systems Group, College Park, Maryland

CHANGLONG GUAN

Physical Oceanography Laboratory, Ocean University of China, Qingdao, Shandong, China

(Manuscript received 7 September 2016, in final form 6 January 2017)

ABSTRACT

Track, intensity, and, in some cases, size are usually used as separate evaluation parameters to assess numerical model performance on tropical cyclone (TC) forecasts. Such an individual-parameter evaluation approach often encounters contradictory skill assessments for different parameters, for instance, small track error with large intensity error and vice versa. In this study, an intensity-weighted hurricane track density function (IW-HTDF) is designed as a new approach to the integrated evaluation of TC track, intensity, and size forecasts. The sensitivity of the TC track density to TC wind radius was investigated by calculating the IW-HTDF with density functions defined by 1) asymmetric, 2) symmetric, and 3) constant wind radii. Using the best-track data as the benchmark, IW-HTDF provides a specific score value for a TC forecast validated for a specific date and time or duration. This new TC forecast evaluation approach provides a relatively concise, integrated skill score compared with multiple skill scores when track, intensity and size are evaluated separately. It should be noted that actual observations of TC size data are very limited and so are the estimations of TC size forecasts. Therefore, including TC size as a forecast evaluation parameter is exploratory at the present. The proposed integrated evaluation method for TC track, intensity, and size forecasts can be used for evaluating the track forecast alone or in combination with intensity and size parameters. As observations and forecasts of TC size become routine in the future, including TC size as a forecast skill assessment parameter will become more imperative.

1. Introduction

The track, intensity, and size of tropical cyclones (TCs) have been used as evaluation parameters in assessing TC forecasts or the performance of TC numerical forecast models since the first attempts were made at forecasting TCs in the Atlantic region in the 1870s (Sheets 1990). For instance, Neumann and Pelissier (1981) analyzed Atlantic tropical cyclone forecast errors in track and intensity, separately. Liu and Xie (2012) used errors in track, intensity, and size to assess how well the scale-selective data assimilation (SSDA) approach improved tropical cyclone forecasts in a limited-area model. More recently, Landsea and Franklin (2013) estimated the uncertainty of the Atlantic hurricane database in terms of the errors in track, intensity, and size. Forecast errors in track, intensity, and size are also analyzed individually in annual and seasonal forecast verifications at the National Hurricane Center (NHC; available online at http://www.nhc.noaa.gov/verification/verify2.shtml).

Corresponding author e-mail: Lian Xie, xie@ncsu.edu, lianxie3@gmail.com

DOI: 10.1175/WAF-D-16-0161.1

TC track is considered to be a primary assessment variable in TC forecast verification. The way to assess TC track forecasts is through the differences between the predicted and observed tracks (i.e., TC track forecast error), as well as the error relative to a low-skill climatological forecast (i.e., TC track forecast skill). TC track forecast error is usually defined as the great-circle distance between the TC location in the best-track data or other observation data and the forecast TC center valid at the forecast verification time. In advisory products, the TC center is usually defined by the location of minimum wind or minimum pressure at the surface. The track forecast error can be calculated through (Neumann and Pelissier 1981; Powell and Aberson 2001).

$$e(\text{n mi}) = 60.0\cos^{-1}[\sin\varphi_0 \sin\varphi_f + \cos\varphi_0 \cos\varphi_f \cos(\lambda_0 - \lambda_f)], \qquad (1)$$

where φ_0 and λ_0 are the latitude and longitude, respectively, of the TC center in the best-track or observation data and φ_f and λ_f are the latitude and longitude of the forecast TC center. Track forecast skill, representing a normalization of the forecast error against some standard or baseline, is given by

$$s_f(\%) = 100(e_b - e_f)/e_b, \qquad (2)$$

where e_b is the error of the baseline model and e_f is the error of the forecast being evaluated (Wilks 2006; Liu and Xie 2012; Cangialosi and Franklin 2015). It is clearly seen that skill is positive when the forecast error is smaller than the baseline error, and skill increases as the forecast error decreases. Track errors from the climatology and persistence statistical model (CLIPER5) are often used as the baseline e_b for evaluating the track forecast skill of other numerical forecast models and the official forecast (Neumann 1972; Aberson 1998). CLIPER5, originally developed in 1972, is a statistical track forecast model based on climatology and persistence and is used primarily as a benchmark for evaluating the degree of skill in a set of track forecasts, rather than as a forecast aid.

Forecast intensity (usually maximum wind speed) is assessed by using intensity forecast error, relative error, or root-mean-square error, as well as the intensity forecast skill (e.g., Neumann 1972; Neumann and Pelissier 1981; DeMaria and Kaplan 1994; Feser and Von Storch 2008; Xie et al. 2010; Liu and Xie 2012; Cangialosi and Franklin 2015). Intensity forecast error is defined as the absolute value of the difference between the forecast intensity and best-track intensity or observation data at the forecasting verification time. Intensity forecast skill can be evaluated by Eq. (2), and the baseline error e_b is from Decay-SHIFOR5, which is a version of the Statistical Hurricane Intensity Forecast (SHIFOR5) model including a weakening component that occurs when TCs move inland (DeMaria et al. 2006). SHIFOR5 is a climatology and persistence model for the intensity that is analogous to the CLIPER5 model for the track (Jarvinen and Neumann 1979; Knaff et al. 2003).

In addition to the track and intensity of a TC, the TC size is another significant structure parameter in the forecast, because the impact of a TC, such as storm surge, is also affected by TC size (Powell and Reinhold 2007; Maclay et al. 2008; Irish et al. 2008). Also, TC size has a direct influence on the extent of evacuations, ship rerouting, along-track timing of the arrival of storm conditions, and the duration of high winds at a given location (Hill and Lackmann 2009). TC size can be defined in several ways, including the distance from the TC center to the outermost closed sea level isobar; the radius of gale, tropical storm, or hurricane-force winds; the radius of the maximum wind speed (RMW); and the breadth of the satellite-observed cloud shield (Hill and Lackmann 2009; Spencer and Braswell 2001). Most often, the extent of the 34 kt ($17\,\text{m}\,\text{s}^{-1}$) winds is used to indicate the TC size (Merrill 1984). TCs are usually asymmetric (Chen and Yau 2003; Xie at al 2011), and among the parameters that can be used to describe asymmetric wind structure are the wind radii in the four quadrants (i.e., northeast, southeast, southwest, and northwest) relative to the TC center. The size error, defined as the absolute value of the difference between the forecast and the best-track data or observation data at the forecast verification time, and the root-mean-square error are typically used to assess the size forecast (Bell and Ray 2004; Demuth et al. 2004; Demuth et al. 2006).

Although the individual-parameter evaluation approach is accepted widely, it has limitations in some practical applications. First, there is a need for an unambiguous and unified assessment for TC forecasts either based on numerical or statistical models or a blend of the two. The individual-parameter evaluation method has difficulty obtaining a reasonable and integrated comprehensive assessment for the performance of a numerical forecast model or an individual TC forecast when contradicting evaluation indexes for different parameters are encountered, for instance, small track error with large intensity error and vice versa. Second, forecast error and forecast skill are usually used at the same time, and may provide conflicting assessment results, such as small track error with low track forecast skill. This may be confusing for some users (e.g., untrained users or the public) if they do not understand the different meanings of these two parameters. Third, track, intensity, and size forecasts are interdependent, especially when a TC makes landfall (DeMaria et al. 2009), so utilizing the individual-parameter evaluation approach would not be appropriate. In addition, distinguishing good forecasts from poor ones by comparing track, intensity, and size forecast error or skill, separately, is quite subjective. Another issue that should receive attention is that the traditional assessment approach does not provide the verification information of TC trajectory (track shape or moving direction) due to its one-dimensional metrics. The TC trajectory has a direct and important effect on the life of residents living near the coast as well as ships sailing at sea. For example, failing to predict a hurricane landfall as a result of an incorrect moving direction forecast will likely result in life and property losses.

In this study, an intensity-weighted hurricane track density function (IW-HTDF) is designed as a new evaluation criterion for assessing TC forecasts. This method combines hurricane track, intensity, and size data into a single parameter. It uses the concept of a cyclone track density field first proposed by Anderson and Gyakum (1989) to transform the one-dimensional track, intensity, and size into a two-dimensional field to obtain a single assessment score. It should be noted that although it is possible to create a single error (or skill score) parameter by simply combining track, intensity, and size forecast errors (or skill scores), it is a nontrivial task as the three parameters are of different dimensions. Furthermore, a simple combination of track, intensity, and size forecast errors (or skill score) does not provide an assessment of track shape.

The primary objective of this study is to introduce an integrated approach to assessing TC track, intensity, and size forecasts. The rest of the paper is organized as follows. Section 2 gives a description of the construction of IW-HTDF. A structural similarity (SSIM) index based on the comparison of observed and predicted IW-HTDFs is introduced as an integrated track, intensity, and size forecast score (IW-HTDF score) for assessing Official NHC forecasts (OFCL). The test results of the IW-HTDF approach and comparisons with the traditional evaluation criterion are given in section 3, followed by a discussion in section 4 and conclusions in section 5.

2. IW-HTDF assessment

Track, intensity, and size are the main characters of TCs and are used as the assessment variables in evaluating TC forecasts or numerical model performance on TC forecasts. The commonly adopted evaluation approach is based on assessing individual parameters separately, which may result in an inconsistent assessment, as discussed in the previous section. The motivation for the proposed new approach is to find a more concise way to assess TC forecasts or the performance of TC forecast models or methods. The IW-HTDF assessment approach contains two steps: construction of the IW-HTDF and evaluation of its applicability to TC forecast assessments.

a. Construction of the IW-HTDF

The IW-HTDF, which is a time–space function converting track, intensity, and size data from discrete hurricane data into a regularly gridded two-dimensional field in time and space, is derived from the work of Anderson and Gyakum (1989). They proposed a cyclone track density field to study the interannual and intraseasonal track variability of cold season extratropical cyclones in the Pacific basin. Xie et al. (2005) defined an HTDF based on the cyclone track density field of Anderson and Gyakum to analyze the spatial and temporal variability of North Atlantic hurricane tracks. Keith and Xie (2009) established a statistical model for predicting Atlantic tropical cyclone seasonal activity using an HTDF as a predictor selection criterion. By modifying the HTDF, which is only related to the hurricane track, the IW-HTDF is designed for individual hurricane, integrating hurricane track, intensity, and size data.

The function is defined as

$$C(\mathbf{X},t)=\sum_{i=1}^{n} W(\mathbf{X}-\mathbf{X}_i, t-t_i)\cdot P_i\cdot\alpha_i \Big/ \sum_{i=1}^{n}\alpha_i, \qquad (3)$$

where

$$W(\Delta\mathbf{X},\Delta t)=\begin{cases}\cos^2\dfrac{|\Delta\mathbf{X}|}{S_x}\cos^2\dfrac{|\Delta t|}{S_t}, & \text{if } \dfrac{|\Delta\mathbf{X}|}{S_x}<\dfrac{\pi}{2} \text{ and } \dfrac{|\Delta t|}{S_t}<\dfrac{\pi}{2};\\ 0, & \text{otherwise};\end{cases} \qquad (4)$$

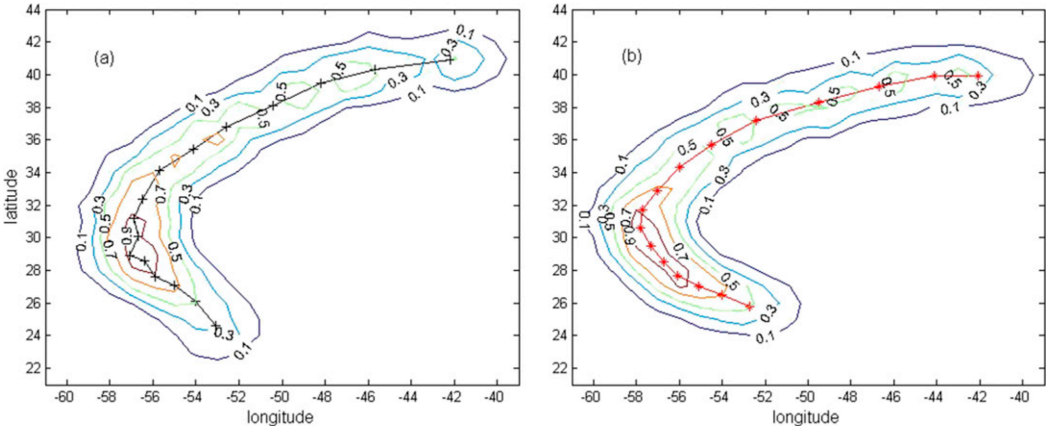

FIG. 1. Hurricane period integrated IW-HTDF result of Hurricane Edouard (2014), calculated through (a) 48-h OFCL data and (b) best-track data. The black line with crosses in (a) is the 48-h OFCL track; the red line with asterisks in (b) is the best track.

$$P_i = \frac{V_i}{V_{\text{his}}}; \tag{5}$$

$$\alpha_i = \begin{cases} 1/n, & \text{if } \dfrac{|\Delta \mathbf{X}|}{S_x} < \dfrac{\pi}{2} \quad \dfrac{|\Delta t|}{S_t} < \dfrac{\pi}{2}; \\ 0, & \text{otherwise}; \end{cases} \tag{6}$$

where $\mathbf{X}_i$ is defined as the position of ith observation or forecast of a hurricane taken at time t_i, and the grid point being estimated is at position $\mathbf{X}$ and time t. The term $W(\Delta\mathbf{X}, \Delta t)$ is a weighting function for defining the space and time smoothed hurricane track density. The spatial resolution S_x is defined as the influence radius of a TC and can be set to TC size. The temporal resolution is S_t and is set to $24/\pi$. The term P_i describes a hurricane intensity weighting factor defined as the maximum wind speed V_i at t_i normalized by the historical hurricane maximum wind speed V_{his} (which is set to 160 kt, where 1 kt = 0.51 m s^{-1}, the peak sustained wind speed of Hurricane Wilma in 2005). We set α_i as a weighted average coefficient.

IW-HTDF describes TC track, intensity, and size in time and space through the term $W(\Delta\mathbf{X}, \Delta t)$, the term P_i, and the parameter S_x, respectively, and provides a specific value of the perceived effects of a hurricane on its surroundings. To assess an integrated TC forecast over a specific forecast cycle, IW-HTDF can be integrated by time throughout the forecast cycle to obtain a two-dimensional space field $C_{\text{int}}(\mathbf{X})$:

$$C_{\text{int}}(\mathbf{X}) = \sum_{j=1}^{n} C(\mathbf{X}, t_j). \tag{7}$$

An example of this integrated field is shown in Fig. 1, which provides the accumulated IW-HTDF (S_x is set to 160 n mi; 1 n mi = 1.852 km) results [obtained from Eq. (7)] integrated over the life of Hurricane Edouard (2014) using the 48-h OFCL forecast (Fig. 1a) and the best-track data (Fig. 1b). The accumulated IW-HTDF field can be considered to be the spatial distribution of the potential influence of a hurricane. Large values appear in places closer to the track. Moreover, the value of the integrated IW-HTDF is higher in the area where the hurricane moves at a slower speed.

b. Comparison algorithm

As shown in Fig. 1, the integrated IW-HTDF fields depict the spatial pattern of the TC track, weighted by storm intensity and size. The question then is how to compare the two-dimensional IW-HTDF fields between the forecast and the best track. Several methods have been proposed for quantitative spatial verification (e.g., Gilbert 1884; Ebert and McBride 2000; Casati et al. 2004; Wang et al. 2004; Davis et al. 2006; Roberts and Lean 2008). In this study, we chose the SSIM (Wang et al. 2004) as a concise score to assess the IW-HTDF fields of TC forecasts. A similar method, known as the fractions skill score (FSS) and defined by Roberts and Lean (2008), can also be used for reference in the quantitative spatial verification of IW-HTDF fields. The FFS and its assessment results of IW-HTDF fields are described in the appendix.

Wang et al. (2004) proposed the SSIM for image quality assessment under the assumption that the human visual perception system is adapted for extracting structural information from a scene, which is widely used in the field of image and video quality assessment (Allam and Abdel-Ghaffar 2004; Coskun and Sankur 2004; Chikkerur et al. 2011). The two-dimensional IW-HTDF fields with attributes of luminance, contrast, and structure are similar to image signals; therefore, SSIM is applicable in the assessment of IW-HTDF fields and can provide a comprehensive score that is useful in the verification of the performance of a numerical weather forecast model.

In this study, SSIM is introduced as the following. A function for luminance comparison of two two-dimensional fields (X and Y) is defined:

$$L(X,Y)=\frac{2\mu_X\mu_Y+C_1}{\mu_X^2+\mu_Y^2+C_1}, \tag{8}$$

where X and Y denote integrated IW-HTDF from TC forecasts and observations, respectively. The mean values of X and Y are μ_X and μ_Y:

$$\mu_X=\frac{1}{N}\sum_{i=1}^{N}X_i \quad \text{and} \tag{9a}$$

$$\mu_Y=\frac{1}{N}\sum_{i=1}^{N}Y_i. \tag{9b}$$

Then, a contrast comparison function is introduced:

$$C(X,Y)=\frac{2\sigma_X\sigma_Y+C_2}{\sigma_X^2+\sigma_Y^2+C_2}, \tag{10}$$

where σ_X and σ_Y are the standard deviations of X and Y,

$$\sigma_X=\left[\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\mu_X)^2\right]^{1/2} \quad \text{and} \tag{11a}$$

$$\sigma_Y=\left[\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\mu_Y)^2\right]^{1/2}. \tag{11b}$$

Further, a structure comparison function is defined as

$$S(X,Y)=\frac{\sigma_{XY}+C_3}{\sigma_X\sigma_Y+C_3}, \tag{12}$$

where σ_{XY} is the covariance of X and Y:

$$\sigma_{XY}=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\mu_X)(Y_i-\mu_Y). \tag{13}$$

Additionally, C_1, C_2, and C_3 are all small constants that are included to avoid instability when $\mu_X^2+\mu_Y^2$, $\sigma_X^2+\sigma_Y^2$, or $\sigma_X\sigma_Y$ is close to 0.

Finally, the SSIM index between X and Y is defined as

$$\mathrm{SSIM}(X,Y)=[L(X,Y)]^\alpha\cdot[C(X,Y)]^\beta\cdot[S(X,Y)]^\gamma, \tag{14}$$

where α, β, and γ are positive parameters used to adjust the relative importance of the three components (i.e., L, C, and S). In this study, we set $\alpha=\beta=\gamma=1$, for simplicity.

The structure component $S(X, Y)$ is the correlation coefficient between X and Y, which measures the degree of linear correlation between X and Y, and has a theoretical range from −1 to 1. The best value, 1, is obtained when $Y = aX + b$, where a and b are constants and $a > 0$. Theoretically, as shown in Fig. 2d, if the track error is large enough (e.g., >130 n mi in Fig. 2d), the S term will become negative, and the confidence level will not exceed 90%. We consider this forecast to be very poor and to have no skill. Therefore, for those forecasts whose correlation coefficients between the predicted and observed IW-HTDF fields are negative or the confidence levels do not exceed 90%, we set $S = 0$. Even if X and Y are linearly related, there still might be relative distortions between them, which are evaluated in the components of $L(X, Y)$ and $C(X, Y)$. The luminance component $L(X, Y)$, with a value range of [0, 1], measures how close the mean luminance is between X and Y. The value 1 is gained if and only if $\overline{X} = \overline{Y}$ (i.e., $\mu_X=\mu_Y$). The terms σ_X and σ_Y can be viewed as an estimate of the contrast of X and Y, respectively, so the contrast component $C(X, Y)$ measures how similar the contrasts of the two-dimensional fields are. $C(X, Y)$ also has a value range of [0, 1], where the best value, 1, is achieved if and only if $\sigma_X = \sigma_Y$. The SSIM index gives a score (the IW-HTDF score) that measures how well the integrated track, intensity, and size forecasts perform compared with the best-track data or observations, and its range is [0, 1]. A perfect forecast (i.e., $X = Y$) has a score of 1. The relationship between the IW-HTDF score and more conventional quality metrics (e.g., errors in track, intensity, and size) is introduced in section 2c.

c. *Idealized examples*

Figure 2 shows an idealized situation in which a hurricane, moving to the north, with an intensity of 140 kt and a size of 160 n mi, is predicted with different track errors. The intensities, sizes, and moving directions of the ''observed'' and ''forecast'' hurricanes are identical; only the distance between them varies. The forecast in Fig. 2a with a relatively small track error of 32.4 n mi, obtains a relatively high IW-HTDF score of 0.89. The forecast in Fig. 2c, with a relatively large track error of 120.7 n mi, receives a low IW-HTDF score of 0.07. Figure 2d depicts the curves of three components of SSIM (i.e., L, C, and S) and the IW-HTDF score in this idealized situation. The luminance L and contrast C components are the same and constant (both equal to 1), while the structure component S decreases from 1 to 0 with the track error. So, the IW-HTDF score has the same curve with S. The smaller the track forecast error, the higher the IW-HTDF score.

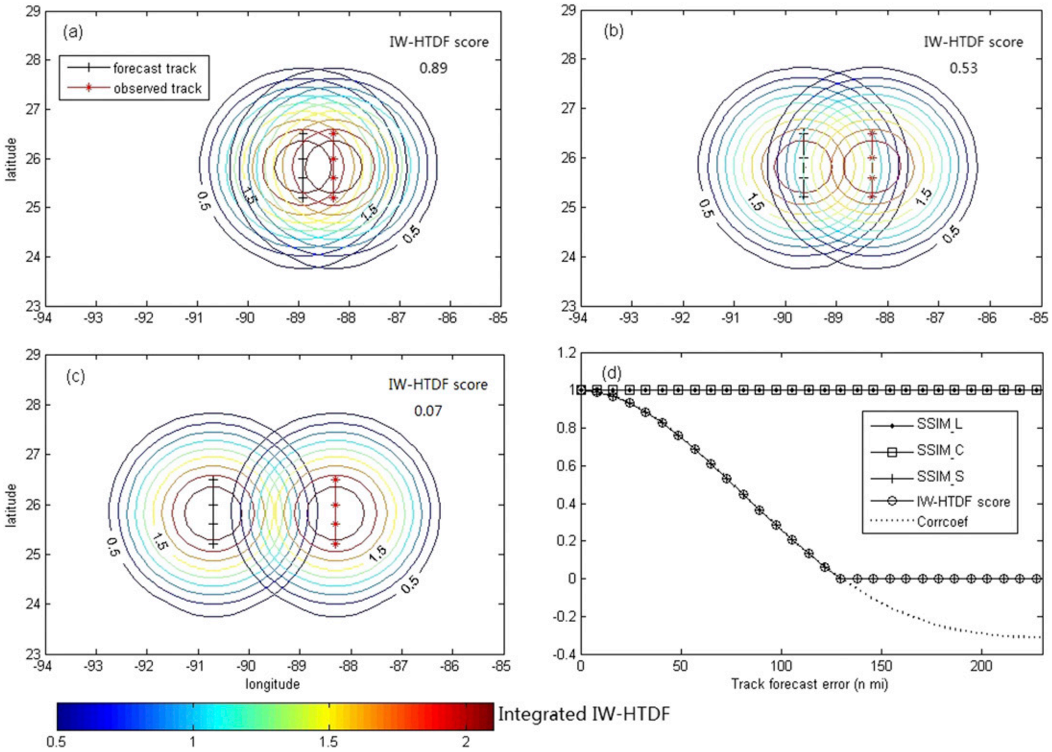

FIG. 2. Idealized situation in which a hurricane, moving to the north, with an intensity of 140 kt and a size of 160 n mi, is predicted with different track errors: (a) 32.4, (b) 73.0, and (c) 120.7 n mi. (d) The solid lines with different icons show the variations of three components of SSIM (i.e., L, C, and S) and the IW-HTDF score in the idealized situation, and the dashed line is the curve of the correlation coefficient between the observed and forecast integrated IW-HTDF fields.

For the other idealized example, the track errors (32.4 n mi), sizes (160 n mi), and moving directions (north) of the observed and forecast hurricanes are identical; only the intensity between them varies. Figure 3a shows that the structure component S is constant, unaffected by the intensity error; while the luminance and contrast components decrease as the intensity error increases in this idealized situation. So the IW-HTDF score is lower with larger intensity error.

As shown in Fig. 3b, all of the three SSIM components (i.e., L, C, and S) as well as the IW-HTDF score obtain lower values with larger size error in the idealized situation in which the track errors (32.4 n mi), intensities (140 kt), and moving directions (north) of the observed and forecast hurricanes are identical; only the size between them varies. It is also clearly seen that the structure term S decreases slowly (compared with L and C) with size error in this idealized situation.

3. Experimental results

In this study, we apply the new IW-HTDF assessment approach to evaluate the 48-h hurricane forecasts from OFCL, and compare the results with those from the traditional individual-parameter evaluation method.

NHC has been issuing forecasts of hazardous tropical weather such as hurricanes and tropical storms since 1954. NHC utilizes several models [e.g., GFDL, HWRF, the GFS Aviation Ontime (AVNO), SHIPS] as guidance during the preparation of official track and intensity forecasts. The official forecasts contain intensity (i.e., maximum 1-min surface wind speed), central pressure, position (i.e., latitude and longitude of storm center), and size (i.e., the maximum extent of winds of 34, 50, and 64 kt in each of the four quadrants about the center) of tropical or subtropical cyclones. These forecasts are issued every 6 h (at 0300, 0900, 1500, and 2100 UTC), and each contains projections valid for 12, 24, 36, 48, 72, 96, and 120 h after the forecast's nominal initial time (i.e., 0000, 0600, 1200, and 1800 UTC). These data are available from the Automated Tropical Cyclone Forecast System (ATCF).

The best-track dataset created by NHC, used as the benchmark by both the IW-HTDF assessment approach and individual-parameter evaluation method, is a poststorm analysis of each tropical cyclone's intensity,

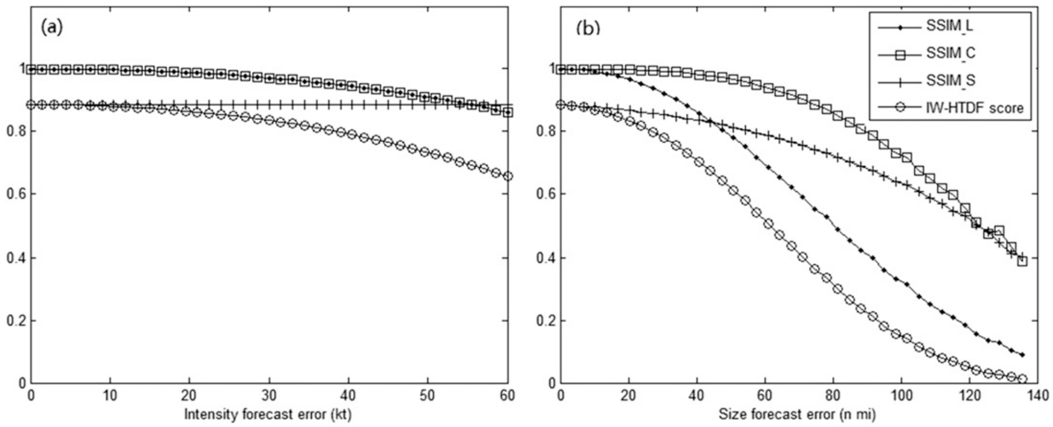

FIG. 3. The curves of three components of SSIM (i.e., L, C, and S) and the IW-HTDF score in the idealized situations: (a) in which a hurricane, moving to the north, with a track error of 32.4 n mi and size of 160 n mi, is predicted with different intensity errors, and (b) in which a hurricane, moving to the north, with track error of 32.4 n mi and intensity of 140 kt, is predicted with different size errors.

central pressure, position, and size (Jarvinen et al. 1984; Landsea and Franklin 2013). This analysis makes use of all available observations, including those that may not have been available in real time. This dataset contains 6-hourly information on the location, maximum wind, central pressure, and size of all known tropical cyclones and subtropical cyclones. Because wind radii were not included in the poststorm best-track database until 2004, tropical cyclones from 2004 to 2015 were chosen to construct the IW-HTDF on a 0.2° gridded domain. During this period, there were 193 tropical cyclones, including 15 tropical depressions, 91 tropical storms, and 87 hurricanes generated in the Atlantic. One of the significant properties of IW-HTDF is the asymmetrical wind structure of tropical cyclones, so we only consider ''hurricane'' cases that have a well-defined eye and four-quadrant wind radius data. Finally, for statistical significance, we chose 54 hurricane cases whose active periods were the same (consistent) between the 48-h OFCL forecast results and the baseline data (best track), and longer than 24 h.

a. *Assessing track, intensity, and asymmetric size*

In the first experiment, S_x is defined as the influence radius of a hurricane and set to 34-kt wind radii in four quadrants (i.e., northeast, southeast, southwest, and northwest) surrounding the hurricane as shown in Fig. 4a. Based on the traditional individual-parameter assessment approach, errors in track, intensity, and 34-kt wind radii in four quadrants (hereafter, asymmetric size) are used to evaluate the forecast results of the 54 hurricane cases. The results show that track forecast errors

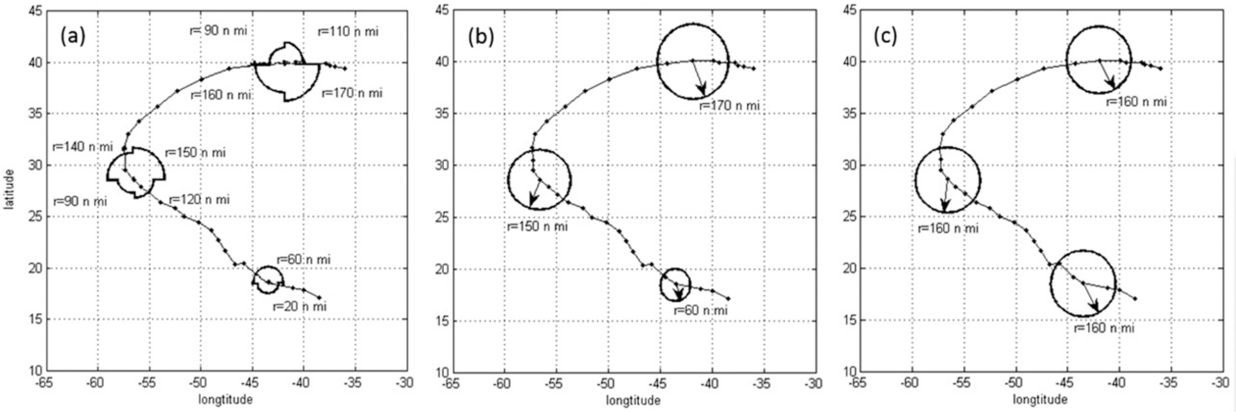

FIG. 4. Hurricane Edouard (2014) influence radius of (a) experiment 1, where S_x is set to 34-kt wind radii in four quadrants (i.e., northeast, southeast, southwest and northwest) surrounding the hurricane; (b) experiment 2, where S_x is set to the maximum extent of the 34-kt wind radius; and (c) experiment 3, where S_x is set to a constant 160 n mi.

TABLE 1. Pearson correlation coefficients *r* between IW-HTDF scores in three experiments and for each component error, i.e., track, intensity, asymmetric size (34-kt wind radii in four quadrants), and symmetric size (maximum 34-kt wind radii) error, as well as the product of the normalized errors. Values of *r* with $p < 0.05$ are shown in boldface. Here, $p < 0.05$ is the traditional indicator of statistical significance. The product of the normalized errors is calculated by multiplying normalized errors in track, intensity, and size by each other. (In experiment 3, only the track and intensity normalized errors are multiplied by each other to calculate the product of normalized errors.)

Coeff *r*	IW-HTDF score in expt 1	IW-HTDF score in expt 2	IW-HTDF score in expt 3
Track error	**−0.67**	**−0.64**	**−0.78**
Intensity error	**−0.27**	−0.26	−0.21
Asymmetric size error	−0.21	—	—
Symmetric size error	—	−0.23	—
Product of normalized errors	**−0.42**	**−0.38**	**−0.59**

range from 27.1 to 195.6 n mi with an average error of 74.4 n mi; intensity forecast errors range from 2.5 to 31.1 kt with a mean value of 13.9 kt, and errors in asymmetric size range from 48.6 to 389.0 n mi with an average of 117.2 n mi. IW-HTDF scores of the 54 hurricane cases have a range from 0.00 to 0.94, and the mean value is 0.59. Table 1 lists Pearson correlation coefficients *r* between IW-HTDF scores and each component error: track, intensity, and asymmetric size error, as well as the total product of three normalized errors. Note that errors in track, intensity, and size cannot be summed or averaged directly because of their differences in dimensions. Therefore, the product of the normalized errors in track, intensity, and size is used as the overall error of a hurricane forecast for comparison. Normalized error is defined as the error divided by the maximum error. Theoretically, a negative correlation exists between the skill score and the forecast error. The coefficients in Table 1 show that the IW-HTDF score and each component error as well as the product of the normalized errors are all negatively correlated ($r < 0$). Moreover, the IW-HTDF score and track error have a relatively close correlation ($r = -0.67$, $p < 0.05$) compared with intensity error ($r = -0.27$, $p < 0.05$) and size error ($r = -0.21$, $p > 0.05$).

To make a detailed comparison between the new IW-HTDF assessment approach and the traditional evaluation method, we divide 54 hurricane cases into four zones—A1, B1, C1, and D1—based on the average errors of the storm track and intensity, as shown in Fig. 5.

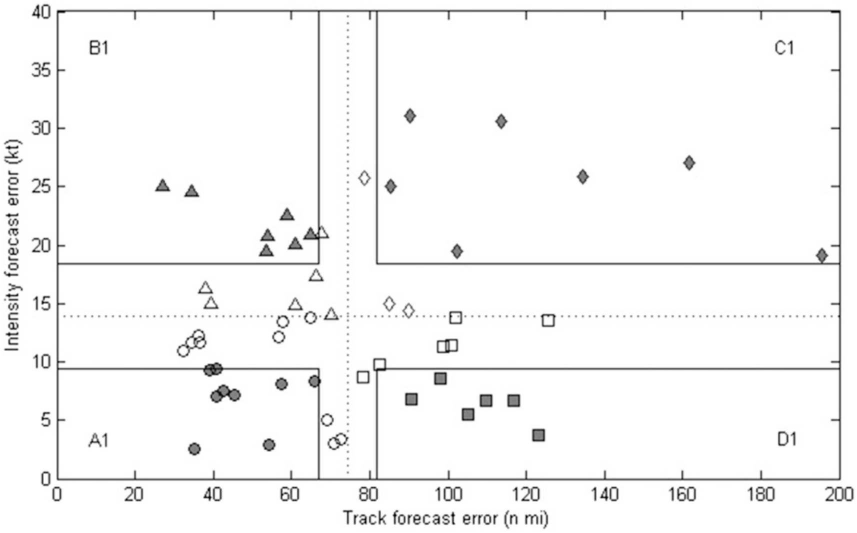

FIG. 5. Track forecast error and intensity forecast error. Circles show track and intensity forecast errors are both under the average errors; triangles show track forecast errors are under the average track error, but intensity forecast errors are above the average intensity error; diamonds show track and intensity forecast errors are both above the average errors; squares show track forecast errors are above the average track error, but intensity forecast errors are under the average intensity error. The total sample quantity is 54, and the number of analyzable samples (icons filled with gray) that are beyond the scope of the track and intensity error uncertainty is 29.

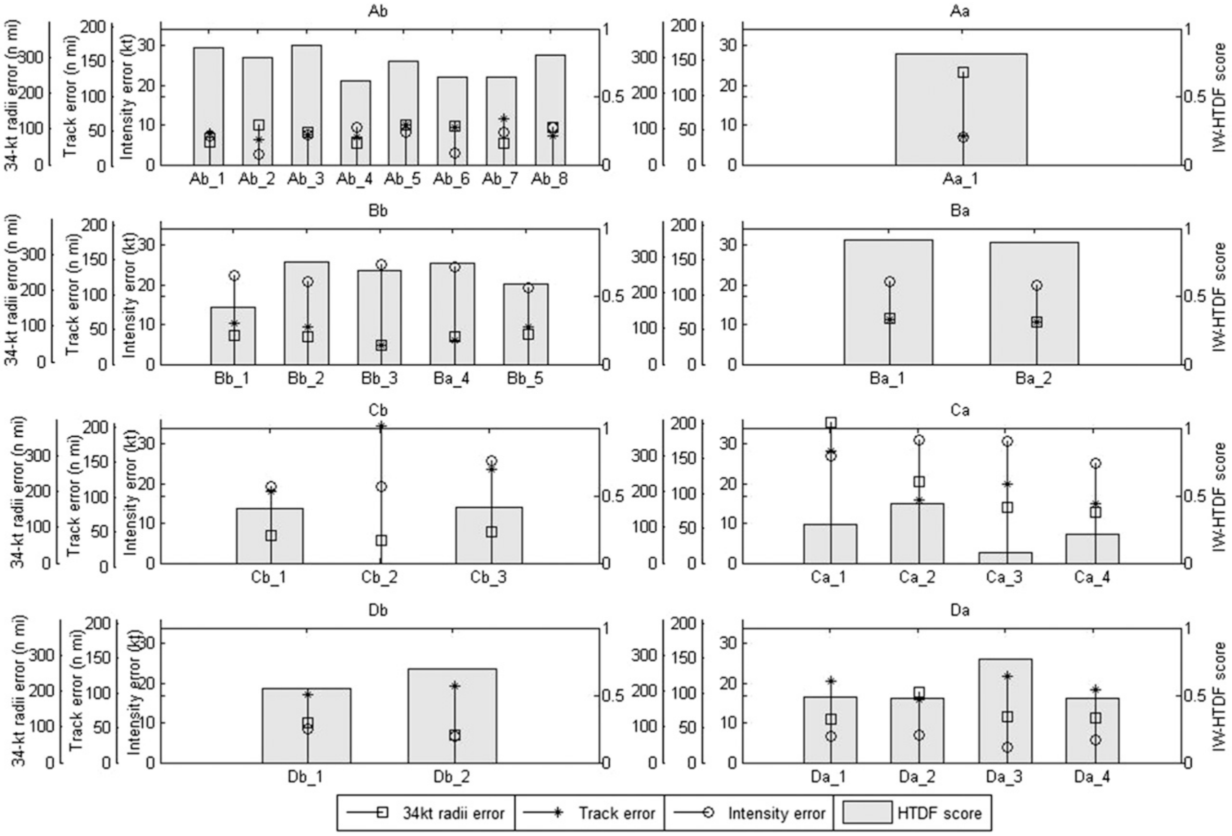

FIG. 6. Errors in track, intensity, and 34-kt wind radii maximum extent in four quadrants and the IW-HTDF scores of experiment 1 in the zones A_b, A_a, B_b, B_a, C_b, C_a, D_b, and D_a.

Meanwhile, because of the best-track uncertainty, which was estimated to be about 14.9 n mi for track and 9.1 kt for intensity (Landsea and Franklin 2013), hurricanes with forecast errors close to the average errors are not analyzed to avoid confusion. Therefore, 29 hurricane cases are obtained for analysis. Then, each zone (e.g., A1, B1, C1, D1) is divided into two parts based on the average error of the asymmetric sizes to form eight subzones: A_b, A_a, B_b, B_a, C_b, C_a, D_b, and D_a (subscript b or a represents the asymmetric size errors of hurricane cases are below or above the average value). Because of the limited number of available hurricane cases, hurricane cases whose size errors are close to the mean value are not removed. Finally, a total of 29 hurricanes that are beyond the scope of the track and intensity error uncertainty are analyzed in this study. Subzone A_b describes hurricanes whose forecast errors in track, intensity, and asymmetric size are all less than the average errors, meaning a good forecast zone. Subzone C_a, as opposed to A_b, is a poor forecast zone with each component error above the average error. The remaining subzones (i.e., A_a, B_b, B_a, C_b, D_b, and D_a) contain hurricanes whose forecast errors in track, intensity, and asymmetric size are mixed, depicting inconsistent forecast zones. For instance, subzone A_a shows only one hurricane (case A_a_1 in Fig. 6) whose track and intensity forecast errors (45.6 n mi, 7.1 kt) are both below the average errors (74.4 n mi, 13.9 kt), but the error of the asymmetric size (256.7 n mi) is above the average error (117.2 n mi).

Figure 6 presents the IW-HTDF assessment scores and the forecast errors in track, intensity, and asymmetric size of the 29 analyzed hurricane cases in the eight subzones. In the traditional good forecast zone A_b, the hurricane cases' IW-HTDF scores are all above the average IW-HTDF score (0.59). For example, case A_b_1 (Hurricane Jeanne, 2004), obtains a relatively high IW-HTDF score of 0.86, and its forecast errors in track, intensity, and asymmetric size are 45.4 n mi (<74.4 n mi), 7.1 kt (<13.9 kt), and 61.8 n mi (<117.2 n mi), respectively. On the contrary, in the traditional poor forecast zone C_a, all the hurricane IW-HTDF scores are below the average IW-HTDF score. For instance, case C_a_3 (Hurricane Epsilon, 2005) earns a low IW-HTDF score of 0.07 with forecast errors of 113.7 n mi (track), 30.6 kt (intensity), and

TABLE 2. Average track forecast error, average intensity forecast error, average asymmetric size (34-kt wind radii in four quadrants) error, and average IW-HTDF score in experiment 1 of hurricanes located in zones A_b, A_a, B_b, B_a, C_b, C_a, D_b, and D_a. The ranges of these parameters are given within parentheses.

	Zone A_b	Zone A_a	Zone B_b	Zone B_a
Track error (n mi)	47.6 (35.3–65.9)	40.8	45.6 (27.1–58.8)	62.9 (60.8–65.0)
Intensity error (kt)	6.9 (2.5–9.4)	7.1	22.4 (19.4–25.0)	20.4 (20.0–20.8)
Asymmetric size error (n mi)	88.2 (58.3–111.3)	256.7	74.4 (52.5–84.4)	124.3 (118.9–129.6)
IW-HTDF score	0.75 (0.62–0.88)	0.82	0.64 (0.42–0.75)	0.90 (0.89–0.91)
	Zone C_b	Zone C_a	Zone D_b	Zone D_a
Track error (n mi)	144.2 (102.4–195.6)	112.7 (85.4–161.5)	103.9 (98.2–109.6)	108.9 (90.5–123.1)
Intensity error (kt)	21.5 (19.2–25.8)	28.4 (25.0–31.1)	7.6 (6.7–8.6)	5.7 (3.8–6.8)
Asymmetric size error (n mi)	75.2 (61.7–88.3)	227.8 (141.2–389.0)	94.5 (76.7–112.3)	142.8 (122.5–196.8)
IW-HTDF score	0.27 (0–0.41)	0.25 (0.07–0.44)	0.62 (0.55–0.69)	0.55 (0.47–0.76)

153.8 n mi (asymmetric size). In the inconsistent zones B_b, C_b, D_b, and D_a, hurricane cases receive in-between IW-HTDF scores. For example, case D_b_1 (Hurricane Ivan, 2004) earns a 0.55 IW-HTDF score, with forecast errors of 98.2 n mi (track), 8.6 kt (intensity), and 112.3 n mi (asymmetric size). Meanwhile, hurricane cases in traditional inconsistent zones A_a and B_a have different results, earning relatively high IW-HTDF scores like those located in good forecast zone A_b. The reason for this will be discussed in section 4. In the subzone C_b, case C_b_2 (Hurricane Omar, 2008) obtains a zero score, mainly because of the very high track error (195.6 n mi), which results in only a small overlap between the observed and forecast integrated IW-HTDF fields (overlap rate 35.5%, which is defined as the percentage of overlap grid points between the observed and forecast IW-HTDF fields divided by the total points) and a value of 0 for the S term in SSIM.

On average (Table 2), hurricanes located in zone A_b with smaller forecast errors (i.e., errors in track, intensity, and asymmetric size) earn a high mean IW-HTDF score of 0.75. The hurricanes located in zone C_a with larger forecast errors obtain a low mean IW-HTDF score of 0.25. Additionally, hurricanes in the inconsistent forecast zones B_b, C_b, D_b, and D_a, earn moderate IW-HTDF scores of 0.64, 0.27, 0.62, and 0.55, respectively. Hurricanes in the inconsistent forecast zones A_a and B_a obtain relatively high mean IW-HTDF scores of 0.82 and 0.90, respectively.

b. *Assessing track, intensity, and symmetric size*

In some practical applications, the maximum 34-kt wind radius is appropriate for describing the hurricane size. So, in the second experiment, S_x is set to the maximum 34-kt wind radius (hereafter, symmetric size) which is symmetric and changes over time, as shown in Fig. 4b. Symmetric size errors of the 54 cases selected in this study range from 0 to 125 n mi, and the mean value is 32.6 n mi. IW-HTDF scores range from 0.12 to 0.95, the mean value is 0.67. The Pearson correlation coefficients between the IW-HTDF score and each component error are all negative (shown in Table 1), similar to results obtained in experiment 1. The IW-HTDF score and track error have a relatively close correlation compared with the intensity and size error ($r = -0.64$, $p < 0.05$), and the total product of the normalized error and IW-HTDF score is significantly correlated with $r = -0.38$ and $p < 0.05$.

The 29 hurricane cases extracted during the first experiment are also divided into eight subzones (i.e., A_b, A_a, B_b, B_a, C_b, C_a, D_b, and D_a) by the mean error of the symmetric size. Figure 7 presents the IW-HTDF assessment scores, forecast errors in track, intensity, and symmetric size of these analyzed hurricane cases in the eight subzones. In the traditional good forecast zone A_b, all the hurricane cases obtain relatively high IW-HTDF scores. For example, the case A_b_6 (Hurricane Edouard, 2014) IW-HTDF score is 0.89, and its forecast errors are 40.8 n mi (track), 9.4 kt (intensity), and 23.1 n mi (symmetric size). In contrast, in the traditional poor forecast zone C_a all the hurricane IW-HTDF scores are relatively low. For instance, case C_a_1 (Hurricane Alex, 2004) receives a low IW-HTDF score of 0.35, with forecast errors of 161.5 n mi (track), 27.0 kt (intensity), and 125.0 n mi (symmetric size). In the inconsistent zones A_a, B_b, C_b, D_b, and D_a, hurricane cases receive moderate IW-HTDF scores. Such as case B_b_5 (Hurricane Danny, 2015) whose IW-HTDF score is 0.68, and the forecast errors in track, intensity, and symmetric size are 53.7 n mi, 19.4 kt, and 20.0 n mi, respectively. On the other hand, hurricane cases in inconsistent zone B_a have different results, earning relative high IW-HTDF scores like those located in good forecast zone A_b. This inconsistent result was also obtained in the first

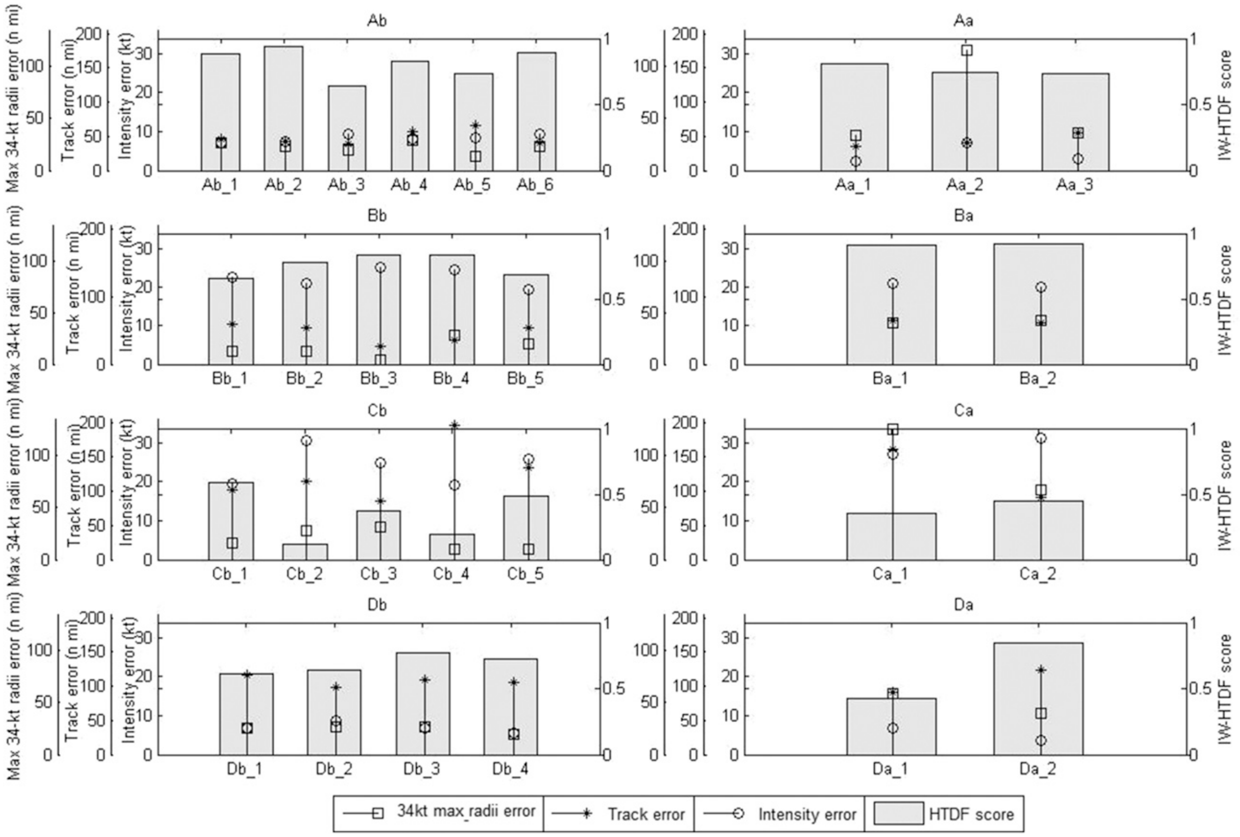

FIG. 7. Errors in track, intensity, and maximum 34-kt wind radii and the IW-HTDF scores of experiment 2 in the zones A_b, A_a, B_b, B_a, C_b, C_a, D_b, and D_a.

experiment. The reason is the same and will be discussed in section 4.

On average (Table 3), hurricanes located in zone A_b with smaller forecast errors in track, intensity, and symmetric size earn a high mean IW-HTDF score of 0.82. The hurricanes located in zone C_a with larger forecast errors obtain a low mean IW-HTDF score of 0.40. Hurricanes in the inconsistent forecast zones A_a, B_b, D_b, and D_a earn in-between mean IW-HTDF scores: 0.76, 0.76, 0.68, and 0.63, respectively. Hurricanes in the inconsistent forecast zone B_a (C_b) receive a relatively high (low) mean IW-HTDF score of 0.91 (0.35).

c. *Assessing track and intensity only*

It should be noted that actual observations and best-track data of TC size are very limited, having very large room for improvement (Landsea and Franklin 2013), and so are the estimations of TC size forecasts. Therefore, including TC size as a forecast evaluation parameter is exploratory at the present. The proposed

TABLE 3. As in Table 2, but for experiment 2.

	Zone A_b	Zone A_a	Zone B_b	Zone B_a
Track error (n mi)	48.5 (39.2–65.9)	43.5 (35.3–54.2)	45.6 (27.1–58.8)	62.9 (60.8–65.0)
Intensity error (kt)	8.3 (7.1–9.4)	4.1 (2.5–7.1)	22.4 (19.4–25.0)	20.4 (20.0–20.8)
Symmetric size error (n mi)	22.3 (13.3–28.8)	61.5 (33.8–115.0)	15.4 (3.8–28.0)	41.6 (40.6–42.6)
IW-HTDF score	0.82 (0.64–0.94)	0.76 (0.74–0.80)	0.76 (0.66–0.83)	0.91 (0.91–0.91)
	Zone C_b	Zone C_a	Zone D_b	Zone D_a
Track error (n mi)	126.3 (85.4–195.6)	125.9 (90.4–161.5)	107.4 (98.2–116.9)	106.8 (90.5–123.1)
Intensity error (kt)	24.0 (19.2–30.6)	29.1 (27–31.1)	6.9 (5.6–8.6)	5.3 (3.8–6.8)
Symmetric size error (n mi)	18.9 (10.0–31.3)	96.1 (67.2–125.0)	24.4 (20–26.7)	49.5 (40.0–59.1)
IW-HTDF score	0.35 (0.12–0.59)	0.40(0.35–0.44)	0.68 (0.61–0.77)	0.63 (0.42–0.84)

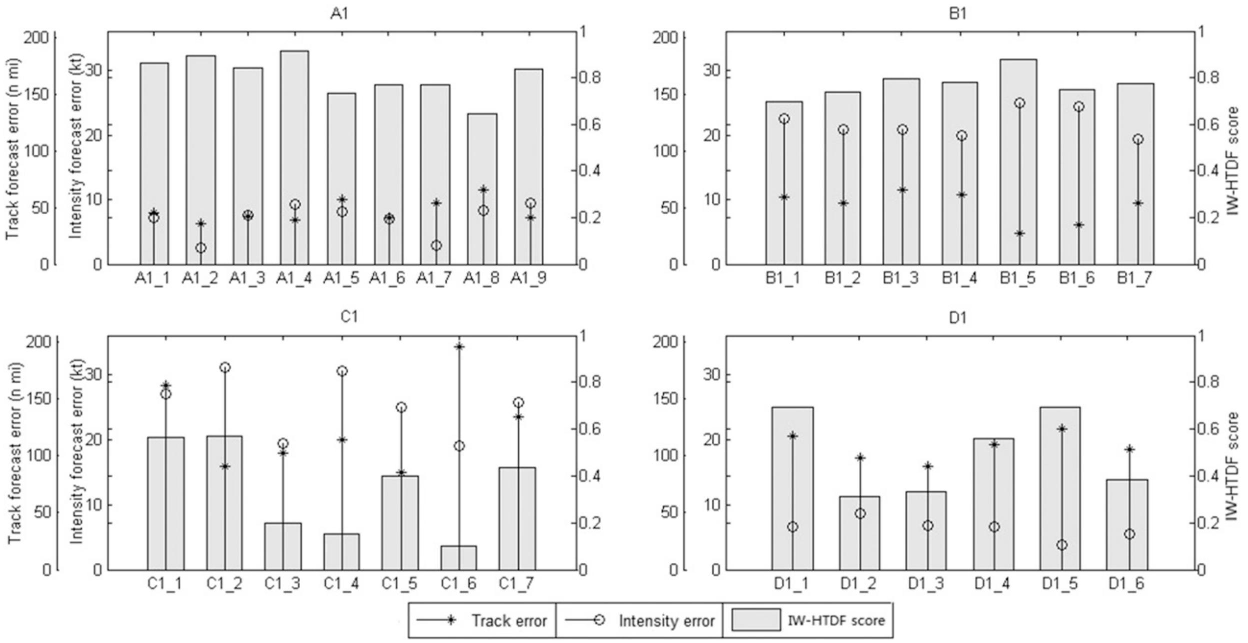

FIG. 8. Errors in track, intensity, and IW-HTDF scores of experiment 3 in the zones A1, B1, C1, and D1.

integrated IW-HTDF can be reduced for track forecast separately or in combination with intensity and size parameters. In the third experiment, we assess track and intensity forecasts without dealing with size forecasts by setting S_x to a constant of 160 n mi.

As shown in Fig. 5, 29 analyzed hurricane cases are divided into four zones—A1, B1, C1, and D1—based on the average errors of track and intensity. Zone A1 describes hurricanes whose track and intensity forecast errors are both below the average errors, meaning it is a good forecast zone. Conversely, zone C1 describes hurricanes whose forecast errors of track and intensity are both above the average errors, meaning this is a poor forecast zone. Zone B1 (D1) describes hurricanes whose track forecast errors are below (above) the average track error but whose intensity forecast errors are above (below) the average intensity error, meaning this is an inconsistent forecast zone.

To further examine the forecast skill for each individual hurricane, the IW-HTDF scores are plotted for all 29 hurricanes (Fig. 8). The IW-HTDF scores in the third experiment range from 0.10 to 0.91, and the mean value is 0.65. Pearson correlation coefficients (Table 1) show that IW-HTDF score and track error have a relatively close correlation compared with intensity error ($r = -0.78$, $p < 0.05$). And the total product of the normalized errors (in track and intensity) and the IW-HTDF score is significantly correlated with $r = -0.59$ and $p < 0.05$. In the traditional good forecast zone A1, all the hurricane cases obtain relatively high IW-HTDF scores. For example, case A1_4 (Hurricane Michael, 2012) receives the highest IW-HTDF score of 0.91 with forecast errors of 39.2 n mi (track) and 9.3 kt (intensity). All of the seven hurricanes located in zone C1 (the poor forecast zone) receive lower IW-HTDF scores than the mean IW-HTDF score. For instance, case C1_3 (Hurricane Rita, 2005) earns a low IW-HTDF score of 0.20 with a track forecast error of 102.4 n mi, and an intensity forecast error of 19.4 kt. In the inconsistent zones B1 and D1, hurricane cases receive in-between IW-HTDF scores. However, hurricanes located in zone B1 obtain relatively higher IW-HTDF scores than those located in zone D1. For example, case B1_5 (Hurricane Rina, 2011) located in zone B1 earns an IW-HTDF score of 0.87, with a track forecast error of 27.1n mi and an intensity forecast error of 25.0 kt, while D1_6 (Hurricane Cristobal, 2014) located in zone D1 obtains a low IW-HTDF score 0.38 and its forecast errors in track and intensity are 105.1 n mi and 5.5 kt. The reason will be discussed in section 4.

On average (Table 4), hurricanes located in good forecast zone A1 obtain a high mean IW-HTDF score of 0.81, while hurricanes located in poor forecast zone C1 obtain a low mean IW-HTDF score of 0.34. The mean IW-HTDF scores of hurricanes located in inconsistent zones B1 and D1 are 0.77 and 0.50, respectively. These results are consistent with the expectation that when both track and intensity forecast errors are low (high), the IW-HTDF scores are high (low), and when the track

TABLE 4. Average track forecast error, average intensity forecast error, and average IW-HTDF score in experiment 3 of hurricanes located in zones A1, B1, C1, and D1. The ranges of these parameters are given within parentheses.

	Zone A1	Zone B1	Zone C1	Zone D1
Track error (n mi)	46.8 (35.3–65.94)	50.5 (27.1–65.0)	126.2 (85.4–195.6)	107.2 (90.5–123.0)
Intensity error (kt)	6.9 (2.5–9.4)	21.9 (19.4–25.0)	25.5 (19.2–31.1)	6.3 (3.8–8.6)
IW-HTDF score	0.81 (0.65–0.91)	0.77 (0.69–0.87)	0.34 (0.10–0.57)	0.50 (0.31–0.69)

and intensity forecast errors are inconsistent, the IW-HTDF scores are in between.

In addition to forecast errors, forecast skill is an alternative measure used widely to assess hurricane forecasts. Figure 9 presents the track forecast skill relative to CLIPER5 and intensity forecast skill relative to Decay-SHIFOR5 of the official NHC forecast for the 29 analyzed hurricane cases. Track forecast skill levels range from −11.4% to 89.0%, and the average skill is 60.0%. Intensity forecast skill levels range from −132.6% to 78.7%, and the mean value is 15.6%. Figure 9 is divided into four zones (A2, B2, C2, and D2) based on the average track skill and average intensity skill. Zone A2 describes hurricane cases whose track and intensity forecast skills are both above the average skill. On the contrary, zone C2 includes cases when track and intensity forecast skill levels are both below the average. Zone B2 (D2) describes hurricane cases whose track forecast skill levels are above (below) the average skill, but intensity forecast skill levels are below (above) the average.

Figure 10 shows the IW-HTDF scores of the 29 analyzed hurricane cases located in the four zones: A2, B2, C2, and D2. On average (Table 5), hurricane cases located in zone A2 have a high mean IW-HTDF score of 0.79, while hurricanes located in zone C2 have a low mean IW-HTDF score of 0.47. And the mean IW-HTDF scores of hurricane cases located in zone B2 and D2 are 0.60 and 0.49, respectively. There are 11 hurricane cases located in zone A2 (the good forecast zone), producing higher IW-HTDF scores than the mean IW-HTDF score. Case A2_4 (Hurricane Michael,

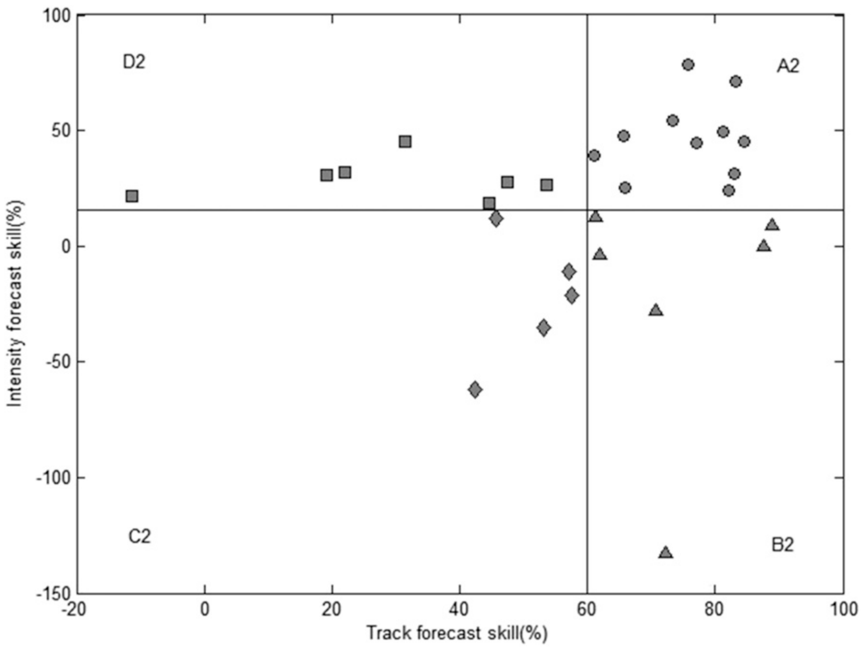

FIG. 9. Track forecast skill relative to CLIPER5 and intensity forecast skill relative to Decay-SHIFOR5. Circles show track and intensity forecast skill levels are both above the average; triangles show track forecast skill levels are above the average track skill, but intensity forecast skill levels are below the average intensity skill; diamonds show that the track and intensity forecast skill levels are both below the average skill; and squares show track forecast skill levels are below the average track skill, but intensity forecast skill levels are above the average intensity skill.

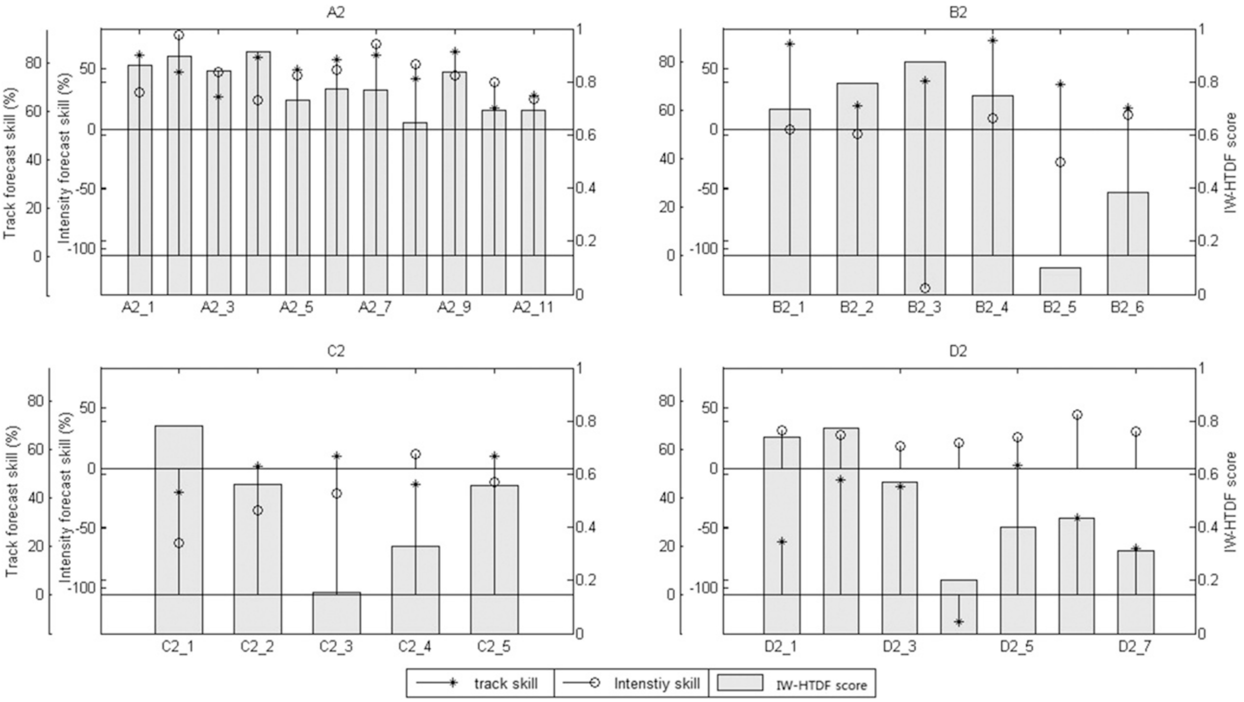

FIG. 10. Track and intensity forecast skill levels and IW-HTDF scores in the zones A2, B2, C2, and D2.

2012) earns the highest IW-HTDF score of 0.91, with a track forecast skill of 82.1% and an intensity forecast skill of 24.2%. In zone C2 (the poor forecast zone), hurricane cases obtain relatively low IW-HTDF scores. For example, zone C2_3 (Hurricane Epsilon, 2005) earns a low IW-HTDF score of 0.15, with forecast skill levels of 57.5% (track) and −21.3% (intensity). There are four hurricanes located in zone B2, which means better track forecast skill and poorer intensity forecast skill. Overall, cases in B2 received relatively high IW-HTDF scores, such as B2_3 (Hurricane Rina, 2011), whose IW-HTDF score is 0.87, and forecast skill levels in track and intensity are 72.2%, −132.6%, respectively. While the other two hurricanes obtain relatively lower IW-HTDF scores, such as case B2_5 (Hurricane Omar, 2008), whose IW-HTDF score is 0.10, the track forecast skill is 70.66% and the intensity forecast skill is −27.7%. In zone D2, which suggests poorer track forecast skill and better intensity forecast skill, five out of seven hurricanes earn relatively low IW-HTDF scores compared with the mean IW-HTDF score. For example, D2_7 (Hurricane Ivan, 2004) receives a low IW-HTDF score of 0.31 with a track forecast skill of 19.1% and an intensity forecast skill of 30.6%.

Analogous to the traditional assessment method performed by assessing track and intensity forecast skill levels, IW-HTDF skill scores of most hurricanes located in zone A2 indicate that those hurricane forecasts are ‘‘good,’’ and IW-HTDF scores in zone C2 indicate that those hurricane forecasts are ‘‘poor.’’ Meanwhile, hurricane cases in zones B2 and D2, where individual track and intensity forecasts contradict each other, receive mixed IT-HTDF scores between those in zones A2 and C2.

d. *Assessing OFCL forecasts at different forecast hours*

In this study, the IW-HTDF scores of OFCL forecasts at different forecast times (i.e., 12, 24, 36, 48, 72, 96, and 120 h) are calculated to test the feasibility of the

TABLE 5. As in Table 4, but for zones A2, B2, C2, and D2.

	Zone A2	Zone B2	Zone C2	Zone D2
Track forecast skill (%)	75.7 (61.1–84.6)	73.7 (61.2–89.0)	51.2 (42.4–57.5)	29.6 (from −11.4 to 53.5)
Intensity forecast skill (%)	46.4 (24.2–78.7)	−23.8 (from −132.6 to 12.3)	−23.5 (from −62.0 to 11.8)	28.9 (18.8–45.0)
IW-HTDF score	0.79 (0.65–0.91)	0.60 (0.10–0.87)	0.47 (0.15–0.78)	0.49 (0.20–0.77)

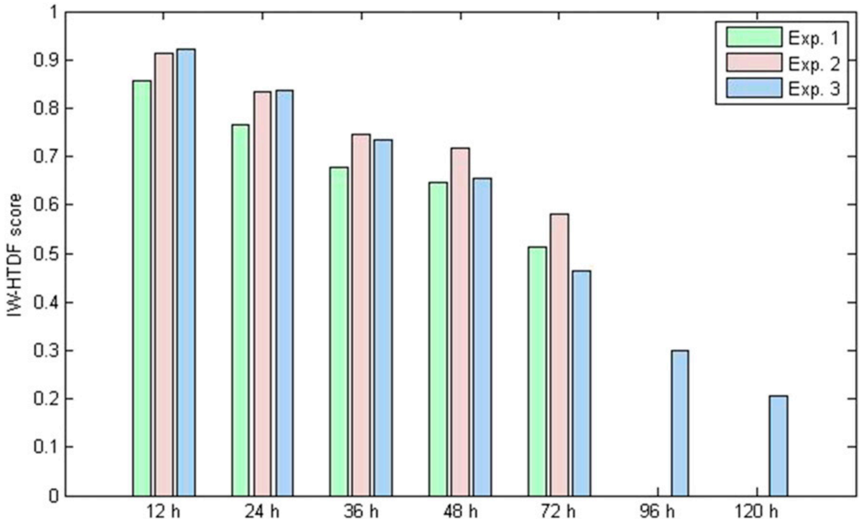

FIG. 11. Mean IW-HTDF score of OFCL forecasts at different forecast time: 12, 24, 36, 48, 72, 96, and 120h. Green bars show the mean IW-HTDF score in experiment 1 setting S_x = asymmetric size, red bars show the mean IW-HTDF score in experiment 2 setting S_x = symmetric size, and blue bars show the mean IW-HTDF score in experiment 3 setting S_x = constant. (The 96- and 120-h OFCL forecasts of wind radii are unavailable, so the IW-HTDF scores in experiment 1 and experiment 2 are not calculated.)

IW-HTDF assessment approach in evaluating the performance of the TC numerical forecast models. Figure 11 shows the mean IW-HTDF scores of OFCL forecasts at each forecast time: 12, 24, 36, 48, 72, 96, and 120 h. In each experiment, 12-h OFCL forecasts earn the highest mean IW-HTDF score, and the mean IW-HTDF score decreases with increasing forecast time as a result of the increasing forecast errors in track, intensity, and size (shown in Fig. 12). These results are consistent with the expectation that the performance of numerical forecast models degrades over time. The difference in the IW-HTDF scores between experiments 1 and 2 versus experiment 3 increases with forecast time, which is somehow related to the influence of size. In experiment 3, only track and intensity are considered, while in experiments 1 and 2, in addition to track and intensity, the size is also regarded as a factor on the IW-HTDF score. On one hand, the additional size error makes a contribution to the lower IW-HTDF scores in experiments 1 and 2 than that in experiment 3 (at forecast times of 12, 24, and 36 h in experiment 1; 12 and 24 h in experiment 2). On the other hand, the size and the overlap between the observations and the forecast have a positive influence on the IW-HTDF score (discussed in section 4). Compared with the constant S_x of 160 n mi in experiment 1, the average maximum 34-kt wind radius increases from 145 to 172 n mi with forecast time. The differences in the overlap rate between experiments 1 and 2 versus experiment 3 increase from −9.6% to 0.1%, and from −6.0% to 1.3%, respectively. Therefore, the differences in the IW-HTDF score between experiments 1 and 2 versus experiment 3 increase from negative to positive.

4. Discussion

In this study, an integrated TC track, intensity, and size forecast evaluation parameter (IW-HTDF) has been designed using three different forecast variable setups. To demonstrate this new approach, we used the IW-HTDF score (SSIM index) to assess OFCL hurricane forecasts and compared the results with those of the traditional individual-parameter evaluation method.

The results show that the IW-HTDF assessment approach is feasible and has some merit over the traditional individual-parameter assessment approach. Currently, TC forecast errors, such as errors in track and intensity, are used widely to assess TC forecast models or methods, which can provide mixed evaluation results, for instance, small track error with large intensity error and vice versa. In this situation, some users (e.g., modelers interested in determining a model's performance or untrained users) may feel confused, and a concise and integrated assessment score can be more useful. For example, Hurricane Emily's (2005) track forecast error (53.7 n mi) is below the average error (74.4 n mi) and the intensity forecast error (20.8 kt) is above the average error (13.9 kt). The

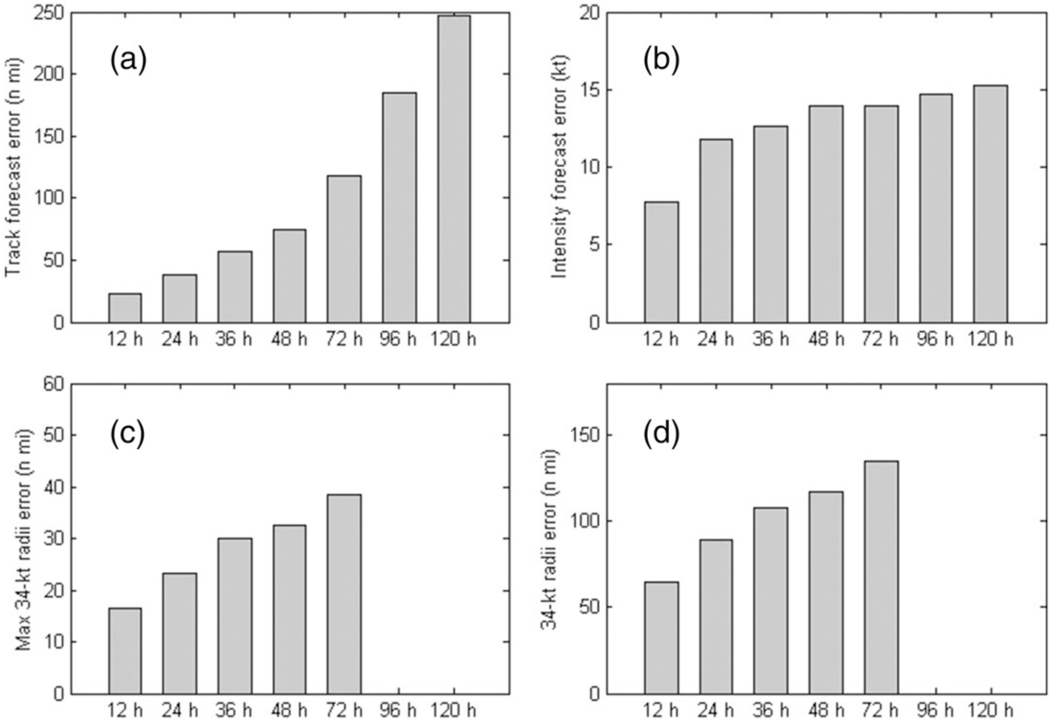

FIG. 12. (a) OFCL mean track forecast error at different forecast times. (b) OFCL mean intensity forecast error at different forecast hours. (c) OFCL mean maximum 34-kt wind radii error at different forecast hours. (d) OFCL mean errors of 34-kt wind radii maximum extent in four quadrants at different forecast hours. The 96- and 120-h OFCL forecasts of wind radii are unavailable.

forecast errors indicate that Hurricane Emily's track forecast is ''good'' but the intensity forecast is ''poor.'' Meanwhile Hurricane Emily's IW-HTDF score of 0.74 (in experiment 3) is higher than the mean IW-HTDF score of 0.65, which suggests an integrated evaluation that the forecast is good.

There are a number of issues that are worth discussing with regard to the IW-HTDF approach. First, it is necessary to study the influence of S_x and the overlap (between the IW-HTDF fields of observations and forecasts) on the IW-HTDF score. In experiments 1 and 2, the IW-HTDF approach seems appealing in traditional inconsistent zones. For example, hurricanes located in zones $\mathrm{B_a}$ (in experiments 1 and 2) earn relatively high IW-HTDF scores like those located in the good forecast zone $\mathrm{A_b}$, which is a result of the monotonicity of the IW-HTDF score with S_x. As shown in Fig. 13, IW-HTDF score increases as S_x increases. On the other hand, the overlap between the IW-HTDF fields of the observations and the forecast has a significantly positive effect on the IW-HTDF score (Fig. 14). The Pearson correlation coefficients r between the IW-HTDF score and the overlap rate in experiments 1, 2, and 3 are 0.83, 0.82, and 0.83, respectively, and are all significant with $p < 0.001$, which indicates that these two indices (i.e., IW-HTDF score and overlap rate) are closely correlated. Hurricanes in zone $\mathrm{B_a}$ (in experiments 1 and 2) have larger sizes (the mean maximum 34-kt radii are 240.5 and 240 n mi, respectively) and higher overlap rates (77.1% and 79.7%) than those in good forecast zone $\mathrm{A_b}$ (for experiment 1, the mean

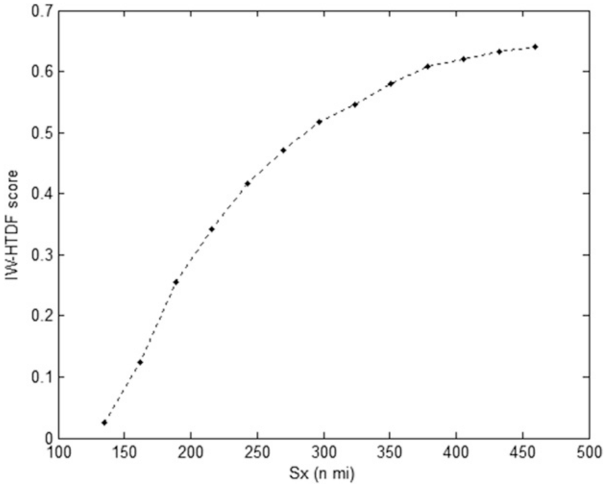

FIG. 13. IW-HTDF score of Hurricane Rita (2005) with different values of S_x.

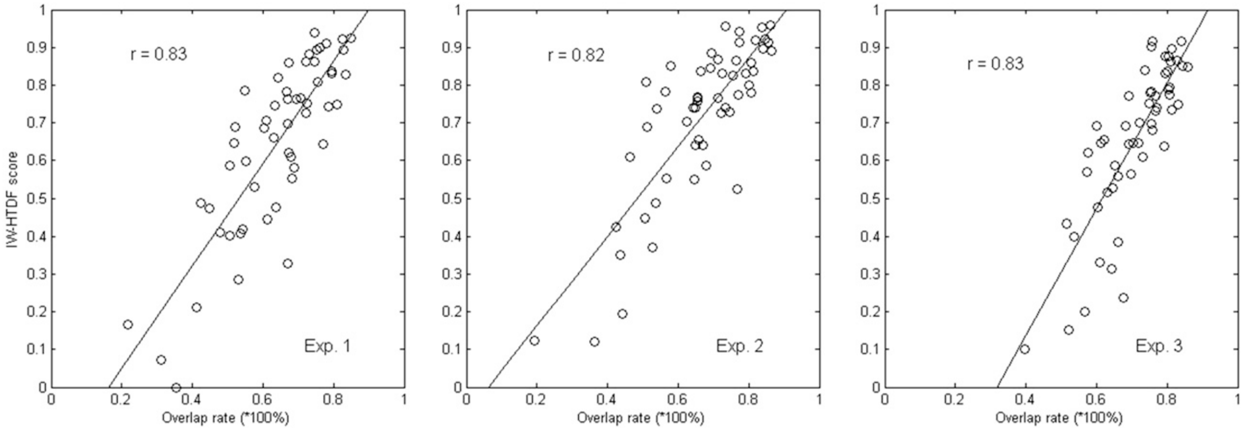

FIG. 14. IW-HTDF scores vs rates between the IW-HTDF fields of observations and forecasts in experiments 1–3 with different S_x settings, i.e., asymmetric, symmetric and constant sizes.

maximum 34-kt radius is 143 n mi and the overlap rate is 68.1%; for experiment 2, the mean maximum 34-kt radius is 147 n mi and the overlap rate is 74.1%). In these cases, large S_x with large overlap plays a more significant role in determining the IW-HTDF score than the errors of track and intensity. Therefore, these hurricanes obtain relatively high scores.

Second, the current IW-HTDF score algorithm [for simplicity, we set $\alpha = \beta = \gamma = 1$ in Eq. (14) in this study] is more sensitive to track than intensity. As shown in Table 1, the absolute values of the correlation coefficients between the IW-HTDF scores and track errors are all higher (in three experiments) than those between the IW-HTDF scores and the intensity errors. The IW-HTDF score increases quickly with the decrease in track error, but increases slowly with the decrease in intensity error. For example, as shown in Fig. 15 for Hurricane Rita (2005), if its track forecast error improves by 50%, the IW-HTDF score will increase by 244%. On the other hand, if its intensity forecast error improves by 50%, the IW-HTDF score will only increase by 1.8%. So, in the inconsistent zones B1 and D1 (in experiment 3), though hurricane cases receive moderate IW-HTDF scores, hurricanes located in zone B1 obtain relatively higher IW-HTDF scores than those located in zone D1. Table 6 lists Pearson correlation coefficients between each SSIM component [i.e., luminance L, contrast C, and structure S] and each traditional forecast error (i.e., track, intensity, and size error) in three experiments. The S term is significantly correlated to the track error ($r = -0.68$, -0.63, and -0.75, respectively in experiments 1, 2 and 3; $p < 0.05$). If we consider the influence of size in experiments 1 and 2, the L term has a close correlation with the size error ($r = -0.62$ and -0.66,

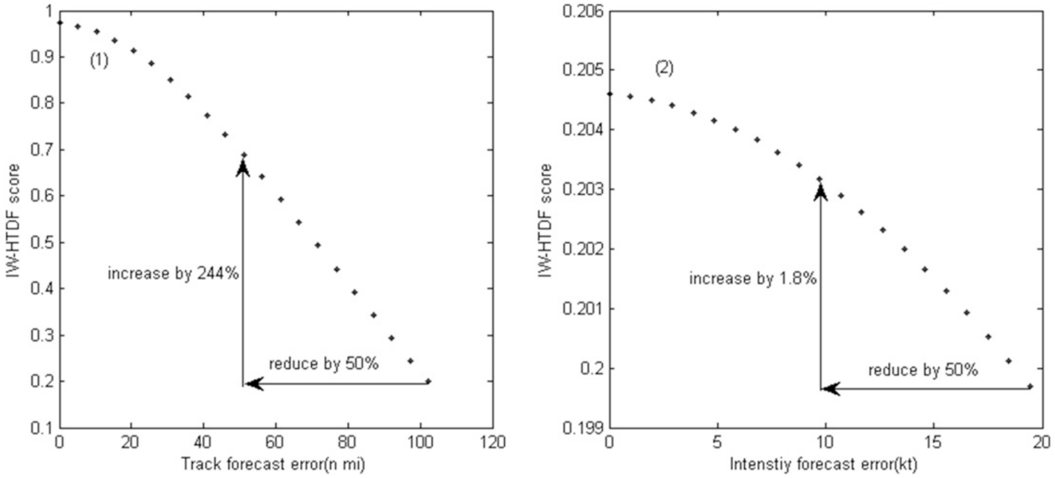

FIG. 15. Variations of IW-HTDF scores with the improvements in track and intensity simulation errors of Hurricane Rita (2005) during experiment 3.

TABLE 6. Pearson correlation coefficients r between each SSIM component, i.e., L, C, and S, and each component error, i.e., track, intensity, asymmetric size (34-kt wind radii in four quadrants), and symmetric size (maximum 34-kt wind radii) error. Values of r with $p < 0.05$ are shown in boldface.

	L	C	S
Expt 1			
Track error	−0.15	−0.24	**−0.68**
Intensity error	0.02	**−0.38**	−0.27
Asymmetric size error	**−0.62**	**−0.49**	−0.007
Expt 2			
Track error	−0.15	−0.31	**−0.63**
Intensity error	0.026	**−0.33**	−0.24
Symmetric size error	**−0.66**	**−0.48**	0.12
Expt 3			
Track error	−0.32	−0.32	**−0.75**
Intensity error	**−0.51**	**−0.51**	−0.14

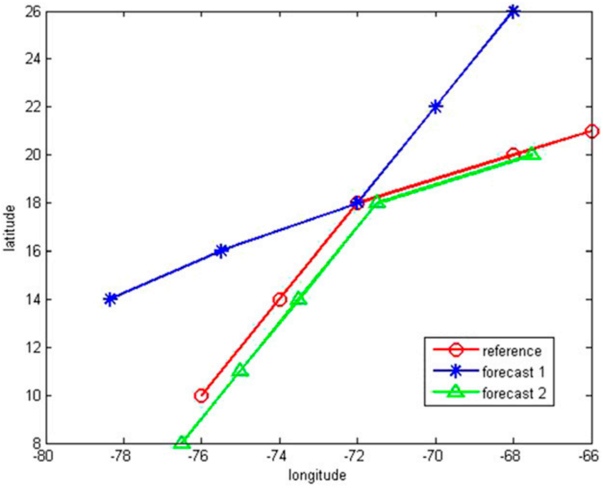

FIG. 16. Two forecast cases with the same track forecast error but different trajectories.

$p < 0.05$); and the term C is mainly affected by the intensity and size. In experiment 3 without dealing with size, the intensity error has a significant correlation with terms L and C ($r = -0.51$ and -0.51, $p < 0.05$), while track error is much more closely correlated to the term S ($r = -0.75$, $p < 0.05$). Therefore, it is possible to adjust the weightings of the track, intensity, and in some cases size by setting α, β, and γ in Eq. (14) to different values.

We propose the IW-HTDF approach for assessing TC forecasts by combining TC track, intensity, and size data into a single parameter. Following the same idea, other approaches may emerge. For instance, one could construct a score function that combines forecast errors in track, intensity, and size after assessing the forecast error of each parameter separately. A simple weighted combination of the track, intensity, and size error can be designed as

$$\mathrm{err}_W = F(\mathrm{err}_T, \mathrm{err}_I, \mathrm{err}_S), \tag{15}$$

where err_T, err_I, and err_S are the errors in track, intensity, and size, respectively, and err_W is the weighted-combined error of these three errors by some transformation function F. It should be noted that errors in track, intensity, and size have different dimensions (i.e., L, L/T, L, where L is length and T is time), which cannot be summed or averaged directly. So, a simple form of the transformation function F can be some form of multiplication. Meanwhile, the forecast error is a one-dimensional parameter and cannot provide the verification information of the TC trajectory forecast (track shape or moving direction). In practical applications, an accurate trajectory forecast is extremely important because of its influence on ship rerouting, hurricane landfall location, etc. If two TC forecasts have identical errors in track, intensity, and size, separately, while their trajectory forecasts are different, for example, one is similar to the observation, whereas the other is opposite the observation, then the simple weighted-combined error assessment would fail to distinguish these two forecasts from each other with the same calculated result through Eq. (15). Intuitively, the forecast with a similar trajectory to the observation is better than the other. The integrated IW-HTDF is a two-dimensional field, which is derived from the track, intensity, and size directly rather than from the errors, providing the information of the TC track pattern as well as the storm intensity and size. Therefore, the IW-HTDF assessment approach can provide more comprehensive verification information containing trajectory that is not considered in a simple weighted combination of track, intensity, and size errors. For a specific case, as shown in Fig. 16, there are two forecasts with the same track error of 180 n mi, but their trajectories are different. The trajectory of forecast 1 is opposite to the reference track, while the trajectory of forecast 2 is similar to the reference track. To assess these two forecasts via the IW-HTDF approach, forecast 1 obtains an IW-HTDF score of 0.07, whereas forecast 2 earns an IW-HTDF score of 0.65. Obviously, the IW-HTDF approach shows that forecast 2 performs better than forecast 1, which is consistent with the expectation that when the trajectory forecast is more similar to the reference track, the track forecast is better, while a distance error-based assessment approach like that in Eq. (15) does not distinguish between the cases.

In our study, we choose SSIM (Wang et al. 2004) to compare the two-dimensional IW-HTDF fields between the forecasts and the observations. The previous literature has explored the subject of quantitative spatial verification (e.g., Gilbert 1884; Ebert and McBride 2000; Casati et al. 2004; Davis et al. 2006; Roberts and Lean 2008). The Gilbert skill score (Gilbert 1884) is a function of counts of forecast–observation (yes–no) pairs, which is a typical verification approach; however, it only focuses on the spatial accuracy of the forecasts, removing the impact of any bias in the intensity forecasts, such as rainfall amounts. Ebert and McBride (2000) introduced an objected-oriented verification procedure for gridded quantitative precipitation forecasts. They decomposed the total mean squared error between the observed and forecast fields into components as a result of 1) location, 2) rain volume, and 3) pattern. Casati et al. (2004) developed an intensity-scale approach for the verification of spatial precipitation forecasts. The technique allows the skill to be diagnosed as a function of the scale of the forecast error and the intensity of the precipitation events. The mean-square error (MSE) skill score is obtained from binary images of observed and forecasts fields and is equivalent to the Heidke skill score or Peirce skill. Davis et al. (2006) produced an object-based verification methodology, which is complementary to Ebert and McBride's approach. The method first targeted rain areas by performing a convolution and thresholding operation. Once rain areas were identified, their attributes, including centroid location, size, orientation, curvature, and intensity distribution, were computed. They performed a statistical comparison of these attributes of the forecasts and the observations. Roberts and Lean (2008) introduced a scale-selective verification method for examining whether improved model resolution alone is able to produce more skillful precipitation forecasts on useful scales. This method is based on binary fields that are converted from observed and forecast rainfall fields by suitable thresholds. However, this processing (similar to Casati et al. 2004) removes the effect of any bias in rainfall amounts.

The two approaches introduced by Ebert and McBride (2000) and Davis et al. (2006) are in the same category: object-based verification methodology, which assesses the forecasts from different aspects (e.g., location, rain volume, size) by transforming the spatial fields to one-dimensional errors. If one wants to assess TC forecasts from different components (e.g., track, intensity, and size), the traditional metrics are easier and more direct than the methods of Ebert and McBride (2000) and Davis et al. (2006), which have a complicated order of operations on the spatial fields. The verification methods proposed by Gilbert (1884), Casati et al. (2004), and Roberts and Lean (2008), which are based on binary (0 and 1) fields, fail to account for the magnitudes of errors in cases in which the forecasts are concerned with several degrees of intensity of a phenomenon. Meanwhile the MSE skill score (introduced by Casati et al. 2004) and the fraction skill score (defined by Roberts and Lean 2008), which are similar in essence, provide a comprehensive score, calculated relative to the MSE of a random forecast or reference and can be used for reference in the quantitative spatial verification of IW-HTDF fields. The FSS and its assessment results of IW-HTDF fields are described in the appendix. The FSS and the IW-HTDF score (SSIM index) are strongly correlated ($r = 0.99$, $p < 0.00001$). Similar to the IW-HTDF score, FSS will also give very low scores for small storms with little overlap. It is not easy for small storms to receive high scores. This problem is an intrinsic limitation of comparisons between two-dimensional fields.

IW-HTDF also has some potential advantages as compared to current separate track, intensity, and size forecasts. It should be remembered that the effects of a hurricane can be experienced well away from the hurricane center. IW-HTDF, incorporating the track, intensity, and size information, provides the spatial and temporal distribution of the perceived effects of a hurricane on its surroundings. On one hand, it can provide more specific and accurate information in space and time for hurricane watches and warnings when considering the asymmetric wind structure. On the other hand, it is useful for sophisticated users such as government officials and other decision-makers in cost-benefit analyses or damage assessments by providing more detailed two-dimensional forecast error assessments.

5. Conclusions

In this study, an integrated IW-HTDF has been designed as a new evaluation criterion for assessing TC forecasts. The results from the forecast verification analyses of 29 hurricane cases show that the advantages of the IW-HTDF-based forecast verification are twofold: 1) providing an integrated track, intensity, and size forecast skill score for each TC forecast, thus avoiding the confusion arising from contradictory assessments among track, intensity, and size forecasts when they are evaluated separately, and 2) providing a unique assessment of forecast or model performance based on the two-dimensional spatial similarity of all three facets of TC forecasts, namely, track, intensity, and size, rather than examining the forecast along a single forecast track.

Although the IW-HTDF assessment approach is exploratory, it shows an integrated way of assessing TC forecasts or the performances of TC models for some users who need a comprehensive assessment in determining a model's performance. In the current IW-HTDF assessment, track is dominant, and it is possible

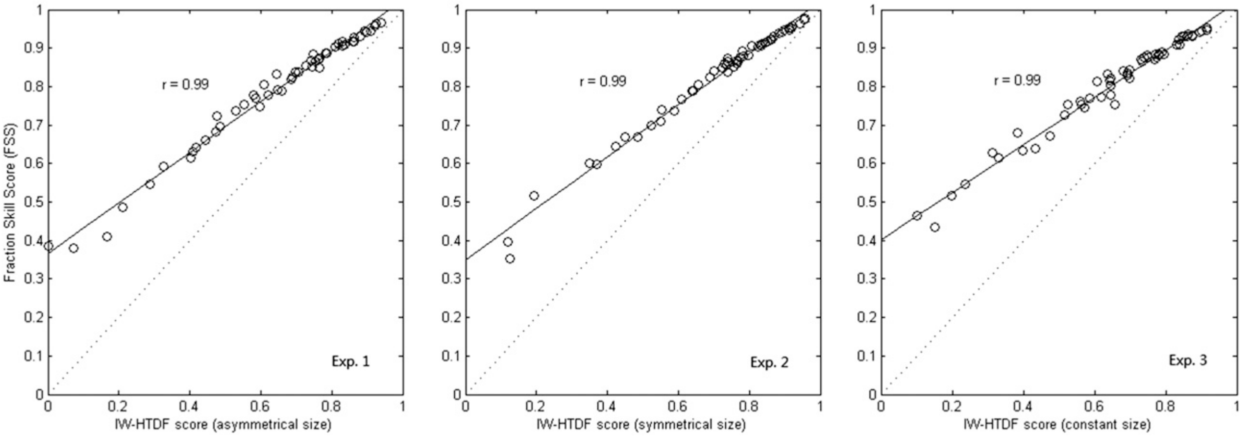

FIG. A1. FSSs and IW-HTDF scores in experiments 1–3 with different S_x settings: asymmetric, symmetric, and constant sizes. (Dashed lines are 1–1 lines.)

to adjust the weightings of track, intensity, and in some cases size for different applications or purposes. The construction of a track density function and comparison method for the spatial fields is not without its own limitations, and other approaches or processes may be introduced to provide a remedy.

It is should also be noted, however, that TC size (wind radii) in the best-track data as well as the forecast archive contains large uncertainty at the present. Therefore, it is likely to distort the assessments when TC size is included as a forecast evaluation parameter. Until reliable TC size data become available, it is recommended that the IW-HTDF evaluation approach presented in this study be used for track and intensity forecasts only.

Acknowledgments. The authors appreciate Ms. Dianna Francisco for the English proofreading. We also thank Ms. Xiaoping Zhang for her helpful discussion and suggestion on the evaluation of IW-HTDF. In addition, we also wish to thank the anonymous reviewers for their useful comments. CG appreciates the financial supports from the Ministry of Science and Technology of China (2016YFC140140005) and from the National Natural Science Foundation of China (U1406401). This work was partially supported by the China Scholarship Council (201406330010).

APPENDIX

Fractions Skill Score

Roberts and Lean (2008) defined a fractions skill score (FSS) to examine whether improved model resolution alone is able to produce more skillful precipitation forecasts on useful scales. Our IW-HTDF score is

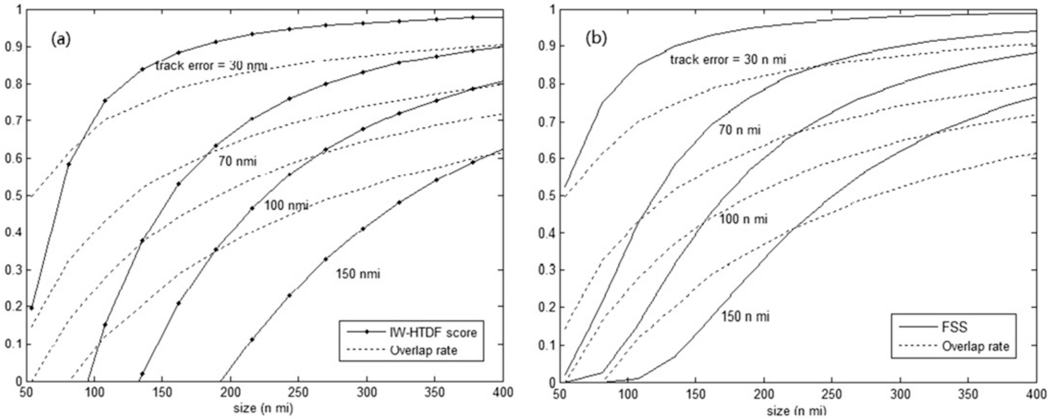

FIG. A2. Variations of the (a) IW-HTDF score and (b) FSS, with the overlap rate (dashed lines) against storm size with different track errors shown.

compared with FSS, which is also a comprehensive parameter for assessing two-dimensional spatial fields. FSS is calculated as

$$\mathrm{FSS} = \frac{\mathrm{MSE}_f - \mathrm{MSE}_{\mathrm{ref}}}{\mathrm{MSE}_{\mathrm{perfect}} - \mathrm{MSE}_{\mathrm{ref}}} = 1 - \frac{\mathrm{MSE}_f}{\mathrm{MSE}_{\mathrm{ref}}}, \qquad \text{(A1)}$$

where MSE_f is the mean-square error (MSE) for the observed and forecast fields, given by

$$\mathrm{MSE}_f = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2. \qquad \text{(A2)}$$

The MSE of a perfect forecast is $\mathrm{MSE}_{\mathrm{perfect}} = 0$. The reference MSE is defined as

$$\mathrm{MSE}_{\mathrm{ref}} = \frac{1}{N}\left(\sum_{i=1}^{N}X_i^2 + \sum_{i=1}^{N}Y_i^2\right), \qquad \text{(A3)}$$

where X and Y are the two-dimensional spatial fields of the forecast and observation, respectively.

Figure A1 shows the FSS results of the integrated IW-HTDF fields (derived from hurricane track, intensity, and size) between the forecasts and observations. We found that FSSs and IW-HTDF scores are strongly correlated. All the Pearson correlation coefficients r between FSSs and IW-HTDF scores at different S_x settings (i.e., S_x = asymmetric, symmetric, and constant size) are approximately 0.99 with $p < 0.00001$. FSS is higher than the IW-HTDF score and has a positive truncation error against the IW-HTDF score.

FSS and IW-HTDF score have different definitions for zero skill of a forecast. Based on Eqs. (A1)–(A3), if and only if X and Y do not overlap each other (overlap rate = 0), MSE_f is equal to $\mathrm{MSE}_{\mathrm{ref}}$ and then FSS obtains a value of zero, meaning zero skill (Fig. A2b). While the IW-HTDF score is zero when the correlation coefficient between the forecast and observed IW-HTDF fields becomes zero or does not exceed the confidence level of 90% (mainly affected by the track error), at this time the overlap rate usually is small but above zero (Fig. A2a). This condition is more easily met than that of FSS. Therefore, the IW-HTDF score is more rigorous for poor forecasts than FSS. These two assessment scores will both severely penalize cases where the forecast and reference IW-HTDF fields do not overlap. It is not easy for small storms to receive high scores, especially when there is little overlap between the predicted and observed track density functions. This problem is an intrinsic limitation of both FSS and the IW-HTDF score in the comparison between two-dimensional fields.

REFERENCES

Aberson, S. D., 1998: Five-day tropical cyclone track forecasts in the North Atlantic basin. *Wea. Forecasting*, **13**, 1005–1015, doi:10.1175/1520-0434(1998)013<1005:FDTCTF>2.0.CO;2.

Allam, M. E., and E. A. Abdel-Ghaffar, 2004: JPEG 2000 performance evaluation. *Proc. Int. Conf. on Electrical, Electronic and Computer Engineering*, Cairo, Egypt, IEEE, 389–392, doi:10.1109/ICEEC.2004.1374477.

Anderson, J. R., and J. R. Gyakum, 1989: A diagnostic study of Pacific basin circulation regimes as determined from extratropical cyclone tracks. *Mon. Wea. Rev.*, **117**, 2672–2686, doi:10.1175/1520-0493(1989)117<2672:ADSOPB>2.0.CO;2.

Bell, K., and P. S. Ray, 2004: North Atlantic hurricanes 1977–99: Surface hurricane-force wind radii. *Mon. Wea. Rev.*, **132**, 1167–1189, doi:10.1175/1520-0493(2004)132<1167:NAHSHW>2.0.CO;2.

Cangialosi, J. P., and J. L. Franklin, 2015: 2014 National Hurricane Center forecast verification report. National Hurricane Center, 82 pp. [Available online at http://www.nhc.noaa.gov/verification/pdfs/Verification_2014.pdf.]

Casati, B., G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.*, **11**, 141–154, doi:10.1017/S1350482704001239.

Chen, Y., and M. K. Yau, 2003: Asymmetric structures in a simulated landfalling hurricane. *J. Atmos. Sci.*, **60**, 2294–2312, doi:10.1175/1520-0469(2003)060<2294:ASIASL>2.0.CO;2.

Chikkerur, S., V. Sundaram, M. Reisslein, and L. J. Karam, 2011: Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. Broadcast.*, **57**, 165–182, doi:10.1109/TBC.2011.2104671.

Coskun, B., and B. Sankur, 2004: Robust video hash extraction. *Proc. 12th Conf. on Signal Processing and Communications Applications*, Vienna, Austria, IEEE, 292–295.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methods and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, doi:10.1175/MWR3145.1.

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

——, J. A. Knaff, and J. Kaplan, 2006: On the decay of tropical cyclone winds crossing narrow landmasses. *J. Appl. Meteor Climatol.*, **45**, 491–499, doi:10.1175/JAM2351.1.

——, ——, R. Knabb, C. Lauer, C. R. Sampson, R. T. DeMaria, 2009: A new method for estimating tropical cyclone wind speed probabilities. *Wea. Forecasting*, **24**, 1573–1591, doi:10.1175/2009WAF2222286.1.

Demuth, J. L., M. DeMaria, J. A. Knaff, and T. H. Vonder Haar, 2004: Evaluation of Advanced Microwave Sounding Unit tropical-cyclone intensity and size estimation algorithms. *J. Appl. Meteor Climatol.*, **43**, 282–296, doi:10.1175/1520-0450(2004)043<0282:EOAMSU>2.0.CO;2.

——, ——, and ——, 2006: Improvement of Advanced Microwave Sounding Unit tropical cyclone intensity and size estimation algorithms. *J. Appl. Meteor. Climatol.*, **45**, 1573–1581, doi:10.1175/JAM2429.1.

Ebert, E. E., and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.*, **239**, 179–202, doi:10.1016/S0022-1694(00)00343-7.

Feser, F., and H. Von Storch, 2008: Regional modelling of the western Pacific typhoon season 2004. *Meteor. Z.*, **17**, 519–528, doi:10.1127/0941-2948/2008/0282.

Gilbert, G. K., 1884: Finley's tornado predictions. *Amer. Meteor. J.*, **1**, 166–172.

Hill, K. A., and G. M. Lackmann, 2009: Influence of environmental humidity on tropical cyclone size. *Mon. Wea. Rev.*, **137**, 3294–3315, doi:10.1175/2009MWR2679.1.

Irish, J. L., D. T. Resio, and J. J. Ratcliffe, 2008: The influence of storm size on hurricane surge. *J. Phys. Oceanogr.*, **38**, 2003–2013, doi:10.1175/2008JPO3727.1.

Jarvinen, B. R., and C. J. Neumann, 1979: Statistical forecasts of tropical cyclone intensity for the North Atlantic basin. NOAA Tech. Memo. NWS NHC-10, 22 pp. [Available online at http://www.nhc.noaa.gov/pdf/NWS-NHC-1979-10.pdf.]

——, ——, and M. A. S. Davis, 1984: A tropical cyclone data tape for the North Atlantic Basin, 1886–1983: Contents, limitations, and uses. NOAA Tech. Memo. NWS NHC-22, 24 pp. [Available online at http://www.nhc.noaa.gov/pdf/NWS-NHC-1988-22.pdf.]

Keith, E., and L. Xie, 2009: Predicting Atlantic tropical cyclone seasonal activity in April. *Wea. Forecasting*, **24**, 436–455, doi:10.1175/2008WAF2222139.1.

Knaff, J. A., M. DeMaria, B. Sampson, and J. M. Gross, 2003: Statistical, five-day tropical cyclone intensity forecasts derived from climatology and persistence. *Wea. Forecasting*, **18**, 80–92, doi:10.1175/1520-0434(2003)018<0080:SDTCIF>2.0.CO;2.

Landsea, C. W., and J. L. Franklin, 2013: Atlantic hurricane database uncertainty and presentation of a new database format. *Mon. Wea. Rev.*, **141**, 3576–3592, doi:10.1175/MWR-D-12-00254.1.

——, ——, and J. Beven, 2015: The revised Atlantic hurricane database (HURDAT2). NOAA/National Hurricane Center, 6 pp. [Available online at http://www.nhc.noaa.gov/data/hurdat/hurdat2-format-atlantic.pdf.]

Liu, B., and L. Xie, 2012: A scale-selective data assimilation approach to improving tropical cyclone track and intensity forecasts in a limited-area model: A case study of Hurricane Felix (2007). *Wea. Forecasting*, **27**, 124–140, doi:10.1175/WAF-D-10-05033.1.

Maclay, K. S., M. Demaria, and T. H. V. Haar, 2008: Tropical cyclone inner-core kinetic energy evolution. *Mon. Wea. Rev.*, **136**, 4882–4898, doi:10.1175/2008MWR2268.1.

Merrill, R. T., 1984: A comparison of large and small tropical cyclones. *Mon. Wea. Rev.*, **112**, 1408–1418, doi:10.1175/1520-0493(1984)112<1408:ACOLAS>2.0.CO;2.

Neumann, C. J., 1972: An alternate to the HURRAN (Hurricane Analog) tropical cyclone forecast system. NOAA Tech. Memo. NWS SR-62, 32 pp.

——, and J. M. Pelissier, 1981: An analysis of Atlantic tropical cyclone forecast errors, 1970–1979. *Mon. Wea. Rev.*, **109**, 1248–1266, doi:10.1175/1520-0493(1981)109<1248:AAOATC>2.0.CO;2.

Powell, M. D., and S. D. Aberson, 2001: Accuracy of United States tropical cyclone landfall forecasts in the Atlantic basin (1976–2000). *Bull. Amer. Meteor. Soc.*, **82**, 2749–2767, doi:10.1175/1520-0477(2001)082<2749:AOUSTC>2.3.CO;2.

——, and T. A. Reinhold, 2007: Tropical cyclone destructive potential by integrated kinetic energy. *Bull. Amer. Meteor. Soc.*, **88**, 513–526, doi:10.1175/BAMS-88-4-513.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.

Sheets, R. C., 1990: The National Hurricane Center—Past, present, and future. *Wea. Forecasting*, **5**, 185–232, doi:10.1175/1520-0434(1990)005<0185:TNHCPA>2.0.CO;2.

Spencer, R. W., and W. D. Braswell, 2001: Atlantic tropical cyclone monitoring with AMSU-A: Estimation of maximum sustained wind speeds. *Mon. Wea. Rev.*, **129**, 1518–1532, doi:10.1175/1520-0493(2001)129<1518:ATCMWA>2.0.CO;2.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, 2004: Image quality assessment: Form error visibility to structural similarity. *IEEE Trans. Image Process.*, **13**, 600–612, doi:10.1109/TIP.2003.819861.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Elsevier, 627 pp.

Xie, L., T. Yan, L. J. Pietrafesa, J. Morrison, and T. Karl, 2005: The climatology and interannual variability of regional landfall hurricane frequency and its association with North Atlantic hurricane tracks. *J. Climate*, **18**, 5370–5381, doi:10.1175/JCLI3560.1.

——, B. Liu, and S. Peng, 2010: Application of scale-selective data assimilation to tropical cyclone track simulation. *J. Geophys. Res.*, **115**, D17105, doi:10.1029/2009JD013471.

——, H. Liu, B. Liu, and S. Bao, 2011: A numerical study of the effect of hurricane wind asymmetry on storm surge and inundation. *Ocean Modell.*, **36**, 71–79, doi:10.1016/j.ocemod.2010.10.001.