Evaluation and Error Analysis of Official Tropical Cyclone Intensity Forecasts during 2005–2018 for the Western North Pacific

Xiaogang HUANG

College of Meteorology and Oceanography, National University of Defense Technology, China

Xudong PENG

Unit 31010, PLA, China

Jianfang FEI, Xiaoping CHENG, Juli DING

College of Meteorology and Oceanography, National University of Defense Technology, China

and

Dandan YU

Unit 31010, PLA, China

(Manuscript received 23 February 2020, in final form 16 October 2020)

Abstract

This study systematically evaluates the accuracy, trends, and error sources for western North Pacific tropical cyclone intensity forecasts between 2005 and 2018. The study uses homogeneous samples from tropical cyclone (TC) intensity official forecasts issued by the China Meteorological Administration (CMA), Joint Typhoon Warning Center (JTWC), and Regional Specialized Meteorological Center Tokyo-Typhoon Center (RSMC-Tokyo). The TC intensity forecast accuracy performances are as follows: 24–48 h, JTWC > RSMC-Tokyo > CMA; 72 h, JTWC > CMA > RSMC-Tokyo; and 96–120 h, JTWC > CMA > RSMC-Tokyo. Improvements in TC intensity forecasting are marginal but steady for all three centers. The 24–72 h improvement rate is approximately 1–2 % yr⁻¹. The improvement rates are statistically significant at the 95 % level for almost half of the verification times from 0–120 h. The three centers tend to overestimate weak TCs over the northern South China Sea, but strong TCs are sometimes underestimated over the area east of the Philippines. The three centers generally have higher skill scores associated with forecasting of rapid weakening (RW) events than rapid intensification (RI) events. Overall, the three centers are not skillful in forecasting RI events more than three days in advance. Fortunately, RW events could be forecasted five days in advance with an accuracy order of CMA > RSMC-Tokyo > JTWC.

Keywords tropical cyclone intensity forecast; accuracy analysis; trend analysis; error source

Citation Huang, X., X. Peng, J. Fei, X. Cheng, J. Ding, and D. Yu, 2020: Evaluation and error analysis of official tropical cyclone intensity forecasts during 2005–2018 for the western North Pacific. J. Meteor. Soc. Japan, 99, 139–163, doi:10.2151/jmsj.2021-008.

Corresponding author: Jianfang Fei, College of Meteorology and Oceanography, National University of Defense Technology, No. 60, Shuanglong Road, Nanjing, 211101, China

E-mail: feijf@sina.com

J-stage Advance Published Date: 19 November 2020

©The Author(s) 2021. This is an open access article published by the Meteorological Society of Japan under a Creative Commons Attribution 4.0 International (CC BY 4.0) license (https://creativecommons.org/licenses/by/4.0).
1. Introduction

Over the past two decades, the tropical cyclone (TC) track forecast skill has improved significantly as a result of tremendous progress in numerical model guidance and the adoption of consensus forecasting techniques (Elsberry et al. 2007; Peng et al. 2017). However, TC intensity (i.e., the maximum sustained 10-m wind speed or MSW) forecasting progress still lags behind track forecast progress, especially for 24–48 h lead times (DeMaria et al. 2014).

One key reason for the modest improvement in intensity forecasting is that TC intensity changes are controlled by complex physical processes over a wide range of temporal and spatial scales. Large-scale environmental effects such as vertical wind shear (Emanuel et al. 2004; Wang and Wu 2004) and humidity (Black et al. 2002; Chen et al. 2006; Hendricks et al. 2010; Tang and Emanuel 2012), as well as underlying surface factors such as the sea surface temperature and air–sea interactions (Emanuel 1986; Shay et al. 2000; Chen et al. 2007; Lin et al. 2013; Ito et al. 2015), have been proven to be important. These dynamic processes are less well understood, and thus, they make intensity less predictable. Another key reason is that there are many deficiencies associated with TC intensity change modeling in the current generation of intensity forecast guidance tools. These deficiencies include inaccurate TC vortex initialization (Holland 1980; Kurihara et al. 1993; Pu and Braun 2001; Hendricks et al. 2013), physical parameterization uncertainties (Nolan et al. 2009; Green and Zhang 2014; Andreas et al. 2015), insufficient model resolution (Fierro et al. 2009; Davis et al. 2011; Jin et al. 2014), and a lack of TC intensity change observations (Wang et al. 2015; Creasey and Elsberry 2017).

With the help of statistical models, statistical-dynamic models, and dynamic models, TC intensity forecasts covering up to 120 h are routinely issued at the three major operational centers that are responsible for TC monitoring and prediction over the western North Pacific (WNP). These include the Joint Typhoon Warning Center (JTWC), Regional Specialized Meteorological Center Tokyo-Typhoon Center (RSMC-Tokyo), and China Meteorological Administration (CMA). The centers usually used the same “general” types of model guidance over the years. The details of the models that the three centers use for intensity forecasting can be found in their annual reports (e.g., Chen et al. 2019; Japan Meteorological Agency 2019; Joint Typhoon Warning Center 2019). Also, a few “center-specific” models such as the Coupled Ocean–Atmosphere Mesoscale Prediction System for Tropical Cyclones (COAMPS-TC) model (Doyle et al. 2012), Typhoon Intensity Forecasting scheme based on SHIPS (TIFS; Yamaguchi et al. 2018), and the WNP TC intensity prediction scheme (WIPS, Chen et al. 2011) are used by different centers, which may contribute to differences in their TC intensity forecast performance. To facilitate the application of these official forecasts and to provide better guidance that coastal communities can use to make appropriate preparations, it is necessary to evaluate their performance and analyze their error characteristics.

Most operational centers analyze and verify the results of their TC intensity forecasts annually (Chen et al. 2019; Japan Meteorological Agency 2019; Joint Typhoon Warning Center 2019). In addition, trend analyses of official TC intensity forecasts have been performed by a number of researchers. However, there are some conflicting results on the progress of TC intensity forecasting in verification studies. For example, DeMaria et al. (2007) assessed long-term TC intensity forecast trends issued by the US National Hurricane Center (NHC) between 1985 and 2005, and their results showed that the improvements at 24 h and 48 h for the Atlantic are fairly small (~ 0.1 kt yr\(^{-1}\); 1 kt = 0.514 m s\(^{-1}\)). However, DeMaria et al. (2014), who updated their verification period to 1996–2012, concluded that there are significant improvements in NHC forecasts for 48–96 h leads. Recent statistics on intensity errors published online by NHC (https://www.nhc.noaa.gov/verification/verify5.shtml) suggested that improvements had been slow for the period between 1990 and 2010 and fast for the last decade or so. For the WNP, a similar picture emerges from the JTWC official intensity forecast verifications [Joint Typhoon Warning Center 2019 (cf. Fig. 6-6)]. This pace of progress in intensity forecasting can be attributed to the availability of skillful intensity guidance such as the statistical-dynamic models and dynamic models in the last decade, as discussed by DeMaria et al. (2014). To the best of the authors’ knowledge, a detailed assessment of WNP TC intensity forecast accuracy was last presented in the refereed literature by DeMaria et al. (2014). An update to that information is provided here.

Furthermore, most previous studies have focused on evaluating the accuracies of official intensity forecasts issued by individual operational centers. However, occasional studies have compared the accuracies of TC intensity forecasts issued by different operational centers. Objectively comparing TC intensity forecasting skill between centers remains necessary for several
reasons. First, the samples and best-track data (BTD) in annual reports and previous studies are often different. This does not meet the principles of sample homogeneity and error calculation consistency required by the World Meteorological Organization (WMO; World Meteorological Organization 2013). Second, there are differences between the wind speed averaging periods that various agencies use when reporting the TC MSW. For example, the JTWC uses a 1-min average sustained 10-m wind speed, the CMA uses 2-min average values, and the RSMC-Tokyo uses 10-min average values (the WMO standard). Though the exact ratio of 10-min average to 1-min average wind speeds is variable and situation-dependent, the 10-min average wind speed is approximately 88% of the 1-min average wind speed (Atkinson 1974). In addition to the differences in wind averaging periods, the methods used to derive these winds also vary among centers. All centers use the Dvorak technique to help in assessing storm intensity by converting the Dvorak current intensity (CI) number directly to a maximum near-surface wind speed (Dvorak 1984; Velden et al. 2006; World Meteorological Organization 2012). The JTWC has applied the original, objective (Velden et al. 1998), and advanced objective Dvorak techniques (Olander and Velden 2007) using a conversion table with 1-min wind speeds. However, many centers often modify the mapping table to convert the winds to a different averaging period (Knapp and Kruk 2010). The RSMC-Tokyo uses a unique table that transfers CI numbers directly to TC intensities described by 10-min wind speeds (Koba et al. 1989, 1991). The Koba et al. (1989, 1991) table specifies higher (lower) wind speeds for higher (lower) CI numbers relative to the original Dvorak table (Nakazawa and Hoshino 2009; Knapp and Kruk 2010). The CMA used a simplified Dvorak technique in the past (World Meteorological Organization 2012), and it has been applying the classical Dvorak technique (1984 version) since 2012 (see https://www.wmo.int/pages/prog/www/tecp/documents/1.2_SatTC-AnalysisInOperations-Changes_CMA%20_ChunyiXIANG.pdf). In light of these noted differences in intensity estimation techniques applied by the three operational centers, wind speeds must be standardized prior to comparison.

The current study attempts to address the following four questions:

1) Which of the CMA, JTWC, and RSMC-Tokyo performs best in TC intensity forecasting?
2) Which center has achieved the largest recent TC intensity forecasting improvement?
3) Are large TC intensity prediction errors in particular regions caused by systematic or random biases?
4) If these errors can be attributed to systematic bias, what types of TCs drive such bias?

To answer these questions, we use BTDs from the RSMC-Tokyo as reference standards for evaluation and error analysis of 0, 24, 48, 72, 96, and 120 h intensity forecasts issued by the CMA, JTWC, and RSMC-Tokyo between 2005 and 2018. First, on the basis of the sample selection principle from the WMO (World Meteorological Organization 2013), this study selects 0, 24, 48, 72, 96, and 120 h forecast homogeneous samples from the three centers after standardization. Then, all wind speeds are converted to 10-min averages using a new method. On this basis, we analyze homogeneous sample errors and trends from each center using a variety of statistical metrics. This study also compares the three centers’ forecasts for periods of rapid TC intensity changes (RCs).

The organization of this paper is as follows. Sections 2 and 3 describe the data and methods, respectively, used in this study. Section 4 describes the evaluation results and error characteristics with respect to each forecasting center. Section 5 is a summary of the research.

2. Data

Official 2005–2018 CMA, JTWC, and RSMC-Tokyo TC intensity forecast data for TCs over the WNP area are selected. The official forecast data used in this paper are obtained from a weather and climate data archive provided by Iowa State University (https://mtarchive.geol.iastate.edu/YYYY/MM/DD/text/Severe/, where YYYY, MM, and DD represent the four-digit year, two-digit month, and two-digit day of the official forecast issue date, respectively). Because the RSMC-Tokyo began issuing five-day TC intensity forecasts in 2019 (Ono et al. 2019), only homogeneous samples for the 96 h and 120 h forecasts issued by the JTWC and CMA are evaluated. The JTWC has been providing 96 h and 120 h forecasts since 2003, and the CMA started to issue 96 h and 120 h forecasts in 2008 and 2010, respectively. Therefore, the 96 h and 120 h forecasts cover 2008–2018 and 2010–2018, respectively. It is worth noting that the limited 96 h (2008–2018) and 120 h (2010–2018) datasets introduce discontinuities between these two forecasting periods and shorter forecasting periods (0–72 h). This discontinuity will be discussed later.

Three BTDs are selected. They are referred to as CMA-BTD (http://tcdata.typhoon.org.cn/en/zjllsj_/zlhq.html), JTWC-BTD (https://www.metoc.navy.mil/jtwc/jtwc.html), and RSMC-BTD (http://www.jma.
go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/best track.html). These data contain information regarding the TC center location, MSW, and central minimum pressure at 6 h intervals. JTWC-BTD and CMA-BTD intensity values start with the tropical disturbance and tropical depression (TD) categories, respectively. Like the JTWC-BTD and CMA-BTD, the RSMC-BTD also analyzes weak tropical cyclones (i.e., TD). However, the RSMC-BTD does not provide the MSW analysis information for such weak systems though it does provide the minimum sea level pressure. Because the intensity values of TDs are all recorded as zero in the RSMC-BTD, we examine only concurrent TCs with MSWs larger than 17.2 m s\(^{-1}\) or 35 kt. All of the three BTDs are collected between 2005 and 2018.

Only TCs recorded by all of the three official forecasts and BTDs can be used as homogeneous samples. A large number of homogeneous samples from the three centers are established and are shown in Table 1.

### 3. Methodology

#### 3.1 Wind speed conversion

Many studies have attempted to convert operational center TC best-track wind speeds to a normalized scale. The first method is the use of a multiplicative factor to convert wind speeds that are averaged over different periods into wind speeds based on a single average time. Two commonly used linear relationships were developed by Atkinson (1974) and Harper et al. (2010). However, Song et al. (2010) suggested that the relationship between any two BTDs is nonlinear and obtained nonlinear conversion relationships via least squares regression fitting. Furthermore, Knapp and Kruk (2010) proposed a remapping method that converted wind speed values back to CI numbers and then derived homogenous wind speed values from these CI numbers by applying a single conversion table to all datasets. Knapp and Kruk (2010) showed that the remapping method was more effective at reducing the differences between intensities from different agencies than the linear method.

Because all three centers use the Dvorak technique to help in assessing storm intensity (World Meteorological Organization 2012), converting wind speed values to CI numbers seems to be a viable method of homogenizing data between agencies. However, CMA analysts occasionally subjectively adjust the converted wind speed value during operational TC intensity analysis (Xu et al. 2015). Because this study did not assess the scale of such subjective conversion efforts, the remapping method could not be applied to CMA wind speed data.

Inspired by the methods of Knapp and Kruk (2010) and Song et al. (2010), a new wind speed conversion mapping table based on average values is established by this paper. This conversion method, which depends solely on historical BTD intensity relationships, is more direct than using CI conversion and is not subject to any possible CI scale inaccuracies. We can consider the conversion between 1-min and 10-min winds. First, we mapped wind speed from JTWC-BTD against the wind speed from RSMC-BTD for identical cyclones between 2005 and 2018. In the second step, the mean value of the concurrent RSMC-BTD winds was calculated for a given JTWC-BTD wind value (in 5 kt intervals). This average value was treated as the 10-min wind speed that corresponded to the 1-min wind speed. Because JTWC and RSMC-Tokyo TC intensities were recorded to the nearest 5 kt, the average value was adjusted to the nearest 5 kt for consistency. Using this method, MSWs from JTWC (1-min, kt) and CMA (2-min, m s\(^{-1}\)) forecasts were converted into 10-min winds (kt) based on their mapping relationship with the RSMC-Tokyo data (10-min, kt). The mapping relationships are shown in Tables 2 and 3.

#### 3.2 Conventional statistical metrics

The mean error (ME) and mean absolute error
(MAE) are statistical metrics that are commonly used to evaluate TC intensity forecast performance. The ME is defined as the mean difference between the forecasted \( f_i \) and observed \( o_i \) results, where \( i \) is the sample index. The MAE is similar to the ME but is used for absolute differences. The two quantities are described in Eqs. (1) and (2).

\[
ME = \frac{1}{N} \sum_{i=1}^{N} (f_i - o_i), \quad (1)
\]

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |f_i - o_i|, \quad (2)
\]

where \( N \) is the sample size.

### 3.3 Decomposition of errors

A TC can be simply classified as weak (WT, \(< 64 \) kt) or strong (ST, \( \geq 64 \) kt). The total MAE (TE) of the intensity forecast is defined as \( \sum_{i=1}^{N} |f_i - o_i| \) and can be decomposed into four categories: overestimation of a strong TC (OST), underestimation of a strong TC (UST), overestimation of a weak TC (OWT), and underestimation of a weak TC (UWT):

\[
OST = \sum |f_i - o_i|, \quad \text{when } (o_i \geq 64) \text{ and } (f_i > o_i), \quad (3)
\]

\[
UST = \sum |f_i - o_i|, \quad \text{when } (o_i \geq 64) \text{ and } (f_i < o_i), \quad (4)
\]

\[
OWT = \sum |f_i - o_i|, \quad \text{when } (o_i < 64) \text{ and } (f_i > o_i), \quad (5)
\]

\[
UWT = \sum |f_i - o_i|, \quad \text{when } (o_i < 64) \text{ and } (f_i < o_i). \quad (6)
\]

| JTWC (1 min, kt) | RSMC-Tokyo (10 min, kt) |
|------------------|-------------------------|
| 25               | 30                      |
| 30               | 30                      |
| 35               | 35                      |
| 40               | 40                      |
| 45               | 45                      |
| 50               | 45                      |
| 55               | 50                      |
| 60               | 55                      |
| 65               | 60                      |
| 70               | 65                      |
| 75               | 65                      |
| 80               | 70                      |
| 85               | 75                      |
| 90               | 75                      |
| 95               | 80                      |
| 100              | 80                      |
| 105              | 85                      |
| 110              | 90                      |
| 115              | 90                      |
| 120              | 95                      |
| 125              | 95                      |
| 130              | 100                     |
| 135              | 100                     |
| 140              | 105                     |
| 145              | 110                     |
| 150              | 110                     |
| 155              | 110                     |
| 160              | 110                     |
| 165              | 120                     |
| 170              | 125                     |

| CMA (2 min, m s\(^{-1}\)) | RSMC-Tokyo (10 min, kt) |
|---------------------------|-------------------------|
| 13                        | 30                      |
| 15                        | 30                      |
| 18                        | 35                      |
| 20                        | 40                      |
| 23                        | 45                      |
| 25                        | 50                      |
| 28                        | 55                      |
| 30                        | 60                      |
| 33                        | 60                      |
| 35                        | 65                      |
| 38                        | 70                      |
| 40                        | 75                      |
| 42                        | 80                      |
| 45                        | 85                      |
| 48                        | 85                      |
| 50                        | 90                      |
| 52                        | 90                      |
| 55                        | 95                      |
| 58                        | 100                     |
| 60                        | 105                     |
| 62                        | 105                     |
| 65                        | 110                     |
| 68                        | 115                     |
| 70                        | 115                     |
| 72                        | 120                     |
| 75                        | 120                     |
| 78                        | 125                     |
3.4 Assessment of RC event forecasting skill

TC intensity RCs can be divided into rapid intensification (RI) and rapid weakening (RW). RI is usually defined as the 95th percentile of all 24 h intensity changes (Kaplan and DeMaria 2003).

In this study, RI and RW are defined as intensity increases and decreases, respectively, of at least 25 kt within a 24 h period. Because the physical processes that govern TC intensity changes after landfall are quite different from those that operate over water, this study considers only RCs that occur over water. All RC cases in which the forecasted or BTD TC moved over land were removed from the sample. Intensity changes of −25 kt and 25 kt represent the 7th and 92nd percentiles, respectively, in the 2005–2018 RSMC-BTD dataset. These are not the same as the operational definition, which refers to the 5th and 95th percentiles. The lower threshold was chosen to increase the number of RC events and improve the statistics.

Occurrence and nonoccurrence of RC events can be treated as deterministic binary processes, the evaluation of which is typically performed with the help of a contingency table (Table 4). In this table, hits (a) are RC events that were both forecasted and observed, false alarms (b) are RC events that were forecasted but not observed, misses (c) are RC events that were not forecasted but were observed, and correct rejections (d) are RC events that were not forecasted and did not happen. The hit (H) and false alarm (F) rates can then be calculated:

\[
H = \frac{a}{a + c}, \quad (7)
\]

\[
F = \frac{b}{b + d}, \quad (8)
\]

where \(H\) is the ratio of the number of hits to the total number of RCs that occur and \(F\) is the ratio of the number of false alarms to the total number of non-RCs observed. \(H\) and \(F\) range from 0 to 1. This \(F\) should not be confused with the false alarm ratio, which is the ratio of the number of false alarms to the total number of forecasts (Barnes et al. 2009). Additionally, the RC events are based on the intensity forecasts converted to the RSMC-Tokyo scale. Thus, the number of events that qualify as “forecasted RC” in this study is less than the number of RC events that might otherwise be considered as forecasted by the JTWC and CMA (using the +/- 25 kt criterion) based on their own intensity scales (because their intensity scale ranges are longer).

RI or RW occur within the forecast period covered by 7% of all intensity forecasts issued, which makes them rare events. Rare event evaluation measures require certain desirable properties, including equitability, propriety, and nondegeneracy (Gandin and Murphy 1992). The hit and false alarm rates lack most of these desirable properties. Therefore, a more appropriate evaluation measure, the symmetric external dependence index (SEDI), was also used in this study (Ferro and Stephenson 2011). The SEDI is calculated as follows:

\[
SEDI = \frac{\log F - \log H - \log(1 - F) + \log(1 - H)}{\log F + \log H + \log(1 - F) + \log(1 - H)}.
\]

(9)

This index can be used to evaluate rare events [see Table 3.4 of Jolliffe and Stephenson (2012) for details]. The SEDI ranges from −1 to 1, with 0 being the expected score for a random forecasting system and negative scores indicating forecasts that are worse than random.

4. Results

Typically, there are TC intensity discrepancies between the BTDS provided by different centers (Knapp and Kruk 2010; Song et al. 2010). Performance evaluations verified using different BTDS are expected to be dissimilar (Yu et al. 2012). This study used the CMA-BTD, JTWC-BTD, and RSMC-BTD separately as references against which to calculate the MAEs. This study found that although the evaluation results verified with different BTDS differed in values, the overall performances were consistent with each other, i.e., the overall relative performance of the operational centers’ intensity forecasts was not impacted by selection of a different baseline BTD. On the basis of this consideration and because of space limitations, we show only the RSMC-BTD verification results because the 10-min wind speed is the WMO standard (World Meteorological Organization 2017).

4.1 Overall assessment

a. MAEs and MEs of TC intensity forecasts

As shown in Fig. 1 and Table 5, the MAE of the

| RC Observation | RC Forecast |
|----------------|-------------|
| yes            | hits (a)    |
| no             | misses (c)  |
| false alarms (b) |                   |
| correct rejections (d) |               |
RSMC-Tokyo forecasts is significantly smaller than those from the other two centers at 0 h. Here, significance means that the MAE differences are significantly positive (negative) at the 5% significance level with a two-sided bootstrap significance test (Jolliffe and Stephenson 2012). This is an expected result. For the RSMC-Tokyo, the MAE value represents the differences between “real-time” intensity analysis and the best-track intensities at initialization time. For the CMA and JTWC, the differences between each center’s “real-time” intensity analysis and RSMC-Tokyo’s “real-time” intensity analysis also contribute to the MAEs. Overall, JTWC forecasts compare favorably to those from the other two centers through 72 h and to those from the CMA through 120 h. It is noteworthy that as the lead time increases, the CMA forecast accuracy gradually improves to approach that of the RSMC-Tokyo, even surpassing it at 72 h. Upon aggregating the 24, 48, and 72 h lead times, the MAEs of the JTWC and RSMC-Tokyo forecasts are approximately 13% and 3% less, respectively, than those of the CMA forecasts.

Figure 1b shows the forecast ME relative to the RSMC-BTD. It is clear that the CMA and RSMC-Tokyo forecasts tend to overestimate the actual TC intensity and that longer forecast lead times are associated with more overestimation. However, the JTWC forecasts slightly underestimate the actual TC intensity at 0 h. This gradually transitions to slight overestimation. A t-test ($\alpha = 0.05$) of the ME data from each center was performed at each forecast time to test for significance. The results are shown in Table 5. All three centers show significant TC intensity forecast biases regardless of the lead time. The only exception is the JTWC at a lead time of 48 h.

Assessment and analysis of the MAE and ME
indicate that the JTWC performs best among the three operational forecasting centers, followed by the RSMC-Tokyo and CMA. The CMA and RSMC-Tokyo forecasts tend to overestimate TC intensities (not conditional bias). If the 0–48 h forecasts are considered as a whole, the proportions of over (under) estimated forecasts for the CMA, JTWC, and RSMC-Tokyo centers are 46 (32), 36 (41), and 38 % (27 %), respectively. These differences in proportions show that the largest positive bias occurs at the CMA, followed by the RSMC-Tokyo. The JTWC has a negative overall bias (Table 5).

b. Conditional distribution analysis

The joint distribution verification method introduced by Moskaitis (2008) was used to reveal the overall TC intensity prediction performances of the three centers (not shown). Similar to that study’s findings for NHC and model intensity forecasts, our analysis identified a conditional bias that grows with the forecast lead time: the intensity forecasts of the three centers are generally too low for strong TCs and too high for weak TCs. This is especially evident for forecasts issued by the RSMC-Tokyo.

To better show this conditional bias, the deviation of a mean forecast \( \bar{f} \) from an observed value is calculated, as shown in Fig. 2. Here, a set of dots represent \((\bar{f}, o)\) for all values of \(o\) (interval of 5 kt). In each panel, the solid black line represents the main diagonal, and the dashed black line marks a representative value of the mean forecast, as described in the text.

Fig. 2. Conditional bias comparative scatterplots for lead times of (a) 0, (b) 24, (c) 72, and (d) 120 h. For a given lead time, a set of dots is plotted for each of the (red) CMA, (green) JTWC, and (blue) RSMC-Tokyo forecasts. The dots in each set mark \((\bar{f}, o)\) for all values of \(o\) (interval of 5 kt). In each panel, the solid black line represents the main diagonal, and the dashed black line marks a representative value of the mean forecast, as described in the text.
clarity. Figure 2 shows that the forecast performances of the three centers are quite similar. There is little conditional bias at 0 h (Fig. 2a), but the dots migrate closer to the dashed black line as the forecast lead time increases. The magnitude of strong TC underestimation is larger than that of weak TC overestimation. This trait is most pronounced at 72 h, especially for the RSMC-Tokyo data.

c. Long-term trend analysis

Figure 3 shows the annual MAEs of the CMA, JTWC, and RSMC-Tokyo intensity forecasts. The average yearly homogeneous sample sizes are 410, 360, 281, 213, 155, and 117 at 0, 24, 48, 72, 96, and 120 h, respectively.

Linear regression of the MAE as a function of the year was performed to determine time trends, and $F$ tests ($\alpha = 0.05$) were performed on the trends. In addition, the improvement rate per year as introduced by DeMaria et al. (2014) was used (Table 6). The CMA intensity forecast shows statistically significant improvements at 0, 72, and 96 h, with improvement rates of 1.10, 1.38, and 2.07 % per year, respectively. The JTWC forecasts exhibit significant downward trends at 24, 48, and 72 h, and the improvement rates are almost twice those of the CMA forecasts for the same periods. However, the RSMC-Tokyo forecasts exhibit no significant improvement at 0–72 h. This result is confirmed in Fig. 3c, where the trend lines for the RSMC-Tokyo forecasts are flat. It is worth noting that the MAE of the RSMC-Tokyo official intensity forecast decreases substantially in 2017. This can be attributable to recent advances at the RSMC-Tokyo as TIFS has been used in trial mode since 2016 (Yamaguchi et al. 2018) and became fully operational in 2019 (Ono et al. 2019).

As described in Section 2, there are discontinuities between 96 h and 120 h and shorter lead times (0 h to 72 h). We speculate that the MAEs of 96 h and 120 h in Table 5 would be higher if a full 2005–2018 dataset of forecasts for those lead times was available because the MAE decreases over time as shown in Fig. 3 and Table 6.

For the purpose of comparison, Table 6 also shows improvement rates of TC track forecasts issued by the CMA, JTWC, and RSMC-Tokyo for the same periods. All of the track improvement trends for the

| Forecast center     | Lead time (h) | 0       | 24      | 48      | 72      | 96       | 120  |
|---------------------|--------------|---------|---------|---------|---------|----------|------|
| CMA                 |              | 1.10 (−0.05) | 0.91 (3.50) | 0.67 (3.38) | 1.38 (3.15) | 2.07 (4.88) | 0.14 (2.86) |
| JTWC                |              | 0.17 (−1.11) | 1.86 (2.42) | 1.83 (3.12) | 2.10 (3.39) | 1.66 (5.53) | −2.05 (2.54) |
| RSMC-Tokyo          |              | −0.10 (−1.95) | 0.70 (2.72) | 0.55 (2.96) | 0.64 (2.51) | *        | *    |
CMA, JTWC, and RSMC-Tokyo forecasts are statistically significant at 24–72 h. The rates of CMA and RSMC-Tokyo TC intensity forecast improvements are only one-quarter to one-half of those of track out to 72 h. The rates of JTWC forecast intensity improvements are nearly one-half of those associated with the track. However, the official intensity forecast improvement rates are significant at 0, 72, and 96 h at the CMA and at 24, 48, and 72 h at the JTWC.

Although the above results indicate that the intensity improvements remain limited from 2005 to 2018, the CMA, JTWC, and RSMC-Tokyo intensity forecasts appear to make some obvious improvements in recent years, as shown in Fig. 4. Figure 4 depicts MAE differences at each center between 2005 and 2011, as well as between 2012 and 2018. Selecting these longer periods reduces the impact of TC activity interannual variability. It is worth noting that the first period with MAE differences at 96 h are from 2008 to 2011 as 96 h forecasts have been produced routinely only since 2008. The first period with MAE differences at 120 h is from 2010 to 2011 for similar reasons.

At the CMA, the multiyear MAEs at 0, 24, 48, 72, 96, and 120 h during the first period are 5.2, 10.1, 13.1, 14.5, 16.5, and 17.0 kt, respectively. However, the multiyear MAEs at the corresponding lead times during the second period are reduced to 4.8, 9.3, 12.3, 13.6, and 14.9 kt, respectively (Fig. 4a). This indicates that the five-day CMA intensity forecasts made during the second period are better than those made during the first period (Fig. 4d), especially at 96 h and 120 h. During the first period, the multiyear MAE of the CMA is larger at each lead time than those of the other two centers. During the second period, the multiyear MAE of the CMA is smaller than that of the RSMC-Tokyo at 48 h and 72 h but still larger than that of the JTWC. The overall improvement in JTWC forecasts is similar to the improvement in CMA forecasts (Fig. 4b). Interestingly, the multiyear average in Fig. 4 indicates a decreasing trend in JTWC forecast errors for forecast hour 120, whereas the linear regression line in Fig. 3 indicates an increasing trend. This seeming paradox is explained by the fact that there are more JTWC cases in the “more accurate forecast” years between 2012 and 2018 than there are for the “less accurate forecast” years. In contrast to the obvious progress made by the CMA and JTWC, the RSMC-Tokyo intensity forecast accuracy only improves at 0, 24, and 48 h, with slight progress at 72 h (Fig. 4d).
In general, both Figs. 3 and 4 indicate that the CMA, JTWC, and RSMC-Tokyo have made slow but steady progress in WNP TC intensity forecasting. From 24 h to 72 h, the JTWC makes the largest accuracy improvement. Its prediction accuracy is better than those of the other two centers over the past seven years. The CMA offers the second fastest improvement rate but still has substantial room for further improvement. The RSMC-Tokyo has made progress but the least among the three centers. At 96 h and 120 h, the CMA forecasts improve more than those from the JTWC. TC intensity forecast performance can change significantly when new operational techniques are used (e.g., the MAE of the RSMC-Tokyo in 2017 in Fig. 3), as DeMaria et al. (2014) also noted. Moreover, Fig. 4 suggests that more rapid gains in intensity forecast performance have been evident since 2012. Therefore, the previously accepted statement that intensity forecasts “aren’t really improving much”, which can be found in many papers drafted or published before 2012 or so (e.g., Qian et al. 2012; Mohapatra et al. 2013), is no longer accurate. Also, because our verification period is from 2005 to 2018, the verification results presented in this study may not necessarily reflect the latest forecast performances of the three centers.

**d. Spatial distribution characteristics**

We analyze spatial distributions of errors (MAEs and MEs) by binning the WNP (0°–60°N, 100°–180°E) into 2° latitude × 2° longitude grid boxes and averaging the error values within each box (Figs. 5–7). This analysis is based on the positions of the observed storms at each verification time, not the initialization time. Shading is limited to grid cells with ≥ 5 samples.

For brevity, this section primarily analyzes the spatial distribution of the MAEs at 24 h, as shown in Figs. 5a–c. The MAEs are mainly distributed in three general regions:

1) the northern portion of the South China Sea and the islands of Taiwan and Luzon (10°–26°N, 104°–122°E; hereafter referred to as region 1);
2) east of the Philippines (10°–20°N, 122°–150°E; hereafter referred to as region 2); and
3) east of Taiwan Island and south of Japan (20°–36°N, 122°–150°E; hereafter referred to as region 3).

The CMA and RSMC-Tokyo’s grid average MAE distribution characteristics are similar, but the CMA’s MAE is larger. The proportions of grids with average MAEs that are greater than 10 kt exceed 36 % for the CMA and 25 % for the RSMC-Tokyo. The number of grids with larger average MAE of JTWC is much less than those of the CMA and RSMC-Tokyo, which are mainly concentrated in regions 1 and 2. Only 15 % of the JTWC’s grid average MAE values exceed 10 kt.

By applying a t-test (Figs. 5d–f) with α = 0.05 to determine the significance of the average ME spatial distribution, we find that the sources of errors at the three centers are different. In regions 1 and 3, the CMA usually has a significant positive ME, with almost half of the grids having positive MEs that exceed 6 kt. In addition, the CMA has a significant negative ME in region 2. The ME distribution pattern of the RSMC-Tokyo is similar to that of the CMA. For the JTWC, significant positive and negative MEs are identified in regions 1 and 2, respectively, but the number of grids with significant MEs is smaller than that for the CMA and RSMC-Tokyo. In region 3, small significant positive and negative MEs appear for the JTWC. Combining Figs. 5a–c and Figs. 5d–f, it is apparent that when a TC approaches region 1, the three centers tend to overestimate the TC intensity. In region 2, the three centers tend to underestimate the TC intensity. In region 3, the CMA and RSMC-Tokyo tend to overestimate the TC intensity.

Finally, we calculate the grid average MAE differences (CMA–RSMC-Tokyo, JTWC–RSMC-Tokyo, and CMA–JTWC). We assess the significance of these results by applying a bootstrap test at the 95 % level in each of the grid boxes. When the null hypothesis is rejected, the MAEs of paired centers in the grid box are considered nonidentical at the 95 % level. Figs. 5g–i show the results for CMA–RSMC-Tokyo, JTWC–RSMC-Tokyo, and CMA–JTWC, respectively. The relative performance of paired centers for each of the regions can be determined by the number of grids with a significant MAE difference value in that region. There are seven grid boxes with a significant positive value in region 1 in Fig. 5g, suggesting that CMA MAEs are significantly higher than RSMC-Tokyo MAEs in region 1. Furthermore, Fig. 5h depicts that JTWC MAEs are comparable to RSMC-Tokyo MAEs in region 1 as there are no grid boxes with a statistically significant MAE difference value in that region. CMA MAEs are also significantly higher than JTWC MAEs for region 1, as supported by Fig. 5i. Overall, in region 1, JTWC and RSMC-Tokyo are comparable in accuracy, with the CMA ranking last. In the same way, we conclude that in region 2, the JTWC exhibits the best forecasts among the three centers, with the CMA comparable to the RSMC-Tokyo. For region 3, the JTWC performs best again, followed by the RSMC-Tokyo and CMA. The grid average MAE and ME distribution characteristics at 48 h and 72 h (Figs.
After understanding the relationships between TC intensity forecast errors and the location in the sea, the following two questions still remain to be answered:

1) If the error is large, what type of TC is produced?
2) Do the intensity forecasts tend to overestimate or underestimate intensity for this type of TC?

4.2 Classification assessment

In this section, we examine how different conur-
rent TC intensities might relate to different intensity forecast error characteristics. Each intensity forecast error was categorized according to the TC wind scale of the Typhoon Committee based on the best-track observations at each verification time, rather than the initialization time. The samples were divided into three categories: tropical storms (TS, 34–47 kt), severe tropical storms (STS, 48–63 kt), and typhoons (TY, > 64 kt) (World Meteorological Organization 2019). The proportion of each category and the MEs and MAEs of the intensity forecasts from the three centers up to 120 h were used in the statistical analysis, and they are shown in Fig. 8.

TSs typically range from 17 % to 33 % of all forecasts (Fig. 8a). All three centers have positive MEs, and the ME magnitude increases with the forecast lead time, with the 120 h forecast ME being close to 20 kt (Fig. 8d). The MAE characteristics are similar to those of the MEs. Comparison of the MAE values of the three centers indicates that the average MAE magni-

Fig. 6. Same as Fig. 5 but for a 48 h lead time.
tudes for the three centers are almost the same, except at 0 h (Fig. 8g). This indicates that the three centers face the same TS intensity prediction difficulties.

A total of 20% of all homogenous samples are STS (Fig. 8b). The ME and MAE are less than those for TS at corresponding lead times, but the trends are similar. All three centers tend to overestimate STS intensities, and the ME increases with the forecast lead time (Fig. 8e), resulting in an increase in the MAE.

TYs were dominant, accounting for 40% to 60% of all homogenous samples (Fig. 8c). In all three centers, the TY forecast performance is different from the TS and STS performance. All three centers tend to underestimate the TY strength. The trend of MAE growth with forecast lead time is not as clear as that with the TS and STS intensity forecasts.

Overall, the three centers perform better in forecasting strong TCs (TY) than weak TCs (TS and STS). All three centers almost always overestimate the intensities of weak TCs (TS and STS), resulting in large
forecast errors. When a TC strengthens (TY), the three centers tend to underestimate its intensity.

4.3 Decomposition of total MAEs

Because the difference in overall MAEs for the three centers is smallest at 24 h (Fig. 1) and owing to space limitations, the study focuses on analysis of 24 h MAE decomposition results. Figure 9 shows the spatial distribution of the MAE components at 24 h. Note that grid cells are shaded with the cumulative values. Each center occupies a column in Fig. 9 that represents the TE and its four components: OST, UST, OWT, and UWT, as described in Section 3.3. Their relationship can be expressed as TE = OST + UST + OWT + UWT.

The spatial distributions of the TE and its four components for the three centers share considerable similarities. The TEs have different sources in regions 1, 2, and 3. In region 1, the TE is derived primarily from the OWT error components in the north of the South China Sea and the South China coastal areas (Figs. 9j–l). The TE in region 2 can be traced to the UST error component in the waters east of the Philippines (Figs. 9g–i). In region 3, the OST and UST error components east of Taiwan are dominant (Figs. 9d–f), followed by the OWT error components in the sea south of Japan (Figs. 9j–l).

The results of the 48 h and 72 h total MAE decompositions (Figs. 10, 11) are similar to those of the 24 h decomposition. The TE in region 1 is caused primarily...
Fig. 9. Spatial distribution of error components (kt) at 24 h for the CMA, JTWC, and RSMC-Tokyo. (a)–(c) Total MAE. (d)–(f) OST error. (g)–(i) UST error. (j)–(l) OWT error. (m)–(o) UWT error. Grid cells are shaded with cumulative values.
Fig. 10. Same as Fig. 9 but for a lead time of 48 h.
Fig. 11. Same as Fig. 9 but for a lead time of 72 h.
by overestimation of weak TCs. The TE in region 2 is caused primarily by underestimation of strong TCs, and the TE in region 3 is caused primarily by missetimation of strong TCs and overestimation of weak TCs.

To quantitatively analyze the contributions of various components in different areas of the sea, we computed the percentage of these components in the TE (Fig. 12). Overall, the contribution rates of each TE component at each center are similar across regions. In region 1, OWT components are the main error source, and the OWT proportion is approximately 10–20%. The other components account for about 5% (Figs. 12a, d, j). In region 2, the UST component is the main source of error, but the proportion decreases with the lead time after 48 h. The error composition in region
3 is quite different from those in regions 1 and 2. The proportion of OST is slightly larger than or equal to that of UST at the CMA and RSMC-Tokyo, but the opposite is true at the JTWC (Figs. 12c, f). All three centers overestimate the intensities of weak TCs in region 3 at all forecast lead times, except for 0 h (Fig. 12i).

In summary, error decomposition enables one to better identify error sources and their contributions to total MAEs. The three centers exhibit considerable similarities as follows:

1) When a TC (especially a weak TC) moves into the South China Sea, the intensity forecasts of the three centers are prone to overestimation (e.g., for Linfa (2015) from 00:00 UTC on July 3 to 00:00 UTC on July 9 in Figs. 13a–d). This shows that the three centers experience great difficulty in forecasting weak TCs.

2) When a strong TC approaches the sea to the east of the Philippines, the three centers (especially the RSMC-Tokyo) tend to underestimate its intensity (e.g., Muifa (2011) around 00:00 UTC on July 30 in Figs. 13e–h).

3) When a strengthening TC moves into the area east of Taiwan, the three centers (especially the CMA) tend to overestimate its intensity (e.g., Muifa (2011) from 00:00 UTC on August 1 to 00:00 UTC on August 7 in Figs. 13e–h).

The relationship between TC intensity forecast error and offshore distance suggests that the three centers may adopt a “better mistake than miss” protocol for strong TCs. When a strong TC moves close to mainland China or the islands, all three centers (especially the CMA) tend to maintain or overestimate strong TC intensities in their forecasts. This may be the case because an underestimated strong TC intensity forecast could lead to inadequate disaster preparedness, neglected preventive measures, and substantial economic losses. Bhatia and Nolan (2013) also found that there was overestimation of strong TCs by the NHC in their error analysis of north Atlantic hurricane intensity forecasts.

It is interesting that region 2 used in our study is similar to the RI zone (longitudes of 121–143°E and
latitudes of 9°E–21°N) defined by Fudeyasu et al. (2018), where RI was frequently observed in their study. This similarity suggests that the errors in region 2, which are associated with negative biases, may be caused primarily by RI-TC underforecasting. Therefore, it is necessary to evaluate the skill of the three centers in forecasting rapid changes in TC intensities.

4.4 RC event forecast capabilities

Rapid intensity change is considered one of the most challenging TC intensity forecasting subjects (Elsberry et al. 2007). In a further analysis, the hit rate, false alarm rate, and SEDI score were used to evaluate RC forecast accuracy. To concisely present the results, we consider the relative operating characteristic, which is a plot of the hit rate against the false alarm rate. On this plot, we overlie SEDI contours, which represent the skill score.

The thresholds described in Section 3.4 were selected to define RC events. On the basis of these definitions, we obtained the RC cases from the dataset based on the best-track intensity changes during every 24 h of forecasts. That is, the total number of RC cases is a count of the number of RC cases observed in the RSMC-Tokyo BTD for all 24-h increments, i.e., between lead time $t - 24$ h and verification time $t$ for any given BTD data point. Table 7 and Fig. 14 present the assessments of the RC forecast performance of each center. As shown in Fig. 14a, the SEDI scores of the three centers for the 24 h lead time are not very high. The JTWC has the highest score (0.27), followed by the RSMC-Tokyo (0.26) and CMA (0.15). This is because the JTWC has the highest number of hits (28) and the second largest number of false alarms (33). Although the RSMC-Tokyo has the fewest hits (17), it also has the fewest false alarms (only one-half to one-quarter of the false alarms produced by the CMA and JTWC). Therefore, the RSMC-Tokyo SEDI score remains quite close to that of the JTWC. The CMA has only one more hit than the RSMC-Tokyo, but it has the lowest SEDI score because it has the highest number of false alarms (33).

As the forecast lead time increases, the three centers begin to exhibit differences in forecasting performance. At 48 h, the CMA SEDI score increases (0.28) because of its higher hit rate and lower false alarm rate (Fig. 14a). Although the false alarm rate of the JTWC decreases, its hit rate decreases even faster, resulting in a large decline in its SEDI score (0.22). There is little change in the hit rate of the RSMC-Tokyo, but there are only three false alarms. This low false alarm rate eventually produces the highest final SEDI score.

| Forecast center  | Hits   | Misses   | False alarms | SEDI   |
|------------------|--------|----------|--------------|--------|
| 24 h             |        |          |              |        |
| CMA              | 18 (49)| 424 (116)| 50 (12)      | 0.15 (0.67) |
| JTWC             | 28 (41)| 414 (124)| 33 (16)      | 0.27 (0.61) |
| RSMC-Tokyo       | 17 (50)| 425 (115)| 15 (16)      | 0.26 (0.62) |
| 48 h             |        |          |              |        |
| CMA              | 11 (33)| 225 (110)| 13 (14)      | 0.28 (0.59) |
| JTWC             | 10 (40)| 226 (103)| 24 (25)      | 0.21 (0.60) |
| RSMC-Tokyo       | 9 (33) | 227 (110)| 3 (8)        | 0.36 (0.62) |
| 72 h             |        |          |              |        |
| CMA              | 5 (20) | 91 (110) | 10 (27)      | 0.30 (0.41) |
| JTWC             | 0 (8)  | 96 (122) | 2 (10)       | * (0.32)  |
| RSMC-Tokyo       | 0 (9)  | 96 (121) | 3 (8)        | * (0.35)  |
| 96 h             |        |          |              |        |
| CMA              | 0 (8)  | 47 (79)  | 2 (14)       | * (0.31)  |
| JTWC             | 0 (4)  | 47 (83)  | 0 (11)       | * (0.21)  |
| 120 h            |        |          |              |        |
| CMA              | 0 (4)  | 12 (62)  | 0 (1)        | * (0.41)  |
| JTWC             | 0 (0)  | 12 (66)  | 0 (6)        | * ( * )   |

Table 7. Metrics used in skill evaluation of RI (intensification by more than 25 kt within 24 h) forecasting for the three major centers: hits, misses, false alarms, and SEDI scores. The numbers in parentheses are the corresponding metrics used to evaluate RW events (weakening by more than 25 kt within 24 h).
At 72 h, neither the RSMC-Tokyo nor the JTWC can successfully predict RI events so they both have no SEDI scores. Only the CMA has some forecasting skill (the SEDI score is 0.30) at 72 h. These results show that the three centers are not skillful in forecasting RI events more than three days in advance. Interestingly, for RI, the ratio of hits to misses seems very low, suggesting that the centers forecast RI infrequently. Take the CMA 24 h forecasts for example, the ratio is about 0.04, which is consistent with that by Na et al. (2018) (see their Fig. 1c). The high number of misses means that the RI cases observed are generally underestimated in the forecasts.

In addition to considering RIs, the RW event forecasting skills of the three centers are also evaluated. The results are summarized in Table 7 and Fig. 14b. For short lead times (24 h and 48 h), the three centers all have high hit rates. However, the hit rate declines continuously as the lead time becomes longer, resulting in SEDI scores that decline as the lead time increases. The forecasting skills of the three centers are nearly the same at lead times of less than 48 h. At 72–120 h, the CMA has the best performance and highest SEDI score. The JTWC has the lowest forecasting SEDI score because of its low hit rate.

Combining Figs. 14a and 14b reveals that for short-time forecasts (24–48 h, which is the most important time for early warning information issuance), the SEDI scores are higher for RW events than for RI events. The main factor limiting RI event forecasting skill is the low hit rate (< 0.1). The three centers issue very few forecasts, including RI events more than 72 h in advance. The situation is somewhat better for RW events, which can sometimes be forecasted five days in advance. Of the three centers, the CMA performs best in RW forecasting, followed by the RSMC-Tokyo and JTWC.

5. Conclusions

An extensive and comprehensive evaluation of homogeneous samples from the official 2005–2018 WNP TC intensity forecasts issued by the CMA, JTWC, and RSMC-Tokyo was performed. This study not only quantified accuracy metrics but also decomposed error into four components. Furthermore, the performances of the three centers were evaluated with regard to long-term intensity improvement trends and their ability to forecast RC events. The main conclusions are summarized below.

The JTWC had the highest accuracy at 24–120 h lead times. At 24 h and 48 h, the accuracy of the RSMC-Tokyo was higher than that of the CMA. At 72 h, the CMA’s accuracy exceeded that of the RSMC-Tokyo. The 72 h total MAEs of the JTWC and
RSMC-Tokyo were 13 % and 3 % less than that of the CMA, respectively.

Contrary to the prevailing view that no significant progress has been made in TC intensity forecasting, this study found evidence of slow but steady progress by the CMA, JTWC, and RSMC-Tokyo. Although the improvement rate of 24–72 h predictions was only 1–2 % per year, half of these rates were statistically significant at the 95 % confidence level. The accuracies of the JTWC forecasts improved the most at 24–72 h lead times over the period studied. The CMA and RSMC-Tokyo had the next largest improvements.

After decomposing total MAEs into four independent components (OST, UST, OWT, and UWT), the contributions of these four components to the TE and the spatial and lead time characteristics of the errors and their components were explored. In the northern waters of the South China Sea (region 1), operational forecasting centers tended to overestimate weak TCs, resulting in large errors. In the waters east of the Philippines (region 2), operational forecasting centers tended to underestimate strong TCs, and the corresponding regional error distribution was more concentrated. In the sea east of Taiwan and south of Japan (region 3), the three centers found it difficult to determine TC intensities, and forecasts were prone to misestimate. The magnitudes of the total MAEs and error components of the three centers were similar, which indicates that although the overall forecasting performance of each center was different, the three centers may face the same difficulties in forecasting TCs with different strengths.

Finally, the abilities of the three centers to forecast rapid changes in TC intensity were assessed. The SEDI scores of the three centers were higher for RW events than for RI events at 24 h and 48 h. The three centers issued very few forecasts more than 72 h in advance. Fortunately, RW events could be forecasted five days in advance. The CMA performed best in RW forecasting, followed by the RSMC-Tokyo and JTWC.

It should be noted that the verification results and conclusions of this study are based on the official TC intensity forecasts from 2005 to 2018, so they do not necessarily reflect the latest forecast performance of the three centers.

Acknowledgments

We acknowledge the CMA, JTWC, and RSMC-Tokyo for providing TC best-track datasets and Iowa State University for providing official forecast data. We thank the editor, Dr. Wataru Yanase, and two anonymous reviewers for their constructive comments and suggestions. This work was supported by the National Natural Science Foundation of China (Grant Nos. 41675058 and 41775055).

References

Andreas, E. L., L. Mahrt, and D. Vickers, 2015: An improved bulk air–sea surface flux algorithm, including spray-mediated transfer. Quart. J. Roy. Meteor. Soc., 141, 642–654.

Atkinson, G. D., 1974: Investigation of gust factors in tropical cyclones. No. FLEWEACEN/JTWC-TN-74-1, Fleet Weather Central/Joint Typhoon Warning Center, 9 pp. [Available at https://apps.dtic.mil/dtic/tr/fulltext/u2/a064535.pdf.]

Barnes, L. R., D. M. Schultz, F. D. Marks, C. E. Samsury, and H. E. Willoughby, 2002: Eastern Pacific Hurricanes Jimena of 1991 and Olivia of 1994: The effect of vertical shear on structure and intensity. Mon. Wea. Rev., 130, 2291–2312.

Black, M. L., J. F. Gamache, F. D. Marks, C. E. Samsury, and H. E. Willoughby, 2002: Eastern Pacific Hurricanes Jimena of 1991 and Olivia of 1994: The effect of vertical shear on structure and intensity. Mon. Wea. Rev., 130, 2291–2312.

Chen, G., X. Zhang, L. Bai, and R. Wan, 2019: Verification on forecasts of tropical cyclones over western North Pacific and South China Sea in 2017. Meteorol. Monthly, 45, 577–586 (in Chinese).

Chen, P. Y., H. Yu, and J. C. L. Chan, 2011: A western North Pacific tropical cyclone intensity prediction scheme. Acta Meteor. Sin., 25, 611–624.

Chen, S. S., J. A. Knaff, and F. D. Marks, 2006: Effects of vertical wind shear and storm motion on tropical cyclone rainfall asymmetries deduced from TRMM. Mon. Wea. Rev., 134, 3190–3208.

Creasey, R. L., and R. L. Elsberry, 2017: Tropical cyclone center positions from sequences of HDSS sondes deployed along high-altitude overpasses. Wea. Forecasting, 32, 317–325.

Davis, C. A., W. Wang, S. Cavallo, J. Done, J. Dudhia, S. Fredrick, J. Michalakes, G. Caldwell, T. Engel, and R. Torn, 2011: High-resolution hurricane forecasts. Comput. Sci. Eng., 13, 22–30.

DeMaria, M., J. A. Knaff, and C. Sampson, 2007: Evaluation of long-term trends in tropical cyclone intensity forecasts. Meteor. Atmos. Phys., 97, 19–28.

DeMaria, M., C. R. Sampson, J. A. Knaff, and K. D. Murphy, 2014: Is tropical cyclone intensity guidance
improving? Bull. Amer. Meteor. Soc., 95, 387–398.

Doyle, J. D., Y. Jin, R. M. Hodur, S. Chen, H. Jin, J. Moskaitis, A. Reinecke, P. Black, J. Cummings, E. Hendricks, T. Holt, C.-S. Liou, M. Peng, C. Reynolds, K. Sasaki, J. Schmidt, and S. Wang, 2012: Real-time tropical cyclone prediction using COAMPS-TC. Adv. Geosci., 28, 15–28.

Dvorak, 1984: Tropical cyclone intensity analysis using satellite data. NOAA Technical Report NESDIS, No.11, 45 pp. [Available at https://repository.library.noaa.gov/view/noaa/19322.]

Elsberry, R. L., T. D. B. Lambert, and M. A. Boothe, 2007: Accuracy of Atlantic and eastern North Pacific Tropical Cyclone intensity forecast guidance. Wea. Forecasting, 22, 747–762.

Emanuel, K. A., 1986: An air–sea interaction theory for tropical cyclones. Part I: Steady-state maintenance. J. Atmos. Sci., 43, 585–605.

Emanuel, K., C. DesAutels, C. Holloway, and R. Korty, 2004: Environmental control of tropical cyclone intensity. J. Atmos. Sci., 61, 843–858.

Ferro, C. A. T., and D. B. Stephenson, 2011: Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. Wea. Forecasting, 26, 699–713.

Fierro, A. O., R. F. Rogers, F. D. Marks, and D. S. Nolan, 2009: The impact of horizontal grid spacing on the microphysical and kinematic structures of strong tropical cyclones simulated with the WRF-ARW model. Mon. Wea. Rev., 137, 3717–3743.

Fudeyasu, H., K. Ito, and Y. Miyamoto, 2018: Characteristics of tropical cyclone rapid intensification over the western North Pacific. J. Climate, 31, 8917–8930.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. Mon. Wea. Rev., 120, 361–370.

Green, B. W., and F. Zhang, 2014: Sensitivity of tropical cyclone simulations to parametric uncertainties in air–sea fluxes and implications for parameter estimation. Mon. Wea. Rev., 142, 2290–2308.

Harper, B. A., J. D. Kepert, and J. D. Ginger, 2010: Guideline for converting between various wind averaging periods in tropical cyclone conditions. Technical Report, WMO/TD-No.1555, World Meteorological Organization, 64 pp. [Available at https://www.wmo.int/pages/prog/www/tcp/documents/WMO_TD_1555_en.pdf.]

Hendricks, E. A., M. S. Peng, B. Fu, and T. Li, 2010: Quantifying environmental control on tropical cyclone intensity change. Mon. Wea. Rev., 138, 3243–3271.

Hendricks, E. A., M. S. Peng, and T. Li, 2013: Evaluation of multiple dynamic initialization schemes for tropical cyclone prediction. Mon. Wea. Rev., 141, 4028–4048.

Holland, G. J., 1980: An analytic model of the wind and pressure profiles in hurricanes. Mon. Wea. Rev., 108, 1212–1218.

Ito, K., T. Kuroda, K. Saito, and A. Wada, 2015: Forecasting a large number of tropical cyclone intensities around Japan using a high-resolution atmosphere–ocean coupled model. Wea. Forecasting, 30, 793–808.

Jin, H., M. S. Peng, Y. Jin, and J. D. Doyle, 2014: An evaluation of the impact of horizontal resolution on tropical cyclone predictions using COAMPS-TC. Wea. Forecasting, 29, 252–270.

Japan Meteorological Agency, 2019: Annual Report on the activities of the RSMC-Tokyo Typhoon Center 2018. Japan Meteorological Agency, 118 pp. [Available at https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/AnnualReport/2018/Text/Text2018.pdf.]

Jolliffe, I. T., and D. B. Stephenson, 2011: Forecast Verification: A Practitioner’s Guide in Atmospheric Science. John Wiley & Sons, 292 pp.

Joint Typhoon Warning Center, 2019: Annual tropical cyclone report 2018. Joint Typhoon Warning Center, 147 pp.

Kaplan, J., and M. DeMaria, 2003: Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic basin. Wea. Forecasting, 18, 1093–1108.

Knapp, K. R., and M. C. Kruk, 2010: Quantifying inter-agency differences in tropical cyclone best-track wind speed estimates. Mon. Wea. Rev., 138, 1459–1473.

Koba, H., S. Osano, T. Hagisawa, S. Akashi, and T. Kikuchi, 1989: Determination of the intensity of typhoons passing over the Philippines. J. Meteor. Res. Japan, 41, 157–162 (in Japanese).

Koba, H., T. Hagisawa, S. Osano, and S. Akashi, 1991: Relationships between CI number and minimum sea level pressure/maximum wind speed of tropical cyclones. Geophys. Mag., 44, 15–25.

Kurihara, Y., M. A. Bender, and R. J. Ross, 1993: An initialization scheme of hurricane models by vortex specification. Mon. Wea. Rev., 121, 2030–2045.

Lin, I.-I., P. Black, J. F. Price, C.-Y. Yang, S. S. Chen, C.-C. Lien, P. Harr, N.-H. Chi, C.-C. Wu, and E. A. D’Asaro, 2013: An ocean coupling potential intensity index for tropical cyclones. Geophys. Res. Lett., 40, 1878–1882.

Mohapatra, M., B. K. Bandyopadhyay, and D. P. Nayak, 2013: Evaluation of operational tropical cyclone intensity forecasts over North Indian Ocean issued by India Meteorological Department. Nat. Hazards, 68, 433–451.

Moskaitis, J. R., 2008: A case study of deterministic forecast verification: Tropical cyclone intensity. Wea. Forecasting, 23, 1195–1220.

Na, W., J. L. McBride, X.-H. Zhang, and Y.-H. Duan, 2018: Understanding biases in tropical cyclone intensity forecast error. Wea. Forecasting, 33, 129–138.

Nakazawa, T., and S. Hoshino, 2009: Intercomparison of Dvorak parameters in the tropical cyclone datasets over the western North Pacific. SOLA, 5, 33–36.

Nolan, D. S., J. A. Zhang, and D. P. Stern, 2009: Evaluation
of planetary boundary layer parameterizations in tropical cyclones by comparison of in situ observations and high-resolution simulations of Hurricane Isabel (2003). Part I: Initialization, maximum winds, and the outer-core boundary layer. *Mon. Wea. Rev.*, **137**, 3651–3674.

Olander, T. L., and C. S. Velden, 2007: The advanced Dvorak technique: Continued development of an objective scheme to estimate tropical cyclone intensity using geostationary infrared satellite imagery. *Wea. Forecasting*, **22**, 287–298.

Ono, M., S. Notsuhara, J. Fukuda, Y. Igarashi, and K. Bessho, 2019: Operational use of the typhoon intensity forecasting scheme based on SHIPS (TIFS) and commencement of five-day tropical cyclone intensity forecasts. Technical Review, No. 21, RSMC Tokyo – Typhoon Center, 17 pp. [Available at https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/techrev/text21-2.pdf.]

Peng, X., J. Fei, X. Huang, and X. Cheng, 2017: Evaluation and error analysis of official forecasts of tropical cyclones during 2005–14 over the western North Pacific. Part I: Storm tracks. *Wea. Forecasting*, **32**, 689–712.

Pu, Z.-X., and S. A. Braun, 2001: Evaluation of bogus vortex techniques with four-dimensional variational data assimilation. *Mon. Wea. Rev.*, **129**, 2023–2039.

Qian, C. H., Y. H. Duan, and S. H. Ma, 2012: The current status and future development of China operational typhoon forecasting and its key technologies. *Adv. Meteor. Sci. Technol.*, **2**, 36–43 (in Chinese).

Shay, L. K., G. J. Goni, and P. G. Black, 2000: Effects of a warm oceanic feature on Hurricane Opal. *Mon. Wea. Rev.*, **128**, 1366–1383.

Song, J.-J., Y. Wang, and L. Wu, 2010: Trend discrepancies among three best track data sets of western North Pacific tropical cyclones. *J. Geophys. Res.*, **115**, D12128, doi:10.1029/2009JD013058.

Tang, B., and K. Emanuel, 2012: Sensitivity of tropical cyclone intensity to ventilation in an axisymmetric model. *J. Atmos. Sci.*, **69**, 2394–2413.

Velden, C. S., T. L. Olander, and R. M. Zehr, 1998: Development of an objective scheme to estimate tropical cyclone intensity from digital geostationary satellite infrared imagery. *Wea. Forecasting*, **13**, 172–186.

Wang, Y., and C.-C. Wu, 2004: Current understanding of tropical cyclone structure and intensity changes—A review. *Meteor. Atmos. Physics*, **87**, 257–278.

World Meteorological Organization, 2012: *Proceedings of the international workshop on satellite analysis of tropical cyclones*. WMO Report, TCP-No. 52, 99 pp.

World Meteorological Organization, 2013: *Verification methods for tropical cyclone forecasts*. WWRP 2013-7, 98 pp.

World Meteorological Organization, 2017: *Global guide to tropical cyclone forecasting*. WMO-No. 1194, 399 pp [Available at https://cyclone.wmo.int/pdf/Global-Guide-to-Tropical-Cyclone-Forecasting.pdf.]

World Meteorological Organization, 2019: *Typhoon Committee operational manual. Meteorological component*. WMO/TD-No. 196, 115 pp.

Xu, Y., L. Zhang, and C. Xiang, 2015: Typhoon intensity estimation technique and its operational application: With example of Dvorak technique. *Adv. Meteor. Sci. Technol.*, **5**, 22–34 (in Chinese).

Yamaguchi, M., H. Owada, U. Shimada, M. Sawada, T. Iriuchi, K. D. Musgrave, and M. DeMaria, 2018: Tropical cyclone intensity prediction in the western North Pacific basin using SHIPS and JMA/GSM. *SOLA*, **14**, 138–143.

Yu, H., S. T. Chan, B. Brown, M. Kunitsugu, E. Fukada, S. Park, W. Lee, Y. Xu, P. Phalla, B. Sysouphanthavong, S. W. Chang, C. G. Ismail, R. B. Paciente, P. Ee, C. Kaeochada, and H. V. Van, 2012: Operational tropical cyclone forecast verification practice in the western North Pacific region. *Trop. Cyclone Res. Rev.*, **1**, 361–372.