Evaluation of Tropical Cyclone Intensity Forecasts from Five Global Ensemble Prediction Systems During 2015–2019

XIN Jia-jie (辛佳洁), YU Hui (余 喜) 1, CHEN Pei-yan (陈佩燕) 2, 3
(1. Chinese Academy of Meteorological Sciences, Beijing 100081 China; 2. Shanghai Typhoon Institute of China Meteorological Administration, Shanghai 200030 China; 3. Key Laboratory of Numerical Modeling for Tropical Cyclones of China Meteorological Administration, Shanghai 200030 China)

Abstract: This study presented an evaluation of tropical cyclone (TC) intensity forecasts from five global ensemble prediction systems (EPSs) during 2015–2019 in the western North Pacific region. Notable error features include the underestimation of the TC intensity by ensemble mean forecast and the under-dispersion of the probability forecasts. The root mean square errors (brier scores) of the ensemble mean (probability forecasts) generally decrease consecutively at long lead times during the five years, but fluctuate between certain values at short lead times. Positive forecast skill appeared in the most recent two years (2018–2019) at 120 h or later as compared with the climatology forecasts. However, there is no obvious improvement for the intensity change forecasts during the 5-year period, with abrupt intensity change remaining a big challenge. The probability forecasts show no skill for strong TCs at all the lead times. Among the five EPSs, ECMWF-EPS ranks the best for the intensity forecast, while NCEP-GEFS ranks the best for the intensity change forecast, according to the evaluation of ensemble mean and dispersion. As for the other probability forecast evaluation, ECMWF-EPS ranks the best at lead times shorter than 72 h, while NCEP-GEFS ranks the best later on.

Key words: tropical cyclone; intensity; ensemble forecast; evaluation; intensity change
CLC number: P457.8 Document code: A
https://doi.org/10.46267/j.1006-8775.2021.020

1 INTRODUCTION

Tropical cyclones (TCs) are one of the disasters with the most severe impacts to life and properties. In recent years, economies in coastal areas have become increasingly prosperous, and the economic losses caused by TCs are increasing rapidly (Chen et al. [3]; Wu et al. [3]; Yu and Chen [4]). The prediction of TC track and intensity always draws significant attention due to its importance in TC disaster preparedness and prevention. In the past decades, the capability of TC track forecasts has been significantly improved, while the progress in the skill of intensity forecasts is relatively slow (Alley et al. [5]; Chen et al. [6]).

The significant progress of TC track forecast capability is mainly related to the development of numerical weather prediction (NWP) techniques. However, current NWP models cannot accurately describe the core structures of TCs or the physical processes related to intensity changes, resulting in the dominance of statistical and statistical-dynamic models utilizing climatology and / or persistence factors as predictors (Zhang et al. [7]), including the linear statistical model, nonlinear statistical prediction model, and statistical-dynamic models.

The accuracy of TC intensity prediction is closely related to the configuration of the NWP (Aksoy et al. [8]; Zhang et al. [9]; Cavallo et al. [10]), the initial intensity (Emanuel and Zhang [11, 12]; Bhatia et al. [13]; Qin et al. [14]), and the different synoptic-scale environmental parameters (Yu et al. [15]). Compared with deterministic forecasts, ensemble forecasts contain multiple possible future developments that may increase the predictability of the atmosphere and the reliability of the forecasts, which is of great significance for disaster prevention and reduction in operational predictions (Bauer et al. [15]; Titley et al. [17]). There is a recent trend for the ensemble prediction systems (EPSs) being utilized more and more often in operational weather forecasting and warning practice, including TC track forecasting (Qian et al. [18]; Zhang and Yu [19]).

However, how the EPS can help improve TC intensity forecast skill remains a question that needs further study. To address this issue, verification of the TC intensity forecast from the main EPSs has gradually been carried out in recent years. According to Yang et
al. [20], the EPS of European Centre for Medium-Range Weather Forecasts (ECMWF-EPS) generally underestimates TC intensity during 2009 and 2015 and the accuracy of forecasts gradually decreases with an increase in TC intensity. Hodges et al. [21] found that the mean error (ME) in 0–7 days of the ensemble mean intensity forecast of the UK Met Office EPS (UKMO-EPS) has been greater than that of the deterministic forecast and the ensemble is severely underdispersive. Chen et al. [22] compared the forecast performance between the new and old versions of the National Centers for Environmental Prediction Global Ensemble Forecast System (NCEP-GEFS) and found that the new generation of NCEP-GEFS has significantly higher forecast accuracy in both track and intensity predictions. Chen et al. [23] evaluated the TC intensity forecasts in 2016 from several EPSs, including the ECMWF-EPS, the Japan Meteorological Agency Weekly EPS (JMA-WEPS), the JMA Typhoon EPS (JMA-TEPS), the Meteorological Service of Canada Canada Ensemble System (MSC-CENS), and the NCEP-GEFS, and they found that the Brier score (BS) differences among different EPSs narrowed at longer lead times. Magnusson et al. [24] gave a comparison between the 18 km operational ensemble and a 5 km ensemble intensity forecast for Irma of ECMWF, and the results show that the forecast error of each member with higher horizontal resolution is significantly smaller.

Above listed research and several others not mentioned about here [Lei et al. [25]; Yu et al. [26]] focused on either one specific EPS or one year performance of several EPSs. Furthermore, the efforts have never been stopped to improve the EPS at major forecasting centers. This paper aims to present updated information on the performance of TC intensity forecast from several major global EPSs for the most recent five years (2015–2019) in the western North Pacific (WNP) region. Special attention will be paid to the abrupt intensity change process and very strong TCs, both of which have been recognized as major challenges in operational forecast [Demaria et al. [27]; Na et al. [28]]. Section 2 will introduce the five global EPSs being evaluated and the sources of data. The performance of the EPSs will be evaluated from both the deterministic (ensemble mean) and probabilistic (ensemble spread) aspects in Section 3 and Section 4, respectively. Conclusions and discussions will be presented in Section 5.

2 DESCRIPTION OF THE EPSS AND DATA

2.1 Basic information of the global EPSs

A total of five global EPSs are selected for this study, including the ECMWF-EPS, Japan Meteorological Agency Global Ensemble Forecast System (JMA-GEFS) (JMA-WEPS for data before 2017, JMA-GEFS for 2017 and after), MSC-CENS, NCEP-GEFS, and UKMO-EPS. Some basic information of these EPSs has been summarized in Table 1, including the major updates over the past decade since 2015.

ECMWF-EPS has 51 ensemble members, with the horizontal resolution improved from 32 km to 18 km in 2016 [Haiden et al. [29]], and 91 vertical levels. The initial perturbation method is a combination of ensemble data assimilation with a singular vector (EDA-SV) [Buizza et al. [30, 31]] and the stochastic perturbed parametrization tendencies-stochastic kinetic energy backscatter (SPPT-SKEB) method [Palmer et al. [32]; Berner et al. [33]].

<table>
<thead>
<tr>
<th>Name</th>
<th>Horizontal resolution (km) / vertical levels (year.month)</th>
<th>Ensemble size</th>
<th>Initial perturbation</th>
<th>Model uncertainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECMWF-EPS</td>
<td>32/91 (2015–2016.3)</td>
<td>51</td>
<td>EDA-SV</td>
<td>SPPT-SKEB</td>
</tr>
<tr>
<td></td>
<td>18/91 (2016.3–2019)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JMA-GEFS</td>
<td>40/60 (2015–2017.6)</td>
<td>27</td>
<td>SV</td>
<td>SPPT</td>
</tr>
<tr>
<td></td>
<td>40/100 (2017.6–2019)</td>
<td></td>
<td>SVs-LETKF</td>
<td></td>
</tr>
<tr>
<td></td>
<td>66/40 (2015–2015.11)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSC-CENS</td>
<td>50/40 (2015.11–2018.9)</td>
<td>21</td>
<td>EnKF</td>
<td>SPPT-SKEB</td>
</tr>
<tr>
<td>NCEP-GEFS</td>
<td>34/64 (2015–2019)</td>
<td>21</td>
<td>EnKF</td>
<td>STTP</td>
</tr>
<tr>
<td></td>
<td>30/70 (2017.7–2019)</td>
<td>36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

JMA-GEPS has a horizontal resolution of 40 km with 27 ensemble members from 2015 to 2019. Its vertical levels increased from 60 to 100 in 2017, together with a change of initial perturbation scheme from SV to SV-local ensemble transform Kalman filter (LETKF) [Hunt et al. [34]]. The model uncertainty

Table 1. Evolution of the five EPSs in the past five years.
method is SPPT. The horizontal resolution of MSC-CENS updated from 66 km to 50 km in 2015 and 39 km in 2018, respectively, with the vertical levels increased from 40 to 45 in 2018. It has 21 ensemble members generated by the ensemble Kalman filter (EnKF) (Whitaker et al. [35]) and SPPT-SKEB methods.

NCEP-GEFS has no major updates in 2015–2019. It has 21 ensemble members, and its horizontal resolution (vertical levels) is 34 km (64), respectively. The initial perturbation method is EnKF, and the model uncertainty method is stochastic total tendency perturbation (STTP) (Hou et al. [36]).

UKMO-GEPS has 36 ensemble members and a horizontal resolution of 50 km since 2015 with 70 vertical levels (Walters et al. [37]); the horizontal resolution changed to 30 km in 2017. The initial perturbation method is the ensemble transform Kalman filter (ETKF) (Bowler et al. [38]) and the model uncertainty method is the SPPT-SKEB.

2.2 Data description

All the TCs reaching or above tropical depression (TD) intensity in the WNP region from 2015 to 2019 are considered in the study, with the intensity data downloaded from the best-track database of the China Meteorological Administration (CMA) http://tddata.typhoon.org.cn/zljysj_zlhz.html (Ying et al. [39]; Lu et al. [40]). The TC intensity data include both the maximum surface wind speed \( V_{\text{max}} \) and the minimum central pressure \( P_{\text{min}} \). In most cases, the variation characteristics of \( V_{\text{max}} \) and \( P_{\text{min}} \) during 2015–2019 are basically the same, and \( V_{\text{max}} \) is given more attention to compare with \( P_{\text{min}} \). Therefore, the evaluation of strong TCs and intensity change is implemented for \( V_{\text{max}} \) and the results for \( V_{\text{max}} \) will be presented if not stated specifically.

TC intensity forecast data of the five global EPSs are from the THORPEX Interactive Grand Global Ensemble (TIGGE) database (http://tparc.mri-jma.go.jp/cxml/data/dataset/). Forecasts with at least four ensemble members and corresponding best-track data are evaluated for a lead time of 168 h (7 days) at a 12-h interval. The total sample size of the five EPSs ranges between 1500 and 2500 for 24 h \( V_{\text{max}} \) forecasts in each year, which decreases with lead time to be between 200 and 800 for 168 h forecasts (Fig. 1a). The sample size for each EPS is different but most of them varies between 200 (50) and 600 (150) for the 24 h (Fig. 1b) (168 h) forecasts each year.

A statistical probabilistic intensity forecast scheme named by probability climatology-based intensity forecast (PCIF) is selected as a reference method to understand the skill of the EPSs as relative to climatology-based forecasts. The PCIF scheme is proposed by Chen et al. [45] and it is a probabilistic baseline approach used by the CMA to assess the skill of TC intensity probability forecasts. The PCIF scheme operates by searching for analogous historical cases that meet the criteria developed from climatological persistence predictors. The ensemble size of PCIF varies between 20 and 51, which is determined by the sample size of historical TCs meeting the requirement of PCIF.

![Figure 1](image_url)

**Figure 1.** The sample size of the five EPSs for \( V_{\text{max}} \) verification at different lead times: (a) Total sample size for all the five EPSs at different lead times in each year; (b) Annual sample size of each EPS for 24 h forecasts.

3 PERFORMANCE OF THE ENSEMBLE MEAN

3.1 Verification methods

The ensemble mean refers to the forecast result of the equally weighted average of the ensemble members. Here, we select ME and Root Mean Square Error (RMSE) to measure the difference between the ensemble mean forecasts and observations as defined in Equations (1) and (2). ME is the abbreviation for mean error or bias. RMSE is the abbreviation for root mean square error. Skill score (SK) defined in Equation (3) represents the forecast skill of the ensemble mean of EPSs relative to that of PCIF.
According to the mean MEs of 2019 (Fig. 1m s$^{-1}$), the difference between the largest ME and the second stage occurred in 2017 (Fig. 2c), it can be seen that the W_MEs for intensity change ($\Delta V_{\max}$) keep almost stable at 24 h during the five years, with a value varying between −3 to −1 m s$^{-1}$. The W_MEs are outstanding in 2016 with a more severe underestimation at longer lead times. In the other four years, especially in 2018 and 2019, the underestimation of intensity change becomes weaker with the lead time and turns to be overestimation at longer lead times. A closer look at the MEs of 2018–2019 (Fig. 2d) shows that the MSC-CENS has no bias for intensity change at 12 h, better than the other four EPSs which all have negative bias. The largest negative bias at 12 h occurs for ECMWF-EPS and NCEP-GEFS (−3.5 m s$^{-1}$). All the five EPSs have MEs for intensity change increasing with lead time, which become positive at certain changing point. For example, the change point of ECMWF-EPS is 120 h. ECMWF-EPS underestimates the intensity change before that time but overestimates after that. The W_MEs enlarge from −2.0 to 6.9 m s$^{-1}$ during the 168 h with the changing point at around 72 h. The difference between the largest and smallest MEs becomes larger with lead time from 12 h to 72 h (7 m s$^{-1}$), but turns to be stable around 7–8 m s$^{-1}$ after 72 h.
Based on previous verification studies, it is quite normal for an EPS to have a significant underestimation of TC intensity at short lead time, in accordance with the results for 2018–2019 here. However, the weighted mean bias is very weak at short lead times during 2015–2017 (Fig. 2a). Further examination on each EPS (Fig. 3a) tells us that the MEs of NCEP-GEFS and MSC-CENS in 2015–2017 are abnormally large and there is an inconsistency between $V_{\text{max}}$ and $P_{\text{min}}$ (Fig. 3b) as a result of a poor representation of the wind-pressure relationship of these two EPSs during 2015–2017 (Fig. 3c, d). The significant bias in the wind-pressure relationship has been adjusted during 2018–2019.

**Figure 2.** ME of ensemble mean forecasts for $V_{\text{max}}$ (a, b) and the change of $V_{\text{max}}$ (c, d): (a, c) W_ME during 2015–2019; (b, d) W_ME (black dots) and ME of each EPS (colored lines) during 2018–2019. The gray bars represent the ranges of maximum and minimum MEs.

**Figure 3.** ME of the ensemble mean forecasts at 24 h: (a) $V_{\text{max}}$; (b) $P_{\text{min}}$ and the wind-pressure relationships at 72 h during 2015–2019: (c) NCEP-GEFS; (d) MSC-CENS. The black dots and colored curves in (a, b) are similar to those in Fig. 2b. The gray dots in (c, d) represent observations.
3.3 RMSE for intensity and intensity changes

RMSE is another index frequently used in the forecast verification. From Fig. 4a, it can be seen that the largest improvement occurs at 168 h with W_RMSE of \( V_{max} \) consecutively decreasing from 18 m/s\(^2\) in 2015 to about 6 m/s\(^2\) in 2019, the decreasing rate is about 2.672 m/s\(^2\) per year. However, the W_RMSE at 24 h fluctuates between 11 m/s\(^2\) and 16.4 m/s\(^2\) in 2015 and 2019.

During 2018–2019 (Fig. 4b), the ECMWF-EPS has the best performance at all the lead times except for 168 h, with RMSEs varying between 7 and 12 m/s\(^2\) at different lead times. The difference between the largest RMSE and the smallest RMSE of the five EPSs decreases with the lead time from 10 m/s\(^2\) to 5 m/s\(^2\). The UKMO-EPS is the one with RMSEs closest to the weighted mean values. The skill relative to climate forecast (Fig. 5a) shows that there is a significant negative skill of all the five EPSs at short lead times, which turns to be positive at longer lead times. The weighted mean skill changes from negative to positive at 120 h. The ECMWF-EPS has the largest skill among the five EPSs at lead times shorter than 72 h. After 72 h, the NCEP-GEFS is the most skillful one. SK of ECMWF-EPS and NCEP-EPS changes from negative to positive at 60 and 72 h, respectively.

As for the intensity change during 2015–2019, the W_RMSEs at 24 h are significantly smaller than those at 72, 120 and 168 h in all the five years, fluctuating between 7 m/s\(^2\) and 11 m/s\(^2\). The W_RMSEs at 72 h and above are close to each other, being larger than 12 m/s\(^2\) for most of the time with a maximum value reaching 19 m/s\(^2\), comparable to the W_RMSEs for \( V_{max} \) (Fig. 4a). The NCEP-GEFS has the best performance for intensity change during 2018–2019 (Fig. 4d), except for the short lead times at 12 and 24 h. Its RMSEs increase quickly from 6.9 m/s\(^2\) (12 h) to 12.0 m/s\(^2\) (72 h) and remain stable after 72 h with a slight increase to 13.3 m/s\(^2\) at 168 h. The performance of different EPSs is quite close to each other at lead times shorter than 36 h and then the discrepancy in maximum and minimum RMSEs enlarges significantly to be about 8 m/s\(^2\) at 144 h. In Fig. 5b, the W_SKS of intensity change are positive before 48 h and turn to negative after 48 h. The NCEP-GEFS (ECMWF-EPS) has better forecast skill than the weighted mean skill and the positive forecast skills continue until 132 (96) h.

![Figure 4](https://example.com/fig4.png)

Figure 4. The same as Fig. 2, but for the RMSE.
A big challenge of TC intensity forecast is the abrupt intensity change. Thus, further analyses are performed on the forecasts for different rates of intensity change in 2018–2019 (Fig. 6). It can be seen that the bias is generally weak (−3–1 m s⁻¹) when the observed intensity change is between 5–5 m s⁻¹. The RMSEs for that group are less than 7.5 m s⁻¹ at all the lead times, smaller than those for larger intensity change. The EPSs tend to underestimate the intensity change when the observed intensity change exceeds 5 m s⁻¹, which becomes worse with the increase of the intensity change value.

If an increase of \( V_{max} \) beyond 20 m s⁻¹ over a 24 h period is defined to be a rapid intensification (RI) case, the value of \( W_{ME} \) and the \( W_{RMSE} \) of the intensity change are similar at 0–24 h. For example, the \( W_{ME} \) of 20–25 m s⁻¹ band is about −21 m s⁻¹ with RMSE being similar. For weakening TCs, there are also an overestimation of intensity change values but at a lesser extent as compared with the intensifying cases. For intensity change at different lead times, the difference of the \( W_{MEs} \) (\( W_{RMSEs} \)) between different lead times is generally smaller than 5 (10) m s⁻¹ when the observed intensity change is less than 20 m s⁻¹. It enlarges to more than 10 (20) m s⁻¹ when the observed intensity change is larger than 20 m s⁻¹, and the error at longer lead times is smaller (figures not shown). Such a result is in accordance with the better performance of \( V_{max} \) at longer lead times as shown by Fig. 2b and Fig. 4b.

![Figure 5](image1.png)

Figure 5. SK of the ensemble mean intensity forecasts: (a) \( V_{max} \); (b) change of \( V_{max} \). The black dots and colored curves are similar to those in Fig. 2b.

![Figure 6](image2.png)

Figure 6. The performance of the ensemble mean for different rates of intensity change at 0–24 h in 2018–2019: (a) ME; (b) RMSE. The black dots and colored curves are similar to those in Fig. 2b.

4 PERFORMANCE OF THE PROBABILITY FORECASTS

4.1 Verification methods

Probabilistic forecast evaluations include BS, Brier skill score (BSS), the ratio of ensemble spread and RMSE (SPR/RMSE), reliability diagram, and relative operating characteristics (ROC) diagram (Nurmi[2]).
This article evaluates the probability forecast for each category of TCs, as shown in Table 2, in which $V_{max}$ follows the national standard of China ICS 07.060. Unless it is specified as a certain category, the evaluation result is the average of the evaluation values of the 6 categories. For example, the BS evaluation result we finally gave is the average of BS value of 6 levels of TD, TS, STS, TY, STY and Super TY.

<table>
<thead>
<tr>
<th>Category</th>
<th>Abbr</th>
<th>$V_{max}$ (m s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tropical depression</td>
<td>TD</td>
<td>10.8–17.1</td>
</tr>
<tr>
<td>Tropical storm</td>
<td>TS</td>
<td>17.2–24.4</td>
</tr>
<tr>
<td>Severe tropical storm</td>
<td>STS</td>
<td>24.5–32.6</td>
</tr>
<tr>
<td>Typhoon</td>
<td>TY</td>
<td>32.7–41.4</td>
</tr>
<tr>
<td>Severe typhoon</td>
<td>STY</td>
<td>41.5–50.9</td>
</tr>
<tr>
<td>Super typhoon</td>
<td>SuperTY</td>
<td>$\geq 51.0$</td>
</tr>
</tbody>
</table>

The BS is one of the most commonly used probabilistic forecast verification methods, and its calculation formula is as follows:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (P_i - Q)^2$$

(5)

where $N$ is the total number of ensemble members, $P_i$ is the forecast probability of ensemble members and $Q$ is the observation probability. When an event occurs, $Q$ is set to 1, and when it does not occur, $Q$ is set to 0. For example, for the BS evaluation of the TD level, a certain EPS member forecasts the TD level at a lead time, but the observation is not the TD level. At this situation, $P_i$ is 1, and $Q$ is 0. Therefore, the lower the BS is, the more accurate the probability forecast is (Wilks [43]).

The BS is one of the most commonly used probabilistic forecast verification methods, and its calculation formula is as follows:

$$BS = 1 - \frac{BS}{BS_c}$$

(6)

where BS represents the BS value of the climate forecast, which is PCIF here.

The ensemble spread (SPR) as defined in Equation (7) can describe the uncertainty in forecasts. Its magnitude is equivalent to the RMSE and the ratio SPR/RMSE is usually used to measure whether the ensemble spread is reasonable (Ma et al. [44]). An ideal EPS is expected to have SPR/RMSE close to 1. If SPR/RMSE is smaller than 1, the ensemble forecast is underdispersive, and vice versa.

$$SPR = \frac{1}{N} \sum_{i=1}^{N} (F_i - FM)^2$$

(7)

where $N$ is the total number of ensemble members, $F_i$ is the forecast value of each member, and $FM$ is the forecast value of ensemble mean.

Reliability diagrams are used to measure the degree of matching between forecast probability and observation frequency. The $x$-axis and $y$-axis of a reliability diagram represent forecast probability and observation frequency (%), respectively. The forecast probability can be set to multiple intervals between 0–1. In this article, we calculate the observation frequency of samples in the interval [0, 0.0–0.1, 0.2–0.3, ... 0.9–1.0] for the prediction probability. If the forecast probability is consistent with the frequency of the event, the plot points will be distributed at 45° on the diagram, which indicates that the results of the probabilistic forecast system are credible.

The ROC chart is an effective way to compare the skill of a probabilistic forecast with random forecasts. The abscissa and ordinate of the ROC chart represent the false rate and hit rate of a forecast, respectively. The calculation formulas for hit rate and false rate are as shown in Equation (8) and (9) respectively, where $a$ is the number of hits reports, $c$ is the number of missed reports, and $b$ is the number of false reports. We set some thresholds artificially and the ones set in this article are $0, 0.1, 0.2...1.0$. When the probability of a forecast is greater than a certain threshold, we consider the forecast to occur. If it is less than the threshold, it will not happen, and the hit rate and false alarm rate can be calculated. Each threshold can determine a point in the ROC chart. If the verification result is located in the lower-left corner of the chart, the system fails to predict the event; if the result is on the 45° diagonal, no forecast skills are indicated; if the verification result is in the upper-left corner, all forecasts are hits, and there is no false report. The closer the verification result is to the upper-left corner of the chart, the higher the skill of the forecast is. The area of the ROC curve and the $x=0, x=1, y=0$ (AUC) can be used to represent the level of the forecast skills. In a perfect forecast, the AUC value equals 1.

$$H = a/(a + c)$$

(8)

$$F = b/(a + b)$$

(9)

4.2 Brier score

According to Fig. 7a, W_BS generally decreases consecutively at long lead times, with an improving rate of 0.0144 at 168 h per year. However, it fluctuates between 0.18 and 0.22 at 24 h during 2015–2019. These are similar to the results of W_RMSE. During 2018–2019, maximum W_BS is 0.19 at 12 h and the minimum value is 0.13 at 156 h. The gaps between the five EPSs are the largest at 0 h, implying a significant discrepancy in initial fields of different systems. The difference between the maximum and minimum BSs reaches more than 0.1. Such a discrepancy reduces quickly with the lead time being less than 0.02 after 72 h. The ECMWF-EPS and NCEP-GEFS have the smallest BSs before and after 84 h, respectively, being 0.05–0.6 less than the W_BSs (Fig. 7b). Fig. 7c provides the skills of the probability forecasts relative to the climate forecasts at different lead times during 2018–2019. It can be seen...
that the W_BSS is basically negative at lead times shorter than 120 h and becomes positive after that. ECMWF-EPS has the biggest BSSs at 0–72 h and a small difference exists among the five EPSs for the forecasts longer than 72 h.

The BS and BSS for TCs at different intensity levels are further analyzed for 2018 and 2019, with a special attention on the strong TCs. According to Fig. 8a, TD and STS have the largest sample size which decreases with the lead time from 1500 (0 h) to 200 (168 h); it is the opposite for Super TY. In Fig. 8b, the Super TY and STY have the smallest W BS, and the largest W_BS values are associated with the STS. This result may be caused by the small number of samples of STY and Super TY, and most of the samples are TCs with weaker intensity. The accuracy rate is higher when determining whether a weak TC sample is STY or Super TY, which leads to the lowest W_BS values of STY and Super TY. But it does not indicate that the EPS has strong probability forecast ability for strong TCs.

For Super TY (Fig. 8d), the gap between the five EPSs is small, and the JMA-GEPS has the smallest BS among them. But for STY (Fig. 8c), the difference between the five EPSs is larger at longer lead times; JMA-GEPS and NCEP-GEFS have the best probability forecast accuracy at 0–84 h and 84–168 h, respectively. From the BSS in Fig. 9a, it can be seen that the W_BSSs at different levels have an increasing trend at shorter lead times and are all smaller than 0 at 0–48 h. Although Super TY and STY have the highest probability forecast accuracy, the W_BSSs of TY, STY and Super TY are negative at each lead time in 2018–2019, which means the forecast skills are inferior to the climate forecast, especially for Super TY. And the W_BSSs of TD (TS/STS) turn to positive at 50 (120) h, indicating a more comparable skill to the climate forecasts at longer lead times. The BSS gaps between the five EPSs at Super TY level are small, except that the BSSs of JMA-GEPS are larger than those of other EPSs at 96–168 h (Fig. 9b).
Ensemble dispersion and reliability analyses

Improvements of the probability forecast during the five years are also clearly embodied by the SPR/RMSE, Reliability, and AUC. Only the results for 2018–2019 (Fig. 10) are presented in this section to present the most recent status. It can be seen that the W_SPR/RMSEs are between 0.13–0.45, implying a significant underdispersion. The increase of W_SPR / RMSEs with lead time indicates a lessening of underdispersion at longer lead times. The ECMWF-EPS has the most reasonable dispersion at all the lead times except for 168 h, with SPR/RMSEs varying between 0.22 and 0.57. The gaps between the five EPSs at longer lead times are larger than those at shorter lead times. The SPR/RMSEs of JMA-GEPS is close to the weighted mean values at 0–120 h and higher than the weighted mean values over

---

**Figure 8.** BS for TCs in different intensity categories in 2018–2019: (a) sample numbers in the observations; (b) W_BS; (c) BS for STY; (d) BS for Super TY. The black dots and colored curves in (c, d) are similar to those in Fig. 7b.

**Figure 9.** BSS for TCs in different intensity categories in 2018–2019: (a) W_BSS; (b) BSS for Super TY. The black dots and colored curves in (b) are similar to those in Fig. 7b.

4.3 Ensemble dispersion and reliability analyses

Improvements of the probability forecast during the five years are also clearly embodied by the SPR/RMSE, Reliability, and AUC. Only the results for 2018–2019 (Fig. 10) are presented in this section to present the most recent status. It can be seen that the W_SPR/RMSEs are between 0.13–0.45, implying a significant underdispersion. The increase of W_SPR / RMSEs with lead time indicates a lessening of underdispersion at longer lead times. The ECMWF-EPS has the most reasonable dispersion at all the lead times except for 168 h, with SPR/RMSEs varying between 0.22 and 0.57. The gaps between the five EPSs at longer lead times are larger than those at shorter lead times. The SPR/RMSEs of JMA-GEPS is close to the weighted mean values at 0–120 h and higher than the weighted mean values over
120 h with a range of 0.02–0.17.

As for the reliability in Fig. 10b, the observation frequency tends to be underestimated when the forecast probability is less than 20% and tends to be overestimated when the forecast probability is greater than 30%. The reliability increases with increasing lead times. The reliability curve of ECMWF-EPS is closer to the perfect reliability than that of other EPSs at 24 h (Fig. 10c), while no obvious difference appears at longer lead times (figures not shown). According to Fig. 10d, the W_AUCs are generally larger than 0.5, implying positive forecast skills compared with random forecasts. The W_AUCs increase from 0.61 to 0.71 during 12–132 h, and keep stable at 132–168 h. ECMWF-EPS has the largest AUCs at lead times shorter than 60 h with a range of 0.66–0.81. At 60–132 h, the NCEP-GEFS is the most skillful one with a range of 0.67–0.75. After 132 h, the JMA-GEPS has the best performance with a range of 0.76–0.78.

![Figure 10](image_url)

**Figure 10.** SPR/RMSE, reliability and AUC for 2018–2019: (a) SPR/RMSE; (b) W. Reliability; (c) Reliability at 24 h; (d) AUC. The black dots are the weighted mean. In (a, d), the gray bars represent the ranges of the five EPSs. In (b, c), the colored lines correspond to the ordinate on the left (observation frequency), and the histogram corresponds to the ordinate on the right (the number of samples).

## 5 CONCLUSIONS AND DISCUSSION

In this study, the intensity ensemble forecasts from five operational global EPSs (including the ECMWF-EPS, JMA-GEPS, MSC-CENS, NCEP-GEFS, and UKMO-EPS) were evaluated in the WNP over a 5-year period (2015–2019), with the intention of presenting the change characteristics of the weighted mean of the five EPSs in the five years at different lead times and the current performance of the five EPSs on the samples in the last two years (2018–2019) at 0–168 h lead times. The evaluation involved performance verification of both the ensemble mean and probability forecast; the abrupt intensity change performances of the ensemble mean and strong TCs performance of the probability forecast were also assessed. The conclusions and discussion are outlined below:

1. For TC intensity forecast, the weighted mean errors of ensemble mean and probability forecast, as represented by RMSE and BS, decrease consecutively at long lead times during 2015–2019, but fluctuate between certain values at short lead times. The TC intensity tends to be underestimated by the ensemble mean forecasts, and the ensemble is severely underdispersive. The degrees of underestimation and under-dispersion decrease significantly as the lead time increases. For ensemble mean and dispersion, ECMWF-EPS behaves better than other EPSs, while for the other probability
forecasts, ECMWF-EPS ranks the best at lead times shorter than 72 h, while NCEP-GEFS ranks the best later.

(2) The EPS has poor forecast skills at short lead times but becomes more comparable to the climate forecasts at longer lead times. And the forecast skill of ensemble mean and probability forecast become positive at 120 h and above in 2018–2019. This shows that the EPSs already have reference values for operational forecasts at longer lead times and are likely to make greater contributions to the improvement of TC intensity forecast capabilities in the future.

(3) As for intensity change, no obvious improvement appears in the five years. And with the lead time increases, the ME changes from underestimation to overestimation and the RMSE increases first and then remains stable in 2018–2019. Positive forecast skill only appears at shorter lead times. The difference between the ME and RMSE is small for abrupt intensity change, especially at 24 h. The error of RI shows that the ensemble mean cannot provide an accurate forecast for the appearance of RI. NCEP-GEFS has the best performance among the five EPSs in intensity change forecast in 2018–2019.

(4) TC at STY and Super TY levels has the highest forecast accuracy but has the negative forecast skill compared with climate forecast at each lead time, indicating that strong TCs are still a big challenge for EPS to forecast.

As indicated in the introduction, the evaluations can help forecasters make better use of the EPS. Encouraging results of the performance improvement in TC intensity for EPSs have been obtained during the past few years, which may be attributed to the development of high-resolution EPSs (Magnusson et al. [24], Zhang [40]). However, the performances in abrupt intensity changes and strong TC are not yet satisfactory. In this article, only a simple preliminary assessment of the intensity changes has been carried out. Classification of intensity changes and further examination of the abrupt intensity change in probability forecast are needed. Moreover, the influence of model update on intensity forecast is also desired.

It can be seen from the evaluation results that obvious systematic deviations exist in the intensity forecast of the current EPSs, so we can make post-correction to the ensemble forecast to improve the TC intensity forecast ability. In future plans, the correlation between the ensemble mean error and the characteristics of the TC itself, the large-scale environmental field, and the underlying surface characteristics will be analyzed. Furthermore, the statistical correction scheme for the TC intensity ensemble forecast may also be investigated to further improve the TC intensity forecast skill.

REFERENCES


