Use of the JMA Ensemble Prediction System for Tropical Cyclone Intensity Forecasting

P.Y. Chen* & S.T. Chan


*Shanghai Typhoon Institute, Shanghai, China
Use of the JMA Ensemble Prediction System for Tropical Cyclone Intensity Forecasting

Chen Pei-yan ¹ and Chan Sai-tick ²

¹ Shanghai Typhoon Institute, Shanghai, China
² Hong Kong Observatory, Hong Kong, China

Abstract

In studying the performance of the Japan Meteorological Agency (JMA)'s ensemble prediction system (EPS) in the prediction of tropical cyclone (TC) intensity, it was found that the simple ensemble mean forecasts demonstrated skills in both short and medium range. However, the extent of intensity change as predicted by the EPS was in general smaller than observed.

A procedure based on an artificial neural network (ANN) to calibrate the simple ensemble mean of EPS forecasts is presented in this paper. The procedure successfully reduced TC intensity forecast error in the first 120 hours by more than 50% and 20% respectively in terms of the root mean square errors of minimum pressure and maximum wind at the TC centre.

Another procedure to post-process the probability forecast of TC intensity category, based on the rank histogram calibration method, is also presented. Results showed that the procedure could improve both the resolution and reliability of the forecasts. However, its benefit when compared with forecasts derived directly from the ANN-calibrated intensities was found to be only marginal.

1. Introduction

The prediction of tropical cyclone (TC) intensity remains a challenge despite advances in numerical weather prediction (NWP) capability. Currently, the key methods in use are statistical models based primarily on climatology, persistence, and synoptic-environmental parameters (DeMaria and Kaplan, 1994, 1999; Fitzpatrick, 1997; DeMaria et al., 2005; Knaff et al., 2003, 2005).

For more effective applications of NWP-based guidance in operational weather forecasting, the ensemble technique is becoming increasingly popular such as in the prediction of precipitation and temperature. However, in terms of TC forecasting, the focus is still very much on track and motion prediction. So far, only Weber (2005) has presented a probabilistic prediction of TC intensity using a multi-model ensemble approach.

In this study, we assess the performance of TC intensity forecasts obtained from the One-week Ensemble Prediction System (EPS) operated by the Japan
Meteorological Agency (JMA), and develop procedures to calibrate the deterministic and probability forecasts derived from the system. In summary, the JMA EPS is a low-resolution version of the JMA's Global Spectral Model (GSM) bearing the same dynamical framework and physical processes as GSM except for the horizontal resolution. It runs up to 9 days ahead for medium-range forecasting. The ensemble size including the control run is 25 (expanded to 51 in 2006). The system specifications are as shown Table 1. More details can be found in Japan Meteorological Agency (2002). Under a cooperative research arrangement set up in 2004, the Hong Kong Observatory (HKO) started receiving on a regular basis from JMA the TC position and intensity predictions from all members of the EPS to study their utilization and performance.

Table 1 - Specifications of JMA’s One-week EPS

<table>
<thead>
<tr>
<th>EPS model</th>
<th>JMA global spectral model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency of operation</td>
<td>Once every day at 12 UTC</td>
</tr>
<tr>
<td>Forecast range</td>
<td>216 hours</td>
</tr>
<tr>
<td>Ensemble size</td>
<td>25</td>
</tr>
<tr>
<td>Integration domain</td>
<td>Global from surface to 0.4 hPa</td>
</tr>
<tr>
<td>Horizontal resolution</td>
<td>T106, about 1.125 degree Guassian grid</td>
</tr>
<tr>
<td>Vertical levels</td>
<td>40</td>
</tr>
<tr>
<td>Perturbation generator</td>
<td>Breeding of Growing Modes (BGM) method</td>
</tr>
<tr>
<td>Perturbed area</td>
<td>The Northern Hemisphere and the tropics (20S-90N)</td>
</tr>
</tbody>
</table>

Section 2 of this paper introduces the datasets used in the study. In Section 3, the performance of the EPS intensity forecasts is discussed. Calibration of the deterministic forecasts derived from the EPS is presented in Section 4; while the probabilistic approach is explored in Section 5. Final conclusions and discussion are given in Section 6.

2. Dataset

JMA EPS TC data used in this study runs from 2003 to 2005. The datasets contain forecasts of TC intensity, i.e. minimum pressure (in hPa) and maximum wind speed (in knots, or kt), at the TC centre from each of the 25 ensemble members. The EPS runs were initialized at 12 UTC, with the forecast data output at 6 hourly intervals up to a maximum range of 216 hours. The number of samples in the JMA EPS TC datasets are shown in Fig. 1.
For the purpose of evaluating the JMA EPS performance in TC intensity forecasts in Section 3, HKO’s best track (BT) intensity data are taken as the “ground truth” in the verification process. For the purpose of calibrating the intensity forecasts using an artificial neural network (ANN) in Section 4, the samples in 2003-2004 are used as the training data, and the samples in 2005 are used as an independent dataset.

3. Performance of model intensity forecasts

Since the ensemble mean forecast tends to filter out the components of the forecast that are uncertain, in general it performs better than the control or individual member forecasts. Here we take the ensemble mean intensity (EMI), comprising the ensemble mean maximum wind speed (EMW) and ensemble mean minimum pressure (EMP), as the deterministic forecasts derived from the EPS. Table 2 summarizes the root mean square errors (RMSE) of EMW and EMP:

Table 2 - RMSE of EMW and EMP for various forecast hours during 2003-2004 and 2005 (in parentheses).

<table>
<thead>
<tr>
<th>Forecast Hour</th>
<th>RMSE for EMW (kt)</th>
<th>RMSE for EMP (hPa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T+24 hour</td>
<td>19.2 (23.4)</td>
<td>35.0 (38.4)</td>
</tr>
<tr>
<td>T+48 hour</td>
<td>22.7 (27.0)</td>
<td>39.1 (41.6)</td>
</tr>
<tr>
<td>T+72 hour</td>
<td>25.7 (28.9)</td>
<td>42.6 (43.3)</td>
</tr>
<tr>
<td>T+96 hour</td>
<td>26.6 (29.2)</td>
<td>43.7 (43.0)</td>
</tr>
<tr>
<td>T+120 hour</td>
<td>26.3 (25.1)</td>
<td>42.5 (38.3)</td>
</tr>
</tbody>
</table>

As evident in the mean error plot in Fig. 5, the EPS significantly underestimated the TC intensity across the whole forecast range. The EMI errors can be attributed to two factors: initialization error and forecast error.
a. Initialization error

Given the sparse observations over the oceans, TC structure cannot be adequately resolved in global models. Many NWP centers employ a “bogussing” scheme to force a tropical cyclone vortex into the numerical analysis. For JMA, a TC bogus is constructed from a standard axisymmetric vortex for well-developed tropical cyclones based on several manually-analyzed parameters such as cyclone position, central pressure and radius of gale force wind (Ueno, 1995). Fig. 2 shows the error of EMI at analysis time for the 2003-2005 dataset. In general, the more intense the cyclone (x-axis), the larger are the initial EMI errors both in terms of wind and pressure (y-axis). The initial EMI minimum pressure is predominantly higher than the BT-analyzed minimum pressure for the whole range of TC intensities. For maximum wind, the EMI wind speeds in most cases are larger than the BT-analyzed values for TC of sub-typhoon strength, but are smaller than the BT-analyzed values for most typhoon cases.

![Fig. 2 - Initial errors of the ensemble mean intensity of the JMA EPS TC datasets in 2003-2005](image)

b. Forecast error

Scatter diagrams of BT intensity changes and EMI intensity changes for all samples during 2003-2005 are shown in Fig. 3. JMA EPS predictions demonstrate skills in forecasting the trend of TC intensity changes. Nevertheless, as shown by the bold line of linear regression against the perfect diagonal, the extent of changes has generally been under-predicted by the EPS, particularly for TC cases with significant weakening or intensification.
Fig. 3 - Scatter diagram of BT intensity changes and EPS forecast intensity changes for all samples during 2003-2005: (a) minimum pressure; (b) maximum wind.

4. Calibration of deterministic forecasts

If the initialization errors are corrected, the RMSE of EMW for 2005 at T+24, T+48, T+72, T+96, and T+120 hour forecasts would become 16.5, 23.7, 27.9, 31.8, and 32.4 kt respectively. As shown in Fig. 4(c), removing the initialization errors could reduce the EMW errors in the first 78 hours, whereas error reduction in EMP can be achieved all the way up to T+120 hour.

To cater for the non-linearity of the forecast errors and the correlated nature of the two intensity parameters (minimum pressure and maximum wind speed), a commercially available statistical software with a radial basis function artificial neural network (ANN) (Broomhead and Lowe, 1988; Haykin, 1994) was used to devise a calibration mapping. Radial basis function networks consist of three layers: one for the inputs, one for the outputs, and a single hidden layer in between. Each unit in the hidden layer is represented by a radial basis function. The output units then complete the computation based on a weighted sum of results generated by all hidden units. The excellent approximation capabilities of radial basis function networks have been demonstrated by Park and Sandberg (1991), Poggio and Girosi (1990).

In the construction of the ANN for TC intensity change, the following input parameters are used: (a) forecast hour; (b) initial BT minimum pressure; (c) initial BT maximum wind speed; (d) change of EMP during the forecast period concerned; and (e) change of EMW during the forecast period concerned. The BT initial intensity is used in order to remove the initialization errors. There are two output nodes in the ANN, namely BT-analysed change in minimum pressure and that in maximum wind speed.
The verification results are given in Fig. 4. The ANN successfully reduced both the mean errors of EMP and EMW, especially in the short to medium-range where the bias reductions reached 36 hPa and 15 kt respectively (not shown). This successful bias reduction led to significant improvement in the RMSE as depicted in Fig. 4(c) and (d). The improvements in the first 120 hours exceeded 50% in terms of
the RMSE of minimum pressure and over 20% in terms of the RMSE of maximum wind speed.

By comparison, the performance of ANN at longer forecast range was far less satisfactory. The decrease in the skill of the underlying model could be one reason; the other reason could be due to insufficient training data as the number of samples dropped rapidly at longer forecast range (Fig. 1).

Besides, the calibrated intensity forecasts were still not quite able to forecast the rapid change of TC intensity (both deepening or weakening) as evident in Fig. 4(a) and (b). This could be attributed to model limitations in adequately resolving the TC structure with a coarse grid spacing of 1.125 degrees (i.e. around 120 km).

5. Probabilistic approach

A suite of methods have since been developed for calibrating probability forecasts derived from ensemble systems, such as multiple implementation of single-integration MOS equations (Erickson, 1996), ensemble dressing (Roulston and Smith, 2003) and logistic regression methods (Hamill et al. 2004). Hamill and Colucci (1997, 1998) described a rank histogram calibration based on the reliability of past forecasts. This method has been applied to temperature and precipitation forecasting. In this study, the same approach was tested to post-process the probability forecast of TC intensity category, according to the following classification:

- Low system (LOW) – maximum wind < 22 kt
- Tropical Depression (TD) – 22 kt ≤ maximum wind < 34 kt
- Tropical Storm (TS) – 34 kt ≤ maximum wind < 48 kt
- Severe Tropical Storm (STS) – 48 kt ≤ maximum wind < 64 kt
- Typhoon (TY) – 64 kt ≤ maximum wind

In view of the lack of samples and relatively unsatisfactory performance of the calibrated intensity forecasts in the longer forecast range as discussed in Section 4, probability forecasts as explored in this study will be confined to the first 120 hours.

a. Methodology

The rank histogram calibration method can be divided into two steps: bias correction and calibration.

1) Bias correction

Before constructing the rank histogram, each member forecast was first de-biased to remove any systematic errors in the maximum wind forecasts. Two different methods have been tested. The first method was simple bias removal. The correction \( corr \) to be made was determined as follows:

\[
corr_i = \sum_{j=1}^{n} (OBSW_{i,j} - EMW_{i,j}) 
\]  

where \( i \) is the forecast range (6, 12, ..., 120-h), \( n \) the number of samples, OBSW the BT maximum wind speed and EMW the ensemble mean of the maximum wind speed.
Another method was to correct all member forecasts using the ANN approach described in Section 4. The mean errors of the direct EPS outputs and the corrected member forecasts in 2005 are plotted in Fig. 5.

The mean errors after correction using both methods described above are much reduced, falling within -5 to +5 kt for the whole forecast range.

Fig. 5 - The mean errors of the maximum wind speed forecasts in 2005. DMO: the maximum wind speed as derived from the direct EPS outputs; ANN: correction by ANN described in Section 4; and SBR: correction by simple bias removal.

2) Calibration

The bias correction described above effectively removed the systematic biases but the corrected probability forecasts might still not be reliable. Following Hamill and Colucci (1997), the probability distribution was calibrated using the verification rank histogram. The rank histogram consisting of 24 bins de-limited by the 25 members’ forecasts of maximum wind were sorted in numerical order, with two outlier bins placed at both ends. It was constructed by counting the number of verifying observations falling within each bin. The relative frequency in each bin was then used as the weight to calibrate the probability forecasts.

Unlike temperature or precipitation forecasts in which each EPS member would always output a forecast, some members would not provide any forecast if the TC was forecast to dissipate. Besides, the verifying BT dataset from HKO did not contain any information for LOW (i.e. systems with maximum wind less than 22 kt). For cases when both forecast and observation were not available (i.e. EPS members correctly forecast the dissipation of the TC), the following procedures were adopted to assign the frequency: if there were \( m \) members in total not outputting a forecast, i.e. each of the first \( m \) bins represent a correct forecast, the frequency count will be equally assigned to these \( m \) bins. In other words, \( 1/m \) will be assigned to each of the \( m \) bins.

Data from 2003-2004 were used to construct the rank histograms at various forecast range and the weights obtained were applied to calibrate the forecasts in 2005. The rank histograms for T+24, T+72 and T+120 hours forecasts are shown in Fig. 6.
In general, the two outlier bins were most populated, and such tendency was much more prominent at shorter forecast range when the EPS spread was usually less.

With the rank histograms constructed above, the probability of forecasts was calibrated according to the following procedures.

Suppose $V$ is the verifying TC intensity and $W = \{w_1, w_2, ..., w_{26}\}$ represent the verification rank histogram distribution, i.e. the relative frequency for the first, second, ..., and the 26th bin. For a forecast quantity $q$ that are bounded by the ensemble (Smith 2001), the calibrated probability would be:

$$\Pr(V \leq q) = \sum_{j=1}^{i} w_j + w_{i+1} \frac{q - \bar{x}(i)}{\bar{x}(i+1) - \bar{x}(i)}, \quad \bar{x}(i) \leq q \leq \bar{x}(i+1) \quad \text{(2)}$$

Here the tildes denote the de-biased ensemble members, and the parenthetical subscripts indicate ranking of the ensemble members in ascending order (i.e., $x(i) \leq x(i+1) \leq x(i+2)$, etc.).

For $q$ larger or smaller than all 25 ensemble members, the probability represented in the outlier bins of the rank histogram (i.e., in $w_1$ and $w_{26}$) must be extrapolated in some way. Some assumptions are made here. First, the rank histogram probability is uniformly distributed between the lowest ensemble member and zero. The lower tail is then fitted by (Hamill and Colucci, 1997):

$$\Pr(V \leq q) = \left(\frac{q}{\bar{x}(1)}\right)w_1, \quad 0 < q < \bar{x}(1) \quad \text{(3)}$$

For the upper tail, we assume that the probability beyond the largest ensemble member follows the shape of Gaussian distribution fitted to the ensemble data:

$$\Pr(V \leq q) = \sum_{j=1}^{25} w_j + w_{26} \frac{\Phi[Z_q] - \Phi[Z_{\bar{x}(25)}]}{1 - \Phi[Z_{\bar{x}(25)}]}, \quad q > \bar{x}(25) \quad \text{(4)}$$

where $Z$ indicates standardization by subtraction of the (de-biased) ensemble mean and dividing by the ensemble standard deviation, and $\Phi[\cdot]$ represents the Gaussian cumulative distribution function.

The probability of a TC reaching a certain TC intensity category can then be obtained by replacing $q$ in Eqn (2) – (4) with the maximum wind thresholds of the respective TC intensity category.
The Brier score, BS, (Wilks, 1995) verifies the probability forecast $Y$ against the event observation $O$ ($O = 1$ if the event occurs, $O = 0$ if the event fails to occur):

$$BS = \frac{1}{n} \sum_{k=1}^{n} (y_k - o_k)^2$$

with BS ranging between 0 (perfect forecasts) and 1 (completely wrong forecasts).

Fig. 7 compares the BS for direct model (EPS) output (DMO), forecasts after simple bias correction and calibration (CAL), forecasts after ANN bias correction only (ANN), and forecasts after ANN bias correction and calibration (ANN+CAL). Improvement in the probability forecast as a result of ANN+CAL was most prominent in the TS, STS and TY categories, especially in the first 72 hours for the TY category as the model initialization errors were successfully corrected through ANN. For the LOW and TD categories, the advantage of applying ANN+CAL was not obvious, with BS even poorer than CAL in the longer range beyond T+72 hour. This could be attributed to the tendency of ANN in under-estimating TC weakening (see Fig. 4(b)). In such cases, the systematic errors could have been more effectively removed through the simple bias removal procedure in CAL. As for comparison between ANN and ANN+CAL, the latter in general achieved marginally better BS for most categories except for the STS cases.
Fig. 7 - Brier score of the probability forecasts of TC intensity in 2005: (a) LOW, (b) TD, (c) TS, (d) STS, (e) TY. DMO: direct EPS output; CAL: forecast after simple bias correction and calibration; ANN: forecast after bias correction by ANN only; ANN+CAL: forecast after bias correction by ANN and calibration.

The BS can be decomposed into three components, namely reliability, resolution and uncertainty (Wilks, 1995). Uncertainty depends only on the variability of the observations and is therefore unrelated to the forecasts. The calibration procedures, however, should lead to better BS by improving the reliability or the resolution of the forecasts. As shown in the reliability diagrams at various forecast ranges for the TY category (Fig. 8), DMO had minimal resolution with the outcome of the high probabilities forecast not quite differentiable from the outcome of the low probabilities. On the other hand, ANN and ANN+CAL significantly improved the BS by increasing (i) the resolution, as illustrated in the deeper slopes of the calibrated curves (left panel) and the increased number of forecasts of higher probabilities (right panel); and (ii) the reliability of the probability forecasts, with the points of the calibrated curves on the left panel falling closer to the diagonal line, especially at the T+24 hour range.
Fig. 8 - Reliability diagram of TC intensity probability forecasts based on TY category samples in 2005 (left panels): (a) T+24 hour; (c) T+48 hour; (e) T+72 hour. Number of samples in each probability class (right panels): (b) T+24 hour; (d) T+48 hour; (f) T+72 hour. DMO: direct EPS output; CAL: forecast after simple bias correction and calibration; ANN: forecast after ANN bias correction only; ANN+CAL: forecast after ANN bias correction and calibration.

6. Conclusion

In studying the performance of JMA EPS in TC intensity forecasts, model initialization errors were found to be highly correlated with the initial intensity of the TC. The initial maximum wind speeds from direct EPS outputs were larger than the BT intensity values for most TCs of sub-typhoon strength, but became generally smaller than BT intensity values for typhoons. Although JMA EPS demonstrated skills in forecasting the trend of intensity changes, the extent of changes predicted was far less than actual.
A procedure to calibrate the simple ensemble mean forecasts of JMA EPS using an ANN was developed. The procedure successfully reduced the forecast errors in the first 120 hours to a level with useful operational value. At the Hong Kong Observatory, the ANN calibration procedure has since been put into operational use in 2007.

A further step to post-process the JMA EPS forecasts to generate more reliable probability forecasts of TC intensity category has also been explored. The Brier score analysis showed that the rank histogram calibration procedure could bring about noticeable improvements to the predicted intensity categorization of TS, STS and TY by enhancing both the resolution and reliability of the probability forecasts. While ANN+CAL was marginally the best performer in the independent verification based on 2005 data, forecasts derived directly from the ANN-calibrated intensities delivered a comparable level of improvement in terms of forecast skill.

Acknowledgments. We are most grateful to JMA for providing the EPS data used in this study.

References

Prediction at the Japan Meteorological Agency, Japan Meteorological Agency, Tokyo.


