Progress in Earth and
Planetary Science

RESEARCH ARTICLE Open Access

# Deep learning approach for detecting tropical cyclones and their precursors in the simulation by a cloud-resolving global nonhydrostatic atmospheric model

Daisuke Matsuoka[1,2*], Masuo Nakano[3], Daisuke Sugiyama[1] and Seiichi Uchida[4]

## Abstract

We propose a deep learning approach for identifying tropical cyclones (TCs) and their precursors. Twenty year simulated outgoing longwave radiation (OLR) calculated using a cloud-resolving global atmospheric simulation is used for training two-dimensional deep convolutional neural networks (CNNs). The CNNs are trained with 50,000 TCs and their precursors and 500,000 non-TC data for binary classification. Ensemble CNN classifiers are applied to 10 year independent global OLR data for detecting precursors and TCs. The performance of the CNNs is investigated for various basins, seasons, and lead times. The CNN model successfully detects TCs and their precursors in the western North Pacific in the period from July to November with a probability of detection (*POD*) of 79.9–89.1% and a false alarm ratio (*FAR*) of 32.8–53.4%. Detection results include 91.2%, 77.8%, and 74.8% of precursors 2, 5, and 7 days before their formation, respectively, in the western North Pacific. Furthermore, although the detection performance is correlated with the amount of training data and TC lifetimes, it is possible to achieve high detectability with a *POD* exceeding 70% and a *FAR* below 50% during TC season for several ocean basins, such as the North Atlantic, with a limited sample size and short lifetime.

**Keywords:** Tropical cyclogenesis, Cloud resolving atmospheric model, Deep learning, Convolutional neural network

## Introduction

Tropical cyclones (TCs), also referred to as typhoons, cyclones, and hurricanes, cause significant damage to human life, agriculture, forestry, fisheries, and infrastructure. For example, Typhoon Lionrock in 2016 caused record-breaking heavy rainfall, which resulted in severe floods and the loss of 23 lives in Japan. Moreover, TCs occasionally form very close to and approach countries at low latitudes (e.g., the Philippines) with rapid intensification. Therefore, accurate prediction of TC track and intensity is necessary. Early prediction of TC formation is important not only from an academic but also from a disaster mitigation perspective.

TCs form from convective disturbances in the tropics (Riehl 1954). Dynamic environmental conditions (e.g., small vertical wind shear, low-level cyclonic vorticity, and non-zero planetary vorticity) and thermodynamically favorable environmental conditions (e.g., sea surface temperature > 26 °C, existence of convective instability, and mid-tropospheric moisture) necessary for TC formation were proposed in the pioneering work of Gray (1968, 1975). However, because only a small percentage of convective disturbances in the tropics develop into TCs under favorable environmental conditions (Emanuel 1989), accurate and early prediction of TC formation is still a developing area of research. The Japan Meteorological Agency (JMA) extended the Dvorak method (Dvorak 1975; Dvorak 1984), which estimates TC intensity based on satellite infrared imagery (IR), to tropical depressions (maximum sustained surface wind speed $< 17.5\ \mathrm{m\,s^{-1}}$). This extension, known as early-stage Dvorak analysis (EDA),

* Correspondence: daisuke@jamstec.go.jp
[1]Center for Earth Information Science and Technology (CEIST), Japan Agency for Marine-Earth Science and Technology (JAMSTEC), 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan
[2]PRESTO, Japan Science and Technology Agency (JST), 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan
Full list of author information is available at the end of the article

has been utilized in operational forecasts since 2001 (Tsuchiya et al. 2001), and the JMA issues early warnings on typhoon occurrence 1 day before its formation based on EDA. The National Hurricane Center (NHC) and Central Pacific Hurricane Center (CPHC) also use the advanced Dvorak method for predicting TC genesis with lead time and accuracies of 48 h and 15–57%, respectively (Cossuth et al. 2013). Yamaguchi and Koide (2017) demonstrated that that the predictability of TC genesis could be improved to 35–79% by combining the Dvorak method and multi-model ensemble forecasts. On the other hand, with recent advancements in high-performance computing and numerical weather prediction, TC formation could be simulated 2 weeks in advance in case studies of eight typhoons in August 2004 (Nakano et al. 2015), Hurricane Sandy in 2012, and Super Typhoon Haiyan in 2013 (Xiang et al. 2015).

In recent years, deep learning, a machine learning method based on neural networks, has been receiving increasing attention and is being applied to various pattern recognition tasks (Krizhevsky et al. 2012; Simonyan and Zisserman 2015). In meteorology, several studies have proposed applying deep neural networks to existing TC detection (Liu et al. 2016; Kim et al. 2017), tornado prediction (Trafalis et al. 2014), hurricane pathway prediction (Kordmahalleh et al. 2015), and extreme rain fall prediction (Gope et al. 2016). Although several studies have used deep learning approaches for TCs after their formation, no research has considered this approach for detecting TCs before their formation.

In general, there are two approaches to detecting extreme events such as TCs: the model-driven approach (deductive approach), including numerical simulation, and the data-driven approach (inductive approach), including machine learning. The model-driven approach has the limitation that the prediction error increases with lead time because numerical models are inherently dependent on initial values. On the other hand, machine learning, as a data-driven approach, requires a large amount of high-quality training data. Most related works use reanalysis data and/or satellite observational data and labeled data as TCs or precursors based on the best track data provided by meteorological agencies. However, best track data for a TC's occurrence well ahead of its formation is limited because the best track data is basically generated using the EDA technique and are limited in accuracy and elapsed time. For example, the best track data from the Regional Specialized Meteorological Center (RSMC), Tokyo, captures precursors 60 h before TC formation, on average, whereas simulation data from cloud-resolving global atmospheric models (Kodama et al. 2015) and TC tracking algorithms capture TC formation 107 h ahead, on average.

In this work, we adopted the deep learning approach to detect precursors of TCs before their formation using only two-dimensional (2D) Outgoing Longwave Radiation (OLR) data, which is equivalent to IR and is a good proxy of atmospheric deep convection and cloud cover. In our 2D deep convolutional neural networks (CNNs), we use 30 year cloud-resolving global atmospheric simulation data (20 year data for training and 10 year data for verification) and a TC tracking algorithm for automatic labeling. Although the basic concept, simulation data, and TC tracking algorithm of this work are the same as those in our previous conference paper (Matsuoka et al. 2017), the present study improves the deep learning architecture and investigates predictive ability for various basins, seasons, and elapsed times.

The manuscript is organized as follows. The "Data" section presents the climate simulation data and TC tracking algorithm. The "Method" section explains the training data preparation, deep convolutional neural networks, and evaluation metrics of prediction results. The "Results and discussion" section examines the detection results, including detectability for each ocean basin, spatial detectability, seasonal detectability, and long-term detectability. The "Conclusions" section provides a summary of the main conclusions of the present work.

## Data

### Climate simulation data

Thirty year atmospheric simulation data were produced by the Nonhydrostatic Icosahedral Atmospheric Model (NICAM) with a 14 km horizontal resolution (Kodama et al. 2015). This model employs fully compressible nonhydrostatic equations and guarantees the conservation of mass and energy. Equations were discretized by the finite volume method. One characteristic feature of this model is that it explicitly calculates deep convective circulations without using any cumulus parameterizations. Moist processes are calculated using a single-moment bulk cloud microphysics scheme (NSW6) (Tomita 2008). HadISST (Rayner et al. 2003) is used for lower boundary condition. The seasonal march of TC genesis, TC track, and TC intensity in each basin is well simulated, as described in Kodama et al. (2015). The dataset includes simulated OLR, precipitation, wind velocity, pressure, temperature, water vapor, and cloud (liquid, ice, rain, snow, and graupel) for 30 years since January 1979. OLR and precipitation are output every hour, and other physical quantities are output every 6 h. An example of simulation results of OLR is depicted in Fig. 1a. Three TCs and three precursors are reproduced at 0:00:00 08/17/2008 UTC.

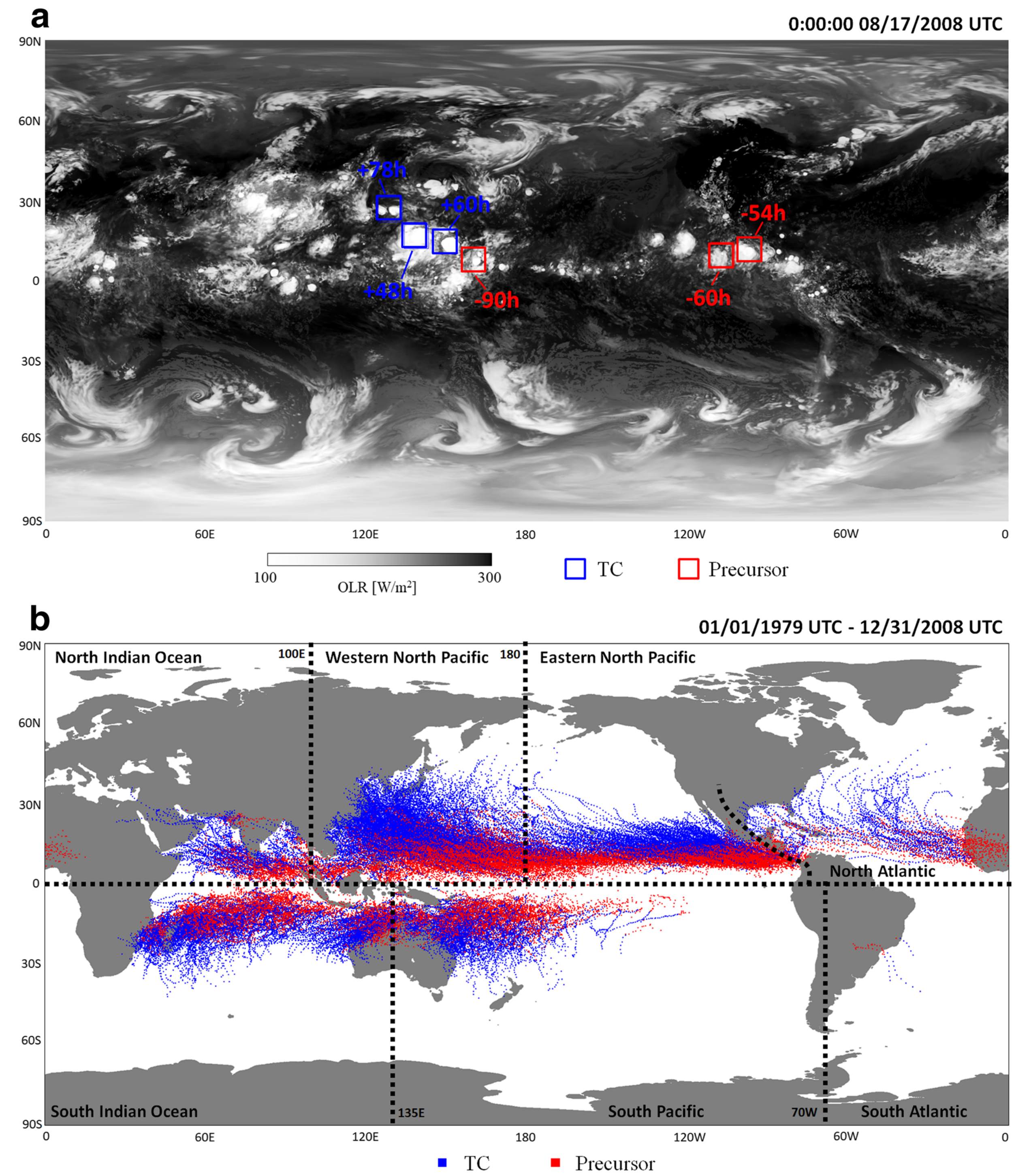

**Fig. 1 a** An example of the simulated OLR. Blue and red squares show TCs and their precursors, respectively. Numeric values next to the squares represent elapsed time (positive numbers: TCs, negative numbers: precursors). **b** Detection results of TCs (blue dots) and their precursors (red dots) during 1979–2008 from NICAM. The definition of basins is based on Fudeyasu et al. (2014)

This model is suitable for reproduction of tropical phenomena such as TCs (Nakano et al. 2015; Nakano et al. 2017a; Nakano et al. 2017b) and the Madden–Julian oscillation (MJO) (Miura et al. 2007). For additional details on this model, please see the original and review papers (Tomita and Satoh 2004; Satoh et al. 2014).

### TC tracking

To detect TCs and precursors, we employed a TC tracking algorithm for six-hourly outputs of the horizontal components of wind, temperature, and sea level pressure (SLP). This algorithm was originally proposed by Sugi et al. (2002) and optimized for NICAM data by Nakano et al. (2015) and Yamada et al. (2017). In the first step, grid points at which the SLP was 0.5 hPa less than the mean of its surrounding area (eight-neighbor grids) were selected as candidate TC centers. In this step, grid points that met the following criteria were considered as "TCs": (i) the maximum wind speed at 10 m is greater than 17.5 m/s, (ii) the maximum relative vorticity at 850 hPa is greater than $1.0 \times 10^{-3}\ \mathrm{s}^{-1}$, (iii) the sum of temperature deviations at 300, 500, and 700 hPa is greater than 2 K, (iv) the wind speed at 850 hPa is greater than that at 300 hPa, (v) the duration of each detected storm is greater than 36 h, and (vi) the TC is formed within a limited range of latitudes (30° S–30° N). In the second step, these grid points were connected with nearest neighbors in time, and tracks of "precursors" (before becoming TCs), "TCs," and "extratropical cyclones" were subsequently obtained.

Figure 1b shows the result of applying the above algorithm to the 30 year NICAM data. TCs and precursors are represented by blue and red dots. In 30 years, 2532 TCs were detected (72–103 TCs per year). The numbers of TC tracks, detected samples, positive samples in training data, and average lifetime in each ocean basin (North Indian Ocean, western North Pacific, eastern North Pacific, North Atlantic, South Indian Ocean, South Pacific, and South Atlantic) are listed in Table 1. The definition of basins is taken from Fudeyasu et al. (2014) and illustrated in Fig. 1b. Detected TCs and precursors were used not only for labeling data, but also for ground truth. Ground truth was provided beforehand as the center point of TCs and precursors to evaluate identification results.

## Methods/Experimental

### Training data preparation

In this work, we performed a CNN-based binary classification that categorizes 2D cloud data (OLR) into "developing TCs and their precursors" or "non-developing depressions." Binary classification is the task of categorizing (or classifying) objects into two groups (positive examples and negative examples) on the basis of classification rules. Before classification, it is necessary to prepare a labeled training data set, which is a series of sample data with the labels (positive or negative).

In the present work, at first glance, it appears natural to categorize the data into the following three classes: TCs, precursors, and non-developing depressions. However, since they are defined by the threshold value of maximum wind speed, it is difficult to identify them from cloud images. Figure 2a indicates a 2D histogram of cloud cover and elapsed time of all detected TCs including precursors by using the TC tracking algorithm. Here, we define cloud cover as $(\mathrm{OLR}_{\max} - \mathrm{OLR}_{\mathrm{mean}})/(\mathrm{OLR}_{\max} - \mathrm{OLR}_{\min})$, where $\mathrm{OLR}_{\mathrm{mean}}$ is the mean value of OLR in 64 × 64 grids, and $\mathrm{OLR}_{\max} = 300.0\ \mathrm{W/m}^2$ and $\mathrm{OLR}_{\min} = 100.0\ \mathrm{W/m}^2$. Figure 2a shows the developing phase and dissipation phase in which the cloud cover increases and decreases over time, respectively. All precursors (elapsed time < 0) were in the developing phase; therefore, we could identify both precursors and TCs in the developing phase (inside yellow dotted line) under the same category "developing TCs and their precursors" (referred to here as "TCs") for labeling supervised data. The range was 30.0–95.0% for cloud cover and was from 10 days before to 7 days after formation, which could cover 97.0% of all precursors of TCs (in Matsuoka et al. 2017, the cloud cover range was 30.0–90.0% and covered 92.0% of all precursors of TCs). The other category, "non-developing depressions" (hereinafter referred as "non-TCs"), are low pressure clouds that were candidates for TCs but did not satisfy criteria (i)–(vi). For the binary classification,

**Table 1** The numbers of TC tracks, detected samples, positive samples in training data, and average lifetime in each basin

| | Number of TC tracks | Average lifetime [day] | | Number of detected samples (number of positive samples in training data) | |
|---|---|---|---|---|---|
| | | TCs | Pre-TCs | TCs | Pre-TCs |
| North Indian Ocean | 169 | 4.6 | 2.9 | 3011 (1422) | 2162 (1184) |
| Western North Pacific | 754 | 6.8 | 3.1 | 21,546 (9549) | 13,514 (8023) |
| Eastern North Pacific | 589 | 4.8 | 7.8 | 11,880 (4478) | 14,392 (7976) |
| North Atlantic | 125 | 4.4 | 4.1 | 2582 (788) | 1767 (758) |
| South Indian Ocean | 525 | 5.5 | 3.7 | 7989 (4757) | 7989 (4148) |
| South Pacific | 367 | 4.2 | 4.0 | 6649 (3503) | 5346 (3193) |
| South Atlantic | 3 | 1.6 | 2.3 | 22 (15) | 27 (26) |
| Total | 2532 | 5.4 | 4.5 | 53,679 (24,512) | 25,488 (25,308) |

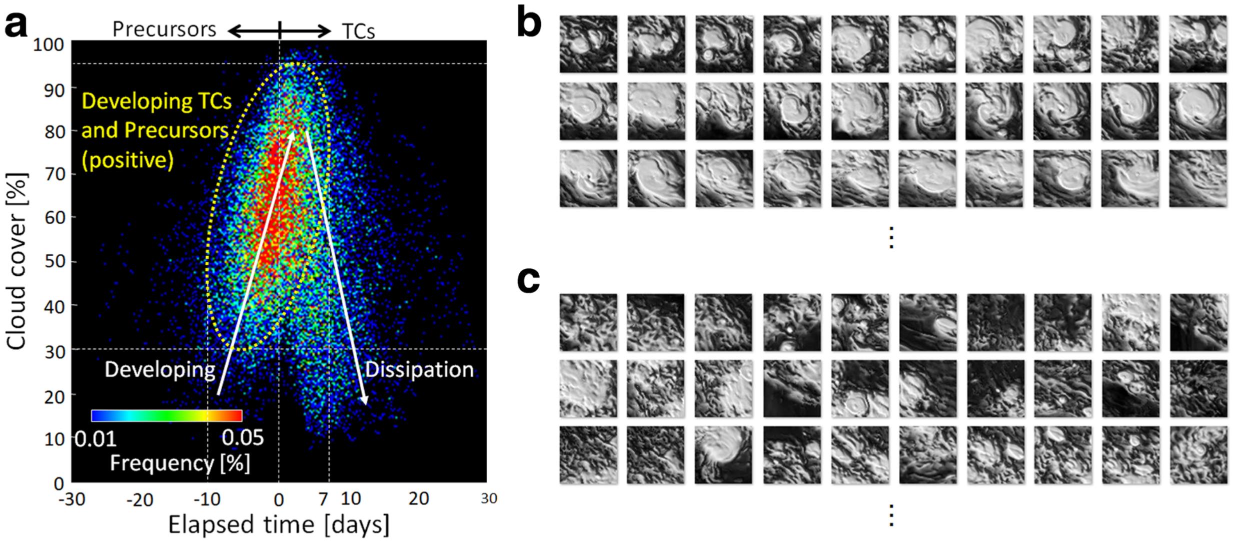

**Fig. 2 a** Histogram of cloud cover and elapsed time for developing TCs and their precursors. Positive and negative values of elapsed time indicate before formation (precursors) and after formation (TCs), respectively. **b**, **c** Examples of training data for binary classification of **b** developing TCs and precursors (positive examples) and **c** non-TCs (negative examples)

we labeled “TCs” and “non-TCs” data as “positive” and “negative” examples, respectively.

Examples of “TCs” and “non-TCs” are shown in Fig. 2b, c. Although these figures are visualized images of OLR, the actual training and test data are single-precision floating point numbers. Their horizontal sizes were approximately 1000 km × 1000 km (64 × 64 grids). For training, 20 year data (1979–1998) were used, and the remaining 10 year data (1999–2008) were used for prediction tests. The numbers of positive data (TCs and precursors) and negative data (non-TCs) for training were approximately 50,000 and 1000,000, respectively (the numbers of positive data in training data in each basin are listed in Table 1). Generally, the numbers of positive and negative data are often set to same number in binary classification. In this work, in order to train the CNN with a vast number of negative data, ten training data sets including the same 50,000 positive data and 50,000 randomly chosen negative data were generated for ten deep CNNs. By combining multiple CNNs, the influence of initial value dependence becomes smaller than when only single CNN is used (Freund and Schapire, 1997; Kearns and Valiant 1989; Breiman 1996; Breiman 2001).

## Training and prediction using deep convolutional neural networks

We used a 2D deep CNN for binary classification (Table 2). CNNs are algorithms of neural networks used for image recognition and classification and for directly learning visual patterns from images. CNNs usually consist of convolutional layers, pooling layers, and fully connected layers (LeCun et al. 1999; Krizhevsky et al. 2012). Convolutional layers extract local features (feature maps) of input images, pooling layers allow spatial invariance by reducing the resolution of the image, and fully connected layers determine which features most correlate to a particular class. Dropout is a regularization technique where randomly selected neurons are ignored during training for preventing overfitting in a neural network.

Our CNN architecture comprises four convolutional layers, three pooling layers, and three fully connected layers. Input data were 64 × 64 grids of OLR data

**Table 2** The architecture of our deep CNN. The parameters of the input layer, convolutional layers, pooling layers, and fully connected layers are denoted as [input data size] (e.g., 64 × 64), [filter size]@[number of feature maps] (e.g., 3 × 3@32), [pooling window size] (e.g., 2 × 2), and [number of units] (e.g., 2048), respectively

| Layer | Shape |
|---|---|
| Input | 64 × 64 |
| Convolution 1 | 3 × 3@32 |
| Convolution 2 | 3 × 3@64 |
| Pooling | 2 × 2 |
| Convolution 3 | 3 × 3@64 |
| Pooling | 2 × 2 |
| Convolution 4 | 7 × 7@128 |
| Pooling | 2 × 2 |
| Fully-connected | 2048 |
| Fully-connected | 2048 |
| Fully-connected | 2 |

consisting of single-precision real numbers, and output was generated in two classes (1 or positive: TCs; 0 or negative: non-TCs). Hyper parameters were optimized on the basis of the cross-validation test, which was conducted to evaluate the performance of the CNN using a random part of the training data. We examined the validation accuracy of 216 combinations of architecture settings: the number of convolutional (1–5) and pooling layers (1–5), the number of parameters in the fully connected layer (128, 256, 512, 1024, 2048), drop out ratio (0.2, 0.3, 0.4, 0.5), size of the convolutional filter (3 × 3, 5 × 5, 7 × 7), and number of feature maps (16, 32, 64, 128). Accordingly, the architecture with the highest level of performance was adopted, as shown in Table 2. The Adam optimizer (Kingma and Ba 2015) was applied to the CNN to update the network parameter to minimize the loss function called binary cross entropy. Batch normalization (Ioffe and Szegedy, 2015) was also applied to the CNN to minimize the initial-value dependence of the parameters.

The source code for deep learning was implemented in Python 3.6.3 with Keras (TensorFlow backend) (Chollet 2015), running on an NVIDIA Tesla P100 (1 node). Training 100,000 data over one epoch consumed approximately 3 min.

The accuracy of ten CNN classifiers using 100,000 data (50,000 for each of the two classes) for training and 5000 data for cross-validation is shown in Table 3. The maximum, minimum, and average values were 90.99%, 89.58%, and 90.30%, respectively (the number of epochs ranged over 19–46). The metric “Accuracy” is defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Here, TP (true positive), TN (true negative), FP (false positive), and FN (false negative) correspond to “correctly predicted positive example as positive,” “correctly predicted negative examples as negative,” “incorrectly predicted negative example as positive,” and “incorrectly predicted positive examples as negative,” respectively. Compared with the average accuracy of ten classifiers in Matsuoka et al. (2017), which was 86.60%, the average accuracy increased by 3.7 percentage point. Note that, although simulation data and the TC tracking algorithm were the same as in Matsuoka et al. 2017, in this study, the target range of cloud cover was expanded from 30.0–80.0% to 30.0–95.0%.

**Table 3** Accuracy values (also used as weights) of the ten classifiers

| Model number: i | Accuracy (weight: $w_i$) |
|---|---|
| Classifier 1 | 0.9099 |
| Classifier 2 | 0.9050 |
| Classifier 3 | 0.9013 |
| Classifier 4 | 0.9014 |
| Classifier 5 | 0.8958 |
| Classifier 6 | 0.9085 |
| Classifier 7 | 0.8987 |
| Classifier 8 | 0.9065 |
| Classifier 9 | 0.9025 |
| Classifier 10 | 0.9007 |

In the present study, in order to effectively train imbalanced data (positive 50,000, negative 1000,000), ten classifiers (*Classifier* 1, 2, …, 10) were generated by training ten sets of 100,000 data on the same neural network, as shown in Fig. 3a. Each classifier was trained with the same 50,000 positive data and randomly selected 50,000 negative data. In this manner, our CNNs could train 50,000 positive examples and 500,000 negative examples simultaneously.

To verify the model’s performance, classifiers trained using the 20 year data were applied to the test data (untrained 10 year simulation data). Candidate regions in the test data to be predicted by applying trained classifiers were clipped with a sliding window, which is widely used for object detection (Kumar 2013). We slid a rectangular area (approximately 1000 × 1000 km) with a 125 km (eight-grid) stride and continued sliding the window over the whole data within latitudes of 30° N to 30° S because three pooling layers of our CNN assumed eight grids of horizontal shift. Furthermore, in order to reduce the number of candidate regions, we set a limit to the cloud cover in the range of 30.0 to 95.0% and 50% or more over sea areas. In this manner, 97.0% of precursors of TCs in the simulation data could be covered.

Our ensemble CNNs output the ensemble average using the weight value of each trained classifier, as shown in Fig. 3b. The weight value given by the accuracy of the ten classifiers is listed in Table 3. The final probability $p$ for detecting the presence of developing TCs and precursors in an arbitrary region is defined as follows:

$$p = \frac{1}{10}\sum_{i=1}^{10}\frac{w_i x_i}{w_i} \tag{2}$$

where $w_i$ is the weight value of classifier $i$, and $x_i$ is the output value obtained by *Classifier* $i$ (0: non-TCs, 1: TC). When the threshold value $p_{\text{th}}$ is given, arbitrary candidate areas that satisfy $p \geq p_{th}$ are regarded as positive. Although we adopt binary classification to facilitate the evaluation of prediction results, we can also output detection results as probabilistic information using $p$.

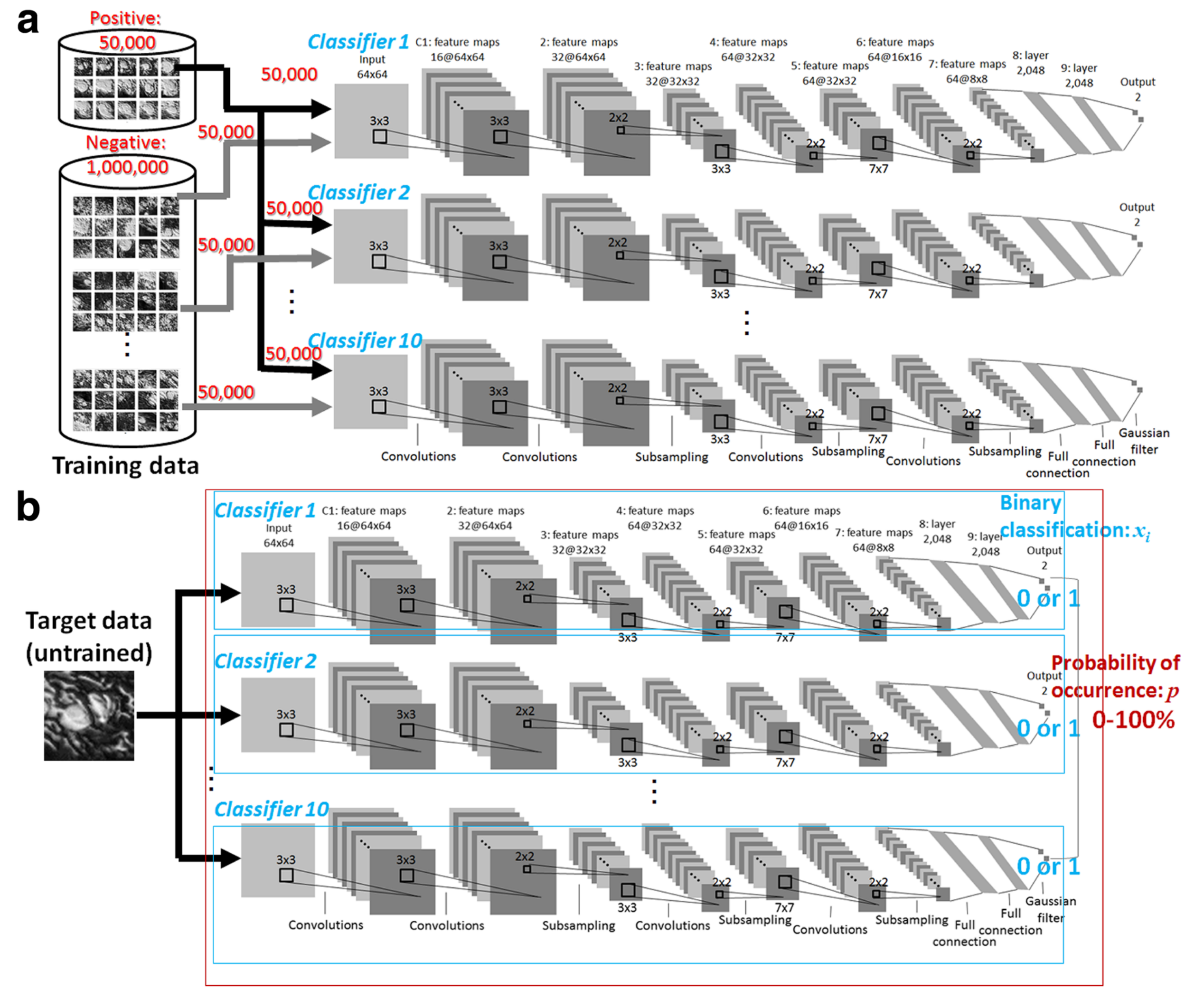

**Fig. 3** The flow charts of the proposed ensemble CNNs: **a** training and cross-validation phase and **b** prediction phase. Hyper parameters of the CNN architecture are listed in Table 2

## Evaluation metrics of prediction results

The false alarm ratio (*FAR*) and the probability of detection (*POD*) are often used as evaluation metrics of prediction results in weather forecasting (Jolliffe and Stephenson, 2003; Wilks 2006; Barnes et al. 2007). *FAR*, incorrectness of positive prediction, is defined using “correctly predicted positive examples as positive (*TP*, true positive)” and “incorrectly predicted negative examples as positive (*FP* false positive)” as follows:

$$FAR = \frac{FP}{TP + FP} \tag{3}$$

As shown in Fig. 4a, when a positive predicted area captures the ground truth, the area is *TP*. Similarly, when a positive predicted area does not capture the ground truth, the area is *FP*. In the example of Fig. 4a, *TP* is 3, *FP* is 1, and *FAR* is 25.0%. Although *FAR* is closely related to $Precision = TP/(TP + FP)$, which is widely used in computer vision and pattern recognition (Forsyth 2011), *Precision* is one minus *FAR* and means correctness of positive prediction.

*POD* is another important metric in prediction; it indicates the amount of ground truth that can be correctly predicted. *POD* is conceptually the same as $Recall = TP/(TP + FN)$ used in computer vision except for cases in which multiple positive predicted areas overlap, as shown in Fig. 4b. This is because the denominator of *POD* is the value of the ground truth, whereas the denominator of *Recall* is the number of predicted areas corresponding to the ground truth. Therefore, *TP* is given to TCs instead of prediction area, and we define the *POD* at multiple areas with the same detected ground truth as follows:

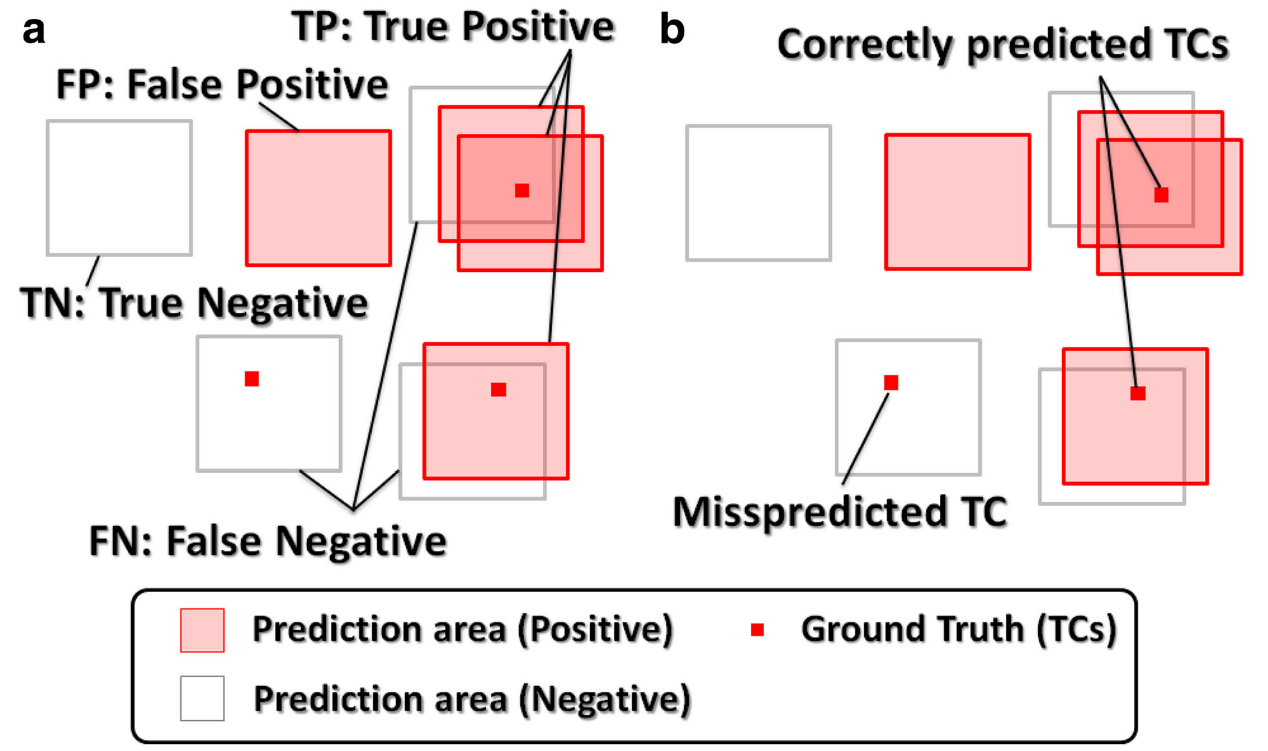

**Fig. 4** Conceptual diagram explaining our evaluation metrics: **a** accuracy and *FAR* and **b** *POD*

$$POD = \frac{\text{number of correctly predicted TCs}}{\text{number of TCs}} \quad (4)$$

Here, the numbers of both TCs and correctly predicted TCs include precursors as mentioned above.

## Results and discussion

### Detection results

This section first introduces one of the best cases of detection results under the condition that the number of TCs and precursors is larger than eight and *POD* is larger than 80.0%. Similarly, one of the worst cases under the condition that the number of TCs and precursors is larger than five is also introduced.

Figure 5 shows the best case of detection results during the 10 year period (October 21, 2003, 18:00:00 UTC) for (a) $p_{th} = 1.0$ and (b) $p_{th} = 0.6$. Red and white boxes represent positive predicted areas (TCs) and negative predicted areas (non-TCs), respectively. Furthermore, blue and red dots represent the central points of actual TCs and precursors (as ground truth) calculated by the TC tracking algorithm, respectively. In Fig. 5a, five developing TCs and three precursors of nine ground truths can be correctly predicted; hence *POD* is 88.9% (= 8/9). Furthermore, 74 of 82 positive prediction areas could be correctly predicted; hence, *FAR* is 9.8% (= 1–74/82). Figure 5b shows the prediction results after decreasing the threshold $p_{th}$ to 0.6. In this case, the correctly predicted area (true positive) increases (*POD* is 100.0%) because the positive predicted area is expanded. However, the false alarm rate also increases (*FAR* is 34.1%).

Representing the worst case, prediction results of August 17, 2006 18:00:00 UTC are shown in Fig. 6, in which many TCs and precursors with less cloud cover were missed and there are numerous false alarms. While the *POD*s range from 20 to 60%, the *FAR*s were high (72.7 to 78.3%).

### Detectability for each basin

The performance for various $p_{th}$ was evaluated on each basin. Figure 8 shows the relationship between *POD* and *FAR* when $p_{th}$ was varied from 10 to 100% in 10% increments. It represents the average value over 10 years (1999–2008) for each basin. The South Atlantic was not considered in this study because the number of TCs that occur in that area is extremely small. Although there are differences in values of detection performance depending on the basin, *POD* and *FAR* exhibited a trade-off relationship for all cases. When $p_{th}$ is increased, the positive prediction area is narrowed down, and both *FAR* and *POD* decrease. In contrast, if $p_{th}$ is decreased, the positive predicted area becomes wider, and both *FAR* and *POD* increase because dropout is reduced.

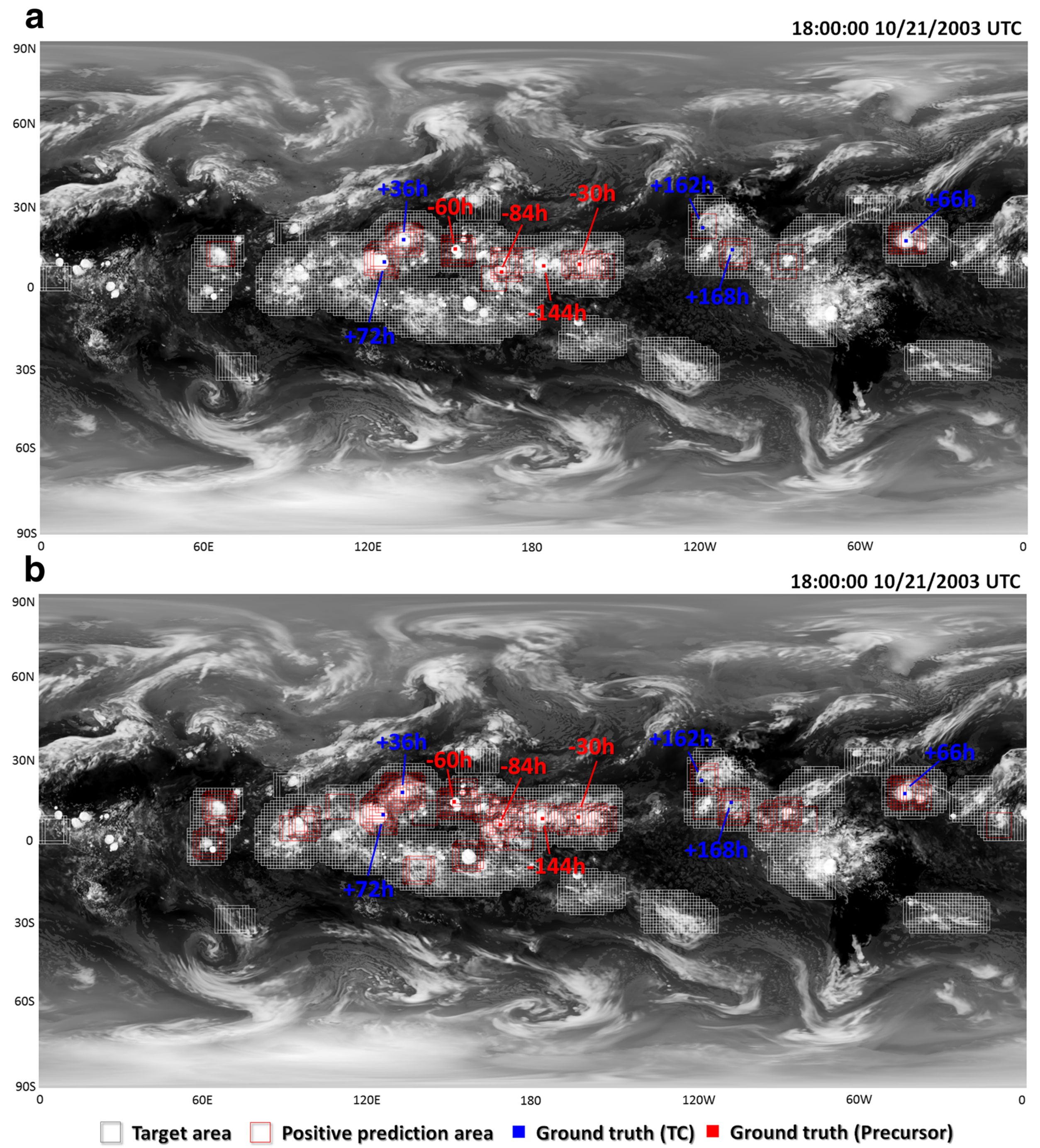

**Fig. 5** Detection results at 18:00:00 UTC 10/21/2003 with **a** $p_{\mathrm{th}} = 1.0$ (*POD* = 88.9% and *FAR* = 9.8%), and **b** $p_{\mathrm{th}} = 0.6$ (*POD* = 100% and *FAR* = 34.1%). Numbers next to the blue or red letters indicate the elapsed time of the ground truth of TC or pre-TC, respectively

There are several reasons for the variation in detection performance for different basins. First, it is known that the pattern of TC genesis is different for each basin (Holland 2008), and the detectability for each pattern may differ (as will be described in the “Conclusion” section, investigation of the detectability for each generation pattern will be undertaken in future studies). Second, in general, the performance of supervised machine learning depends on the number of training data for each pattern. In our results, the correlation coefficient between *POD* and the number of training data is 0.749, and that between the *FAR* and the number of training data is − 0.756. Lastly, since the cloud pattern of TCs is broken over time in the dissipation phase,

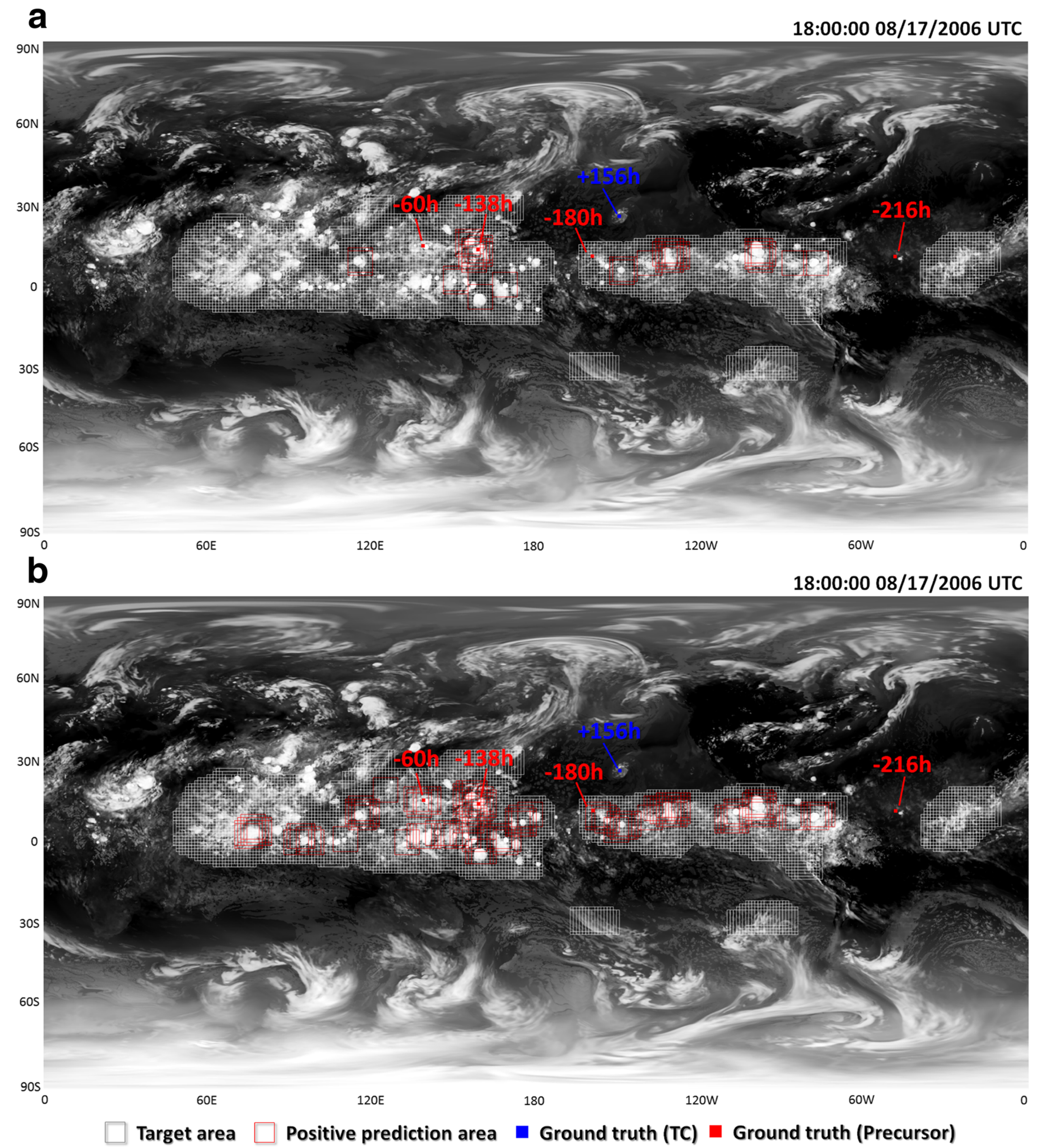

**Fig. 6** Detection results at 18:00:00 UTC 08/17/2006 with **a** $p_{th} = 1.0$ (*POD* = 20.0% and *FAR* = 72.7%) and **b** $p_{th} = 0.6$ (*POD* = 60% and *FAR* = 78.3%). Numbers next to the blue or red letters indicate the elapsed time of the ground truth of the TC or pre-TC, respectively

their detectability should decrease. In other words, the detectability should be high for TCs with long lifetimes. The correlation coefficient between *POD* and average TC lifetime in each basin is 0.821 and that between *FAR* and average lifetime is – 0.802.

For example, as seen in Fig. 7, it was found that the basin with the best detection performance was the western North Pacific and that with the worst detection performance was the North Indian Ocean. In the western North Pacific, the number of training data was the largest (TCs 9549, pre-TCs 8023) and the average lifetime was also the longest (6.8 days). On the other hand, in the North Indian Ocean, the number of training data was relatively small (TCs 1422,

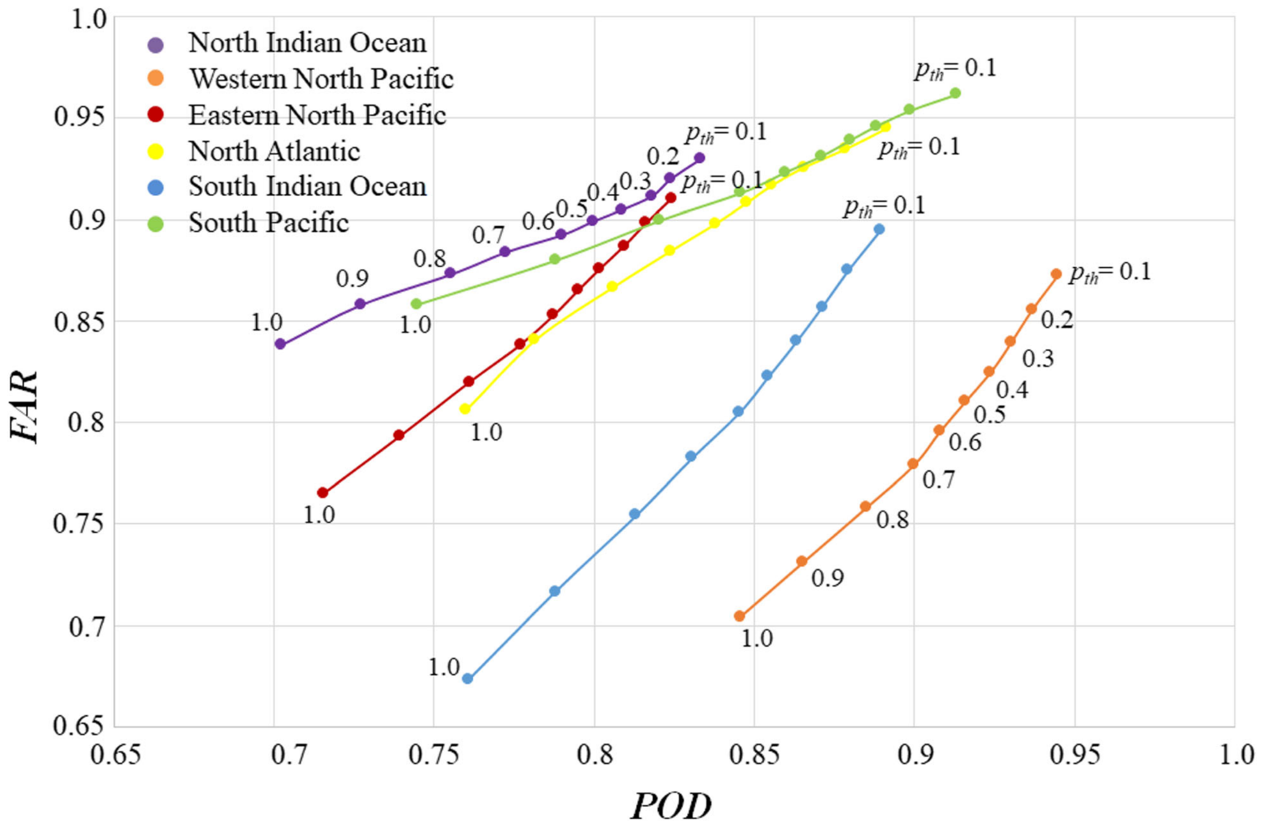

**Fig. 7** Relationship of the *POD* and *FAR* with various $p_{\mathrm{th}}$ for different basins. The *POD* and *FAR* are 10 year average values from 1999 to 2008 for each basin

pre-TCs 1184) and the average lifetime was also relatively short (4.6 days).

### Spatial detectability

This section first shows the spatial distribution of detection performance for a threshold value of $p_{\mathrm{th}} = 1.0$. Figure 8a, b shows the spatial distributions of *POD* and *FAR*, respectively. Here, the spatial distributions of *POD* and *FAR* are calculated as their 10 year means at each grid point and are defined in a 64 × 64 rectangular area centered on the ground truth. Figure 8c shows the count of ground truths for both TCs and pre-TCs in the training data at each grid point. Figure 8d shows the count of positive prediction areas at each grid point. Areas with many TCs and pre-TCs are represented by a gray and white dotted line in Fig. 8a, b, d.

In most basins, both *POD* and *FAR* appeared to be roughly associated with the ground truth count in the training data. Especially in the Indian Ocean and the Pacific, *POD* was higher and *FAR* was lower in regions that had higher ground truth counts. As an exception, *POD* also exceeds 80% and *FAR* falls to 60% in a part of the North Atlantic, despite limited training data. Previous studies reported that there is a pattern of TC genesis unique to the basin (Ritchie and Holland 1999; Yoshida and Ishikawa 2013; Fudeyasu and Yoshida, 2018). For example, Russell et al. (2017) reported that 72% of TCs in the North Atlantic are caused by African Easterly Waves (AEW). Accordingly, our results may indicate that TCs and pre-TCs caused by the AEW are easy to detect using CNNs.

Next, the detection performance of the TC area and pre-TC area are compared. In each basin, pre-TC areas are located at lower latitudes than the TC area. As shown in Fig. 8b, the *FAR* of the pre-TC area was higher than that of the TC area. Although there was no significant difference in *POD* between the TC and pre-TC areas, the count of positive predictions was larger in the pre-TC area than in the TC area. That is, the count of misdetection is larger in the pre-TC area than in the TC area. Intuitively, the pattern of developed TCs is simpler than that before formation, and therefore, it is reasonable that the detectability of TCs is higher than that of pre-TCs.

In the western South Pacific and South Atlantic, although the *FAR* was close to 100%, there were few positive prediction areas. In other words, the number of misdetections (false positives) was small. In contrast,

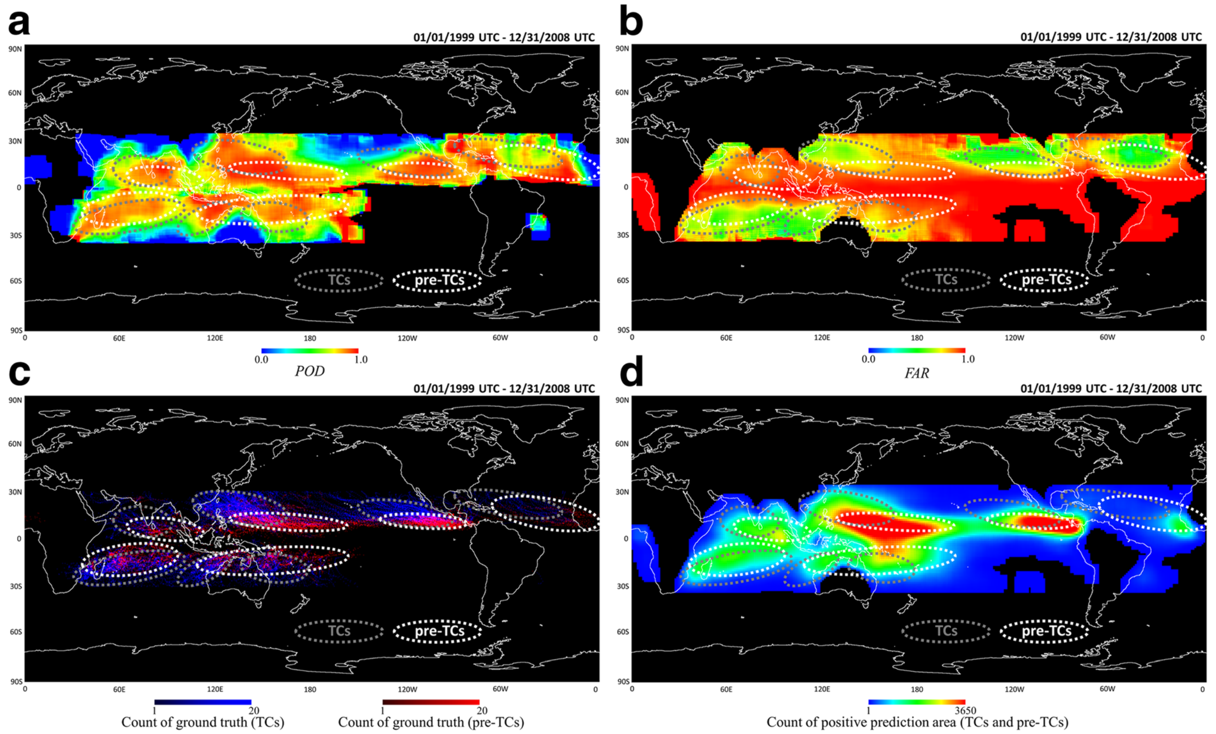
a
01/01/1999 UTC - 12/31/2008 UTC
b
01/01/1999 UTC - 12/31/2008 UTC
c
01/01/1999 UTC - 12/31/2008 UTC
d
01/01/1999 UTC - 12/31/2008 UTC
90N
60N
30N
0
30S
60S
90S
0
60E
120E
180
120W
60W
0
TCs
pre-TCs
0.0
1.0
POD
0.0
1.0
FAR
1
20
Count of ground truth (TCs)
1
20
Count of ground truth (pre-TCs)
1
3650
Count of positive prediction area (TCs and pre-TCs)

**Fig. 8** Ten year means of **a** *POD* and **b** *FAR* with $p_{th} = 1.0$, respectively. **c** Spatial distribution of the count of ground truth in the training data of TCs and precursors. **d** Spatial distribution of the count of positive prediction areas

near the equator of the Indian Ocean and the eastern Pacific, although the *FAR* was also close to 100%, the number of positive prediction areas was large. That is, the number of misdetections was large in these areas.

### Seasonal detectability

Seasonal detectability, monthly mean of *POD*, and *FAR* from 1999 to 2008 in each basin are shown in Fig. 9. Monthly variability of the number of training data (positive) and ground truth in each month and each basin are also shown in the same figure. In each basin, monthly changes in *POD* and *FAR* almost correspond to the number of training data. In other words, seasonal TCs can be detected without generating numerous false alarms. In particular, seasonal TCs in the western North Pacific (from July to November) could be detected with a *POD* of 79.0–89.1% and a *FAR* of 32.8–53.4% ($p_{th} = 1.0$). Similarly, seasonal TCs in the South Indian Ocean (from December to March) could be detected with a *POD* of 76.7–78.0% and a *FAR* of 31.1–40.3%. Furthermore, seasonal TCs in the North Atlantic (from August to November) could be detected with a *POD* of 75.0–78.2% and a *FAR* of 36.7–51.0%.

In contrast, numerous false alarms were generated during seasons with a low frequency of TCs. In the western North Pacific, although *POD* was not very low (65.7–83.0%) during January to May, the *FAR* was remarkably high (75.4–95.3%). It is noteworthy that *POD* in the North Pacific is unlikely to decrease even during seasons with a low frequency of TCs, except for months with extremely small numbers of positive ground truth. In the western North Pacific in December, the *POD* was 78.7% and the *FAR* was 65.9% ($p_{th} = 1.0$). Similarly, in the eastern North Pacific in November, the *POD* was 72.6% and the *FAR* was 68.4% ($p_{th} = 1.0$).

### Long-term detectability

Detectability (*POD*), the number of training data (positive), and ground truth of each elapsed time frame in different basins are shown in Fig. 10. In each basin, the day with the highest *POD* is the day TCs formed (elapsed time = 0–2 days). One of the reasons for this is that this period is the transition period from the developing phase to the dissipation phase for most TCs, as shown in Fig. 2a. Obviously, fully developed TCs are easy to recognize. Another reason is that the training of the CNNs tends to focus on large samples near this period to reduce errors.

The *POD* decreases as lead time is increased because it is difficult to detect the precursors many days in advance. For example, 91.2%, 77.8%, and 74.8% of precursors 2, 5, and 7 days before their formation can be detected in the western North Pacific, respectively (Fig. 10b). Similarly, 91.7%, 76.2%, and 70.1% of

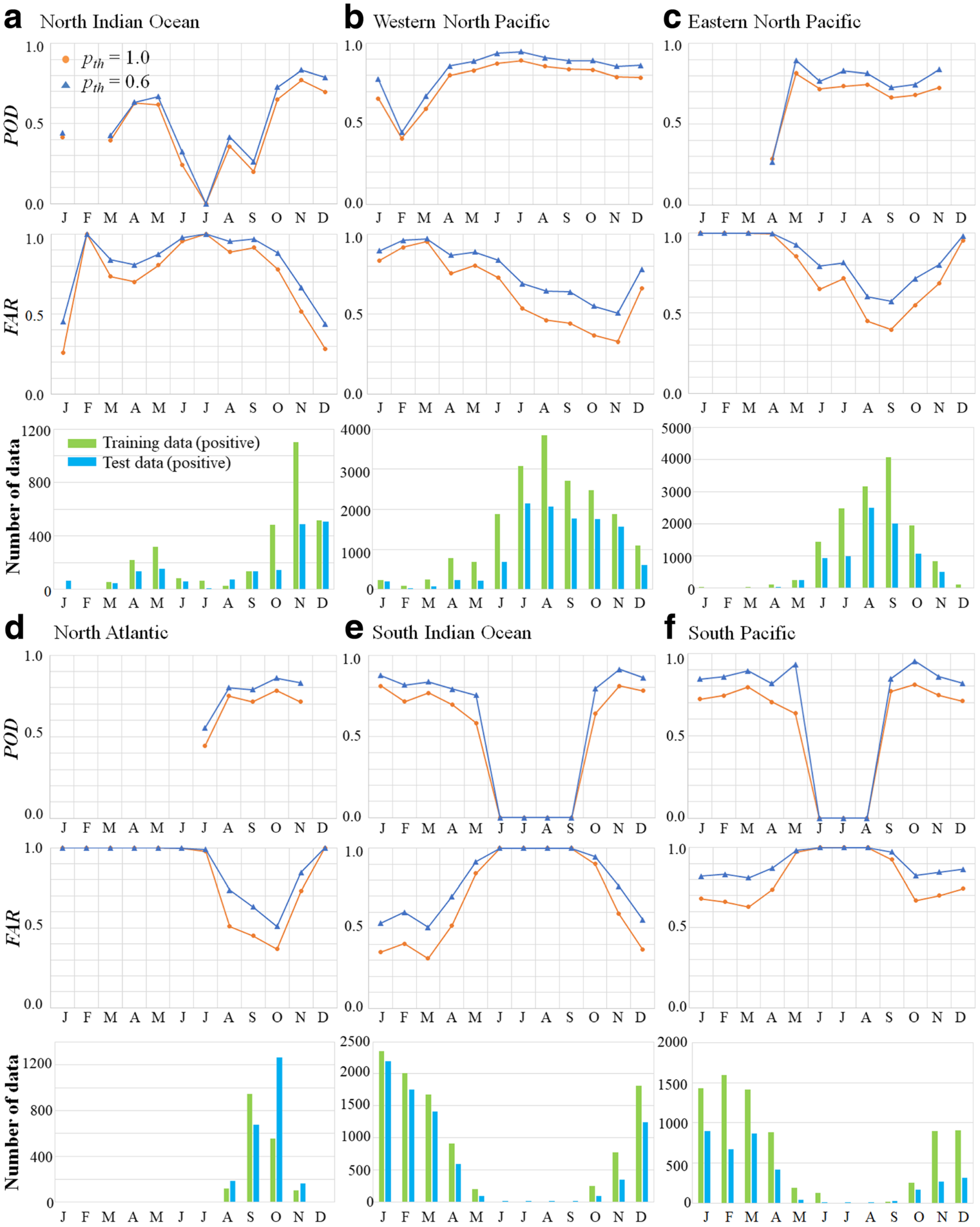

**Fig. 9** *POD* and *FAR* ($p_{\mathrm{th}} = 1.0$ and $p_{\mathrm{th}} = 0.6$) and the number of data (training and test data) for each month in each basin. Note that the *POD* cannot be defined when the number of TCs is zero according to Eq. (2). In the same manner, the *FAR* cannot be calculated when the positive prediction area is zero, according to Eq. (3)

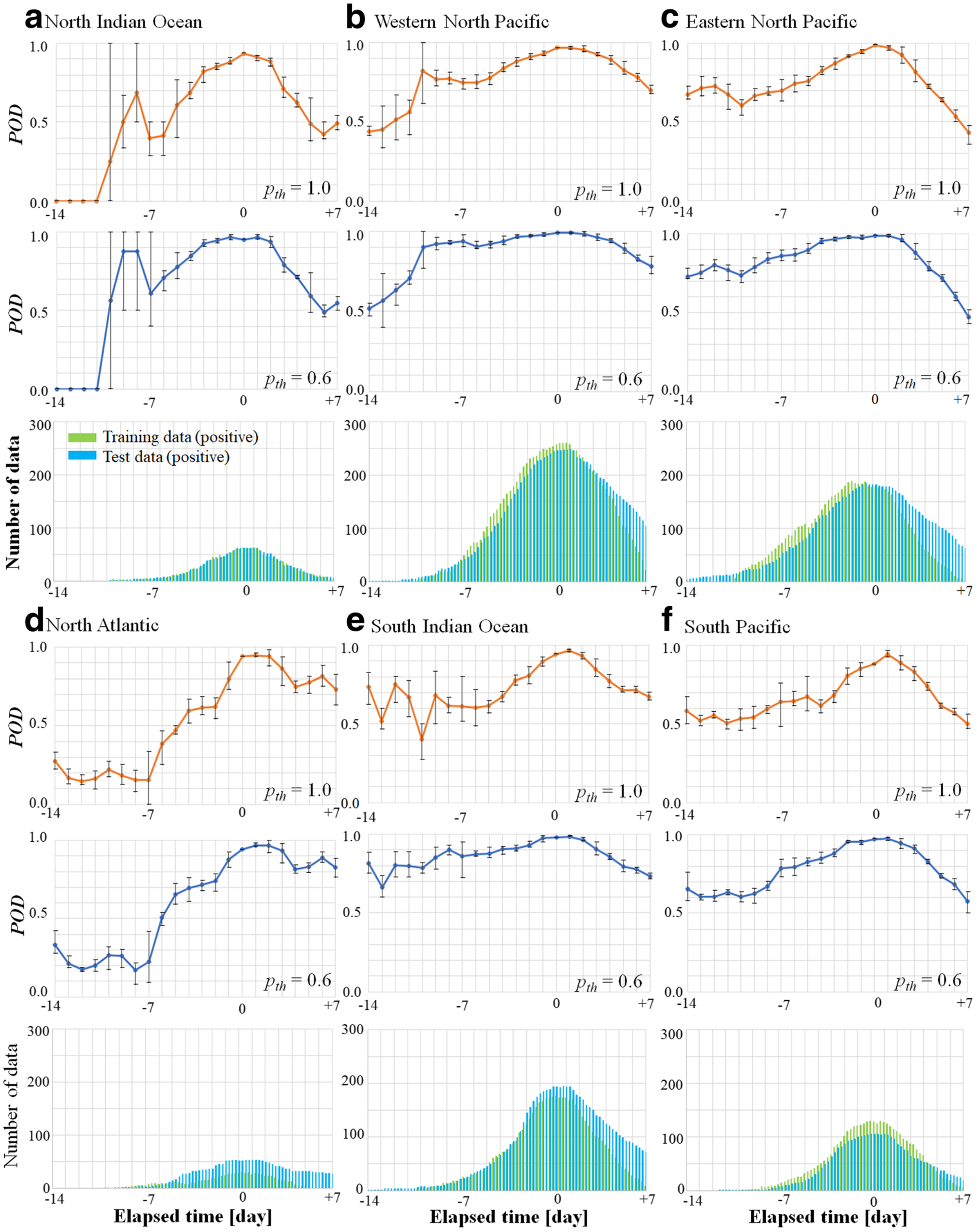

**Fig. 10** Average-max-min charts of *POD* ($p_{\mathrm{th}} = 1.0$ and $p_{\mathrm{th}} = 0.6$) and bar graphs of the number of data (training and test data) for each elapsed time frame in each basin

precursors 2, 5, and 7 days ahead can be detected in the eastern North Pacific, respectively (Fig. 10c). Note that the *FAR* could not be calculated in each elapsed time frame because *TP* could not be defined in each elapsed time frame.

In contrast to these basins, long-term detectability decreases rapidly in basins with less training data, such as the North Indian Ocean (Fig. 10a) and the North Atlantic (Fig. 10d). In most basins, the number of ground truth matches of test data is small from 7 to 14 days before the formation of precursors. Therefore, this time frame involves large errors and decreased reliability of *POD*. Note that these results are reference values because the *POD* and *FAR* could not be evaluated simultaneously.

## Conclusions

In this work, the detectability of TCs and their precursors for each basin, season, and lead time was investigated based on a deep neural network approach using 20 year simulated OLR by the NICAM. From the results of applying the CNNs to untrained 10 year simulated OLR, the following conclusions can be drawn:

- Particularly in the western North Pacific, we could successfully detect TCs and their precursors during July to November with *POD* values of 79.0–89.1% and *FAR* values of 32.8–53.4%. Detection results include 91.2%, 77.8%, and 74.8% of precursors 2, 5, and 7 days before their formation, respectively.
- Although the detection performance was approximately consistent with the number of training data and TC lifetime, the detection performance in the North Atlantic was not relatively low despite limited training data and short lifetimes. In particular, the average *POD* and *FAR* values in the North Atlantic during September to October were 74.8% and 40.9%, respectively.

These results suggest the high potential of the data-driven approach for studying tropical cyclogenesis.

In contrast, the limitations of our framework are as follows:

- Since the candidate regions are narrowed down by the threshold value of cloud cover (30–95%), they cannot be detected when the cloud cover is extremely small (< 30%) or extremely large (> 95%).
- Our method considers developing TCs and precursors as one category. To evaluate the detection performance of only pre-TCs, it is necessary to classify them by improving the CNN model.
- Our CNN classifiers may have model-specific biases arising from training using only NICAM data.

In areas with less data such as the North Atlantic, training data from other basins may have a positive influence on prediction results. However, in areas with sufficient data, such as the North Western Pacific, training according to each basin might improve detectability. In order to verify and improve detection performance, it is necessary to analyze the influence of the data in different basins by training the generation patterns and environmental factors in each basin.

This paper describes the preliminary results of detecting precursors of tropical cyclones using only simulated OLR; we plan to use other proxies of convection such as rain rate and mixing ratio of solid water for improving the detection performance. Furthermore, we also plan to apply our ensemble CNNs to reanalysis data and satellite observation data for practical use. For this purpose, data of different spatial resolutions and different variables or satellite channels must be considered. Furthermore, time-sequence data as well as comparative analyses with the results of the Early Dvorak Method are also required.

### Abbreviations

CNN: Convolutional neural network; NICAM: Nonhydrostatic ICosahedral Atmospheric Model

### Acknowledgements

We are grateful to Dr. C. Kodama for the production of simulation data and Dr. Y. Yamada for the production of best-track data on tropical cyclones. We also thank Editage for the English language review.

### Funding

This work was supported by JST, PRESTO Grant Number JPMJPR1777, JSPS KAKENHI Grant Number 16 K13885, 26700010, and 17 K13010.

### Availability of data and materials

The data used in this study are available from the authors upon request.

### Authors' contributions

DM proposed the topic and conceived and designed the study. MN analyzed the data and helped in their interpretation. DS conducted the experimental study. SU collaborated with the corresponding author in the preparation of the manuscript. All authors have read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interest.

## Publisher's Note

### Author details

[1]Center for Earth Information Science and Technology (CEIST), Japan Agency for Marine-Earth Science and Technology (JAMSTEC), 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan. [2]PRESTO, Japan Science and Technology Agency (JST), 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan. [3]Department of Seamless Environmental Prediction Research, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), 3173-25 Showa-machi, Kanazawa-ku, Yokohama, Kanagawa 236-0001, Japan. [4]Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan.

Received: 28 March 2018 Accepted: 19 November 2018
Published online: 19 December 2018

## References

Barnes LR, Gruntfest EC, Hayden MH, Schultz DM, Benight C (2007) False alarms and close calls: a conceptual model of warning accuracy. Wea. Forecasting 22:1140–1147

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Breiman L (2001) Random forests. Mach Learn 45(1):53–52

Chollet F (2015) Keras, GitHub

Cossuth J, Knabb TD, Brown DP, Hart RE (2013) Tropical cyclone formation guidance using pregenesis Dvorak climatology. Part I: operational forecasting and predictive potential. Wea. Forecasting 28:100–118

Dvorak VG (1975) Tropical cyclone intensity analysis and forecasting from satellite imagery. Mon. Wea. Rev. 103:420–430

Dvorak VG (1984) Tropical cyclone intensity analysis using satellite data. NOAA Technical Report NESDIS 11:1–47

Emanuel KA (1989) The finite-amplitude nature of tropical cyclogenesis. J Atmos Sci 46:3431–3456

Forsyth DA (2011) Computer Vision: A Modern Approach, 2nd edn. Pearson India, Delhi.

Freund Y, Schapire RE (1997) A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci 55:119–139

Fudeyasu H, Hirose S, Yoshioka H, Kumazawa R, Yamasaki S (2014) A global view of the landfall characteristics of tropical cyclones. Tropical Cyclone Research Review 3:178–192

Fudeyasu H, Yoshida R (2018) Western north pacific tropical cyclone characteristics stratified by genesis environment. Mon. Wea. Rev. 146:435–446

Gope S, Sarkar S, Mitra P (2016) In: Banerjee A, Ding W, Dy J, Lyubchich V, Rhines A (eds) Prediction of extreme rainfall using hybrid convolutional-long short term memory networks. Proceedings of the 6th International Workshop on Climate Informatics, Boulder 2016

Gray WM (1968) Global view of the origin of tropical disturbances and storms. Mon. Wea. Rev. 96:669–700

Gray WM (1975) Tropical cyclone genesis. Atmospheric science Paper 234. Colorado State University, Fort Collins

Holland GJ (2008) Tropical cyclones. In: Introduction to Tropical Meteorology, 1st edn. The COMET program, Boulder

Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd international conference on machine learning, Lille Grand Palais, Lille 6–11 July 2015

Jolliffe IT, Stephenson DB (2003) Forecast verification: a practitioner's guide in atmospheric science. Wiley, Hoboken

Kearns M, Valiant L (1989) Cryptographic limitations on learning Boolean formulae and finite automata. Proceedings of the 21st annual ACM Symposium on Theory of Computing, Seattle 14–17 May 1989

Kim SK, Ames S, Lee J, Zhang C, Wilson AC, Williams D (2017) In: Ebert-Uphoff I, Monteleoni C, Nychka D (eds) Massive scale deep learning for detecting extreme climate events. Proceedings of the 7th International Workshop on Climate Informatics, Boulder 2017

Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations (ICLR2015), Hilton San Diego Resort & Spa, San Diego 7–9 May 2015

Kodama C, Yamada Y, Noda AT, Kikuchi K, Kajikawa Y, Nasuno T, Tomita T, Yamaura T, Takahashi HG, Hara M, Kawatani Y (2015) A 20-year climatology of a NICAM AMIP-type simulation. J Meteorol Soc Jpn 93(4):393–424

Kordmahalleh MM, Sefidmazgi MG, Homaifar A, Liess S (2015) In: Dy JG, Emile-Geay J, Lakshmanan V, Liu Y (eds) Hurricane trajectory prediction via a sparse recurrent neural network. Proceedings of the 5th International Workshop on Climate Informatics, Boulder 2015

Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. Paper presented at neural information processing systems (NIPS) 2012, Harrahs and Harveys, Lake Tahoe 3–8 December 2012

Kumar DS (2013) Context and subcategories for sliding window object recognition. LAP LAMBERT Academic Publishing, Saarbrücken

LeCun Y, Haffner P, Bottou L, Bengio Y (1999) Object recognition with gradient-based learning. In: Forsyth DA, Mundy JL, Vd G, Cipolla R (eds) Shape, contour and grouping in computer vision. Lecture Notes in Computer Science, vol 1681. Springer, Berlin, Heidelberg, pp 319–345

Liu Y, Racah E, Prabhat, Correa J, Khosrowshahi A, Lavers D, Kunkel K, Wehner M, Collins W (2016) Application of deep convolutional neural networks for detecting extreme weather in climate datasets. arXiv reprint arXiv:1605.01156

Matsuoka D, Nakano M, Sugiyama D, Uchida S (2017) In: Ebert-Uphoff I, Monteleoni C, Nychka D (eds) Detecting precursors of tropical cyclone using deep neural networks. Proceedings of the 7th International Workshop on Climate Informatics, Boulder 2017

Miura H, Satoh M, Nasuno T, Noda AT, Oouchi K (2007) A Madden-Julien oscillation event realistically simulated by a global cloud-resolving model. Science 318:1763–1765

Nakano M, Kubota H, Miyakawa T, Nasuno T, Satoh M (2017b) Genesis of super cyclone pam (2015): modulation of low-frequency large-scale circulations and the madden-Julian oscillation by sea surface temperature anomalies. Mon. Wea. Rev. 145:3143–3159

Nakano M, Sawada M, Nasuno T, Satoh M (2015) Intraseasonal variability and tropical cyclogenesis in the Western North Pacific simulated by a global nonhydrostatic atmospheric model. Geophys. Res. Lett. 42(2):565–571

Nakano M, Wada A, Sawada M, Yoshimura H, Onishi R, Kawahara S, Sasaki W, Nasuno T, Yamaguchi M, Iriguchi T, Sugi M, Takeuchi Y (2017a) Global 7 km mesh nonhydrostatic Model Intercomparison Project for improving TYphoon forecast (TYMIP-G7): experimental design and preliminary results. Geosci Model Dev 10:1363–1381

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A (2003) Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J Geophys Res 108:4407. https://doi.org/10.1029/2002JD002670

Riehl H (1954) Tropical meteorology. McGraw-Hill, New York

Ritchie EA, Holland GJ (1999) Large-scale patterns associated with tropical cyclogenesis in the Western Pacific. Mon. Wea. Rev. 127:2027–2043

Russell JO, Aiyyer A, White JD, Hannah W (2017) Revisiting the connection between African easterly waves and Atlantic tropical cyclogenesis. Geophys Res Lett 44(1):587–595

Satoh M, Tomita H, Yashiro H, Miura H, Kodama C, Seiki T, Noda AT, Yamada Y, Goto D, Sawada M, Miyoshi T, Niwa Y, Hara M, Ohno T, Iga S, Arakawa T, Inoue T, Kubokawa H (2014) The non-hydrostatic icosahedral atmospheric model: description and development. Prog Earth Planet Sci 1:1–32

Simonyan K, Zisserman Z (2015) Very deep convolutional networks for large-scale image recognition. Paper presented at International Conference on Learning Representation (ICLR) 2015, The Hilton San Diego Resort & Spa, San Diego 7–9 May 2015

Sugi M, Noda A, Sato N (2002) Influence of the global warming on tropical cyclone climatology: an experiment with the JMA global model. J Meteorol Soc Jpn 80(2):249–272

Tomita H (2008) New microphysical schemes with five and six categories by diagnostic generation of cloud ice. J Meteor. Soc. Jpn 86A:121–142

Tomita H, Satoh M (2004) A new dynamical framework of nonhydrostatic global modeling using the icosahedral grid. Fluid Dyn. 1(8):357–400

Trafalis T, Adrianto I, Richman M, Lakshmivarahan S (2014) Machine-learning classifiers for imbalanced tornado data. Comput Manag Sci 11:403–418

Tsuchiya A, Mikawa T, Kikuchi A (2001) Method of distinguishing between early stage cloud systems that develop into tropical storms and ones that do not. Geophys Mag 1-4:49–59

Wilks DS (2006) Statistical methods in the atmospheric sciences, 2nd edn. Academic Press/Elsevier, New York

Xiang B, Lin S-J, Zhao M, Zhang S, Vecchi G, Li T, Jiang X, Harris L, Chen J-H (2015) Beyond weather time-scale prediction for hurricane Sandy and super typhoon Haiyan in a global climate model. Mon. Wea. Rev. 143:524–535

Yamada Y, Satoh M, Sugi M, Kodama C, Noda AT, Nakano M, Nasuno T (2017) Response of tropical cyclone activity and structure to global warming in a high-resolution global nonhydrostatic model. J Clim 30:9703–9724

Yamaguchi M, Koide N (2017) Tropical cyclone genesis guidance using the early stage Dvorak analysis and global ensembles. Wea. Forecasting 32:2133–2141

Yoshida R, Ishikawa H (2013) Environmental factors contributing to tropical cyclone genesis over the Western North Pacific. Mon. Wea. Rev. 141:451–467