Experimental Results (2022), 3, e7, 1–10
doi:10.1017/exp.2022.3

EARTH AND ENVIRONMENTAL SCIENCE
NEGATIVE RESULT

Can machine learning models trained using atmospheric simulation data be applied to observation data?

Daisuke Matsuoka[1,*]

[1]Research Institute for Value-Added-Information Generation (VAiG), Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokohama, Japan
***Corresponding author.** Email: daisuke@jamstec.go.jp

(Received 06 January 2022; Revised 21 January 2022; Accepted 21 January 2022)

Abstract
Atmospheric simulation data present richer information in terms of spatiotemporal resolution, spatial dimension, and the number of physical quantities compared to observational data; however, such simulations do not perfectly correspond to the real atmospheric conditions. Additionally, extensive simulation data aids machine learning-based image classification in atmospheric science. In this study, we applied a machine learning model for tropical cyclone detection, which was trained using both simulation and satellite observation data. Consequently, the classification performance was significantly lower than that obtained with the application of simulation data. Owing to the large gap between the simulation and observation data, the classification model could not be practically trained only on the simulation data. Thus, the representation capability of the simulation data must be analyzed and integrated into the observation data for application in real problems.

Key words: deep convolutional neural network; image classification; numerical simulation; satellite observation

1. Introduction

Deep learning is a machine learning method that uses multilayered neural networks; recently, it has been used to detect objects and structures in the field of atmospheric science. In particular, deep convolutional neural networks (DCNNs) specialized for image pattern recognition have exhibited excellent performance in detecting and/or classifying tropical cyclones (TCs) (Matsuoka et al., 2018), cloud type (Gorooh et al., 2020), weather fronts (Biard & Kunkel, 2019), and atmospheric river (Prabhat et al., 2021) from atmospheric data.

In general, machine learning using DCNNs requires extensive training data to achieve high performance. However, in atmospheric science, the aforementioned targets such as TCs occur infrequently. In addition, the reduction in detection accuracy for extreme phenomena is a limitation owing to the inadequate number of observation cases.

Although numerical simulations employing an atmospheric model can generate data of extreme events under various initial conditions and scenarios, they do not perfectly correspond to the real atmospheric conditions. If the simulated data can interpolate a small number of observed cases, it could contribute toward improving the performance of the machine learning-based models for recognizing extreme events. This paper reports the initial results of applying a classification model developed by training only simulation data to satellite observation data, considering typhoon classification as a simple example.

2. Materials and methods

2.1. Observation and simulation data

The satellite observation data included infrared (IR) from GridSat data, which corresponds to the merging of multiple satellite observations into a grid with a horizontal resolution of 7 km (Knapp et al., 2011). The simulation data included the outgoing longwave radiation (OLR) with a horizontal resolution of 14 km, which was reproduced by the cloud-resolving model NICAM (Kodama et al., 2015).

To detect TCs, we prepared patch images of TCs (positive examples) and non-TCs (negative examples) cropped from the original data using the TC track data in the northwest Pacific Ocean. The size of the patch images was 64 × 64 for the simulation data and 128 × 128 for the observation data, which was approximately 1,000 km square in real scale. The most suitable track data of actual TCs—the International Best Track Archive for Climate Stewardship (IBTrACS) provided by NCAR—were used for the observation data. For the simulation data, the TCs were detected by employing a TC track algorithm (Yamada et al., 2017) on 6-hourly outputs of the horizontal components of wind, air temperature, and sea-level pressure. For the negative example data, the entire area was horizontally scanned in eight grids, and the patch areas depicting portions of clouds were cropped.

Examples of a cloud image from the simulation and observation data are depicted in Figure 1. To match the resolution of the simulation dataset to that of the observation dataset, the observation data were resized to half their original resolution. To identically treat the distinct data, we applied a min–max normalization (also known as min–max scaling) to IR and OLR.

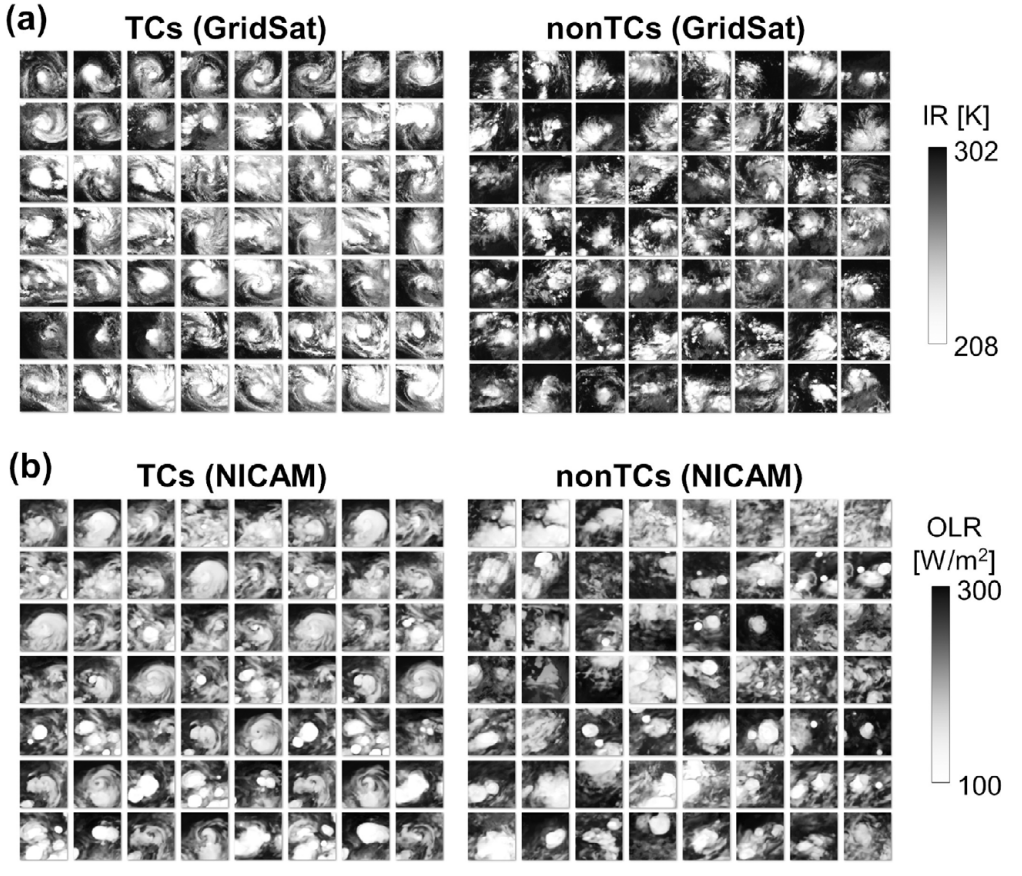

Figure 1. Examples of (a) observation and (b) simulation data; left: TCs; right: nonTCs.

https://doi.org/10.1017/exp.2022.3 Published online by Cambridge University Press

Table 1. Numbers of positive and negative examples in training and test data

	Observation data (GridSat-IR)	Simulation data (NICAM-OLR)
Training data (1980–2003)	Positive: 18,302; Negative: 18,302	Positive: 15,521; Negative: 15,521
Test data (2004–2008)	Positive: 2,683; Negative: 56,896	Positive: 3,644; Negative: 74,043

The numbers of data for the positive and negative examples in the training and test data are listed in Table 1, wherein the classifiers trained on observation and simulation data are referred to as ObsCNN and SimCNN, respectively. Moreover, we randomly sampled a large number of negative examples to construct both the classifiers, such that an equal number of data points were present for the positive and negative examples, as same as the related work (Matsuoka et al., 2018).

2.2. Deep convolutional neural networks

We developed a binary classification model using a DCNN to classify TC and nonTC images. Generally, a DCNN comprises a stack of convolutional layers, pooling layers, and fully connected layers (LeCun & Bengio, 1995). In classification, the output layer outputs a score, $P(0)$ and $P(1)$, corresponding to the probability for each class—negative (non-TCs) and positive (TCs). Ultimately, the final class $\widehat{y}$ was inferred using the following equation employing the threshold value.

$$\widehat{y} = \begin{cases} 1, & (P(1) \leqq Threshold), \\ 0, & (otherwise). \end{cases} \tag{1}$$

Here, $P(0) + P(1) = 1.0$, $0 \leqq P(0) \leqq 1.0$, $0 \leqq P(1) \leqq 1.0$, and $0 \leqq Threshold \leqq 1.0$.

Although several DCNN models have been proposed in related research literature, we conducted experiments in this study using the VGG16, which is known for its high recognition accuracy despite being a relatively lightweight model (Simonyan & Zisserman, 2015). Based on the VGG16 model, we constructed two types of classification models: one trained only on simulated data and the other trained only on observed data for comparison.

Moreover, we used recall and precision as metrics to evaluate the classification performance for the test data. In particular, recall denotes the ratio of correctly classified TCs to those with the correct class as TCs, whereas precision represents the ratio of correctly classified TCs to those with the inferred class as TCs, and they can be represented by the following equation.

$$Recall = \frac{TP}{TP + FN}, \tag{2}$$

$$Precision = \frac{TP}{TP + FP}, \tag{3}$$

where TP denotes true positive, FN represents false negative, and FP denotes false positive.

3. Results

The classification performance of the ObsCNN and SimCNN on the observation data is illustrated in Figure 2a as a precision–recall curve (P–R curve), which was plotted by varying the threshold for the output value of the DCNN. In most cases, the models trained on the observation data delivered higher classification performance than those trained on the simulation data. For a precision of 0.5, the recall of ObsCNN was 0.984, whereas that of SimCNN was 0.829. Similarly, for a precision of 0.7 and 0.9, the recall of ObsCNN was higher than that of SimCNN. This signified that the classification of the simulation data

https://doi.org/10.1017/exp.2022.3 Published online by Cambridge University Press

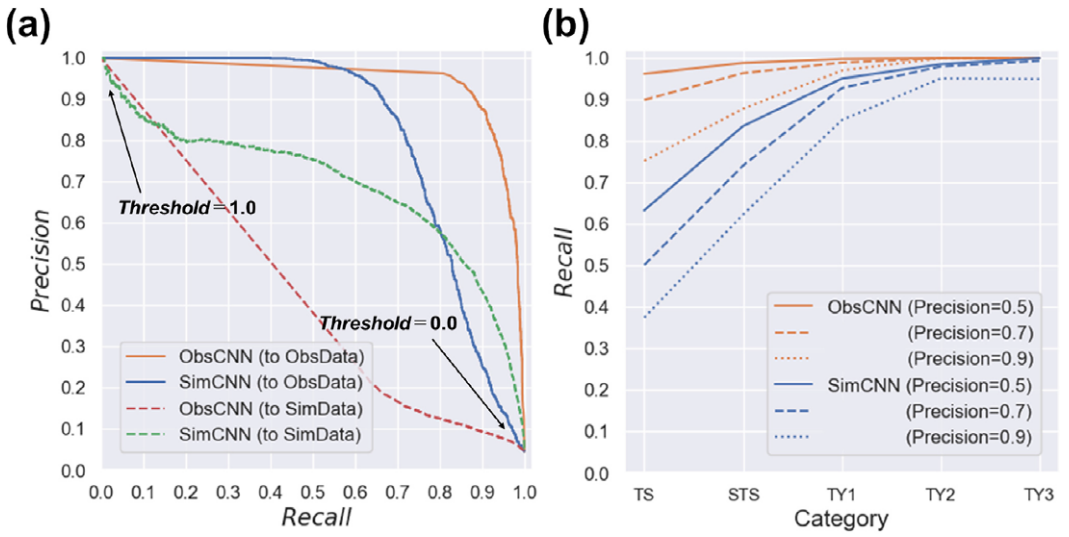

Figure 2. Classification performance of ObsCNN and SimCNN: (a) precision—recall curve and (b) recall for each category.

was more challenging than that of the observation data. For reference, the classification performance of both models when applied to simulation data is also shown in Figure 2a. SimCNN showed better performance than ObsCNN for the simulation data.

The classification performance for each tropical cyclone intensity is presented in Figure 2b. Accordingly, we classified the TCs into the following five categories using their maximum wind speed (10-min average): TS (17–24 m/s), STS (25–32 m/s), TY1 (33–43 m/s), TY2 (44–53 m/s), and TY3 (over 54 m/s). The recall of both ObsCNN and SimCNN for each TC intensity for the precision fixed at a specific value (0.5, 0.7, 0.9) is portrayed in Figure 2b. For both ObsCNN and SimCNN, the recall was higher for a stronger TC intensity. In addition, the performances of the SimCNN and ObsCNN were similar for strong TC intensities. The recalls of ObsCNN and SimCNN for TS were approximately 0.75 and 0.37, whereas they were approximately 1.0 and 0.94 for TY3, implying that the features of the observation and simulation data were distinct for weaker TCs.

4. Discussion

The differences in the properties of ObsCNN and SimCNN were visualized using a technique called Grad-CAM, which is a region visualization method with significant contributions toward CNN prediction (Selvaraju et al., 2020). The important areas in the decision visualized by the Grad-CAM are depicted in Figure 3. The correct inference obtained using ObsCNN and SimCNN for observation data (only TCs) are indicated in Figure 3a,b. The results from ObsCNN revealed that the regions of high importance were clustered around the center of the TC. On the contrary, the SimCNN results indicated that the regions of high importance were clustered a little farther from the center of the TC. Especially in the TY3 example with a clear eye of the TC, the ObsCNN indicated that the eye of the TC was a more important factor in the classification.

The important regions inferred from the simulation data (only TCs) using the SimCNN is visualized in Figure 3c. In any TC category, the pattern slightly outside the center of the TC was recognized to determine it as a TC. Based on these results, with respect to the classification capability of CNNs, the patterns outside the center in the simulation were similar to those in the observation data. The detection performance of strong TCs is high for both ObsCNN and SimCNN shown in Figure 2b, suggesting that the observed TCs also have a characteristic pattern around off-center region.

https://doi.org/10.1017/exp.2022.3 Published online by Cambridge University Press

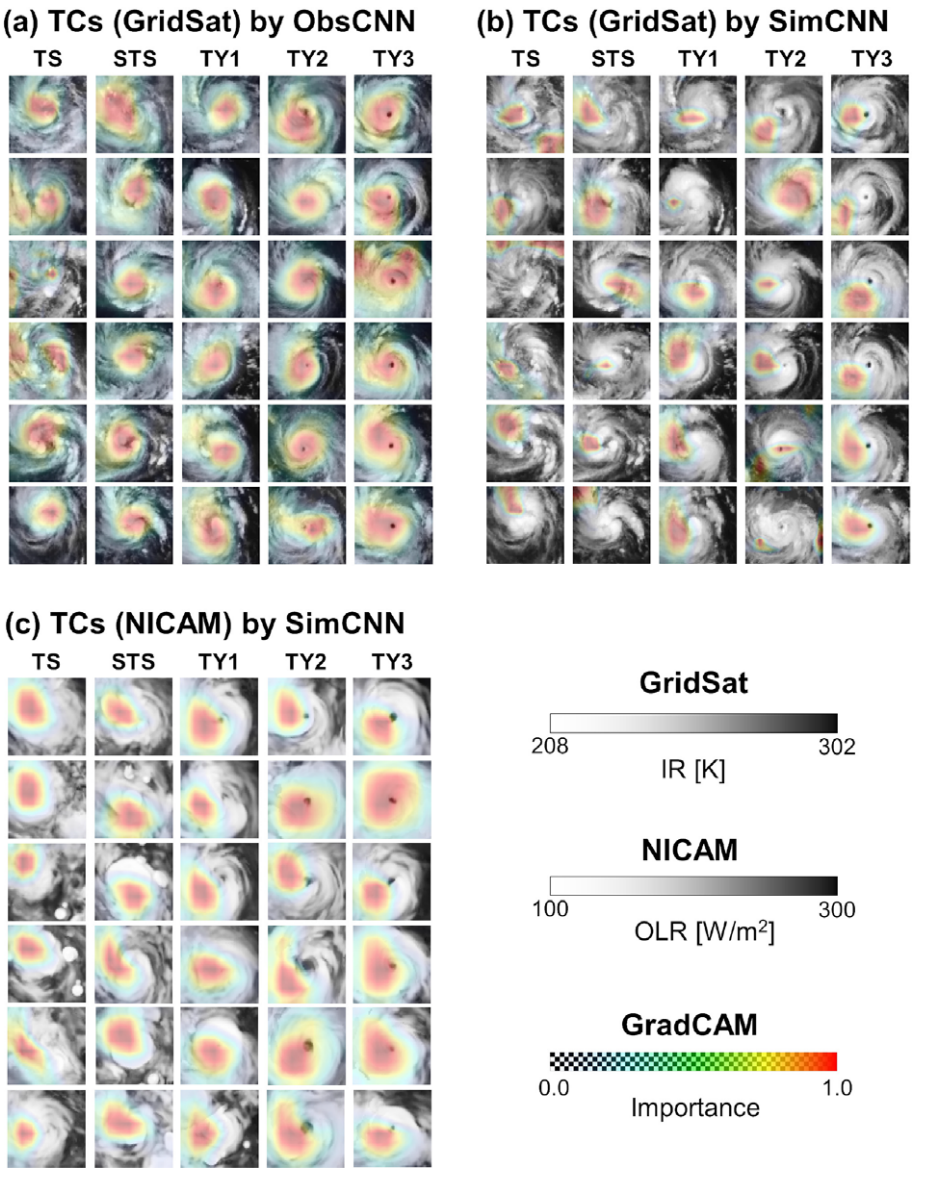

Figure 3. Visualization results of vital regions in the CNN trained on (a) observation data (ObsCNN) and (b) simulation data (SimCNN).

Finally, the observed and simulated TC images (only TY2 and TY3) were dimensionally reduced and mapped into two-dimensional feature space using Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) and t-distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008) as shown in Figure 4a,b. Both results show that some of the observed TCs have similar features to the simulated TCs, while the rest have different features.

https://doi.org/10.1017/exp.2022.3 Published online by Cambridge University Press

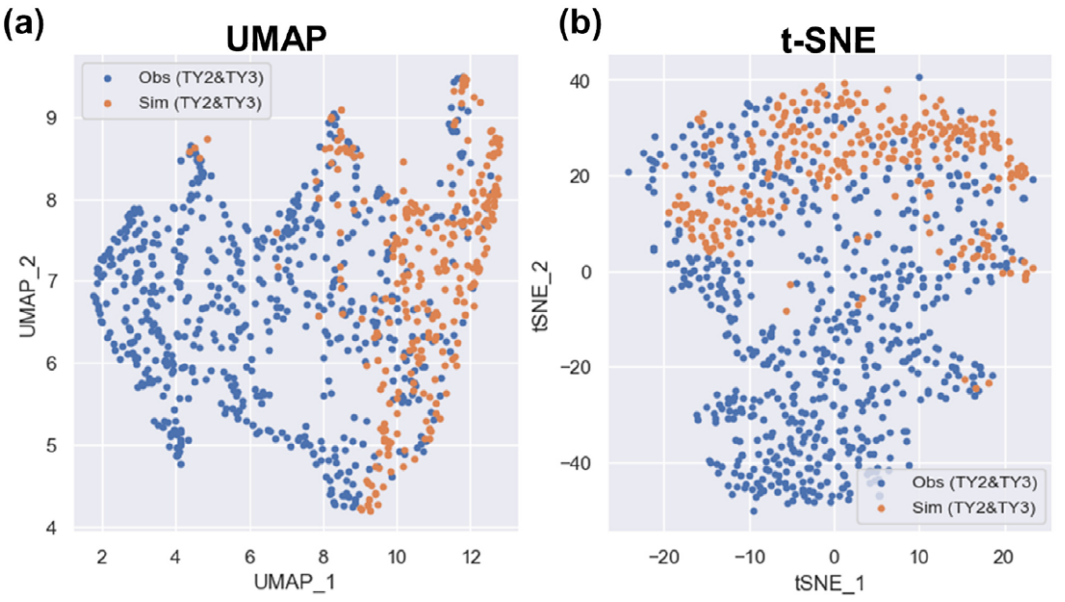

Figure 4. Dimension reduction and two-dimensional projection applied to observed and simulated TCs (only TY2 and TY3) using (a) UMAP and (b) t-SNE.

5. Conclusion

In this study, we developed a CNN-based classifier trained using simulation data and applied it to the observation data.The classification model trained on the simulation data cannot be directly applied to the observation data due to differences in cloud patterns to be recognized. Although negative experimental results were obtained, there is no doubt that the simulation data have great potential. Clarifying the representation capability of both data and integrating the data will lead to advanced machine learning models as well as simulation models.

Acknowledgments. The author thanks Drs. M. Nakano, C. Kodama, and Y. Yamada for producing the training and test data on tropical cyclones.

Data availability statement. Please contact the corresponding author for data requests.

Funding statement. This work was supported by JST, PRESTO (Grant Number JPMJPR1777), and JST, CREST (Grant Number JPMJCR1663).

Conflict of interest. The author has no conflicts of interest to declare.

Authorship Contributions. D.M. designed the study, performed the experiments, analyzed the data, and wrote the manuscript.

References

Biard, J. C., & Kunkel, K. E. (2019). Automated detection of weather fronts using a deep learning neural network. *Advances in Statistical Climatology, Meteorology and Oceanography*, **5**, 147–160. https://doi.org/10.5194/ascmo-5-147-2019

Gorooh, V. A., Kalia, S., Nguyen, P., Hsu, K.-L., Sorooshian, S., Ganguly, S., & Nemani, R. R. (2020). Deep neural network cloud-type classification (DeepCTC) model and its application in evaluating PERSIANN-CCS. *Remote Sensing (Basel)*, **12**, 1–19. https://doi.org/10.3390/rs12020316

Knapp, K. R., Ansari, S., Bain, C. L., Bourassa, M. A., Dickinson, M. J., Funk, C., Helms, C. N., Hennon, C. C., Holmes, C. D., Huffman, G. J., Kossin, J. P., Lee, H.-T., Loew, A., & Magnusdottir, G. (2011). Globally gridded satellite (GridSat) observations for climate studies. *Bulletin of the American Meteorological Society*, **92**, 893–907. https://doi.org/10.1175/2011BAMS3039.1

Kodama, C., Yamada, Y., Noda, A. T., Kikuchi, K., Kajikawa, Y., Nasuno, T., Tomita, T., Yamaura, T., Takahashi, H. G., Hara, M., & Kawatani, Y. (2015). A 20-year climatology of a NICAM AMIP-type simulation. *Journal of the Meteorological Society of Japan*, **93**, 393–424. https://doi.org/10.2151%2Fjmsj.2015-024

LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time-series. In M. A.Arbib (Ed.), *The hand book of brain theory and neural networks*. MIT Press, pages 255–258.

Matsuoka, D., Nakano, M., Sugiyama, D., & Uchida, S. (2018). Deep learning approach for detecting tropical cyclones and their precursors in the simulation by a cloud-resolving global nonhydrostatic atmospheric model. *Progress in Earth and Planetary Science*, **5**, 1–16. https://doi.org/10.1186/s40645-018-0245-y

McInnes, L., Healy, J., Saul, N., Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, **3**, 861. http://doi.org/10.21105/joss.00861

Prabhat, K., Mudigonda, K., Kim, M., Kapp-Schwoerer, S., Graubner, L., Karaismailoglu, A., von Kleist, E., Kurth, L., Greiner, T., Mahesh, A., Yang, A., Lewis, K., Chen, C., Lou, J., Chandran, A., Toms, S., Chapman, B., Dagon, W., Shields, K., … Collins, W. (2021). ClimateNet: an expert-labelled open dataset and deep Learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development*, **14**, 107–124. https://doi.org/10.5194/gmd-14-107-2021

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, **128**, 336–359. https://doi.org/10.1007/s11263-019-01228-7

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR2015)*.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605. https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

Yamada, Y., Satoh, M., Sugi, M., Kodama, C., Noda, A. T., Nakano, M., & Nasuno, T. (2017). Response of tropical cyclone activity and structure to global warming in a high-resolution global nonhydrostatic model. *Journal of Climate*, **30**, 9703–9724. https://doi.org/10.1175/2FJCLI-D-17-0068.1

Cite this article: Matsuoka D (2022). Can machine learning models trained using atmospheric simulation data be applied to observation data? *Experimental Results,* 3, e7, 1–10. https://doi.org/10.1017/exp.2022.3

Peer Reviews

Reviewing editor: Dr. Jacob Carley

NOAA Center for Weather and Climate Prediction, NCEP/Environmental Modeling Center, 5830 University Research Cour, College Park, Maryland, United States, 20740

This article has been accepted because it is deemed to be scientifically sound, has the correct controls, has appropriate methodology and is statistically valid, and has been sent for additional statistical evaluation and met required revisions.

doi:10.1017/exp.2022.3.pr1

Review 1: Can machine learning models trained using atmospheric simulation data be applied to observation data?

Reviewer: Dr. Jili Dong

Date of review: 15 January 2022

Conflict of interest statement. Reviewer declares none

Comment

Comments to the Author: This manuscript investigated and compared the tropical cyclone detection approach with deep learning algorithm by using either satellite observations or model output as training dataset. The authors found out that using model output as training dataset to detect TC is less accurate than using satellite observations. It is generally well written but needs some further clarifications.

Major comments:

1.Are the criteria of detecting/identifying TC the same for satellite observations and NICAM model output? 10-m max. wind can be used in a model to define a TC while the Dvorak technique is commonly used to decide TC categories from cloud patterns of the satellite visible/IR images. If the criteria are different, how will the interpretation of the results be affected?

2.Fig. 2b shows similar recall skill of SimCNN and ObsCNN for strong typhoons (TY3) but Fig. 3b shows a much more off-center pattern of SimCNN for TY3 when compared to ObsCNN in Fig. 3a. How do the authors explain the contradiction?

Minor comments

1.Table1: the positive and negative cases in the training dataset are always the same for both model and obs. Is this a requirement or just an coincidence?

2.Page 3: the denotation of recall/precision. “TN denotes true negative” should be “FN denotes false negative”.

3.Page 4: the third line from bottom “in both the TC categories”. Which two categories do “both” refer to?

Score Card

Presentation

Is the article written in clear and proper English? (30%)	3/5
Is the data presented in the most useful manner? (40%)	3/5
Does the paper cite relevant and related articles appropriately? (30%)	2/5

Context

Does the title suitably represent the article? (25%)	4/5
Does the abstract correctly embody the content of the article? (25%)	4/5
Does the introduction give appropriate context? (25%)	3/5
Is the objective of the experiment clearly defined? (25%)	3/5

Analysis

Does the discussion adequately interpret the results presented? (40%)	3/5
Is the conclusion consistent with the results and discussion? (40%)	3/5
Are the limitations of the experiment as well as the contributions of the experiment clearly outlined? (20%)	3/5

https://doi.org/10.1017/exp.2022.3 Published online by Cambridge University Press

doi:10.1017/exp.2022.3.pr2

Review 2: Can machine learning models trained using atmospheric simulation data be applied to observation data?

Reviewer: Ryuta Miyata

Date of review: 18 January 2022

Conflict of interest statement. Reviewer declares none.

Comment

Comments to the Author: I think more depth of discussion is needed:

(1) As in Fig. 2(a), what is the classification performance of obsCNN when the test SimData is given?

(2) Comparing TY2 and TY3 in Fig. 3(c) indicates that the typhoon's eye in SimData seemed not to work as a feature related to the category while that in ObsData did.

In other words, we can infer that the distribution is different between the SimData and ObsData.

To show the above clearly, I suggest using t-SNE or UMAP to visualize each category distribution of the SimData and ObsData, respectively.

Score Card

Presentation

4.0 /5

Is the article written in clear and proper English? (30%)	4/5
Is the data presented in the most useful manner? (40%)	4/5
Does the paper cite relevant and related articles appropriately? (30%)	4/5

Context

4.0 /5

Does the title suitably represent the article? (25%)	4/5
Does the abstract correctly embody the content of the article? (25%)	4/5
Does the introduction give appropriate context? (25%)	4/5
Is the objective of the experiment clearly defined? (25%)	4/5

Analysis

3.0 /5

Does the discussion adequately interpret the results presented? (40%)	2/5
Is the conclusion consistent with the results and discussion? (40%)	4/5
Are the limitations of the experiment as well as the contributions of the experiment clearly outlined? (20%)	3/5