Digital Typhoon: Long-term Satellite Image Dataset for the Spatio-Temporal Modeling of Tropical Cyclones

Asanobu Kitamoto[1,2] **Jared Hwang**[3,1] **Bastien Vuillod**[4,1] **Lucas Gautier**[5,1]
[1] National Institute of Informatics, Japan
[2] Typhoon Science and Technology Research Center, Yokohama National University, Japan
[3] University of Southern California, USA
[4] Grenoble-INP, Ensimag, France
[5] Université Clermont Auvergne, ISIMA, France

Abstract

This paper presents the official release of the Digital Typhoon dataset, the longest typhoon satellite image dataset for 40+ years aimed at benchmarking machine learning models for long-term spatio-temporal data. To build the dataset, we developed a workflow to create an infrared typhoon-centered image for cropping using Lambert azimuthal equal-area projection referring to the best track data. We also address data quality issues such as inter-satellite calibration to create a homogeneous dataset. To take advantage of the dataset, we organized machine learning tasks by the types and targets of inference, with other tasks for meteorological analysis, societal impact, and climate change. The benchmarking results on the analysis, forecasting, and reanalysis for the intensity suggest that the dataset is challenging for recent deep learning models, due to many choices that affect the performance of various models. This dataset reduces the barrier for machine learning researchers to meet large-scale real-world events called tropical cyclones and develop machine learning models that may contribute to advancing scientific knowledge on tropical cyclones as well as solving societal and sustainability issues such as disaster reduction and climate change. The dataset is publicly available at `http://agora.ex.nii.ac.jp/digital-typhoon/dataset/` and `https://github.com/kitamoto-lab/digital-typhoon/`.

1 Introduction

Tropical cyclones, also known as typhoons and hurricanes in certain regions, have been the critical target of research due to their substantial societal impact [11]. To reduce the impact of tropical cyclones, the meteorological community, along with other earth science communities, has been developing both a theoretical and an empirical understanding of tropical cyclones through efforts such as advancing satellite remote sensing and atmospheric simulation models of higher spatial, temporal, and spectral resolutions for better analysis and forecasting.

Meteorologists have also developed an empirical method, known as the Dvorak technique [10, 53], to estimate the intensity of a tropical cyclone based on time-series observation data collected from worldwide ground sensor networks, meteorological satellites, and reconnaissance flights. This technique consists of a manual procedure to estimate tropical cyclone intensity based on the cloud patterns of satellite images and a temporal model for intensity change. The method was originally developed in the United States in the 1970s and later adopted by meteorological agencies worldwide

Submitted to the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks. Do not distribute.

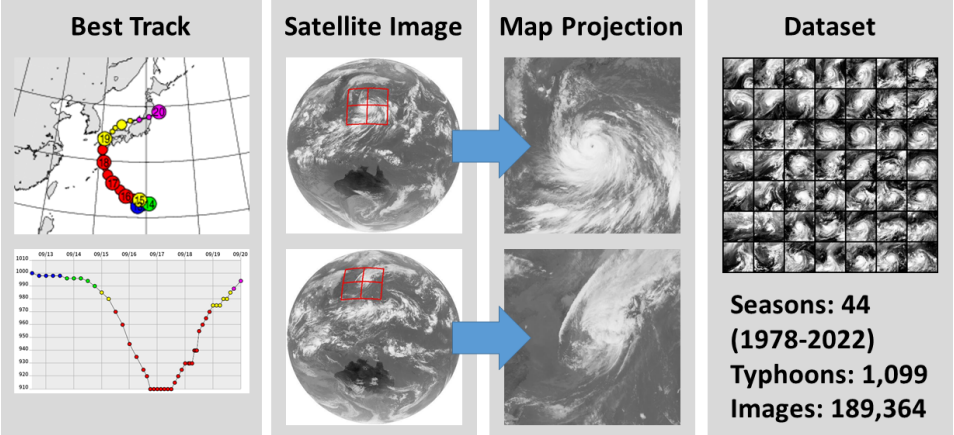


Figure 1: Overview of the Digital Typhoon dataset.

to become the standard procedure. However, experts are aware of its heuristic and subjective nature, as it relies on empirical, rather than theoretical, human interpretation of observation data. Solutions to this problem include more objective and automated versions of the Dvorak technique[38, 37] and a citizen science project to take advantage of collective intelligence [15].

It is clear that the Dvorak technique naturally fits into the machine learning framework by using images as input and intensity values as output. Hence there is a growing interest in both the machine learning community [41, 8, 35, 33] and the meteorology community [18, 5] to take advantage of the big data of tropical cyclones for developing data-driven approaches. One of the authors, Asanobu Kitamoto, started the Digital Typhoon project in 1999 with the aim of applying machine learning to typhoon analysis and forecasting[21, 26]. The first step was to develop a homogeneous satellite image dataset for machine learning as in Figure 1. The second step was to apply machine learning algorithms available at the time, such as SVM [24], Generative Topographic Mapping [22], and content-based image retrieval[23], which is later evolved into deep learning-based models for classification and regression tasks [45, 40], combined with fisheye preprocessing [16]. The third step was to release the website "Digital Typhoon" in 2003 for browsing and searching datasets [25]. In spite of those developments, the Digital Typhoon dataset was only for internal use, and it has been used sporadically only via web-scraping of the dataset (e.g. [52]).

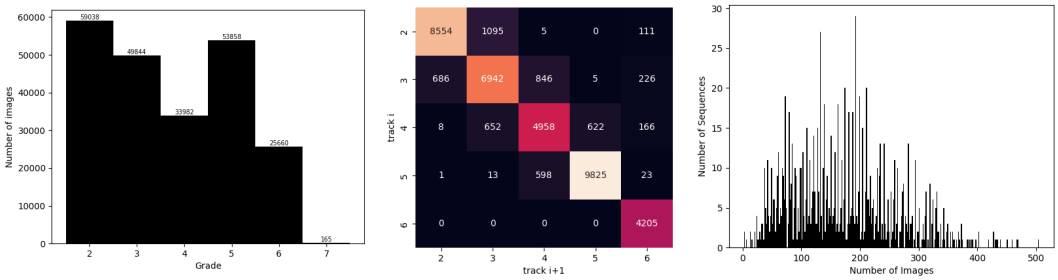
Here we introduce the Digital Typhoon dataset, the *longest* typhoon satellite image dataset. This dataset reduces the burden of researchers to start machine learning on tropical cyclones without solid domain knowledge of meteorology and satellite remote sensing. We also illustrate the variety of tasks so that researchers can concentrate on building and evaluating machine-learning models.

2 Related Work

2.1 Track Datasets

The track data includes the 'annotation' of tropical cyclones, such as location, intensity, and wind circles, based on the interpretation of meteorological experts following the established procedure (e.g. Dvorak Technique). The best estimate, obtained from a retrospective analysis after collecting all the information from the start to the end of life, is called the best track dataset.

The Digital Typhoon dataset targets the Western North Pacific basin, and the Japan Meteorological Agency (JMA) is designated as the regional center to maintain the best track dataset. Globally, the International Best Track Archive for Climate Stewardship (IBTrACS) [28] collects the best track from meteorological agencies worldwide and creates a comprehensive track dataset since 1842.



(a) Number of images for each grade (b) Transition matrix for the grade (c) Number of typhoons by the number of images

Figure 2: Visualization of statistics of the Digital Typhoon dataset.

IBTrACS shows an interesting variation of the best track; namely the location and intensity of the same tropical cyclone show discrepancies across meteorological agencies [47]. This fact suggests that the interpretation of the observation data is not unique, or not the *ground truth* in a strict sense. Nonetheless, we regard the best track as the ground truth for most machine learning tasks, because it is the best estimate available. In a reanalysis task, however, we could critically evaluate the quality of the best track [17].

2.2 Image Datasets

The image dataset has information about the spatial distribution of physical properties such as cloud patterns as grid data. The observation dataset [27, 30] is derived from sensor observation that measures the physical properties of the atmosphere, while the simulation dataset, both typhoon-related [36] and the global atmosphere [43, 1, 3, 42], is generated as the representation of the atmosphere in a simulation model. Observation datasets and simulation datasets are linked through data assimilation, which is a statistical method to integrate observation datasets into a simulation model.

The Digital Typhoon dataset is an observation dataset, and it offers a richer detail of tropical cyclones with higher temporal and spatial resolutions than the simulation dataset. In addition, data quality issues in the observation dataset, such as sensor noise, missing data, and long-term sensor calibration, are handled properly so that machine learning models are not significantly affected by those issues.

3 Digital Typhoon Dataset

3.1 Dataset Overview

The Digital Typhoon dataset is created from the comprehensive satellite image archive of the Japanese geostationary satellite series, Himawari, from Himawari-1 to Himawari-9. Although those images are not copyrighted, some data are not accessible for free, and old satellite images have old formats for which open-source parsers are difficult to find. Hence we developed our own parsers for all generations of satellites, and the workflow to create typhoon-centered images by referring to the best track, as shown in Figure 1.

Using this workflow, we created the Digital Typhoon dataset by integrating metadata and image. The metadata contains hourly best-track data with additional information about the file name and each image's quality. The formatting of the best track data aligns with the original best track data sourced from the JMA. On the other hand, the images feature a 2D array of brightness temperatures around the typhoon's center, formatted in HDF5.

As a result, the dataset comprises a total of 1,099 typhoons and 189,364 images. Figure 2 visualizes some of the statistics of the dataset. It is a complete record of typhoons occurring in the Western North Pacific region (ranging from 100 to 180 degrees east of the northern hemisphere), from the 1978

Table 1: Comparison between the Digital Typhoon dataset and the HURSAT dataset.

Dataset	Digital Typhoon Dataset	HURSAT Dataset
Temporal coverage	1978-2022 (present)	1978-2015
Temporal resolution	one hour	3 hours
Target satellite		
Spatial coverage	Western North Pacific by JMA geostationary satellites	Global by worldwide geostationary satellites
Spatial resolution	5km	8km
Image coverage	512×512 pixels (1250km from the center)	301×301 pixels (1100km from the center)
Spectral coverage	infrared (others on the Website)	visible, infrared, water vapor, near IR, split window
Map projection	Azimuthal equal-area projection	Equirectangular projection
Calibration	Recalibration	ISCCP
Data format	HDF5	NetCDF
Best track	Japan Meteorological Agency	IBTrACS
Data browsing	Digital Typhoon website	Download only

season through the 2022 season, with missing typhoons in 1979 and 1980 due to the unavailability of satellite data. The length of the dataset, spanning 44 typhoon seasons (years), is the *longest* typhoon image dataset. We call it the longest dataset because Japanese geostationary satellite images for typhoons before 1978 were lost forever, and our dataset went back to the oldest satellite image preserved. Hence it provides a unique opportunity to challenge long-term datasets.

The Digital Typhoon dataset can reduce the burden of machine learning researchers to study tropical cyclones. First, it opens up access to tropical cyclone data processed from long-term satellite data. Second, it offers a homogeneous dataset created by the image processing workflow based on expertise in meteorology and satellite remote sensing. Third, it has already done massive computations to process hundreds of terabytes of satellite data. A comprehensive explanation of the workflow for the creation of the dataset is provided in the Appendix.

The Digital Typhoon dataset is available at the official page `http://agora.ex.nii.ac.jp/digital-typhoon/dataset/` with an open data license, namely the Creative Commons Attribution 4.0 International (CC BY 4.0) License.

3.2 Comparison with the HURSAT dataset

Among satellite image datasets of tropical cyclones, Hurricane Satellite Data (HURSAT) dataset [27, 30] from The National Oceanic and Atmospheric Administration (NOAA) is the most notable dataset in size and coverage. Table 1 provides a comparative summary of the Digital Typhoon and HURSAT datasets. There are distinct variations between the two as enumerated below.

Temporal coverage The Digital Typhoon dataset is continually updated, and is the *longest* tropical cyclone image dataset worldwide. On the other hand, the HURSAT dataset stopped in 2015.

Temporal resolution The Digital Typhoon dataset has a temporal resolution of one hour which is higher than the HURSAT dataset's three-hour resolution. A high-frequency change such as rapid intensification is more sensitive to temporal resolution.

Spatial coverage The Digital Typhoon dataset specifically targets the Western North Pacific basin, whereas the HURSAT dataset encompasses all basins.

Spatial resolution The Digital Typhoon dataset possesses a spatial resolution of approximately 5km, superior to the HURSAT dataset's roughly 8km (0.07 degree). A small-scale structure such as the eye of a tropical cyclone is more sensitive to spatial resolution.

Spectral coverage The Digital Typhoon dataset incorporates the infrared (IR) channel, while the HURSAT dataset has more channels. It should be noted, however, that the Digital Typhoon *website* has the same spectral coverage, and the *dataset* can be easily extended to cover these channels.

Map projection The Digital Typhoon dataset utilizes the Lambert azimuthal equal-area projection, maintaining the spherical shape of the tropical cyclone, while the HURSAT dataset employs the equirectangular (lat/long) grid, causing shape distortion in higher latitude or peripheral areas. Figure 1 shows an example of distortion when a typhoon is observed in the north.

Data browsing The Digital Typhoon *dataset* can be browsed via the Digital Typhoon *website*, which offers additional data. In contrast, the HURSAT dataset is solely available for download.

4 Machine Learning Tasks

The Digital Typhoon dataset serves two important roles. First, it offers a practical real-world dataset and tasks for the machine learning community to explore new models and solutions. Second, it provides a tool for meteorologists to apply data-driven approaches in studying tropical cyclones. The following is a summary of tasks in multiple dimensions. Other lists of tasks can be found in the review [7, 54, 44].

4.1 Types of Inference

Analysis The task is to estimate current values using the current and past data. For instance, estimating the intensity of a typhoon falls into an analysis task, as it produces information about the typhoon's intensity using both current and past data. Supervised learning within this task can be further segmented into either a classification task or a regression task, contingent on whether the target variable is categorical or numerical. Additionally, unsupervised tasks can be designed for clustering or identifying typhoons with similar characteristics.

Forecasting The task is to produce predictions based on current and past data. The forecasts can be evaluated with the actual outcomes from the real event which become available over time. The forecasting task has a sub-task called nowcasting, aimed at making short-term forecasting spanning several hours using data-driven extrapolation. Note, however, that weather forecasting is theoretically constrained by the atmosphere's chaotic nature, which states that a minor difference in initial conditions can escalate over time.

Reanalysis The task is to produce the best estimate given all obtainable data. This task is especially relevant to produce a uniform dataset spanning a long period of time, such as detecting trends in tropical cyclone activity to study the effects of climate change. As addressed in Section E, the best track dataset may contain errors due to technological limitations or inconsistencies from different human experts. Machine learning can potentially aid in evaluating the quality of annotated data.

4.2 Targets of Inference

Intensity The task makes inferences on the strength and size of a typhoon. The categorical grade is used to classify both the strength and type of a tropical cyclone. Grades 3, 4, and 5 denote a tropical cyclone, among which grade 5 is the most intense according to the maximum wind speed. Grade 2 signifies a tropical depression, a type of cyclone weaker than a tropical cyclone. Moreover, grade 6 corresponds to an extra-tropical cyclone, a type of cyclone having a different structure from a tropical cyclone. A classification task uses these grades as target variables. On the other hand, the intensity of tropical cyclones is measured numerically by central pressure and maximum wind speed, so an intensity regression task uses either pressure or wind as the target variable. In addition, the metadata includes the radius of the strong wind circle that represents the size of a tropical cyclone, so we can also design a regression task for size using the radius as the target variable.

Track The task makes inferences on the geographical location of a typhoon. The cyclone's center, as estimated by human experts, is represented by latitude and longitude coordinates with a precision of 0.1 degrees. A regression task for predicting the typhoon's location uses these latitude and longitude coordinates as target variables.

Formation The task makes inferences on the birth of a tropical cyclone, which typically occurs in tropical regions. Among the numerous cloud clusters actively evolving in tropical regions, determining which one will evolve into a tropical cyclone presents a challenging forecasting task, making it a target for machine learning applications [6].

Transition The task makes inferences on the transition from a tropical cyclone to an extra-tropical cyclone, which typically occurs in mid-latitude regions. Two types of cyclones are conceptually distinct from a meteorological perspective, but as a natural phenomenon, they are continuous. The data-driven modeling of a continuous transition process connecting two discrete concepts is a machine-learning task.

4.3 Meteorological Analysis

Machine learning can also be applied to analyze meteorological events on tropical cyclones, such as rapid intensification [2], eyewall replacement [14], and overshooting cloud tops [20]. These events may be linked with the forecasting of tropical cyclones, yet their underlying mechanisms are not entirely understood. Data-driven methodologies could potentially provide insights that contribute to the development of a novel theoretical framework for understanding these phenomena.

4.4 Analysis for Societal Impact

The Digital Typhoon dataset represents the atmospheric observation of a tropical cyclone, but its societal impacts are measured by different sources and modalities. For example, hazards are measured by heavy rainfall or strong winds, disasters are measured by landslides and flooding, and damages are measured by human casualties and financial loss. To construct a machine learning model to analyze and forecast the societal impact, real-world datasets from many sources should be integrated with meteorological datasets. This would enable a more comprehensive understanding of the full range of impacts arising from tropical cyclones.

4.5 Analysis for Climate Change

Understanding how a long-term tropical cyclone activity is impacted by climate change is a crucial topic in society [32, 29, 31, 46]. Technological and methodological evolution that occurred during the 40+ years lifespan introduces many types of biases in the dataset. While certain biases may be removed by sensor calibration, others are harder to detect such as annotation errors by human experts. The reanalysis of historical data and the creation of a homogeneous dataset can contribute to advancing our knowledge of the relationship between tropical cyclones and climate change.

5 Benchmarks

Machine learning tasks can be combined to create benchmarks for machine learning. We propose three benchmarks, 1) Analysis, 2) Forecasting, and 3) Reanalysis of the intensity of typhoons.

In meteorological time series, data are auto-correlated and one has to be careful how to split the data before starting to train a new network [48]. At least, a random split for the image level must not be used to avoid overestimating the performance due to data leakage in the same typhoon sequence. Random splits can be applied for the sequence level (split-by-sequence) or the season level (split-by-season), but more complex splits can be designed to reflect the change in data quality in 40 years, such as split by satellite generations (1978-2005, 2005-2015, 2015-2022). These designed splits are especially useful for the reanalysis task in Section 5.3.

Table 2: The result of the pressure regression task for two architectures and three types of input.

RMSE (hPa)	Full (512×512)	Resized (224×224)	Cropped (224×224)
ResNet18	10.51 ($\pm$0.11)	10.47($\pm$0.20)	**10.06** ($\pm$0.09)
ResNet50	11.12 ($\pm$0.41)	11.63 ($\pm$0.35)	10.09 ($\pm$0.04)

Table 3: The result of the wind regression task for two architectures and three types of input.

RMSE (kt)	Full (512×512)	Resized (224×224)	Cropped (224×224)
ResNet18	10.21 ($\pm$0.19)	10.09 ($\pm$0.08)	**9.25** ($\pm$0.25)
ResNet50	10.05 ($\pm$0.26)	10.21 ($\pm$0.14)	9.13 ($\pm$0.11)

To perform those benchmarks, we developed a software library pyphoon2 written in Python and released from the GitHub repository `https://github.com/kitamoto-lab/digital-typhoon/`. It has a data loader and components to help build machine learning pipelines. All the experiments were performed on the internal cluster with 6 GPUs consisting of NVIDIA Quadro RTX 6000, NVIDIA Quadro RTX 8000, and NVIDIA Quadro RTX A6000.

5.1 Analysis for the Intensity

As stated before, the Dvorak technique is the most popular method for the analysis of intensity. A question here is how machine learning can perform this task in comparison to manual analysis. We propose classification tasks, which take an image as input and estimate grade as output, and regression tasks, which take an image as input and estimate a pressure or wind value as output. Note that central pressure in hectopascal (hPa) is available for all grades, whereas maximum wind speed in knot (kt) is available for only grades 3, 4, and 5. In the following, we describe the result of the regression task, and the result of the classification task is described in the appendix.

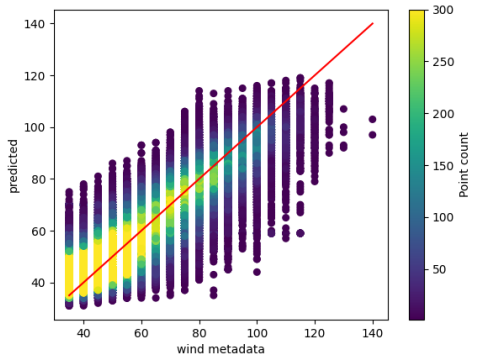
We explored three types of comparisons. First, we compared three architectures, namely VGG ([50], ResNet [13] and Vision Transformer [9]. Second, we compared models trained on 1) full-resolution images (512×512), 2) resized images (224×224), and 3) images cropped around the typhoon center (224×224). Third, we compared two target values, namely pressure and wind.

We used the built-in torchvision ResNet18 and ResNet50 models with a learning rate (LR) of 10^{-4}, batch size of 16, and for 50 epochs. A 80/20 train/test split by sequence was used. The ResNet18 and ResNet50 models were trained five and two times respectively. To evaluate, we measured the root mean square error (RMSE) of the prediction from ground truth and their standard deviations ($\pm$ std).

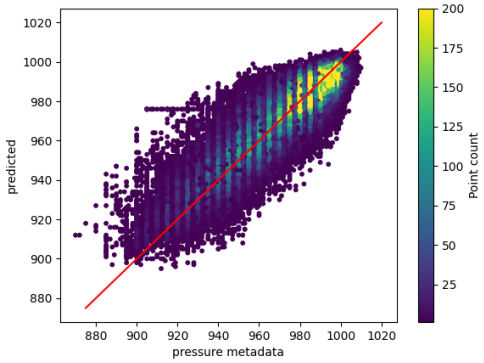
Table 2 and Table 3 summarize the results. Firstly, ResNet50 yielded similar results to ResNet18. Secondly, cropping the images around the typhoon center yielded a lower RMSE than other choices, indicating that cropping may preserve features around the typhoon center better than resizing and that training a model on the full images may be difficult due to their larger size. Furthermore, Figure 3 illustrates prediction plots for the ResNet18 model. They show that the regression performs better for weaker typhoons, but worse for stronger typhoons.

5.2 Forecasting for the Intensity

Forecasting for the intensity is currently performed in computational and empirical approaches. Computational approaches try to represent a typhoon in a simulation model and compute the future based on the theory of the atmosphere. However, this approach has limitations due to spatial and temporal resolutions, and intensity forecasting is still considered a difficult challenge. Instead, meteorologists developed empirical methods, such as SHIPS [12, 55] with linear regression on hand-crafted meteorological features, or a similar approach using XGBoost [4]. A question here is how machine learning can perform this task in comparison to computational or empirical approaches.



(a) ResNet18 predictions plots for wind regression with cropped (224x224) images



(b) ResNet18 predictions plots for pressure regression with cropped (224x224) images

Figure 3: Prediction plots for pressure and wind regressions.

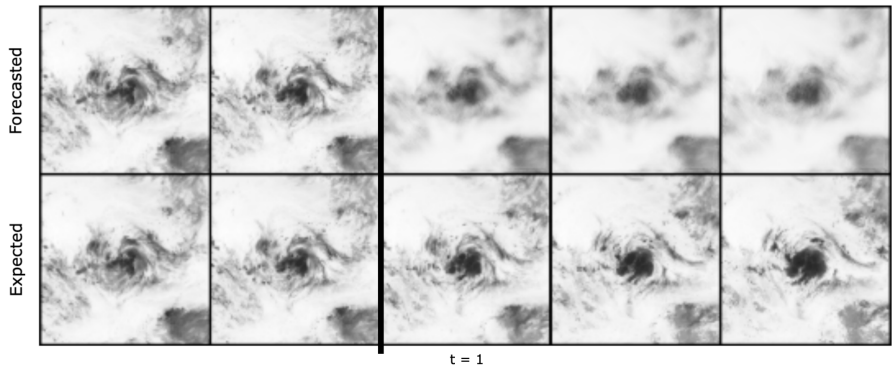


Figure 4: Results of image forecasting by ConvLSTM.

Our previous work used Recurrent Neural Network (RNN) to forecast the pressure directly from images and showed comparable performance with SHIPS [40]. In this paper, we chose another approach using a convolutional LSTM [49] to predict the next n image frames of a typhoon given the previous 12 image frames and analyze the pressure from the predicted image. We adapted an implementation [39], and used a 3-layer ConvLSTM with 128 hidden dimensions. Due to resource limitations, we used 128×128 downsampled images from only the first 24 hours of a given typhoon.

To forecast n hours into the future (starting at $t = 1$), the 12 preceding frames ($t = [-11, 0]$) are passed into the model, which outputs a single image serving as its forecast at $t = 1$. Then, images from $t = [-10, 1]$, including the predicted image, are passed back into the model to get the prediction for $t = 2$. This process is repeated n times to forecast n hours into the future. As a result, Figure 4 shows that the first predicted frame is perceptively blurred and rapidly deteriorates as t advances.

We then trained a ResNet18 model on predicted images, as well as the first 24 images of every typhoon, to predict the pressure given a 128×128 image. As a result, Table 4 shows that intuitively, the model produces a larger RMSE and error as t advances due to the blur of predicted images. A future adaptation may be to train both the ConvLSTM and ResNet in a black box, such that the loss is minimized by image reproduction and pressure prediction.

Both models were trained on the same 80/20 train/test split by sequence. The ConvLSTM was trained once for 230 epochs with a starting LR of 10^{-4}, and used a CosineAnnealing scheduler [34] with 100 steps. The ResNet model used a modified first convolutional layer with a kernel and stride size of (2, 2) and (1, 1). It was trained five times with an LR of 10^{-5} for 34 epochs. These hyperparameters were chosen as they produced more consistent results given the smaller image sizes.

Table 4: Results of pressure forecasting for 12-hours by ResNet18 (values in hPa).

t	1	2	3	6	12
RMSE	10.24 ± 0.73	10.52 ± 0.79	11.00 ± 0.87	12.10 ± 0.85	14.69 ± 0.91

Table 5: The results of regression tasks for each satellite generation (values in hPa).

RMSE	Train the First	Train the Second	Train the Third
Test the First	9.16 ($\pm$0.05)	10.43 ($\pm$0.09)	10.51 ($\pm$0.11)
Test the Second	9.23 ($\pm$0.09)	10.39 ($\pm$0.18)	10.16 ($\pm$0.09)
Test the Third	9.56 ($\pm$0.11)	9.96 ($\pm$0.10)	10.22 ($\pm$0.08)

5.3 Reanalysis for the Intensity

The goal of this paper is to create a homogeneous long-term dataset, and the purpose of the reanalysis task is to identify biases and inconsistencies in the dataset due to factors such as technological evolution arising from satellite sensors, or methodological evolution arising from the improvement of the Dvorak technique to annotate tropical cyclones. One approach to this challenge is to design special data splits to analyze historical factors.

Our previous work studied this task by training the model using recent data and testing on past data to analyze the trend of model performance, indicating that old satellite data may have different characteristics [40]. In this paper, we split the dataset into three buckets by satellite generations, namely the first generation (1978-2005), the second generation (2005-2015), and the third generation (2015-2022), and train and test a ResNet18 model for the regression task.

Input images were resized to 224×224 in the same way as the analysis task. Each bucket was then split into 80/20 train/test sets. Three separate ResNet18 models were trained on each bucket, each for 101 epochs with a batch size of 16 and a learning rate of 10^{-4}; these parameters were chosen for their consistent results. Each of the three models was then tested on the test set of each bucket, such that a model trained on the first bucket was tested on the first, second, and third buckets.

As a result, Table 5 shows that the model trained on the first bucket performed better than the models trained on the second and third data. However, we note that the first bucket contains almost three times more typhoons (531 sequences) than the second bucket (183 sequences) and the third bucket (167 sequences). To acquire more reliable results, we should reduce the effect of dataset size with better historical data splitting.

6 Conclusion

We have introduced the Digital Typhoon dataset for the machine learning community and the meteorology community to promote data-driven research on tropical cyclones. Our dataset offers a unique opportunity to benchmark various types of machine learning models, especially spatio-temporal models for long-term time-series images. A solution to this dataset is not only valuable for machine learning benchmarking but also has the potential to contribute to advancing scientific knowledge on tropical cyclones as well as solving societal and sustainability issues such as disaster reduction and climate change.

Acknowledgments and Disclosure of Funding

Three of the authors, Jared Hwang, Bastien Vuillod, and Lucas Gautier, have been supported by the international internship program of the National Institute of Informatics.

References

[1] Saleh Ashkboos, Langwen Huang, Nikoli Dryden, Tal Ben-Nun, Peter Dominik Dueben, Lukas Gianinazzi, Luca Nicola Kummer, and Torsten Hoefler. ENS-10: A Dataset For Post-Processing Ensemble Weather Forecasts. September 2022.

[2] Kieran T. Bhatia, Gabriel A. Vecchi, Thomas R. Knutson, Hiroyuki Murakami, James Kossin, Keith W. Dixon, and Carolyn E. Whitlock. Recent increases in tropical cyclone intensification rates. *Nature Communications*, 10(1):635, February 2019. Number: 1 Publisher: Nature Publishing Group.

[3] Salva Rühling Cachay, Venkatesh Ramesh, Jason N. S. Cole, Howard Barker, and David Rolnick. ClimART: A Benchmark Dataset for Emulating Atmospheric Radiative Transfer in Weather and Climate Models. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, October 2021.

[4] Ming Hei Kenneth Chan, Wai Kin Wong, and Kin Chung Au-Yeung. Machine learning in calibrating tropical cyclone intensity forecast of ecmwf eps. *Meteorological Applications*, 28(6):e2041, 2021.

[5] Buo-Fu Chen, Boyo Chen, Hsuan-Tien Lin, and Russell L. Elsberry. Estimating Tropical Cyclone Intensity by Satellite Imagery Utilizing Convolutional Neural Networks. *Weather and Forecasting*, 34(2):447–465, April 2019. Publisher: American Meteorological Society Section: Weather and Forecasting.

[6] Rui Chen, Xiang Wang, Weimin Zhang, Xiaoyu Zhu, Aiping Li, and Chao Yang. A hybrid CNN-LSTM model for typhoon formation forecasting. *GeoInformatica*, 23(3):375–396, July 2019.

[7] Rui Chen, Weimin Zhang, and Xiang Wang. Machine Learning in Tropical Cyclone Forecast Modeling: A Review. *Atmosphere*, 11(7):676, July 2020. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[8] Muhammad Dawood, Amina Asif, and Fayyaz ul Amir Afsar Minhas. Deep-PHURIE: deep learning based hurricane intensity estimation from infrared satellite imagery. *Neural Computing and Applications*, 32(13):9009–9017, July 2020.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[10] Vernon F. Dvorak. Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery. *Monthly Weather Review*, 103(5):420–430, May 1975.

[11] Kerry Emanuel. 100 Years of Progress in Tropical Cyclone Research. *Meteorological Monographs*, 59:15.1 – 15.68, 2018. Place: Boston MA, USA Publisher: American Meteorological Society.

[12] Patrick J. Fitzpatrick. Understanding and Forecasting Tropical Cyclone Intensity Change with the Typhoon Intensity Prediction Scheme (TIPS). *Weather and Forecasting*, 12(4):826–846, December 1997. Publisher: American Meteorological Society Section: Weather and Forecasting.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015.

[14] Eric A. Hendricks, Scott A. Braun, Jonathan L. Vigh, and Joseph B. Courtney. A summary of research advances on tropical cyclone intensity change from 2014-2018. *Tropical Cyclone Research and Review*, 8(4):219–225, December 2019.

[15] Christopher C. Hennon, Kenneth R. Knapp, Carl J. Schreck, Scott E. Stevens, James P. Kossin, Peter W. Thorne, Paula A. Hennon, Michael C. Kruk, Jared Rennie, Jean-Maurice Gadéa, Maximilian Striegl, and Ian Carley. Cyclone Center: Can Citizen Scientists Improve Tropical Cyclone Intensity Records? *Bulletin of the American Meteorological Society*, 96(4):591 – 607, 2015. Place: Boston MA, USA Publisher: American Meteorological Society.

[16] Maiki Higa, Shinya Tanahara, Yoshitaka Adachi, Natsumi Ishiki, Shin Nakama, Hiroyuki Yamada, Kosuke Ito, Asanobu Kitamoto, and Ryota Miyata. Domain knowledge integration into deep learning for typhoon intensity classification. *Scientific Reports*, 11(1):12972, June 2021. Number: 1 Publisher: Nature Publishing Group.

[17] Kosuke Ito, Hiroyuki Yamada, Munehiko Yamaguchi, Tetsuo Nakazawa, Norio Nagahama, Kensaku Shimizu, Tadayasu Ohigashi, Taro Shinoda, and Kazuhisa Tsuboki. Analysis and Forecast Using Dropsonde Data from the Inner-Core Region of Tropical Cyclone Lan (2017) Obtained during the First Aircraft Missions of T-PARCII. *Sola*, 14:105–110, 2018.

[18] Neeru Jaiswal, C. M. Kishtawal, and P. K. Pal. Cyclone intensity estimation using similarity of satellite IR images based on histogram matching approach. *Atmospheric Research*, 118:215–221, November 2012.

[19] Viju O. John, Tasuku Tabata, Frank Rüthrich, Rob Roebeling, Tim Hewison, Reto Stöckli, and Jörg Schulz. On the Methods for Recalibrating Geostationary Longwave Channels Using Polar Orbiting Infrared Sounders. *Remote Sensing*, 11(10):1171, January 2019. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[20] Miae Kim, Jungho Im, Haemi Park, Seonyoung Park, Myong-In Lee, and Myoung-Hwan Ahn. Detection of Tropical Overshooting Cloud Tops Using Himawari-8 Imagery. *Remote Sensing*, 9(7):685, July 2017. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[21] Asanobu KITAMOTO. The development of typhoon image database with content-based search. In *Proceedings of the 1st International Symposium on Advanced Informatics (AdInfo)*, pages 163–170, 3 2000.

[22] Asanobu Kitamoto. Evolution Map: Modeling State Transition of Typhoon Image Sequences by Spatio-Temporal Clustering. In Steffen Lange, Ken Satoh, and Carl H. Smith, editors, *Discovery Science*, Lecture Notes in Computer Science, pages 283–290, Berlin, Heidelberg, 2002. Springer.

[23] Asanobu Kitamoto. Spatio-Temporal Data Mining for Typhoon Image Collection. *Journal of Intelligent Information Systems*, 19(1):25–41, July 2002.

[24] Asanobu Kitamoto. Typhoon Analysis and Data Mining with Kernel Methods. In Seong-Whan Lee and Alessandro Verri, editors, *Pattern Recognition with Support Vector Machines*, Lecture Notes in Computer Science, pages 237–249, Berlin, Heidelberg, 2002. Springer.

[25] Asanobu KITAMOTO. Digital typhoon: Near real-time aggregation, recombination and delivery of typhoon-related information. In *Proceedings of the 4th International Symposium on Digital Earth (ISDE)*, page 16 pages, 3 2005.

[26] Asanobu KITAMOTO and Kinji ONO. The construction of typhoon image collection and its application to typhoon analysis. *NII Journal*, (1):7–22, 12 2000. (in Japanese).

[27] Kenneth R. Knapp. Scientific data stewardship of international satellite cloud climatology project B1 global geostationary observations. *Journal of Applied Remote Sensing*, 2(1):023548, November 2008.

[28] Kenneth R. Knapp, Michael C. Kruk, David H. Levinson, Howard J. Diamond, and Charles J. Neumann. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying Tropical Cyclone Data. *Bulletin of the American Meteorological Society*, 91(3):363–376, March 2010.

[29] Thomas R. Knutson, John L. McBride, Johnny Chan, Kerry Emanuel, Greg Holland, Chris Landsea, Isaac Held, James P. Kossin, A. K. Srivastava, and Masato Sugi. Tropical cyclones and climate change. *Nature Geoscience*, 3(3):157–163, March 2010. Number: 3 Publisher: Nature Publishing Group.

[30] James P. Kossin. New global tropical cyclone data set from ISCCP B1 geostationary satellite observations. *Journal of Applied Remote Sensing*, 1(1):013505, February 2007.

[31] James P. Kossin, Timothy L. Olander, and Kenneth R. Knapp. Trend Analysis with a New Global Record of Tropical Cyclone Intensity. *Journal of Climate*, 26(24):9960–9976, December 2013. Publisher: American Meteorological Society Section: Journal of Climate.

[32] Christopher W. Landsea, Bruce A. Harper, Karl Hoarau, and John A. Knaff. Can We Detect Trends in Extreme Tropical Cyclones? *Science*, 313(5786):452–454, July 2006. Publisher: American Association for the Advancement of Science.

[33] Juhyun Lee, Jungho Im, Dong-Hyun Cha, Haemi Park, and Seongmun Sim. Tropical Cyclone Intensity Estimation Using Multi-Dimensional Convolutional Neural Networks from Geostationary Satellite Data. *Remote Sensing*, 12(1):108, January 2020. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[35] Manil Maskey, Rahul Ramachandran, Muthukumaran Ramasubramanian, Iksha Gurung, Brian Freitag, Aaron Kaulfus, Drew Bollinger, Daniel J. Cecil, and Jeffrey Miller. Deepti: Deep-Learning-Based Tropical Cyclone Intensity Estimation System. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:4271–4281, 2020. Conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing.

[36] Daisuke Matsuoka, Chihiro Kodama, Yohei Yamada, and Masuo Nakano. Tropical cyclone dataset for a high-resolution global nonhydrostatic atmospheric simulation. *Data in Brief*, 48:109135, June 2023.

[37] Timothy Olander, Anthony Wimmers, Christopher Velden, and James P. Kossin. Investigation of machine learning using satellite-based advanced dvorak technique analysis parameters to estimate tropical cyclone intensity. *Weather and Forecasting*, 36(6):2161 – 2186, 2021.

[38] Timothy L. Olander and Christopher S. Velden. The Advanced Dvorak Technique: Continued Development of an Objective Scheme to Estimate Tropical Cyclone Intensity Using Geostationary Infrared Satellite Imagery. *Weather and Forecasting*, 22(2):287–298, April 2007. Publisher: American Meteorological Society Section: Weather and Forecasting.

[39] Rohit Panda. Video frame prediction using convlstm network in pytorch, 6 2021.

[40] Clément Playout and Asanobu KITAMOTO. Latent space representation and rnn for image-based typhoon intensity analysis and prediction. In *The 9th International Workshop on Climate Informatics (CI2019)*, pages 47–52, 10 2019.

[41] Ritesh Pradhan, Ramazan S. Aygun, Manil Maskey, Rahul Ramachandran, and Daniel J. Cecil. Tropical Cyclone Intensity Estimation Using a Deep Convolutional Neural Network. *IEEE Transactions on Image Processing*, 27(2):692–702, February 2018.

[42] Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Prabhat, and Christopher Pal. Extreme weather: a large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 3405–3416, Red Hook, NY, USA, December 2017. Curran Associates Inc.

[43] Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A benchmark dataset for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), November 2020. arXiv:2002.00469 [physics, stat].

[44] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195–204, February 2019.

[45] Lucas Rodés-Guirao. *Deep Learning for Digital Typhoon : Exploring a typhoon satellite image dataset using deep learning*. 2019.

[46] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, S. Karthik Mukkavilli,

Konrad P. Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer Chayes, and Yoshua Bengio. Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, 55(2):42:1–42:96, February 2022.

[47] Carl J. Schreck, Kenneth R. Knapp, and James P. Kossin. The Impact of Best Track Discrepancies on Global Tropical Cyclone Climatologies using IBTrACS. *Monthly Weather Review*, 142(10):3881–3899, October 2014. Publisher: American Meteorological Society Section: Monthly Weather Review.

[48] M. G. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194):20200097, February 2021. Publisher: Royal Society.

[49] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. 2015.

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2015.

[51] Tasuku Tabata, Viju O. John, Rob A. Roebeling, Tim Hewison, and Jörg Schulz. Recalibration of over 35 Years of Infrared and Water Vapor Channel Radiances of the JMA Geostationary Satellites. *Remote Sensing*, 11(10):1189, January 2019. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[52] Wei Tian, Wei Huangwei, Xiaolong Xu, and Chao Wang. Tropical Cyclone Maximum Wind Estimation from Infrared Satellite Data with Integrated Convolutional Neural Networks. In *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 575–580, July 2019.

[53] Christopher Velden, Bruce Harper, Frank Wells, John L. Beven, Ray Zehr, Timothy Olander, Max Mayfield, Charles "Chip" Guard, Mark Lander, Roger Edson, Lixion Avila, Andrew Burton, Mike Turk, Akihiro Kikuchi, Adam Christian, Philippe Caroff, and Paul McCrone. The Dvorak Tropical Cyclone Intensity Estimation Technique: A Satellite-Based Method that Has Endured for over 30 Years. *Bulletin of the American Meteorological Society*, 87(9):1195–1210, September 2006.

[54] Zhen Wang, Jun Zhao, Hong Huang, and Xuezhong Wang. A Review on the Application of Machine Learning Methods in Tropical Cyclone Forecasting. *Frontiers in Earth Science*, 10, 2022.

[55] Munehiko Yamaguchi, Hiromi Owada, Udai Shimada, Masahiro Sawada, Takeshi Iriguchi, Kate D. Musgrave, and Mark DeMaria. Tropical Cyclone Intensity Prediction in the Western North Pacific Basin Using SHIPS and JMA/GSM. *Sola*, 14:138–143, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] Described in Section 1 and 3.
   (b) Did you describe the limitations of your work? [Yes] Described in Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Will be discussed in the supplemental material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] Yes
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] In the supplemental material and GitHub.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Described in Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Showed standard deviation in Section 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Described in Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Mentioned the name of the data sources.
   (b) Did you mention the license of the assets? [Yes] Most data are not copyrighted.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Available at http://agora.ex.nii.ac.jp/digital-typhoon/dataset/
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Consent is not necessary
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Table 6: Results of the classification task for three architectures and three types of input.

Acc (%)	Full (512×512)	Resized (224×224)	Cropped (224×224)
VGG	70.0 ($\pm$1.4)	67.2 ($\pm$0.2)	69.0 ($\pm$0.6)
ResNet18	66.6 ($\pm$0.7)	66.3 ($\pm$0.9)	67.8 ($\pm$0.2)
ViT	/	62.0 ($\pm$0.2)	64.9 ($\pm$1.2)

A Additional Benchmark Results

We already introduced benchmark results for three types of tasks in Section 5. The following are additional results of the same tasks, which were omitted due to space limitations.

A.1 Analysis for the Intensity

In comparison to the regression task in Section 5, we summarize the result of the classification task. Here the target value is the categorical value of grades 2, 3, 4, 5, and 6. To compare the performance of the three types of architectures, namely VGG, ResNet18, and Vision Transformer (ViT), we run 2 training sessions on three types of input images. For all experiments, we used a learning rate of 10^{-4}, batch size of 16, a 80/20 train/test split by sequence, the SGD optimizer, and 50 epochs. ViT for the full-size images was not performed due to the lack of memory.

Table 6 summarizes the result of classification in accuracy. VGG is slightly better than ResNet18, while ViT performs worst. This result indicates that the dataset size may not be large enough for models that performed worse. However, we need to work in two directions to explore better models. First, we need to explore deeper architectures, such as ResNet50 or even deeper ResNet, to see how the depth of the network relates to the performance. Second, we need to increase the number of epochs for some models, because the best accuracy of ViT was reached during the last epochs.

Note that the result of Table 6, showing a classification accuracy of 70% at best, is likely to be an underestimate of the actual performance. This is due to the definition of grades, used as the target values of the classification task. As stated earlier in Section 4.2, grades 3, 4, and 5 denote a tropical cyclone with the ascending order of intensity, while grades 2 and 6 denote a tropical depression and an extra-tropical cyclone respectively. These categorical values, however, are continuous and ambiguous as labels for classification.

First, grades 3, 4, and 5 are continuous categories and hence the boundary between categories is not clear-cut. Considering the potential annotation errors in the best track dataset as addressed in Section E, misclassification between neighboring categories should not be considered fatal mistakes. Second, grades 2 and 3 are also neighboring categories that differ only in intensity, whereas grades 3-5 and 6 differ in the structure of the cyclone. This suggests that the categories have semantics that may give different impacts on misclassification. Hence the introduction of a more realistic loss function between categorical transitions is expected to alleviate these problems.

Considering the difficulty in evaluating the classification task, we suggest that the regression task, especially on the central pressure, has a less ambiguous definition of the task. This is the reason that we introduced the result of the regression task in the main text.

A.2 Reanalysis for the Intensity

In Section 5, the reanalysis task was performed on data splits by satellite generations. To remove bias in performance due to different dataset sizes, however, we prepared another data split so that each satellite generation has roughly the same dataset size by sampling 167 typhoons for training and 41 typhoons for testing. We then applied the same methodology used in Section 5.3 to train 3 ResNet18 models for 101 epochs with 6 training sessions.

Table 7: The results of regression tasks for each satellite generation using the same numbers of sequences (values in hPa).

RMSE	Train the First	Train the Second	Train the Third
Test the First	10.04 ($\pm$0.17)	9.92 ($\pm$0.09)	10.03 ($\pm$0.10)
Test the Second	12.80 ($\pm$0.19)	11.05 ($\pm$0.10)	11.17 ($\pm$0.10)
Test the Third	10.34 ($\pm$0.17)	10.03 ($\pm$0.08)	9.94 ($\pm$0.16)

Comparison of Table 7 with Table 5 suggests that higher performance of the first generation in Table 5 was due to the larger dataset size, because performance for each generation is roughly the same in Table 7. To study historical changes in dataset quality, we will try other data splits to take advantage of the whole dataset.

B Data Collection

B.1 Image Dataset

Satellite Generations and Data Formats The satellite image dataset, as introduced in Section 2.2, was built on geostationary meteorological satellites, Himawari-1 to Himawari-9 from JMA and GOES-9 from NOAA. The data is delivered through the following organizations.

1. Japan Meteorological Business Support Center (1978-1995, 2003-2022)
2. Institute of Industrial Science, The University of Tokyo (1995-2003)

Himawari satellite data has three types of data formats.

1. VISSR / S-VISSR format for the first-generation satellite (Himawari 1-5, GOES 9)
2. HRIT format for the second-generation satellite (Himawari 6-7)
3. Himawari Standard format for the third-generation satellite (Himawari 8-9)

Due to the lack of standard open-source tools for parsing data formats of all generations, we built a parser for these data formats from scratch based on the official documentation published by JMA.

Choice of Channels Himawari satellite data has a few channels to observe the atmosphere in different wavelengths. The visible channel (0.6μm) measures the reflection and scattering of the sunlight from the Earth, so it has advantages in observing the detailed structure of the cloud patterns. It has a higher resolution but can be used only in the daytime. The infrared channel (11μm) measures the radiation of the Earth, roughly corresponding to the temperature of the atmosphere. It has a lower resolution but can be used both day and night. Due to the continuity of the observation, we used the infrared channel to create the standard hourly satellite image dataset for typhoon images.

Note that recent satellites have more channels. Since Himawari-5, we have the water vapor channel (6.7μm) to measure the moisture in the atmosphere, and the near-infrared channel (3.9μm) to measure the vegetation on the ground. Images of those channels are processed and provided from the Digital Typhoon website and may be included in the future release of the Digital Typhoon dataset.

B.2 Track Dataset

The track dataset, as introduced in Section E, is based on the best track data from JMA, available at `https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/besttrack.html`. We incorporated most of the best track data, such as grade, location, central pressure, maximum sustained wind, wind circles for the storm wind (50kt) and gale wind (30kt), and the indicator

of landfall or passage. The best track record is available every six hours or at shorter intervals in exceptional cases.

Because the satellite image dataset is an hourly dataset, the best track should be interpolated for missing hours. We designed the interpolation method for each variable so that the interpolation aligns with the design of machine learning tasks that use some of those values as 'ground truth' for supervised learning.

Location Location is interpolated using cubic splines [26] to create a smooth trajectory connecting best track locations.

Grade The grade is persistent throughout the interpolation interval because it is a categorical value.

Central pressure Central pressure is interpolated using a linear function connecting best track pressures. This is a choice to conserve maxima and minima and avoid the shooting of pressures beyond observed values. At the same time, a linear function can reflect continuous changes in the intensity of the typhoon.

Maximum sustained wind The maximum sustained wind is persistent throughout the interpolation interval. The wind is a numerical value, so it matches well with the linear interpolation. However, we should consider the fact that the wind is only available for grades 3, 4, and 5, but not for grades 2 and 6 with the wind value zero. This leads to discontinuity of the wind values at the time of formation (from grade 2 to grade 3-5) and transition (from grade 3-5 to grade 6) of typhoons. To avoid inappropriate linear interpolation at discontinuous points, we decided to keep the value persistent. The same rule applies to wind circles.

C Data Processing Workflow

After data collection, the original satellite image is fed into a data processing pipeline, which has been developed by the author (Asanobu Kitamoto) since 1992. The pipeline consists of software codes written in C and Perl. They are not open-sourced due to the complex structure of software. Note that pyphoon2, which is an open-source software library to work with the "Digital Typhoon dataset" for machine learning tasks, has been released as introduced in Section 5.

The summary of the pipeline is as follows.

1. Parse the original satellite image and create a map-projected 2D array image with digital count as pixel values (Section C.1).
2. Convert each pixel value from digital count (integer value) to brightness temperature (floating-point value). The conversion is done in two steps (Section C.2).
   (a) The first step is to convert the digital count to brightness temperature using the conversion table for each sensor.
   (b) The second step is to convert brightness temperature to recalibrated brightness temperature using inter-calibration parameters from the recalibration project [51, 19].
3. The recalibrated brightness temperature, including a masked pixel value, is saved in the HDF5 file (Section C.3).

In the following, we summarize the details of each step.

C.1 Map projection

The map projection is used to create a typhoon-centered image. When a typhoon is observed from a satellite, the shape of the typhoon is usually distorted from the satellite's viewpoint due to the curvature of the Earth's 3D surface. Here the role of map projection is two-fold. First, the map

projection creates a typhoon image where the typhoon center is located at the center of the image. Second, the map projection reduces the distortion of the typhoon by creating an image from a viewpoint above the typhoon center.

The choice of the map projection method depends on the choice of metric properties that should be preserved at the expense of others. We chose the Lambert azimuthal equal-area projection for the Digital Typhoon dataset to preserve the area and a circular shape around the center. This is in contrast to the choice of map projection in the HURSAT dataset (Section 3.2), where their choice, equirectangular projection (lat/long grid), does not preserve any metric properties. But at least, it is better than a simple image cropping from the original satellite image, because the distortion of the shape is reduced during the map projection.

Next, we consider the size of the typhoon image after map projection. Here, the size has two aspects, either the size of the scene or the size of the image. First, about the size of the scene on the original satellite image, we refer to the maximum circle of gale wind (30kt), whose diameter is 1275nm (nautical mile) or 2361km recorded by Typhoon 199713. Based on this record, we set the diameter of the scene as 2500km. Second, about the size of the image, we decided to use a square image of 512 pixels. This amounts to a resolution of 4.88km/pixel, which is close to the maximum resolution of 4km for the infrared image of old-generation satellites.

The last choice is the geometric transformation for map projection or output-to-input mapping. Some typical choices are bi-linear or bi-cubic interpolation, but they could introduce non-existent pixel values due to the mixture of observation values from multiple pixels, and conversion from digital count to brightness temperature, which will be addressed below, becomes more complicated. To avoid this problem, we chose a simpler nearest-neighbor method that preserves original pixel values. Note that some pixel values are copied more than once when the resolution of the original satellite image is not enough due to the curvature of the Earth.

C.2 Calibration

The Digital Typhoon dataset is the collection of 40+ years of satellite data from 10 different satellites, and calibration for removing biases in each sensor is required to create a homogeneous long-term dataset. The original satellite image is recorded as the collection of the digital count, which is an integer value that records the response of the sensor. Depending on the precision of the sensor, the digital count is represented by 6 bits for a low-precision sensor and 12 bits for a high-precision sensor. From these digital counts before calibration, the following procedure is designed to produce calibrated values.

The first step is to convert the digital count to brightness temperature, which is a physical value measured in Kelvin (K) representing the temperature of the Earth such as cloud top, ground, and ocean surface. The conversion table from the digital count to the brightness temperature is provided for each sensor based on the initial calibration of the sensor.

The second step is to apply a recalibration equation to convert brightness temperature to recalibrated brightness temperature. This recalibration method is effective for inter-calibration across satellite sensors to obtain even better homogeneous long-term observation data for climate change research. This recalibrated brightness temperature is included in the Digital Typhoon dataset.

Note that some machine-learning papers use typhoon images that are scraped from public typhoon information websites, including the Digital Typhoon website. However, the approach of using such scraped images for machine learning research suffers from a number of problems. First, the JPEG format contaminates pixel values with lossy compression. Second, the pixel value of a popular image format does not have a physical meaning. Third, some scraped images contain graphical elements such as the border of the ground and lat/long lines that introduce additional noise. The Digital Typhoon dataset released in this paper can remove all those problematic aspects so that machine learning researchers can focus on designing algorithms and models on a clean dataset.

C.3 Masking

The final step of the pipeline is to save recalibrated brightness temperature to a file. Because each pixel takes a floating point value, we decided to use the HDF5 (Hierarchical Data Format 5), which is a popular data format for scientific applications to deal with a multidimensional array of floating point values.

Here a problem arises with pixels that do not have observation data. This happens in the following situations.

1. A pixel is out-of-frame. It occurs when a typhoon is located at the periphery of a satellite image, and hence a typhoon-centered image overlaps with the boundary of the satellite image. It also occurs when a satellite observation was partial due to scheduled maintenance or emergency operations.
2. A pixel is contaminated by noise. It occurs in old satellites when the sensor malfunctions. These noises can be easily detected when the pixel value takes the minimum or maximum values of the range but is more difficult to detect when the noise is only visible as an irregular spatial pattern.

In HDF5, there is no standard way to represent a pixel value without observation (invalid or null values). Hence we used a temperature of 130.0, which is below the valid brightness temperature (130K or -140 degrees Celsius). They can be used to mask invalid pixels when machine-learning models can properly treat masked pixels.

Note that we can further normalize the brightness temperature as input to machine learning models. In our benchmarking experiments, we used the standard normalization procedure to map a brightness temperature of $[170, 300]$ to a normalized value of $[0, 1]$. In this setting, the masked pixels are mapped to 0 together with other extremely cold pixels.

D Dataset Organization and Versioning

The directory structure of the dataset is summarized in Figure 5. The dataset consists of the image directory, containing the list of HDF5 files for hourly images grouped by each typhoon, and the metadata directory, containing one file for each typhoon with the record of best track data and the quality of the image of the corresponding observation time. At the top level, we also provide the metadata.json file, which contains information about each typhoon in the dataset, such as the season, the number of images, and the name of the typhoon.

The Digital Typhoon dataset will be updated when the best track of new tropical cyclones becomes available. Furthermore, at the end of a typhoon season in winter, the whole best track dataset may be reanalyzed to incorporate a modern understanding of tropical cyclones.

To cope with this updating pattern, we will create an annual dataset by naming the dataset V.YYYY, where YYYY represents the last typhoon season of the dataset. Following this pattern, the current release is V.2022 with the best track of all typhoons by the 2022 season. Currently, the size of the dataset V.2022 is 54GB to download as a ZIP file.

After creating new data for the 2023 season, we will release V.2023.1, V.2023.2, ..., as the diff dataset from V.2022, and finally publish V.2023 that includes all typhoons by the 2023 season. The V.2023 may also include the update of old typhoons due to the reanalysis of the best track dataset. We will add brief comments about which part of the dataset is updated, but will not provide the diff between the two datasets.

E Responsible Use

The dataset should be used responsibly so that machine learning results do not mislead the public in terms of disaster preparedness and response. For example, in Japan, the public announcement of

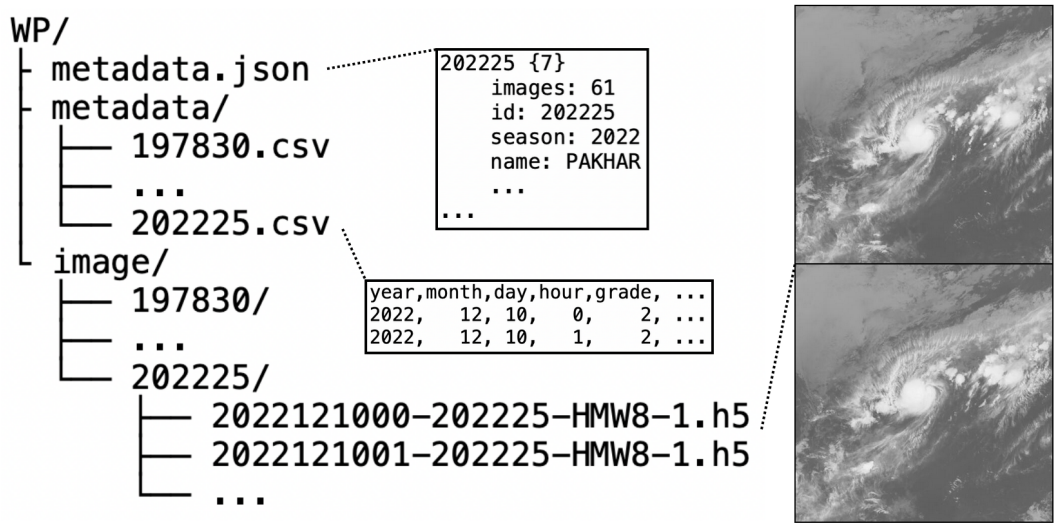


Figure 5: Directory structure of the Digital Typhoon dataset

typhoon forecasting in real time is strictly prohibited by law except for JMA. There may be similar regulations in other countries for public safety. Because typhoon forecasting is critical information for life-and-death decision-making, responsible use is required for real-time forecasting by machine learning. However, this regulation does not apply to machine learning benchmarks for historical datasets.

Another issue for responsible use is aligning with other scientific communities, such as atmosphere and climate sciences when publishing controversial results that might influence the future of human society. As discussed in Section , long-term data has many types of data quality issues such as biases that require careful modeling to produce reliable results. Machine learning research without careful consideration of those issues may lead to misleading conclusions that raise controversies without solid scientific evidence. This problem can be alleviated by communicating with domain experts to cross-check the results from a broader perspective.

F Dataset Availability

The Digital Typhoon dataset will be maintained by the Digital Typhoon project at the National Institute of Informatics. This project started in 1999 and has been continuously maintaining the data processing pipeline for over 20 years. This history proves that we already established a robust system for the maintenance of the dataset.

The official page of the dataset is `http://agora.ex.nii.ac.jp/digital-typhoon/dataset/` as a part of the Digital Typhoon website, which offers a variety of services to explore a wide range of typhoon-related data. Given that satellite images are publicly accessible observational data, we have designated the license for the Digital Typhoon dataset as the Creative Commons Attribution License (CC BY) that requires attribution to our project. An example of the attribution is as follows.

> Digital Typhoon dataset (National Institute of Informatics) http://agora.ex.nii.ac.jp/digital-typhoon/dataset/

Additionally, we credit the original data sources on the official data distribution website. We first acknowledge the JMA, which is responsible for the observations. We also credit organizations that manage data delivery, including the Institute of Industrial Science (IIS), the University of Tokyo.

The GitHub page provides code to work with the dataset to produce machine-learning models used for benchmarking. https://github.com/kitamoto-lab/digital-typhoon/ The code is provided with the MIT license.

In the future, we plan to assign a dataset DOI at DIAS (Data Integration and Analysis System) https://diasjp.net/, which is a Japanese data repository for earth science and environmental datasets.