scientific reports

OPEN

Multi-horizon prediction of tropical cyclone intensity and its interpretability with temporal fusion transformer

Iyan E. Mulia[1,2,3,4]✉, Udai Shimada[5], Naonori Ueda[3,4], Takemasa Miyoshi[3,6] & Madam Taqiyya Maulana[1]

Predicting tropical cyclone (TC) intensity is challenging, involving numerous variables and uncertainty, especially for TC with rapid intensification (RI). One of the frequently used operational methods for such a case relies on statistical–dynamical models subjected to several limitations stemming from linear regression approximation to a complex TC system. This study proposes a new approach using a Temporal Fusion Transformer (TFT) to overcome the limitations attributed to the conventional models. Besides improving accuracy, TFT is also known for its result interpretability, as opposed to the general perception that deep learning is purely a black-box model. Furthermore, the probabilistic nature of TFT facilitates prediction uncertainty quantification, an important feature and advancement to the standard deterministic prediction. We train and evaluate our model using the Western North Pacific basin TC observation and reanalysis datasets from 1996 to 2021. The results show that the prediction of TC intensity by TFT reduces the error of the conventional model by approximately 12% on average for all forecast horizons of up to 72 h, along with estimated uncertainty bands. A higher rate of 14% error reduction is attained specifically for TCs undergoing RI, an intractable phenomenon for the traditional modeling procedure.

A tropical cyclone (TC) is an extreme weather regime that can cause severe compounding and cascading impacts from destructive winds[1], heavy rains[2], lightning[3,4], floods[5,6], and storm surges[7,8]. Forecasting such a catastrophic event, particularly its intensity, remains a challenging topic. Dynamical models[9–14] are the pioneering tool to study and forecast the TC, incorporating a physical understanding of the real-world conditions into the modeling realm. Despite the positive trend of the TC forecast skill improvement by various dynamical models over the past years, they are still generally insufficient to predict the TC intensity. This issue is exacerbated for rapid intensification (RI) cases[15], in which an increase in the maximum sustained winds reaches at least 30 kt within 24 h or above 95th percentiles of over-water 24 h intensity changes[16].

Statistical–dynamical models are another alternative method enhancing the accuracy of TC intensity prediction by the dynamical models alone, which have been implemented in many operational centers[17]. One widely used statistical–dynamical model is the Statistical Hurricane Intensity Prediction Scheme (SHIPS)[17–19]. SHIPS aims to relate various meteorological variables obtained from dynamical models and observations to TC intensity changes through multiple linear regression. As with any linear regression model, the main advantage of such an approach lies in its simplicity and result interpretability. For instance, the contribution of each predictor or independent variable can be straightforwardly determined from its weight. However, the linear approximation also has drawbacks limiting its prediction skill for highly nonlinear TC systems.

Consequently, various deep learning methods for estimating TC intensity have been proposed to improve the traditional methods. Higa et al.[20] used deep learning with the basic structure of convolutional neural networks (CNN) to automate the Dvorak technique[21] to classify TC intensity from satellite imagery. Later, improved CNN versions were introduced[22,23] to estimate the real intensity value instead of categorical targets. Deep learning models with multiple environmental variables and relatively longer lead times are also available. For example,

[1]Hydrography Research Group, Faculty of Earth Sciences and Technology, Bandung Institute of Technology (ITB), Bandung, Indonesia. [2]Research Center for Disaster Mitigation, Bandung Institute of Technology (ITB), Bandung, Indonesia. [3]Prediction Science Research Team, RIKEN Center for Interdisciplinary Theoretical and Mathematical Sciences, Kobe, Japan. [4]Disaster Resilience Science Team, RIKEN Center for Advanced Intelligence Project, Tokyo, Japan. [5]Meteorological Research Institute, Japan Meteorological Agency, Tsukuba, Japan. [6]Data Assimilation Research Team, RIKEN Center for Computational Science, Kobe, Japan. ✉email: iyan.mulia@itb.ac.id

Jiang et al.[24] incorporated a spatial attention module and convolutional long short-term memory (LSTM) in their model configured to forecast TC intensity up to 48 h ahead. Another study by Wenwei et al.[25] utilized a simpler deep learning model based on an MLP for better interpretability with a 24-h forecast lead time. The 24-h lead time was also considered by Boussioux et al.[26], using deep learning in a multimodal framework. A comprehensive review of deep learning applications in TC forecasting can be found in Wang et al.[27].

All the above studies have demonstrated the advantage of deep learning over conventional models in predicting TC intensity. However, the overlooked aspect is mainly related to the opacity of the reasoning behind the acquired solution. One primary factor is regarding the contribution of predictors. Typically, an additional processing or sensitivity analysis is required to interpret deep learning models, as demonstrated in the previous studies[22,23,25]. Therefore, we propose a deep learning method based on the Temporal Fusion Transformer (TFT)[28]. The TFT structure allows for an on-the-fly assessment of not only the contribution of predictors but also the importance of temporal input data points, which are integrated during the training to produce the optimal outcome. Such an interpretation leads to more understandable and transparent results. Moreover, TFT provides a range of solutions owing to its probabilistic prediction feature, which translates into uncertainty levels, a crucial component for the decision-making process in disaster mitigation. Lastly, similar to previous deep learning-based studies[29,30], TFT can significantly improve the prediction accuracy for TCs with RI relative to the reference conventional method.

As TFT is evidently a more advanced method than SHIPS, we compare the result with other equivalent deep learning methods with a similar input and output data structure to TFT using Time-series Mixer (TSMixer)[31] and encoder-decoder bidirectional LSTM (biLSTM)[32,33] commonly utilized for time series forecasting. Additionally, we apply a simple machine learning model using a multilayer perceptron (MLP) to merely introduce nonlinearity without changing the data structure of the SHIPS model. In this study, we refer to a version of the SHIPS model[34–36] developed by the Japan Meteorological Agency as the benchmark to assess our proposed method.

Materials and methods

Data

The first data source is based on the best track data from the Regional Specialized Meteorological Center Tokyo. In this study, the maximum 10-min sustained wind speed, VMAX, values that are not available for tropical depressions in the dataset are determined to be negatively proportional to a change in the central pressure, similar to the method using the wind-pressure relationship[37]. The second dataset comprising the atmospheric environmental variables is derived from the Japanese 55-year reanalysis (JRA-55)[38]. We also incorporate oceanic reanalysis data from Japan's Meteorological Research Institute multivariate ocean variational estimation (MOVE) system[39]. The last dataset is extracted partly from geostationary meteorological satellite products as described in Bessho et al.[40]. More details on the preprocessing and characteristics of each dataset can be found in Yamaguchi et al.[34] and Shimada et al.[35]. All the variables are concisely defined and tabulated in Table 1.

We use the same data sources as in the previous studies[34,35], but with a longer period from 1996 to 2021. Such a period is selected based on data quality consistency across datasets. While the best track data has been updated regularly, some datasets possess a different level of quality over time. For example, the Himawari-5 geostationary meteorological satellite observation has substantially improved after June 1995[41]. Furthermore, according to the JMA report[42], the MOVE reanalysis dataset has been modified and improved since March 2022. Data homogenization or correction should be conducted beforehand to incorporate data of various qualities, but this is beyond the scope of the study. The current dataset length should suffice to capture meaningful patterns associated with TC characteristics. Thus, it is appropriate to assess the proposed method's performance.

All the datasets have or are resampled into a 6-h interval. Unlike previous studies[34,35], where more data exist for shorter forecast times, we fix the number of data points on each TC event based on lag and maximum forecast times of TFT. Therefore, the number of samples or data points is the same for all forecast times, and some relatively short events are removed. We first separate all TCs in 2021 for a one-year independent testing. Thus, only TCs from 1996–2020 are used in the training, validation, and test sets. We identify TCs with RI when intensity increases at least 30 kt within 24 h, then separate them from non-RI TCs, which results in two datasets. Subsequently, we partition the two datasets for training, validation, and test sets with ratios of 0.8, 0.1, and 0.1, respectively. The final training, validation, and test sets are obtained by combining the two datasets after partitioning and shuffling the sequence. Even though such a scheme does not fully address the imbalance of RI TCs due to the low frequency of occurrence[45], it allows for a more balanced distribution of RI TCs across the sets.

The number of non-RI TCs in the training, validation, and test sets are 249, 31, and 30, respectively. As expected, the number of RI TCs is smaller than non-RI TCs, with only 111 events in the training, 12 in the validation, and 14 in the test set. Figures 1a–c show the distribution of TC tracks after data partitioning. With the 6-h interval, the data points in the training set are 7096 and 3786 for non-RI and RI TCs, respectively. The number of data points for the validation set is 977 (non-RI) and 403 (RI), and for the test set, it is 869 (non-RI) and 468 (RI). The distribution of data points represented by VMAX for the respective dataset is shown in the insets of Figs. 1a–c. The list of TCs (1996–2020), including the 2021 dataset, is tabulated in the supplementary Tables S1 and S2. Distribution of TC tracks for the 2021 dataset is shown in the supplementary figure (Fig. S1).

SHIPS

The initial SHIPS model was proposed by previous studies[17,46]. More recently, an improved SHIPS version with a random forest algorithm has been introduced[47]. SHIPS is configured to forecast the change of TC intensity in terms of VMAX (ΔVMAX) from the start to the specified forecast times. Here, we consider forecast times of six h to 72 h with an interval of six h. We note that our model does not include the prediction of MSLP as in the previous SHIPS version by the JMA[34,35]. Similar to Yamaguchi et al.[34], we use 26 predictors in our SHIPS

No	SHIPS predictor	Description	TFT predictor
1	MSLP	Mean sea level pressure (MSLP$_{t=0}$)	pastcov
2	PERV	12-h change in maximum 10-min sustained wind speeds (VMAX) to the initial time of prediction (VMAX$_{t=0}$ to VMAX$_{t=-12}$)	VMAX (past target)
3	VMA2	Square of VMAX	–
4	VMPE	VMAX times PERV	–
5	POT	Maximum potential intensity derived from sea or land surface temperature (DMPI)[43] minus VMAX averaged over forecast times	DMPI (futcov)
6	POT2	Square of POT	–
7	TCHP	Tropical cyclone heat potential	futcov
8	TCHP2	Square of TCHP	–
9	T200	200 hPa temperature (r = 200 – 800 km)	futcov
10	T250	250 hPa temperature (r = 200 – 800 km)	futcov
11	STMX	Zonal storm motion	futcov
12	RHMD	700 – 500 hPa relative humidity (r = 200 – 800 km)	futcov
13	EPOS	Difference of equivalent potential temperature between lifted surface parcel and environment	futcov
14	SHRD	850–200 hPa vertical wind shear magnitude (r = 0 – 500 km)	futcov
15	SHGC	Generalized vertical shear parameter[44]	futcov
16	SHSH	Square of SHRD	–
17	SHLT	SHRD times sines of latitude (LAT)	LAT (futcov)
18	SHVM	SHRD divided by VMAX	–
19	VMSH	VMAX times SHRD	–
20	Z850	850 hPa vorticity	futcov
21	D200	200-hPa divergence (r = 50–1000 km)	futcov
22	DTWAC	Change in 850 hPa tangential wind (r = 0–500 km) (TWAC) over forecast period ((TWAC$_{t=ft}$ to TWAC$_{t=0}$))	TWAC (futcov)
23	TADV	Temperature advection between 850 and 700 hPa (r = 0–500 km)	futcov
24	TGRD	Magnitude of temperature gradient between 850 and 700 hPa (r = 50–500 km)	futcov
25	PC30	Percent area of infrared (IR) brightness temperature (*Tb*) < – 30 °C (r = 50–200 km) (PC30$_{t=0}$)	pastcov
26	SDIR	Standard deviation of IR *Tb* (r = 0 – 200 km) (SDIR$_{t=0}$)	pastcov

Table 1. List of predictors Predictors without time information in the description column are averaged over forecast times. The time information is relevant only for the SHIPS and MLP models. The TFT model takes all data points of a single predictor within the specified lag and forecast times and categorizes it into the target or past target, past covariates (pastcov), and future covariates (futcov).

model, as indicated in Table 1. These independent variables were selected based on physical considerations and statistical significance analysis[34]. We normalize both the predictand and predictors by subtracting the mean and dividing by the standard deviation. Then, multiple linear regression is applied to predict ΔVMAX from the considered predictors. Twelve SHIPS models are built independently per the specified forecast times. Unlike the other models, SHIPS does not require the validation set.

MLP

Before we apply our proposed method using TFT, we also test a simpler machine learning model based on an MLP. The predictand and predictors of this MLP model are the same as those of SHIPS. This experiment aims to assess how MLP can improve the linear regression used in SHIPS by simply introducing the nonlinearity of the MLP network without changing the data input structure as in TFT. We use the same MLP model as our previous study[48] but with only one hidden layer consisting of 64 nodes. We implement a batch size of 64 and a rectified linear unit (ReLU) activation function in the hidden layers with a 20% dropout rate for better generalization. The validation set is used for an overfitting test, in which the training is stopped when the loss function on the validation set does not improve after 20 epochs. Like SHIPS, our MLP model produces a single solution signifying a deterministic prediction.

TSMixer.

TSMixer is a deep learning method comprises stacked MLPs designed to mix time and feature dimensions for efficient information extraction from heterogeneous inputs. Despite its architectural simplicity, TSMixer is a potential alternative to more complex recurrent- or attention-based sequential deep learning models. The method has been comprehensively tested and achieved state-of-the-art performance on many benchmark datasets[31]. We employ a Python machine learning library for time series called Darts[49] to build our TSMixer model and use the same overfitting test procedure as in MLP, with a batch size of 64. The remaining hyperparameters are optimized using the Optuna software framework[50], with the specified range of values tabulated in supplementary Table S3. The hyperparameters achieved after the optimization are the number of hidden sizes of the second feed-forward layer of 20, the size of the first feed-forward layer of 55, the number of mixer blocks of 2, a dropout of 0.08, and a learning rate of 3e-4.

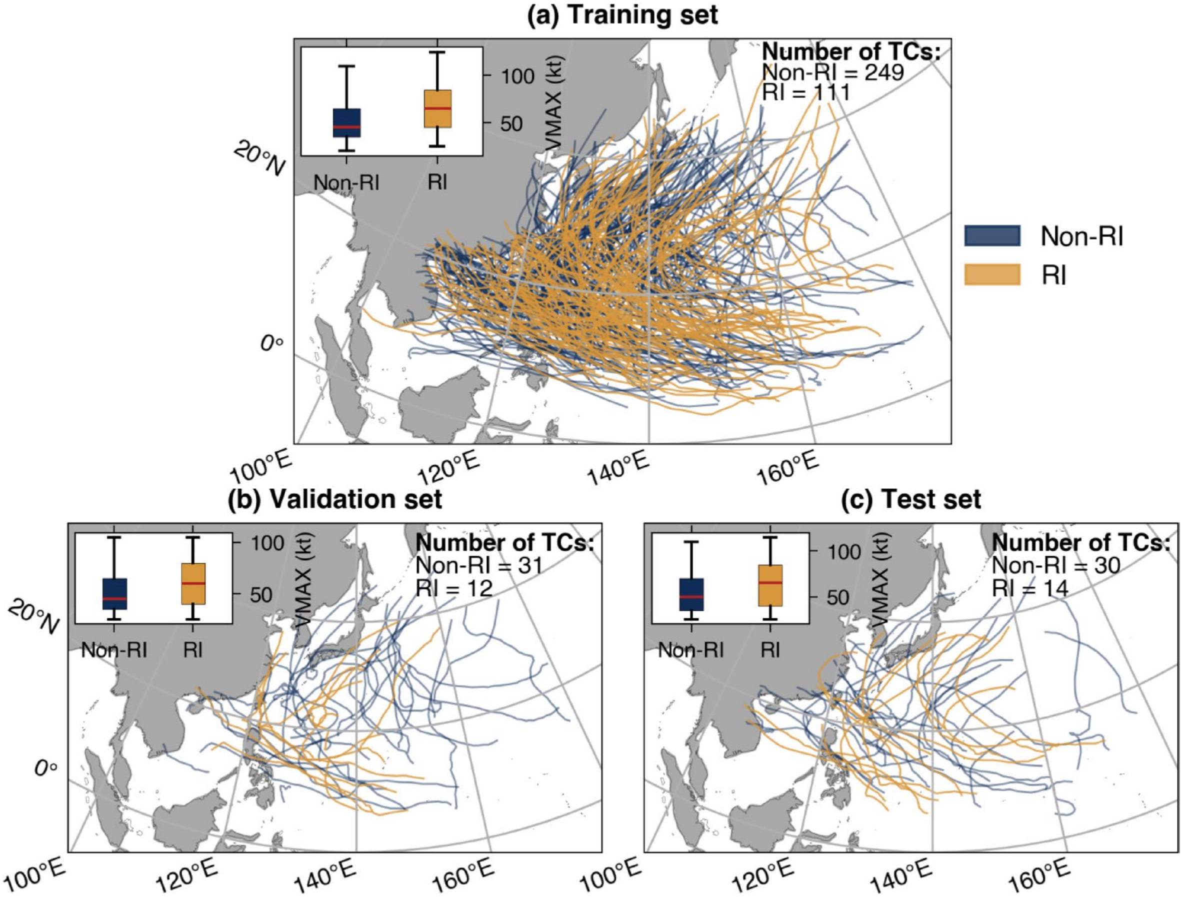

Fig. 1. Distribution of best track for non-RI and RI TCs on the training (**a**), validation (**b**), and test (**c**) sets. VMAX distribution based on the 6-h data interval is indicated in the box plot in the inset of each figure. The map was created using the Cartopy Toolkit (https://scitools.org.uk/cartopy) in Python.

biLSTM

Another deep learning model used as a comparison is based on an encoder-decoder biLSTM[32,33], a type of recurrent neural network that is more similar to TFT. Unlike the traditional LSTM, which processes the data in a single direction, biLSTM processes sequential data in both forward and backward directions to access both past and future context at each time step. Furthermore, the encoder-decoder architecture facilitates handling sequences with varying input and output lengths. We built the biLSTM model in TensorFlow[49] combined with Optuna[50] for hyperparameters optimization. The optimized main hyperparameters are the size of the latent dimension of 10, the size of a feed-forward layer of 30, a dropout of 0.08, and a learning rate of 6e-4 with an adaptive moment estimation (Adam) optimizer. The range of hyperparameters values are shown in Table S4. Initially, we set a maximum epoch of 200, but the algorithm converges earlier due to the overfitting test.

TFT

TFT was introduced by Lim et al.[28] and has been utilized in many applications of time series forecasting[51–53]. We configure our TFT model to directly predict VMAX instead of ΔVMAX as in SHIPS and MLP. Also, TFT combines multi-horizon forecasting in a single model and incorporates all data points within the specified lag and forecast times. Inputs for TFT (see Table 1) are categorized into past target, past covariates, historic future covariates, future covariates, and static covariates. Static covariates indicate characteristics of time series that do not change over time, which is not considered in our model. Past covariates refer to predictors that are unknown in the future, such as those obtained from observation. On the contrary, future covariates are predictors with known future states, which, in our case, are derived from dynamical models. TFT also includes VMAX data prior to the prediction start time as input defined as past target, which is considered as input together with past covariates and historic future covariates. An illustration of these input categories and simplified TFT architecture is depicted in Fig. 2. A combination of multiple predictors or a modification of a single predictor as several inputs in SHIPS is not considered in TFT. Therefore, TFT has only 19 predictors (Table 1).

The main advantage of TFT can be attributed to the variable selection networks and the interpretable multi-head attention components, which are responsible for the solution's interpretability. As shown in Fig. 2, all inputs pass through variable selection networks, suppressing unnecessary features with a minimum contribution to the prediction accuracy indicated in the weights of input features. The multi-head attention captures temporal

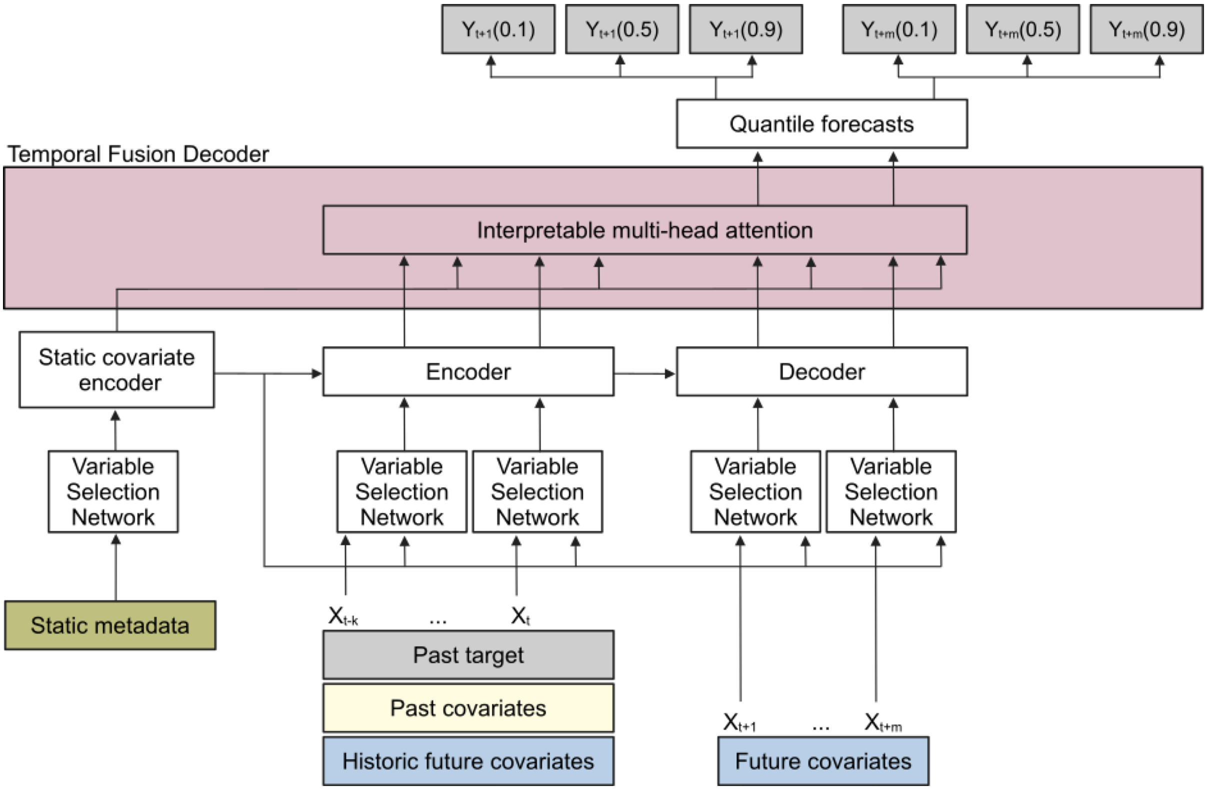

Fig. 2. Simplified TFT architecture. Static covariates are shown as the original TFT architecture but are not considered in this study.

dependencies, which can be used to trace back the most relevant past and future data to the predictions from the attention weights. Furthermore, as a probabilistic model, TFT prediction outputs multiple quantiles, providing a range of possible solutions. More details on the theoretical background and mechanism of TFT can be found in the original article[28].

One of the predictors in SHIPS and MLP, PERV, uses past data of 12 h (Table 1), which is implicitly included in the TFT. Our TFT includes 18 h of past data (four data points including 0 h) of all input variables to predict VMAX at six h up to 72 h of forecast times or horizons (12 data points). We employ the same package, Darts[54], used in the TSMixer model, to build our TFT model. We also use the same overfitting test procedure as in the previous models, with a batch size of 64 and a maximum epoch of 200. The hyperparameters of TFT are also optimized using the Optuna software framework[50], with the specified range of values tabulated in supplementary Table S5. The optimized hyperparameters are the number of hidden states size of 28, the number of recurrent layers of 2, the number of attention heads of 5, the number of hidden sizes for processing continuous variables of 10, a dropout of 0.06, and a learning rate of 7.9e-4 with the Adam optimizer.

Results and discussion

Overview of prediction results

Figure 3 compares prediction results and the improvement rate by TSMixer, biLSTM, and TFT relative to SHIPS on the test set. As TFT produces a simultaneous prediction of various percentiles, we use the median or 50th percentile prediction to evaluate the models. We opt for 100 ensemble samples for predictions of LSTM and TFT, as the variability of the predictions does not significantly change after 100 samples. For all the TCs, all models exhibit comparable performance for up to the 12-h forecast time, except for the TSMixer (Fig. 3a). TSMixer prediction is better than SHIPS only from 42-h forecast time ahead. The improvement by biLSTM and TFT relative to SHIPS is evident from 12-h forecast time onward, with an average improvement over all forecast times of approximately 10% and 12%, respectively. Introducing nonlinearity through the MLP without changing the input data structure of SHIPS has minor effects on the prediction accuracy (Fig. S2).

TFT outperforms the other models for non-RI TCs (Fig. 3b) and maintains its predictive skill for RI TCs (Fig. 3c). However, the best model for RI-TCs prediction is biLSTM, at least up to 48-h forecast time, whereas beyond that, TFT again prevails. On average, the rate of improvement by TFT for RI TCs is 14%, second to biLSTM of 15%, while TSMixer is 11%. The poor performance of TSMixer for non-RI TCs prediction relative to SHIPS (Fig. 3b), particularly at less than 54-h forecast time, suggests that the model tends to overestimate the TC intensity. Such a tendency is not apparent in the TFT and biLSTM prediction results.

Generally, the mean absolute errors (MAEs) increase with forecast time due to poor track forecasts, accumulation of errors, and a wide range of factors contributing to the intensity changes. However, the prediction

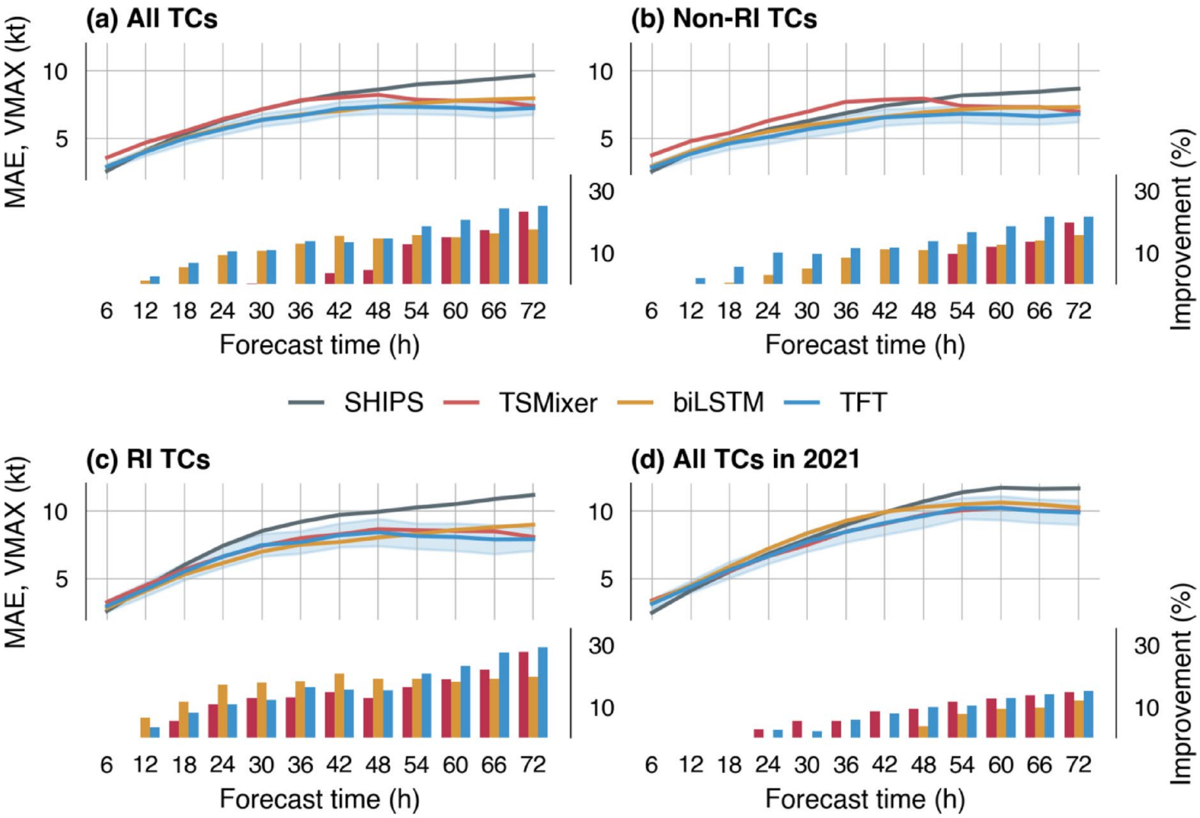

Fig. 3. Comparisons of MAEs for all models, forecast times, and improvement rates by TSMixer, biLSTM, and TFT relative to the SHIPS prediction results. The shaded area with light color in the TFT results indicate the 95% confidence interval from 100 samples, signifying the probabilistic prediction. (**a**) All TCs. (**b**) Non-RI TCs. (**c**) RI TCs. (**d**) All TCs in 2021.

error is higher when the intensity peaks, typically within the second day (see Figs. 4 and 5 as examples). The TC weakening stage is relatively more straightforward to predict, as exhibited in Fig. 6, especially by TFT, owing to its advanced architecture and input structure. Therefore, smaller errors may appear for longer forecast times in TC intensity forecasting. This occurrence was also apparent in the previous studies[34,35,47].

A one-year independent test of 2021 results is shown in Fig. 3d. SHIPS prediction exhibits larger MAEs of more than 10 kt at 48-h forecast time onward, which notably deviates from other models. This pattern is similar to the prediction for RI TCs on the test set (Fig. 3c), indicating the influence of RI TCs on the 2021 dataset. TFT and TSMixer demonstrate comparable performance, with prediction improvement to SHIPS starting from 24-h forecast time. On the other hand, the improvement by biLSTM is only visible after 48 h, similar to MLP (Fig. S2d).

The deep learning models generally surpass the performance of SHIPS and MLP, which can be attributed to the structure of the input data series, despite requiring fewer predictors. The data points averaged over specified forecast times in SHIPS and MLP hinder elucidating the intricate time-dependent relationships among the variables. The deep learning models' performance is comparable, but TFT maintains consistency across datasets and is better for longer forecast times. One of the crucial features of TFT is attention-based temporal processing, a unique and advanced feature of TFT that is not available in the other compared models. This feature is responsible for capturing time-varying dynamics[28], leading to better predictions of the TC intensity. Furthermore, as discussed in detail in the subsequent section, TFT facilitates a more straightforward interpretation of results.

To analyze the performance of the models in the intensifying and weakening stages of TCs, we plot the two-dimensional MAEs distribution of ΔVMAX on the test set resulting from SHIPS and TFT models only (Fig. 6). Figure 6a shows the increase of MAEs by SHIPS for the weakening TCs, with even worse results for the intensifying parts, especially those of longer forecast times. The TFT results exhibit the same patterns beyond the 5th and 95th percentiles, albeit with lower MAEs (Fig. 6b). The difference in MAEs between the two models is depicted in Fig. 6c, which can reach up to 16 kt. This result demonstrates the difficulty in predicting the RI TCs, which, to some extent, can be circumvented by TFT.

Despite the advantages of TFT, we acknowledge that the performance analyses above are based on reanalysis products, which may differ should we use the dynamical models in the operational mode[25]. In such a case, the averaging technique in SHIPS and MLP may help to suppress the uncertainty of operational dynamical models. Nevertheless, the current datasets are sufficient to merely demonstrate TFT's capability against standard

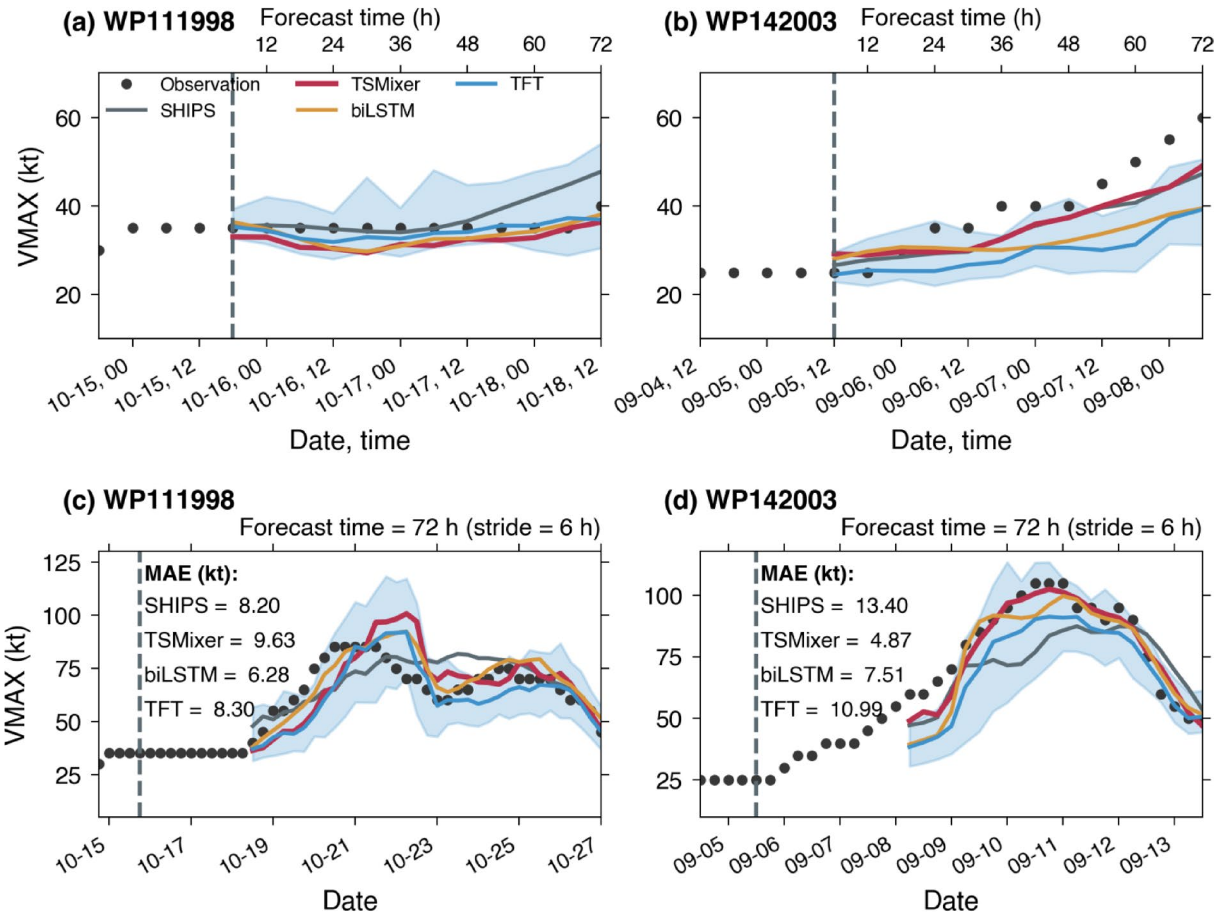

Fig. 4. The best (**a**, **c**) and worst (**b**, **d**) TFT predictions for non-RI TCs on the test set. The blue line depicts the median or 50th percentile of TFT predictions, and the shaded area with lighter color represents the predictions within the 10th and 90th percentiles. The vertical dashed lines indicate the first forecast time.

modeling practices. Therefore, this study primarily contributes to the research and development stage towards the actual operational TC forecasting system.

Another aspect that needs to be considered concerns the model's generalization for basins other than the Western North Pacific. This is challenging as the relative importance of individual predictors is reported to be varied for different basins[55]. In general, TC size information differs across basins[56]. Hence, the sensitivity to environmental conditions favourable for intensification is also different[57]. A small TC is less sensitive to RHMD but is sensitive to SHRD and SHGC. In contrast, a large TC has less sensitivity to SHRD and SHGC, but a higher RHMD is favorable for intensification. Therefore, the level of importance would vary from basin to basin. In addition to appropriately incorporating TC size information, transfer learning and meta-learning[60] can be implemented to handle diverse TC datasets across basins and improve generalization, which is one of the future directions of the study.

Computational cost

We train the model on a single graphics processing unit (GPU) of the NVIDIA GeForce RTX 4060. Based on the above GPU specification, we compare the computing time for training the machine learning models (Table 2). The SHIPS model is excluded from the comparison as it is calculated on the central processing unit instead of the GPU. The probabilistic modeling and the complexity of TFT architecture need to be compensated by its computational cost. Therefore, the TFT training time is 12 min, which is the longest among the other models followed by TSMixer of 2.6 min, and bLSTM of 1.5 min. MLP requires 1.5 min to train, but this is the accumulative time for all forecast horizons, as MLP consists of one model for each forecast horizon, similar to SHIPS. Nonetheless, the inference time of all models is comparable, within seconds.

Best and worst prediction results

For conciseness, we exclude the MLP results in the following analysis. TFT attains the best prediction result for non-RI TCs for TC WP111998 (Babs). The TC track and intensity for this event are shown in Fig. 7, with the observed maximum VMAX of 85 kt. We start the forecast at 1800 Coordinated Universal Time (UTC) on 15 October 1998 for a one-time multi-horizon prediction (Fig. 4a). The median prediction by TFT exhibits the best fit to observation compared to other models, particularly at longer forecast times. We also plot the

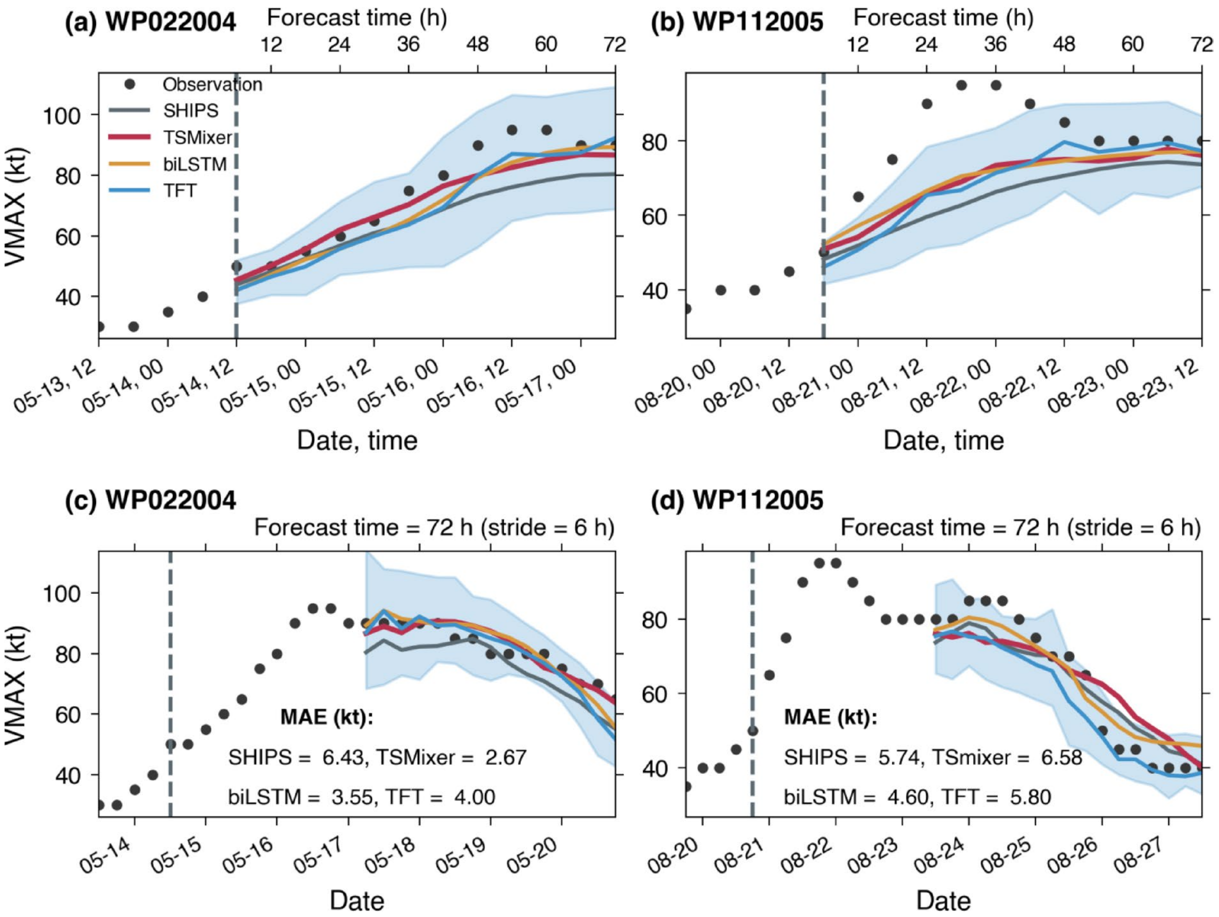

Fig. 5. The best (**a**, **c**) and worst (**b**, **d**) TFT predictions for RI TCs on the test set. The blue line depicts the median or 50th percentile of TFT predictions, and the shaded area with lighter color represents the predictions within the 10th and 90th percentiles. The vertical dashed lines indicate the first forecast time.

prediction results for a single 72-h forecast time with a rolling window or stride of six hours (Fig. 4c). All models, including the median prediction of TFT, fail to predict the maximum VMAX of 85 kt at the early stage of the series between 20–21 October. However, most observed values including during the intensification stage are still within the uncertainty band of TFT. The uncertainty information provides possible maximum intensification rate and possible maximum intensity, which could be useful to construct the second forecast scenario. Using the same statistical evaluation in Fig. 3, the MAE for this 72-h forecast time is 8.20 kt, 9.63 kt, 6.28 kt and 8.30 kt for SHIPS, TSMixer, biLSTM, and TFT, respectively.

To provide a proportionate evaluation, we also show the worst prediction result by TFT for non-RI TCs when applied to TC WP142003, named Maemi (see Fig. 7 for the TC track and intensity). For a multi-horizon prediction started at 1200 UTC on 5 September 2003, the median prediction of TFT yields a more significant discrepancy against observation than other models (Fig. 4b). The upper bound or 90th percentile prediction of TFT better approximates the observed VMAX. However, such a result only represents a single event, which does not reflect the TFT performance entirely, as shown in Fig. 3. Using the 72-h forecast time (6-h stride) as an example of a rolling forecast, the performance of TFT is still better than SHIPS, but worse than TSMixer and biLSTM (Fig. 4d). TSMixer accurately predicts the peak of TC intensity from 10 to 11 September compared to other models, which is confirmed by the statistical evaluation results. The MAE of SHIPS is 13.40 kt, TSMixer is 4.87 kt, biLSTM is 7.51 kt, and TFT is 10.99 kt.

The best result of TFT for RI TCs is demonstrated for the prediction of TC WP022004 (Nida). The genesis and lifespan of this TC are mainly in the sea area, with a VMAX peaked at 95 kt (see Fig. 8). We perform a multi-horizon forecast starting at 1200 UTC on 14 May 2004 (Fig. 5a). The median prediction of TFT agrees with the observation, but with underestimation of the peak of observed VMAX, comparable to TSMixer and biLSTM. The relatively wide uncertainty band of TFT suggests difficulty in RI prediction. The underestimation is worse by SHIPS, particularly from the 36-h forecast time and beyond. Such underestimation by the standard forecasting methods is commonly associated with predictions for RI TCs due to their intricacy[59]. The forecasting skill of TFT remains consistent for the rolling window predictions with the 72-h forecast time (Fig. 5c), as well as TSMixer and biLSTM. Here, SHIPS continue the underestimation trend until the later stage, when the TC weakens.The MAEs for SHIPS, TSMixer, biLSTM, and TFT are 6.43 kt, 2.67 kt, 3.55 kt, and 4.00 kt, respectively.

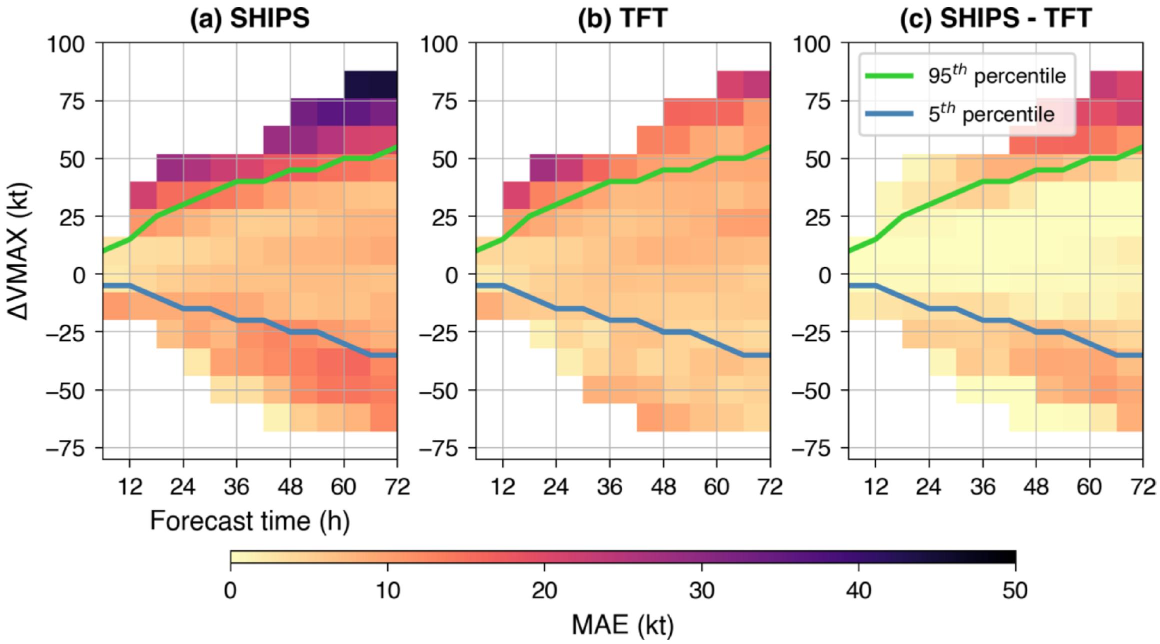

Fig. 6. Two-dimensional distribution of MAEs of ΔVMAX resulting from SHIPS (**a**) and TFT (**b**) on the test set, including the relative MAEs (**c**). The y-axis shows best-track intensity changes from forecast time 0 to each forecast time (x-axis). The solid lines indicate the 5th and 95th percentiles of the test set samples at each forecast time, representing the rapid weakening and intensification, respectively.

No	Model	Computing time (min)
1	MLP	1.5
2	TSMixer	2.6
3	biLSTM	1.5
4	TFT	12.0

Table 2. Training time of all machine learning models

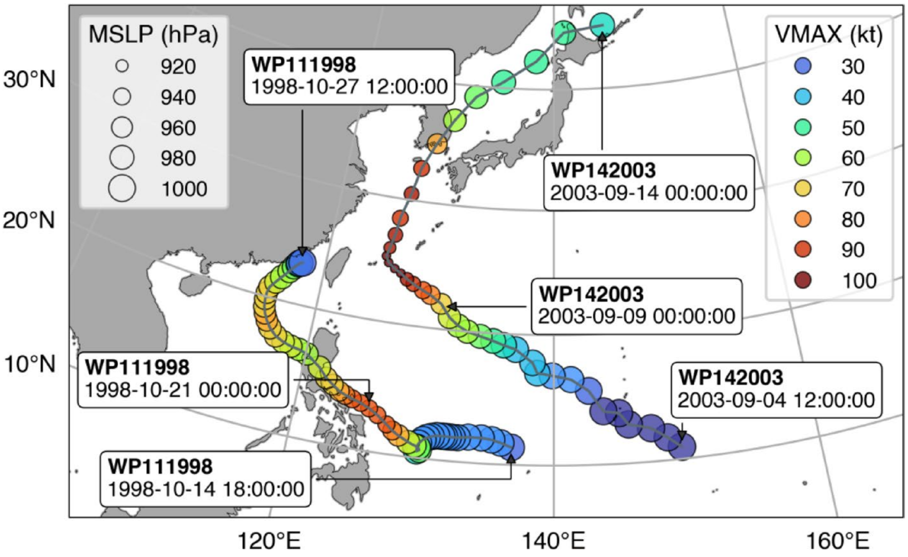

Fig. 7. TC WP111998 (Babs) and WP142003 (Maemi) tracks and intensities. The map was created using the Cartopy Toolkit (https://scitools.org.uk/cartopy) in Python.

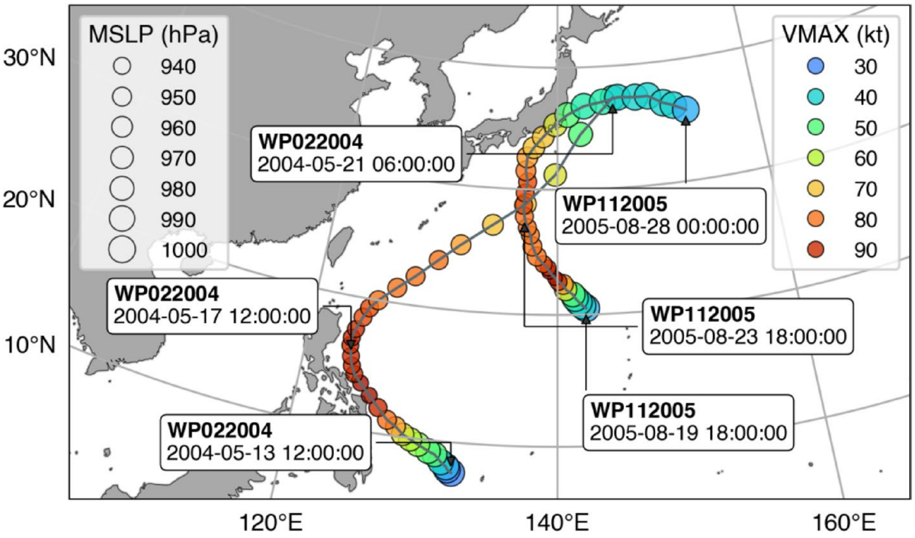

Fig. 8. TC WP022004 (Nida) and WP112005 (Mawar) tracks and intensities. The map was created using the Cartopy Toolkit (https://scitools.org.uk/cartopy) in Python.

Despite the overall satisfying performance depicted in Fig. 3, TFT occasionally fails to accurately predict TC with RI, such as WP112005 named Mawar. The maximum VMAX of this TC was 95 kt, which is similar to Nida, but it underwent RI within a much shorter period. Figure 5b compares the performance of all models for a one-time multi-horizon prediction started at 1800 UTC on 20 August 2005. The sharp increase of VMAX from 0000 UTC 21 August to 0000 UTC 22 August cannot be reproduced by all models, even with the 90th percentile prediction of TFT. This failure suggests that the conditions that WP112005 had were statistically rare. For instance, WP112005 was characterized by its small size (small radius of 30-kt wind speed). According to a study by Shimada[57], RI for small TCs tends to occur under significantly different environmental conditions (e.g., very dry environment, low vertical shear, and low SST) from other typical RI cases. Therefore, we speculate that it is difficult to predict RI for such a specific case without size information, as in our current models.

The predictive skill for the rolling 6-h window prediction using the 72-h forecast time is shown in Fig. 5d. TFT outperforms other models, particularly at the weakening stage of the TC from 26 August onward. At this later stage, the effect of landfall in reducing TC intensity incorporated in the DMPI input is reasonably captured by TFT. However, overall, biLSTM produces the best prediction result. The MAEs level in ascending order is biLSTM of 4.60 kt, SHIPS of 5.74 kt, TFT of 5.80 kt, and TSMixer of 6.58 kt.

Results interpretation

Encoder feature importance contains past target, past covariates, and historic future covariates, while decoder feature importance consists of only future covariates (Fig. 9a). From the encoder feature importance, the past target or VMAX before the prediction start time stands out as the most influential feature, with a contribution level of approximately 40%. This result suggests that the characteristic or trend of TC intensity in the past few hours indicates its likely behaviour shortly after. For instance, a recent intensification may lead to a favourable condition for continued strengthening and vice versa. Our result aligns with the traditional multivariate linear regression methods, in which VMAX-related predictors such as VMA2 and PERV are essential in SHIPS prediction[34,35]. Similarly, machine learning-based models showed a prominent contribution of past intensity values in the predictability of its future state[25,60].

Other notable variables in the encoder are SDIR (11%) and PC30 (5.6%), categorized as past covariates derived from satellite observations. Both variables are based on infrared satellite imagery, which can depict the structure of inner core convection closely linked to the TC intensity[61,62]. Furthermore, the brightness temperature represented in SDIR and PC30 indicates the convection strength determining the TC intensity change[19,35]. Similar to previous studies[57,63], SDIR is weighted more than PC30, indicating the importance of the symmetry of convection over the quantity. Incorporating satellite-based observations in TC forecasting has been a standard practice[34,35,57], although with various degrees of importance, including those by deep learning models[20,25].

The remaining three variables in the encoder's top six importance are historic future covariates of TWAC, TCHP, and RHMD, with contributions of 6%, 5.3%, and 4.8%, respectively, while the subsequent variables have relatively small weights. TWAC constitutes cyclonic mean circulation within 500 km from the center at the lower troposphere, and its historical change can be proportional to the subsequent TC intensity change. TWAC can also exhibit TC size, impacting subsequent intensification. Therefore, Wenwei et al.[25] ranked the tangential wind-related predictors among the top seven importance. TCHP is a measure of ocean heat potential, the energy source for TCs, from the surface to a depth of 26 °C, which is more influential in determining TC intensity change than sea surface temperature[64]. Wahiduzzaman et al.[64] demonstrated the benefit of TCHP in TC intensity forecasting. Lastly, TCs with higher intensity, in general, tend to show larger relative humidity (RHMD) than

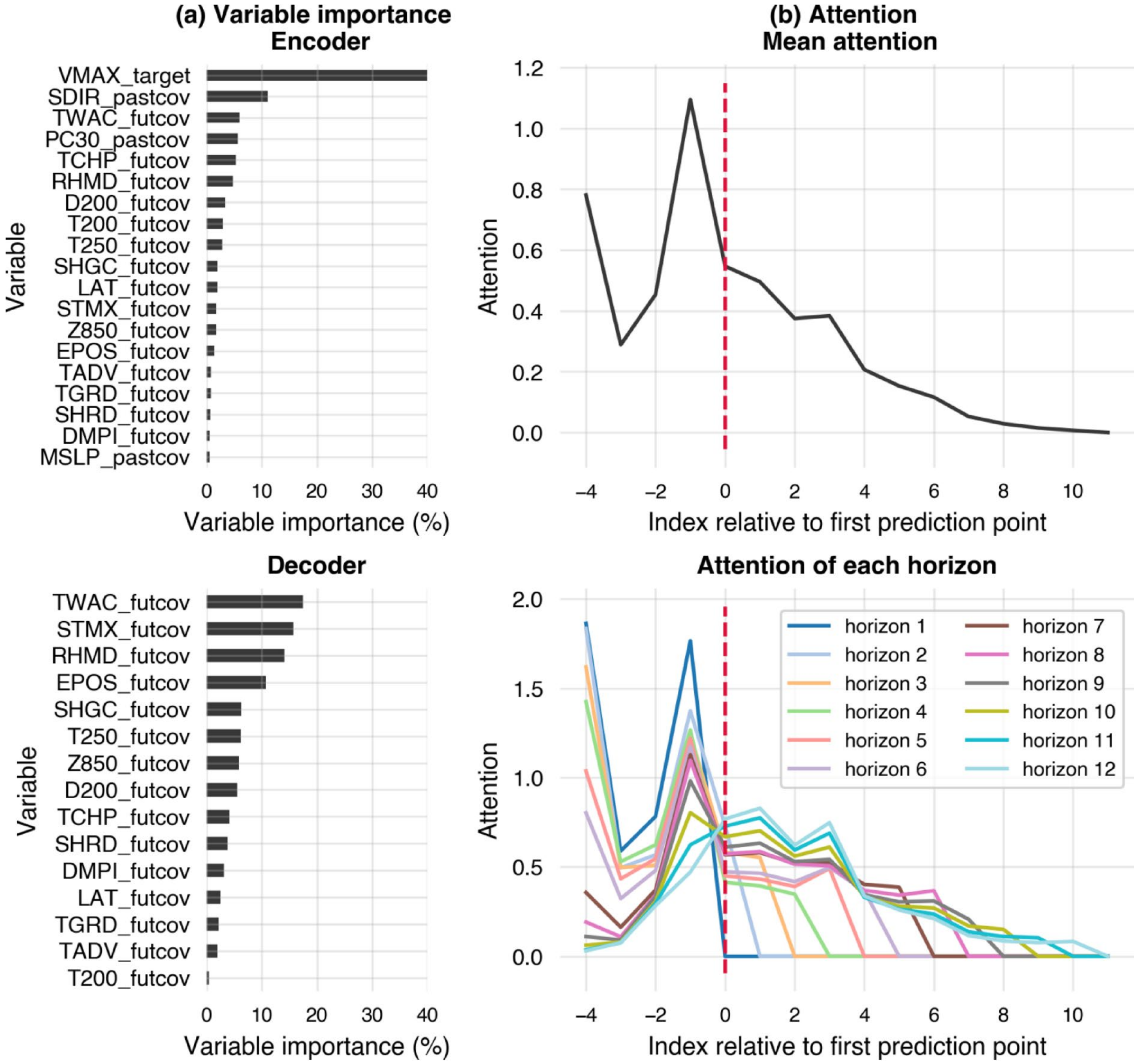

Fig. 9. Explainability of TFT prediction results. (**a**) Variable importance. (**b**) Temporal attention. The red dashed vertical line indicates the first prediction point separating past and future data points or time steps. For example, the index 0 corresponds to the 6-h forecast horizon.

those of lower intensity, even though the relationship is not linear[65]. Such a nonlinear relationship is captured by TFT, ranking RHMD in the top six important variables.

While the variable importance in the encoder is strongly dominated by past VMAX, the level of importance in the decoder is more uniform, with the four largest distinguished contributions of TWAC, STMX, RHMD, and EPOS. The historic future covariate of TWAC is listed among the top three predictors in the encoder, whereas its future covariate is the most essential variable in the decoder, with a contribution of 17.4%. Similarly, RHMD is ranked sixth in the encoder and third in the decoder, with a 14% level of importance. The increasing importance of future TWAC and RHMD makes sense because the sequential effect of environmental conditions on the TC intensity diminishes with forecast times, which is reflected in the attention plot (Fig. 9b). Similar results were obtained in previous studies, in which predictors representing environmental conditions prevailed for longer forecast time[62,66].

The effect of zonal storm motion (STMX) was minimal under the linear modeling framework of SHIPS[34,35], yet STMX is ranked in the top two in the decoder with a 15.7% level of importance. The discrepancy may be related to both models' input data structure and the nonlinearity of TFT, as described previously. Lastly, higher EPOS was reported to be favorable for TC intensification due to heating near the TC center[67]. In this study, EPOS is ranked fourth with 10.7% contribution. However, the small attention weights beyond the first prediction point might negate the level of importance in the decoder (Fig. 9b). TFT may also result in different sequences of variable importance at different runs due to the random initial condition of model parameters. Nonetheless, usually, the first top variables remain the same.

The mean attention plot (Fig. 9b top) depicts the maximum attention weight at the relative index -1 or the initial time at 0-h data point. This result highlights the importance of the 0-h data in the prediction, which is included in the persistence estimate, PERV, used in SHIPS[17,34,35]. The maximum attention at the initial time and the maximum importance of VMAX suggest that TFT predicts intensities mainly based on VMAX at the initial time. This outcome is natural because it is impossible to predict subsequent intensity without the initial VMAX information. The mean attention is also relatively high at the lag time of 18 h, which is longer than the range of PERV in SHIPS and MLP of 12 h. However, we limit the lag time for conciseness and opt for more data or samples. More specifically, from the attention plot of each horizon (Fig. 9b bottom), the contribution of future data increases with forecast horizons. At forecast horizons of 66 h (horizon 11) and 72 h (horizon 12), the attention weight of future data is higher than that of past data. These results are physically relevant and confirm the previous analysis, in which, at relatively longer forecast times, the concurrent influence of environmental conditions is expected to be greater than its sequential consequence.

Model result without future covariates

Based on the results interpretation, the past covariates significantly contribute to the prediction accuracy relative to the future covariates up to 60-h forecast time. However, to further clarify the role of future covariates, we perform an additional experiment by predicting the TC intensity without future covariates and compare the result with the original model (Fig. 10a). We also include a persistence model result as the baseline and additional information for our analysis, in which the VMAX at the initial time represents the predicted VMAX for the entire forecast times.

The deviation of the model without future covariates from the original model with future covariates started at a forecast time of 12 h with an exponential increase of MAEs reaching up to approximately 18 kt at 72 h forecast time. This result suggests that when future covariates are not included as inputs, as indicated by the zero attention (Fig. 10b), the prediction skill of TFT substantially decreases, particularly for longer forecast times of more than 12 h. Furthermore, the difference in MAEs between the model without future covariates and the persistence model of 1.2–6.6 kt is larger than that between the models with future covariates and without future covariates of 0.5–6.1 kt up to 54 h forecast time. This finding implies that within such a range of forecast horizons, the contribution of past information to future intensity change is more significant than that of future information. On the contrary, after the 54 h forecast time, future information contributes more to error reduction than past information, in which the MAEs difference between the model without future covariates and the persistence is 4.7–6.4 kt, and between the models with and without future covariates is 7.6–11.1 kt. This experiment confirms the previous discussion on the result interpretation with more apparent merit of future covariates.

Perturbed VMAX

As VMAX or past target is the essential variable in predicting TC intensity (Fig. 9a top), we perform a sensitivity analysis by introducing perturbation to VMAX and examining its effect on the model accuracy. Synthetic perturbation in the form of Gaussian noise is added to the original VMAX series on the test set. The perturbation level is characterized by the standard deviation of the Gaussian distribution ranging from 5 to 20% of the mean VMAX values. Such a perturbation may represent model and observation errors or real-world changes in the TC environment. We apply this experiment to both SHIPS and TFT models.

Figure 11 shows that when only 5% of the noise is added to the VMAX value, the accuracy of TFT somewhat decreases compared to the prediction without perturbation. However, for 10% or more perturbation levels, the MAEs exponentially increase, especially for shorter forecast times. Such a result aligns with the previous analysis, in which past data (including the past target of VMAX) influence decreases with increasing forecast times. The SHIPS result shows a similar pattern but with a more severe impact, indicating the sensitivity of model prediction to the initial condition. Therefore, a significant perturbation to VMAX will cause an adverse effect on any models, particularly at shorter forecast times. Consistently, under a perfect model framework,

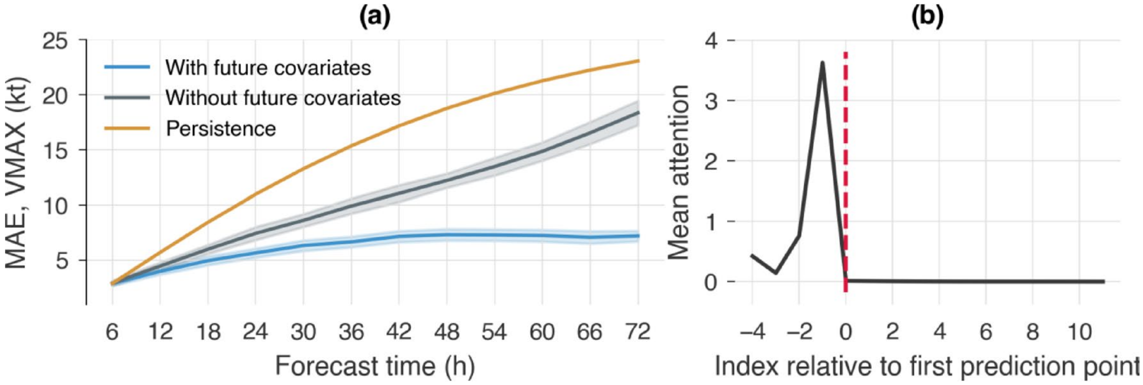

Fig. 10. (**a**) Comparison of MAE for TFT models with and without future covariates including persistence model result. The shaded areas with light colors in the TFT results indicate the 95% confidence interval from 100 samples. (**b**) Temporal attention. The red dashed vertical line indicates the first prediction point separating past and future data points or time steps.

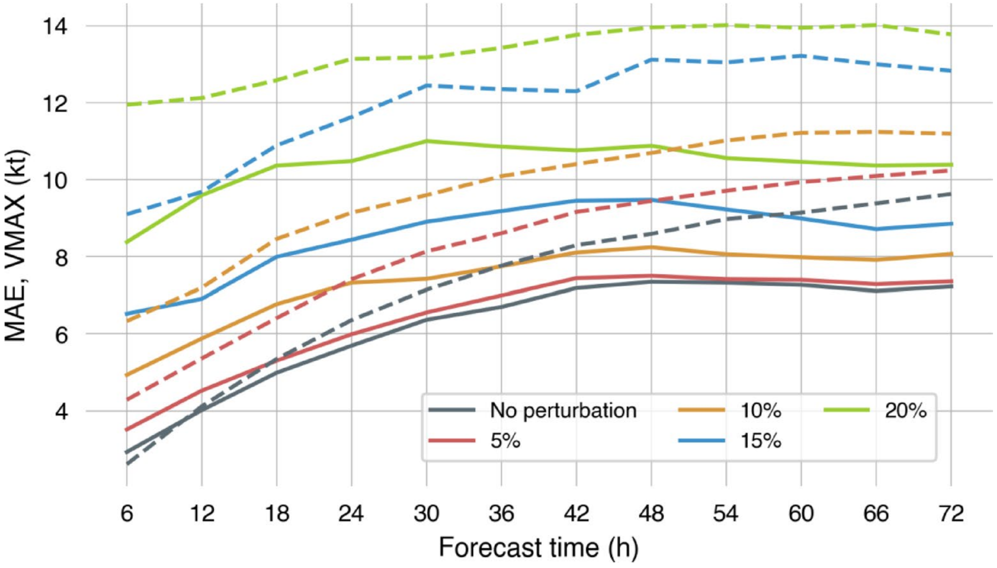

Fig. 11. Comparison of MAEs of median predictions of TFT (solid lines) and SHIPS (dashed lines) with varying perturbation level of VMAX.

Emanuel and Zhang[66] showed that model error at relatively shorter forecast times is greatly influenced by the mischaracterization of the initial intensity, and longer forecast times are more affected by environmental conditions.

Conclusion

We have demonstrated the capability of TFT in predicting TCs with RI, which has the potential to progress the statistical–dynamical modeling framework of SHIPS and other deep learning models. TFT achieves approximately 12% improvement relative to the conventional SHIPS model for non-RI TCs and 14% for RI TCs. Besides the efficiency of multi-horizon forecasting, the main advantage of TFT is attributed to correlating only relevant input features and data points to target future states. Additionally, as a probabilistic model, TFT produces multiple quantile predictions, which provide a range of possible solutions indicating the likely best and worst-case values to facilitate risk management efforts. More importantly, unlike the traditional data-driven approaches with their inherent black-box nature, TFT offers a certain degree of interpretability in its solution. However, we note that TFT's rank of predictor importance may not always constitute feature selection for other models due to data structure differences.

Data availability

The best track data from the Regional Specialized Meteorological Center Tokyo are available on their website (https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/trackarchives.html). The reanalysis products of the JRA-55 can be downloaded from https://jra.kishou.go.jp/JRA-55/index_en.html#jra-55. The MOVE rea nalysis dataset is available at https://search.diasjp.net/en/dataset/MOVEJPN_MRI_2020. The JMA successive geostationary satellite (Himawari) data are available at http://www.cr.chiba-u.jp/japanese/database.html.

Code availability

The TFT model is based on DARTS package downloaded from https://unit8co.github.io/darts/index.html. All figures are produced using the Matplotlib library implemented in Python (https://matplotlib.org/). The biLSTM code is modified from https://github.com/theodorosthd/An-Encoder-Decoder-Deep-Learning-Approach-for-Multistep-Service-Traffic-Prediction.

Received: 12 March 2025; Accepted: 8 August 2025
Published online: 25 August 2025

References

1. Powell, M. D. & Reinhold, T. A. Tropical cyclone destructive potential by integrated kinetic energy. *Bull. Am. Meteorol. Soc.* **88**, 513–526 (2007).
2. Honda, T. & Miyoshi, T. Predictability of the july 2018 heavy rain event in Japan associated with Typhoon Prapiroon and Southern convective disturbances. *SOLA* **17**, 113–119 (2021).
3. Zhang, W., Zhang, Y., Shu, S., Zheng, D. & Xu, L. Lightning distribution in tropical cyclones making landfall in China. *Front. Earth Sci.* **10**, 940205 (2022).
4. Lin, S.-J. & Chou, K.-H. The lightning distribution of tropical cyclones over the western North Pacific. *Mon. Weather Rev.* **148**, 4415–4434 (2020).
5. Woodruff, J. D., Irish, J. L. & Camargo, S. J. Coastal flooding by tropical cyclones and sea-level rise. *Nature* **504**, 44–52 (2013).

6. Rajeev, A. & Mishra, V. On the causes of tropical cyclone driven floods in India. *Weather Clim. Extrem.* **36**, 100432 (2022).
7. Heidarzadeh, M., Iwamoto, T., Šepić, J. & Mulia, I. E. Normal and reverse storm surges along the coast of Florida during the September 2022 Hurricane Ian: Observations, analysis, and modelling. *Ocean Model* **185**, 102250 (2023).
8. Mulia, I. E., Ueda, N., Miyoshi, T., Iwamoto, T. & Heidarzadeh, M. A novel deep learning approach for typhoon-induced storm surge modeling through efficient emulation of wind and pressure fields. *Sci. Rep.* **13**, 7918 (2023).
9. Tallapragada, V. Overview of the NOAA/NCEP Operational Hurricane Weather Research and Forecast (HWRF) Modelling System. in *Advanced Numerical Modeling and Data Assimilation Techniques for Tropical Cyclone Prediction* (eds. Mohanty, U. C. & Gopalakrishnan, S. G.) 51–106 (Springer Netherlands, Dordrecht, 2016). https://doi.org/10.5822/978-94-024-0896-6_3.
10. Mehra, A. et al. Advancing the state of the art in operational tropical cyclone forecasting at Ncep. *Trop. Cyclone Res. Rev.* **7**, 51–56 (2018).
11. Bush, M. et al. The first met office unified model–JULES regional atmosphere and land configuration, RAL1. *Geosci. Model Dev.* **13**, 1999–2029 (2020).
12. Doyle, J. et al. Tropical cyclone prediction using COAMPS-TC. *Oceanography* **27**, 104–115 (2014).
13. Courtney, J. B. et al. Operational perspectives on tropical cyclone intensity change part 2: Forecasts by operational agencies. *Trop. Cyclone Res. Rev.* **8**, 226–239 (2019).
14. Komaromi, W. A., Reinecke, P. A., Doyle, J. D. & Moskaitis, J. R. The naval research laboratory's coupled ocean-atmosphere mesoscale prediction system-tropical cyclone ensemble (COAMPS-TC Ensemble). *Weather Forecast.* **36**, 499–517 (2021).
15. Zhang, Z. et al. A review of recent advances (2018–2021) on tropical cyclone intensity change from operational perspectives, part 1: Dynamical model guidance. *Trop. Cyclone Res. Rev.* **12**, 30–49 (2023).
16. Kaplan, J. & DeMaria, M. Large-scale characteristics of rapidly intensifying tropical cyclones in the North Atlantic Basin. *Weather Forecast.* **18**, 1093–1108 (2003).
17. DeMaria, M. & Kaplan, J. A statistical Hurricane intensity prediction scheme (SHIPS) for the Atlantic Basin. *Weather Forecast.* **9**, 209–220 (1994).
18. DeMaria, M. & Kaplan, J. An updated statistical Hurricane intensity prediction scheme (SHIPS) for the Atlantic and Eastern North Pacific Basins. *Weather Forecast.* **14**, 326–337 (1999).
19. DeMaria, M., Mainelli, M., Shay, L. K., Knaff, J. A. & Kaplan, J. Further Improvements to the statistical Hurricane intensity prediction scheme (SHIPS). *Weather Forecast.* **20**, 531–543 (2005).
20. Higa, M. et al. Domain knowledge integration into deep learning for typhoon intensity classification. *Sci. Rep.* **11**, 12972 (2021).
21. Dvorak, V. F. Tropical cyclone intensity analysis and forecasting from satellite imagery. *Mon. Weather Rev.* **103**, 420–430 (1975).
22. Xu, X.-Y., Shao, M., Chen, P.-L. & Wang, Q.-G. Tropical cyclone intensity prediction using deep convolutional neural network. *Atmosphere* **13**, 783 (2022).
23. Tan, J., Yang, Q., Hu, J., Huang, Q. & Chen, S. Tropical cyclone intensity estimation using Himawari-8 satellite cloud products and deep learning. *Remote Sens.* **14**, 812 (2022).
24. Jiang, S., Fan, H. & Wang, C. Improvement of typhoon intensity forecasting by using a novel spatio-temporal deep learning model. *Remote Sens.* **14**, 5205 (2022).
25. Wenwei, X. et al. Deep learning experiments for tropical cyclone intensity forecasts. *Weather Forecast.* https://doi.org/10.1175/WAF-D-20-0104.1 (2021).
26. Boussioux, L., Zeng, C., Guénais, T. & Bertsimas, D. Hurricane forecasting: A novel multimodal machine learning framework. *Weather Forecast.* **37**, 817–831 (2022).
27. Wang, Z., Zhao, J., Huang, H. & Wang, X. A review on the application of machine learning methods in tropical cyclone forecasting. *Front. Earth Sci.* **10**, 902596 (2022).
28. Lim, B., Arık, S. Ö., Loeff, N. & Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **37**, 1748–1764 (2021).
29. Yang, Q., Lee, C.-Y. & Tippett, M. K. A long short-term memory model for global rapid intensification prediction. *Weather Forecast.* **35**, 1203–1220 (2020).
30. Chen, B., Kuo, Y. & Huang, T. A deep learning ensemble approach for predicting tropical cyclone rapid intensification. *Atmospheric Sci. Lett.* **24**, e1151 (2023).
31. Chen, S.-A., Li, C.-L., Arik, S. O., Yoder, N. C. & Pfister, T. TSMixer: An All-MLP Architecture for Time Series Forecast-ing. *Trans. Mach. Learn. Res.* (2023).
32. Li, S., Wang, T., Li, G., Skulstad, R. & Zhang, H. Short-term ship roll motion prediction using the encoder–decoder Bi-LSTM with teacher forcing. *Ocean Eng.* **295**, 116917 (2024).
33. Theodoropoulos, T., Maroudis, A.-C., Violos, J. & Tserpes, K. An Encoder-Decoder Deep Learning Approach for Multistep Service Traffic Prediction. in *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)* 33–40 (IEEE, Oxford, United Kingdom, 2021). https://doi.org/10.1109/BigDataService52369.2021.00010.
34. Yamaguchi, M. et al. Tropical cyclone intensity prediction in the western north Pacific Basin using SHIPS and JMA/GSM. *SOLA* **14**, 138–143 (2018).
35. Shimada, U. et al. Further improvements to the statistical Hurricane intensity prediction scheme using tropical cyclone rainfall and structural features. *Weather Forecast.* **33**, 1587–1603 (2018).
36. Ono, M., Notsuhara, S., Fukuda, J., Igarashi, Y. & Bessho, K. *Operational Use of the Typhoon Intensity Forecasting Scheme Based on SHIPS (TIFS) and Commencement of Five-Day Tropical Cyclone Intensity Forecasts.* https://www.jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/techrev/text21-2.pdf (2019).
37. Koba, H., Hagiwara, S., Osano, S. & Akashi, S. Relationships between CI number and minimum sea level pressure/maximum wind speed of tropical cyclones. *Geophys. Mag.* **44**, 15–25 (1991).
38. Kobayashi, S. *et al.* The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *J. Meteorol. Soc. Jpn. Ser II* **93**, 5–48 (2015).
39. Usui, N. et al. Meteorological research institute multivariate ocean variational estimation (MOVE) system: Some early results. *Adv. Space Res.* **37**, 806–822 (2006).
40. Bessho, K. *et al.* An Introduction to Himawari-8/9&mdash; Japan&rsquo;s New-Generation Geostationary Meteorological Satellites. *J. Meteorol. Soc. Jpn. Ser II* **94**, 151–183 (2016).
41. Tabata, T., John, V. O., Roebeling, R. A., Hewison, T. & Schulz, J. Recalibration of over 35 years of infrared and water vapor channel radiances of the JMA geostationary satellites. *Remote Sens.* **11**, 1189 (2019).
42. JMA. *Outline of the Operational Numerical Weather Prediction at the Japan Meteorological Agency*. https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline2022-nwp/index.htm (2022).
43. Demaria, M. & Kaplan, J. Sea surface temperature and the maximum intensity of Atlantic tropical cyclones. *J. Clim.* **7**, 1324–1334 (1994).
44. DeMaria, M. Tropical cyclone intensity change predictability estimates using a statistical-dynamical model. in (Amer. Meteor. Soc., Tucson, AZ, 2010).
45. Wang, C., Yang, N. & Li, X. Advancing forecasting capabilities: A contrastive learning model for forecasting tropical cyclone rapid intensification. *Proc. Natl. Acad. Sci.* **122**, e2415501122 (2025).
46. DeMaria, M., Sampson, C. R., Knaff, J. A. & Musgrave, K. D. Is tropical cyclone intensity guidance improving?. *Bull. Am. Meteorol. Soc.* **95**, 387–398 (2014).

47. Shimada, U. Tropical cyclone intensity forecasting with three multiple linear regression models and random forest classification. *J. Meteorol. Soc. Jpn. Ser II* https://doi.org/10.2151/jmsj.2024-030 (2024).
48. Mulia, I. E., Ueda, N., Miyoshi, T., Gusman, A. R. & Satake, K. Machine learning-based tsunami inundation prediction derived from offshore observations. *Nat. Commun.* **13**, 5489 (2022).
49. Abadi, M. *et al.* TensorFlow: A system for large-scale machine learning. in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* 265–283 (2016).
50. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* 2623–2631 (ACM, Anchorage AK USA, 2019). https://doi.org/10.1145/3292500.3330701.
51. Wu, B., Wang, L. & Zeng, Y.-R. Interpretable wind speed prediction with multivariate time series and temporal fusion transformers. *Energy* **252**, 123990 (2022).
52. Zhang, H., Zou, Y., Yang, X. & Yang, H. A temporal fusion transformer for short-term freeway traffic speed multistep prediction. *Neurocomputing* **500**, 329–340 (2022).
53. López Santos, M., García-Santiago, X., EchevarríaCamarero, F., Blázquez Gil, G. & Carrasco Ortega, P. Application of temporal fusion transformer for day-ahead PV power forecasting. *Energies* **15**, 5232 (2022).
54. Herzen, J. et al. Darts: User-friendly modern machine learning for time series. *J. Mach. Learn. Res.* **23**, 1–6 (2022).
55. Kaplan, J., DeMaria, M. & Knaff, J. A. A revised tropical cyclone rapid intensification index for the Atlantic and Eastern North Pacific Basins. *Weather Forecast.* **25**, 220–241 (2010).
56. Chan, K. T. F. & Chan, J. C. L. Global climatology of tropical cyclone size as inferred from QuikSCAT data. *Int. J. Climatol.* **35**, 4843–4848 (2015).
57. Shimada, U. Variability of environmental conditions for tropical cyclone rapid intensification in the Western North Pacific. *J. Clim.* **35**, 4437–4454 (2022).
58. Oreshkin, B. N., Carpov, D., Chapados, N. & Bengio, Y. Meta-learning framework with applications to zero-shot time-series forecasting. *Proc. AAAI Conf. AI* **35**, 9242–9250 (2021).
59. Bhatia, K. et al. A potential explanation for the global increase in tropical cyclone rapid intensification. *Nat. Commun.* **13**, 6626 (2022).
60. Cloud, K. A. et al. A feed forward neural network based on model output statistics for short-term Hurricane intensity prediction. *Weather Forecast.* **34**, 985–997 (2019).
61. Chen, S. & Li, W. A review of Typhoon inner core characteristics and their relationship with intensity changes. *Atmosphere* **15**, 1522 (2024).
62. Zhang, D.-L., Zhu, L., Zhang, X. & Tallapragada, V. Sensitivity of idealized hurricane intensity and structures under varying background flows and initial vortex intensities to different vertical resolutions in HWRF. *Mon. Weather Rev.* **143**, 914–932 (2015).
63. Response, S. & Simulations, N. Nonhydrostatic, three-dimensional perturbations to balanced, Hurricane-like vortices. Part II symmetric response and nonlinear simulations. *J. Atmospheric Sci.* **60**, 2717–2745 (2003).
64. Wahiduzzaman, M., Cheung, K. K., Luo, J.-J. & Bhaskaran, P. K. A spatial model for predicting North Indian Ocean tropical cyclone intensity: Role of sea surface temperature and tropical cyclone heat potential. *Weather Clim. Extrem.* **36**, 100431 (2022).
65. Wu, L. et al. Relationship of environmental relative humidity with North Atlantic tropical cyclone intensity and intensification rate. *Geophys. Res. Lett.* **39**, 2012GL053546 (2012).
66. Emanuel, K. & Zhang, F. On the predictability and error sources of tropical cyclone intensity forecasts. *J. Atmospheric Sci.* **73**, 3739–3747 (2016).
67. Hu, C.-C. & Wu, C.-C. Ensemble sensitivity analysis of tropical cyclone intensification rate during the development stage. *J. Atmospheric Sci.* **77**, 3387–3405 (2020).

Acknowledgements

I.E.M. is financially supported by Hibah DTTP 2025 (Contract No 4770/IT1.B07.1/KU/2025) managed by the Directorate of Research and Innovation (DRI), Institut Teknologi Bandung. T.M. is supported by the JST (Grant Number JPMJCR24Q3 and JPMJSA2109), the COE research grant in computational science from Hyogo Prefecture and Kobe City through Foundation for Computational Science, the JSPS KAKENHI (Grant Number JP24H00021), the Japan Aerospace Exploration Agency, and the RIKEN TRIP initiative (RIKEN Prediction Science). This work was also supported by the RIKEN Pioneering Project "Prediction for Science".

Author contributions

I.E.M. conceived the study and performed the computations. U. S. compiled and processed the data used in SHIPS and involved in the results interpretation. N.U. and T.M. led and supervised the project including results interpretation. M.T.M involved in the atmospheric model analysis and additional statistical and numerical experiments. I.E.M wrote the manuscript together with all authors.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-15522-7.

Correspondence and requests for materials should be addressed to I.E.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.