Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Physics-Guided Multimodal Learning for Flexible Tropical Cyclone Intensity Forecasting

Kuai Dai
Hong Kong University of Science and Technology https://orcid.org/0000-0002-9921-4249
Hui Su
cehsu@ust.hk

Hong Kong University of Science and Technology
Chengxing Zhai
Hong Kong University of Science and Technology
Xiaoyang Zhang
Nanyang Technological University
Zesheng Shi
Harbin Institute of Technology
Chenghao Yang
Zhongnan University of Economics and Law

Research Article

Keywords:

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11618500/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Physics-Guided Multimodal Learning for Flexible Tropical Cyclone Intensity Forecasting

Kuai Dai [1], Hui Su [1*], Chengxing Zhai [1], Xiaoyang Zhang [2], Zesheng Shi [3], Chenghao Yang [4]

[1]The Hong Kong University of Science and Technology, [2]Nanyang Technological University, [3]Harbin Institute of Technology, [4]Zhongnan University of Economics and Law

Corresponding author: cehsu@ust.hk

Abstract

Tropical cyclones (TCs) bring extreme precipitation and strong winds, severely disrupting transportation systems, electricity supply, and public safety. Accurate TC intensity forecasting remains challenging due to the complex and nonlinear dynamics of storm evolution. Although recent AI-based methods have shown promise, they typically rely on best-track data or geostationary satellite observations with generic fusion strategies, without explicitly modeling the consistency between physical variables and visual features. This limits their ability to accurately forecast TC intensity and capture its temporal changes. To address this problem, we propose TCFNet, a physics-guided multimodal learning framework that uses intensity variables to guide visual representation learning and jointly models TC spatial structure and temporal evolution. In addition, with flexible encoding and temporal modeling architectures, TCFNet can handle missing multimodal observations and support continuous-time forecasting, improving its adaptability in practical scenarios. Extensive experiments on two cyclone-active basins, the Western North Pacific and South Pacific, demonstrate that TCFNet outperforms prior baselines, especially in capturing intensity changes. The source code will be made publicly available.

Introduction

Tropical cyclones (TCs), also known as typhoons or hurricanes in certain regions, have garnered significant attention from meteorological researchers worldwide due to their huge destructive impacts on human society (Woodruff, Irish, and Camargo 2013; Emanuel 2018). Accurately forecasting TC intensity, commonly characterized by maximum sustained wind and minimum sea-level pressure, is crucial for assessing and mitigating the associated risks and disasters (Yang et al. 2019; Cangialosi et al. 2020). In general, stronger TCs are associated with higher maximum winds and lower central pressure.

Traditional intensity forecasting approaches usually utilize numerical methods, such as atmospheric simulation models, to make predictions. However, the methods have been challenged by the complex dynamics and multiple environmental factors involved in the evolution process of TCs (Emanuel and Zhang 2016). In the past few years, data-driven approaches have shown promise for TC studies (Li et al. 2025).

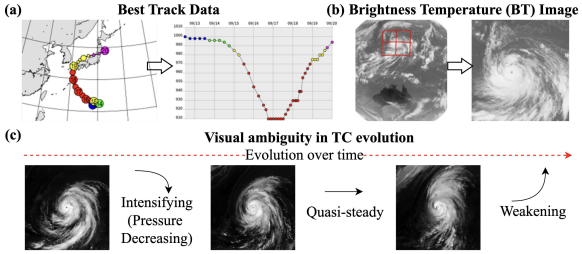


Figure 1: The data sources used in this work and the motivation for TC intensity forecasting. (a) Best track data that records TC intensity, including minimum central pressure and maximum sustained wind speed; (b) Corresponding brightness temperature (BT) images from geostationary satellites, cropped around the cyclone center and covering the spatial extent of the storm; (c) Satellite observations exhibit similar visual patterns while corresponding to distinctly different physical evolution states of TCs.

In particular, several types of data can be exploited to build machine learning forecasting models, including best track data, reanalysis data, such as ERA5 (Hersbach et al. 2020), and geostationary satellite imagery. The best-track data are provided by meteorological agencies worldwide and contain locations, trajectories, and intensities of TCs, which can serve as the ground-truth reference for TC analysis(Knapp et al. 2010). While reanalysis data provide large-scale environmental contexts, their relatively coarse resolution and post-processed characteristics limit their effectiveness for high-resolution and real-time intensity forecasting. In contrast, geostationary satellites continuously capture brightness temperature imagery with high resolution that directly reflects the cloud-top dynamics around the cyclone center, making them naturally suitable for intensity forecasting.

In this work, we study TC intensity forecasting via a multimodal learning perspective by integrating geostationary satellite observations and track data, which are presented in Figure 1 (a) and (b). As a newly emerging research topic, recent works (Li et al. 2025) show the potential of multimodal learning for TC forecasting. However, a fundamental challenge lies in interpreting satellite observations for intensity evolution. As illustrated in Figure 1 (c), satellite observations show that similar visual structures may correspond

*Corresponding author.

to markedly different intensity evolution trends, including intensification, steady, and weakening phases. This inconsistency indicates that similar visual patterns should convey different semantics for various physical conditions, posing challenges to generic feature learning strategies. Moreover, the existing methods require complete and regular historical observations, which cannot be adapted to practical scenarios of missing observations and the demand for forecasting at flexible lead times. To address the problems, we propose a tropical cyclone fusion network (TCFNet), a physics-guided multimodal framework for TC that leverages physical variables to condition visual representation, enabling the model to learn intensity-aware patterns for TC intensity forecasting. Additionally, TCFNet adopts a time-aware encoding strategy coupled with a continuous-time state propagation module to support irregular observations and continuous forecasting. Extensive experiments on two active cyclone basins verify the effectiveness of TCFNet, especially in capturing abrupt intensity changes. The contributions of this work are summarized as follows:

- We propose a physics-guided multimodal framework TCFNet, which guides visual representation learning with physical variables and effectively integrates satellite imagery and track data for TC intensity forecasting.
- We introduce a time-aware encoding strategy with continuous dynamics modeling to handle missing observations and support flexible temporal forecasting.
- Comprehensive experiments conducted on Western North Pacific and the South Pacific demonstrate the advantages of TCFNet over state-of-the-art approaches, particularly in forecasting intensity changes.

Related Work

Unimodal Methods

Best Track Data-based Approaches In the early stage, researchers conduct TC intensity forecasting by building statistical or physical approaches (DeMaria and Kaplan 1994; Chen and Meng 2001; DeMaria 2009) to model the best track data. (DeMaria and Kaplan 1994) designs a SHIPS model to forecast intensity via multi-variable regression analysis, which is the foundation of the statistical forecasting approaches. (DeMaria 2009) proposes an efficient physical forecasting system by representing the intensity change process via some key PDEs. However, due to complex factors involved in TC evolution processes, the development of these mathematical forecasting methods has been stagnant (DeMaria et al. 2014; Emanuel and Zhang 2016). In recent years, researchers have explored machine learning techniques for time series, such as RNN (Pan, Xu, and Shi 2019; Na, Na, and Son 2022) and Transformer (Jiang et al. 2023), to forecast TC intensity with best track data. Specifically, (Pan, Xu, and Shi 2019) introduces a recurrent neural network to predict the TC intensity. (Jiang et al. 2023) employs the multi-head attention mechanism to model the long-term dependency within best track sequences.

Brightness Temperature Images-based Approaches The existing approaches for TC intensity regression with brightness temperature images are roughly categorized into two taxonomies, i.e., traditional methods and machine learning-based methods. First, as a pioneering traditional method, Dvorak Technique (DT) (Dvorak 1975) provides a series of intensity estimation principles based on cloud patterns of satellite images. However, this method is susceptible to the subjectivity of meteorological experts, which limits its reliability and applicability. To alleviate this issue, variants of DT are further proposed, such as the Objective Dvorak Technique (ODT) (Velden, Olander, and Zehr 1998), Advanced Objective Dvorak Technique (AODT) (Olander, Velden, and Kossin 2004), and Advanced Dvorak Technique (ADT) (Olander and Velden 2007), by introducing automatic and fine-grained image processing techniques. Besides the DT-based approaches, another traditional method, Directional Axis Variance (DAV), is proposed for TC intensity regression, in which a quantitative metric is designed to assess TC structure using satellite infrared imagery. Second, the recent machine learning methods utilize advanced vision models such as CNN (Chen, Chen, and Lin 2018; Pradhan et al. 2017; Wang et al. 2021; Ma et al. 2024) and Vision Transformer (Chen et al. 2023; Tian et al. 2023) to extract spatial features from satellite imagery for TC intensity estimation. Specifically, (Chen, Chen, and Lin 2018) designs a rotation-blended CNN model by introducing the rotation-invariance of TCs. (Ma et al. 2024) proposes a multiscale and multilayer feature modeling approach to estimate intensity from satellite images. To overcome the limitation of CNN in capturing global structures of TCs, (Tian et al. 2023) combines rotation equivariant convolution and Transformer to make more accurate intensity estimation.

The previous works mainly concentrate on intensity estimation with single modality, i.e., best track data or satellite brightness temperature imagery. In this work, we aim to effectively integrate the satellite images and best track data and develop a multimodal fusion framework for accurate TC intensity forecasting.

Multimodal Methods

The study for multimodal TC intensity forecasting (Su et al. 2020), especially in integrating track data and satellite imagery, is still at a preliminary stage. Specifically, a recent work (Li et al. 2025) applies large language models (LLMs) such as Gemini (Comanici et al. 2025) and Qwen (Bai et al. 2025) for multimodal TC forecasting by designing a specialized position encoding mechanism PIPE to align the satellite imagery and time series. In addition, SatSolarCast (Dai et al. 2026) proposes a multimodal spatiotemporal forecasting framework that jointly processes brightness temperature satellite imagery and time-series data. Beyond the modality-specific fusion frameworks, the multi-variable meteorological foundation models such as ECMWF-IFS (DOCUMENTATION-Cy40r1 2020), Pangu-Weather (Bi et al. 2023), and GenCast (Price et al. 2025) have demonstrated remarkable success in global-scale trajectory and pattern forecasting. However, their capability in TC intensity prediction, particularly wind speed, remains limited (Shi et al. 2025), due to the coarse spatial resolution and weak sensitivity to mesoscale convective structures (Xu et al. 2025).

However, this difference should be interpreted with caution, as the training objectives are fundamentally different. Foundation models are designed for global-scale meteorological prediction and show strengths in large-scale field modeling and trajectory forecasting, rather than specialized TC intensity estimation. Furthermore, the differences in temporal resolution, output representation, and experimental settings make direct comparison with image-based intensity prediction tasks difficult (Li et al. 2025). Hence, the models are not included as baselines in this work. While the methods provide valuable insights into TC intensity forecasting with multi-modality, the approaches largely overlook the physical consistency between vision patterns and physical variables. To address this limitation, we propose a physics-guided multimodal framework that uses physical variables to guide visual representation learning and employs visual-conditioned fusion for TC intensity forecasting.

Problem Definition

Multimodal TC Intensity Forecasting

In practical forecasting scenarios, meteorological observations are often available at irregular intervals due to varying satellite scanning schedules, delays in post-processing best track data, or missing observations. To represent the flexible setting, we formulate the historical multimodal observations as two sets of time-stamped events $\mathcal{H}^v = \{(X_i^v, \tau_i^v)\}_{i=1}^{T_v}$ and $\mathcal{H}^p = \{(X_j^p, \tau_j^p)\}_{j=1}^{T_p}$. Here, X_i^v denotes the high-dimensional visual observation (i.e., satellite imagery), and X_j^p denotes the low-dimensional physical state (i.e., best track variables such as wind speed and pressure). The timestamps $\tau_i^v, \tau_j^p \in \mathbb{R}$ are not necessarily aligned across modalities. Notably, the time intervals $\Delta\tau_i^v = \tau_{i+1}^v - \tau_i^v$ and $\Delta\tau_j^p = \tau_{j+1}^p - \tau_j^p$ are both irregular, enabling asynchronous sampling within each modality.

Discrete-Time Evaluation Protocol

For a fair comparison, we follow standard evaluation protocols (Kitamoto et al. 2023) and test our model on a regular grid setting. Assume $X^v = \{X_1^v, \cdots, X_T^v\} \in \mathbb{R}^{T\times C\times H\times W}$ with corresponding timestamps $\tau^v = \{\tau_1^v, \cdots, \tau_T^v\}$ represents the historical satellite observations, and $X^p = \{X_1^p, \cdots, X_T^p\} \in \mathbb{R}^{T\times D}$ with timestamps $\tau^p = \{\tau_1^p, \cdots, \tau_T^p\}$ represents the historical best track variables, where T is the history input length. Given $\mathcal{H}^v = (X^v, \tau^v)$, $\mathcal{H}^p = (X^p, \tau^p)$, and model $\mathcal{F}(\cdot)$ with parameters θ, our objective is to produce the future TC intensity sequence $\hat{Y} = \{\hat{Y}_{T+1}, \cdots, \hat{Y}_{T+K}\} \in \mathbb{R}^{K\times D}$ for the next K steps, which should be as close as possible to the ground-truth Y. Formally, the optimization problem is defined as:

$$\hat{Y} = \mathcal{F}(\mathcal{H}^v, \mathcal{H}^p; \theta), \theta^* = \arg\min_{\theta} \mathcal{L}(\hat{Y}, Y). \quad (1)$$

$\mathcal{L}(\cdot,\cdot)$ denotes the training loss function. In particular, though training is performed on regular grid-aligned sequences, our proposed TCFNet explicitly leverages timestamps τ^v and τ^p through temporal encoding and utilizes ordinary differential equations (ODEs) to model the underlying continuous dynamics, improving robustness to irregular sampling and enabling flexible temporal prediction when needed.

Methodology

As illustrated in Figure 2, TCFNet is composed of three integral modules: (1) asynchronous history encoding, which projects high-dimensional satellite imagery and low-dimensional track variables into a unified latent space, explicitly handling irregular sampling via Time-Aware GRUs; (2) continuous-time state propagation, which leverages ODEs to model the continuous-time evolution of physical states of TCs, naturally supporting extrapolation to irregular future timestamps; (3) visual-conditioned asymmetric fusion, a mechanism that utilizes visual context to condition the physical state trajectories at each decoding step.

Asynchronous History Encoding

To address the modality gap and irregularity in historical multimodal observations, we introduce a dual-stream encoding strategy, i.e., a visual stream and a physical state stream. The former stream extracts trend-aware environmental contexts, while the latter one captures low-dimensional temporal intensity features. Both streams utilize time-aware GRUs to robustly process asynchronous observations.

Visual Stream The visual stream is responsible for encoding the sequence of satellite imagery X^v into latent representations. Since satellite images are high-dimensional, we must project them into latent vectors to align with the low-dimensional best track variables. A straightforward approach is to employ a CNN backbone, such as ResNet. However, our goal is to capture the dynamic evolution trend (e.g., intensifying or weakening) from cloud patterns, rather than simply extracting static appearance features.

To achieve this, we design a physics-guided pre-training encoder, PhyMoCo. This module is built upon the momentum contrast (MoCo) paradigm (He et al. 2020), which can learn invariant representations by reducing and increasing distances of positive and negative pairs in the latent space. However, the structure cannot capture the long-term evolution of TC intensity, as the selection of positive sample pairs is based on spatial similarity. Based on this insight, we propose a trend-constrained positive pair sampling strategy, which aligns the visual features in the latent space with intensity trends. Given that central pressure measurements are generally more robust and continuous than wind speed estimates, we utilize pressure gradients as a reliable proxy for TC intensity trends. Let $\mathcal{I} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_M\}$ represent the satellite observations of entire-life TCs and $\mathcal{P}$ denote the corresponding ground-truth pressure sequence. We derive the intensity trend $\mathcal{T}$ from the gradient of $\mathcal{P}$ (intensifying: 1, stable: 0, weakening: -1). Instead of relying solely on temporal proximity, we prioritize trend-consistent positive pairs during sampling $(\mathcal{I}_i, \mathcal{I}_j)$:

$$(\mathcal{I}_i, \mathcal{I}_j) : \mathcal{T}_i = \mathcal{T}_j \quad \text{and} \quad |i-j| <= W, \quad (2)$$

where $\mathcal{W}$ is a dynamic search window. This encourages the encoder $\mathcal{E}$ to learn similar representations for temporally distant images sharing the same intensity trend. Specifically,

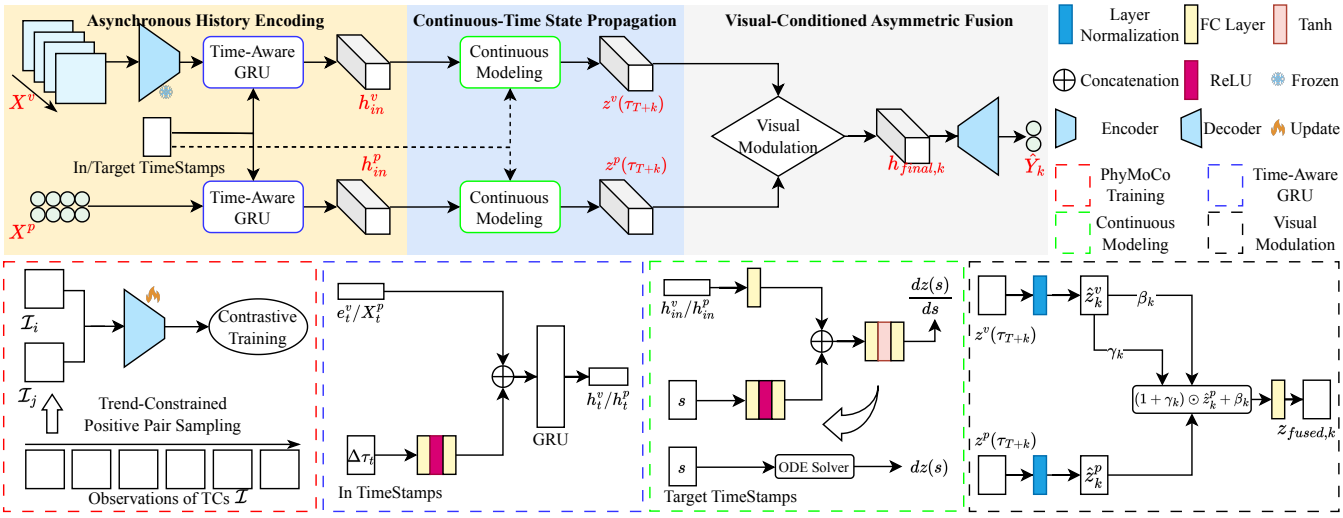


Figure 2: The overall layout of TCFNet. Its three modules, asynchronous history encoding, continuous-time state propagation, and visual-conditioned asymmetric fusion, account for historical observations encoding, temporal modeling, and multimodal fusion, respectively. Notably, we illustrate the procedure for the k-th prediction for simplification.

Algorithm 1: Trend-Constrained Positive Pair Sampling

Require: TC image sequence $\mathcal{I}$, pressure sequence $\mathcal{P}$, scheduler $\mathcal{W}$
Ensure: Positive index pair (i_1, i_2)
$\mathcal{T} \leftarrow \texttt{ComputeTrend}(\mathcal{P})$
$L \leftarrow \mathcal{W}.\texttt{get_value}(), S \leftarrow 3$
Sample $i_1 \sim \mathcal{U}(0, |\mathcal{I}| - 2)$
for $\Delta = L - 1$ down to 1 **do**
  $i_2 \leftarrow \min(i_1 + \Delta, |\mathcal{I}| - 1)$
  if $\mathcal{T}[i_2] = \mathcal{T}[i_1]$ **then**
    return (i_1, i_2)
  end if
end for
Re-sample $i_1 \sim \mathcal{U}(0, |\mathcal{I}| - 2)$
Sample $r \sim \mathcal{U}(1, S)$
$i_2 \leftarrow \min(i_1 + r, |\mathcal{I}| - 1)$
return (i_1, i_2)

Algorithm 1 searches from larger to smaller temporal offsets within a dynamic window L to prioritize trend-consistent positive pairs with greater temporal separation. If no such pair is found, we randomly sample within an empirically chosen short window ($S = 3$) as a fallback, without enforcing trend consistency.

After pre-training, we use the frozen encoder to extract features $e_t^v = \mathcal{E}(X_t^v) \in \mathbb{R}^{512}$ from the t-th historical satellite observation X_t^v. These features and corresponding scalar timestamp τ_t are then fed into a time-aware variant of the GRU (Cho et al. 2014) to encode the visual history. Specifically, the time interval $\Delta\tau_t^v = \tau_t^v - \tau_{t-1}^v$ is first encoded into a vector representation with a lightweight time encoder and then concatenated with visual features and fed into GRU:

$$h_{1:T_v}^v = \text{GRU}_{vis}\big(e_t^v \oplus \psi(\Delta\tau_t^v)\big), \quad t = 1, \ldots, T_v, \quad (3)$$

where the ψ denotes the time encoding function and $\oplus$ is the concatenation operation. We utilize the last hidden states $h_{in}^v = h_{T_v}^v$ as the overall representation of the historical satellite observations, which are robust to irregular sampling and missing frames.

Physical State Stream The physical state stream handles the sequence of best track variables $\mathbf{x}^p$. Since these variables directly lie on the low-dimensional state manifold, we bypass complex feature extraction. To account for asynchronous sampling relative to the visual stream, we employ a separate Time-Aware GRU:

$$h_{1:T_p}^p = \text{GRU}_{phys}\big(X_t^p \oplus \psi(\Delta\tau_t^p)\big), \quad t = 1 \ldots T_p. \quad (4)$$

Similarly, $h_{in}^p = h_{T_p}^p$ serves as the global patterns for the physical dynamics.

Continuous-Time State Propagation

Different from discrete recurrent models that are constrained by fixed time steps, we model the evolution of TC intensity as a continuous physical process governed by ordinary differential equations.

Let $z(s)$ denote the latent state at a relative time s from the last available observation at time τ_{in} (τ_{in} is not equal to τ_T due to possible missing input observations). The evolution of the system is defined by a neural network f_ϕ that explicitly conditions on time: $\mathrm{d}z(s)/\mathrm{d}s = f_\phi(z(s), s)$. Assume $s_k = \tau_{T+k} - \tau_{in}$ is the specific interval at time τ_{T+k}, and we apply this formulation to both visual and physical streams. Taking the last historical states h_{in}^v and h_{in}^p as initial conditions with FC layers, $z^v(0) = FC(h_{in}^v)$, $z^p(0) = FC(h_{in}^p)$. Then, we solve the ODE initial value problem to extrapolate the states

to the target timestamp τ_{T+k}:

$$z^v(\tau_{T+k}) = z^v(0) + \int_0^{s_k} f_\phi^v(z^v(s), \psi(s))\, ds, \quad (5)$$

$$z^p(\tau_{T+k}) = z^p(0) + \int_0^{s_k} f_\phi^p(z^p(s), \psi(s))\, ds. \quad (6)$$

Notably, the continuous-time dynamics are numerically integrated using an adaptive-step ODE solver with explicit time conditioning, enabling stable and efficient state propagation over large time intervals.

Visual-Conditioned Asymmetric Fusion

Through the aforementioned modules, we obtain the evolved latent visual states $z^v(\tau_{T+k})$ and physical states $z^p(\tau_{T+k})$ at each predicted time step $k \in \{1, \ldots, K\}$. To effectively integrate the features from two modalities, we propose an asymmetric fusion strategy. Specifically, we draw inspiration from the feature-wise linear modulation (FiLM) mechanism (Perez et al. 2018), which was originally designed for visual reasoning tasks. However, in the context of TC intensity forecasting, we reinterpret this mechanism through a new perspective. We think that the visual information, such as cloud compactness and eye structure, serves as the environmental context that reflects the fundamental evolution of physical states of TCs.

First, a layer normalization (LN) is applied to stabilize the visual and physical latent states:

$$\hat{z}_k^v = \mathrm{LN}(z^v(\tau_{T+k})), \quad \hat{z}_k^p = \mathrm{LN}(z^p(\tau_{T+k})). \quad (7)$$

Then, we generate the modulation parameters γ (scaling) and β (shifting) from the visual features:

$$\gamma_k = \tanh(W_\gamma \hat{z}_k^v + b_\gamma), \quad \beta_k = W_\beta \hat{z}_k^v + b_\beta. \quad (8)$$

The fused representation $z_{fused,k}$ is obtained by applying an affine transformation to the physical features, conditioned on the visual features:

$$z_{fused,k} = (1 + \gamma_k) \odot \hat{z}_k^p + \beta_k, \quad (9)$$

where $\odot$ denotes the element-wise multiplication. By using the $\tanh$ activation for γ, we constrain the scaling factor to the range $[0, 2]$, preventing numerical instability during the multiplication operation. Finally, the fused features are projected to the hidden dimension:

$$h_{final,k} = W_{proj} z_{fused,k} + b_{proj}. \quad (10)$$

The final prediction $\hat{Y}_k$ at the k-th time step is generated by decoding the fused latent features with a multi-layer perceptron (MLP) layer: $\hat{Y}_k = \mathrm{MLP}_{dec}(h_{final,k})$. This step-wise decoding ensures that the prediction at each future moment is conditioned on its specific continuous dynamical state. At last, we leverage the combination of mean squared error (MSE) and mean absolute error (MAE) loss between the predicted results $\hat{Y}$ and the ground truth Y to optimize TCFNet:

$$\mathcal{L}(\hat{Y}, Y) = \frac{1}{K} \sum_{k=1}^{K} \left(\|\hat{Y}_k - Y_k\|_2^2 + \lambda \|\hat{Y}_k - Y_k\|_1 \right). \quad (11)$$

Experiments

Experimental Setup

Datasets In this work, we conduct experiments on two typical cyclone-active basins, Western North Pacific (WNP) and South Pacific (SP). The corresponding geostationary satellite imagery and best track data are obtained from the latest large-scale open-source dataset DigitalTyphoon (Kitamoto et al. 2023). In particular, the best track data with a temporal interval of 3 hours (interpolated) for TCs in the two areas are provided by the Japan Meteorological Agency (JMA) and Australian Bureau of Meteorology, respectively. Additionally, the satellite imagery is derived from the infrared channel (around 11 μm waveband) of data produced by Himawari series satellites. All images are cropped into 512×512-size ones based on the cyclone center locations (longitude and latitude) from the best track data, resulting in tracking satellite image sequences with a spatiotemporal resolution of 5 km and 1 hour. To align the temporal resolution of the two data sources, the temporal resolution is set to 3 hours. Given the past 12-hour observations, we aim to forecast the future 24-hour pressure and wind speed. Notably, the training, validation, and test datasets are divided at the TC-event level according to the timeline, with ratios of 0.7, 0.15, and 0.15, respectively. Detailed statistics of datasets are presented in the Supplementary Information.

Baselines We adopt Persistence and the operational Typhoon Intensity Forecasting Scheme (TIFS) from JMA (Kitamoto, Dzik, and Faure 2024) as domain-specific baselines; time-series models including LSTM (Hochreiter and Schmidhuber 1997), Transformer (Vaswani et al. 2017), PatchTST (Nie et al. 2023), Mamba (Gu and Dao 2024), TimeMixer (Wang et al. 2024), and TEFN (Zhan et al. 2025); and vision-based methods including ResNet18 (Kitamoto et al. 2023) and CNNLSTM (Donahue et al. 2015) as unimodal baselines. Moreover, we include the state-of-the-art multimodal TC forecasting approaches PIPE-3B (Li et al. 2025) for comparison. Additionally, we consider a more recently proposed multimodal spatiotemporal forecasting framework SatSolarCast (Dai et al. 2026) that can jointly process satellite imagery and auxiliary time-series data. Detailed settings are presented in Supplementary Information.

Implementation Details During the training procedure of PhyMoCo, original satellite images are first center-cropped to a size of 384×384 and then randomly cropped into 224×224-size images. The temporal window length is progressively scheduled over 6, 12 during training, with a 1k-step warm-up. As for the inference phase, the satellite images are center-cropped into 224×224-size ones. Additionally, training batch sizes for PhyMoCo on Western North Pacific and South Pacific datasets are set to 768 and 256. TCFNet is optimized using an Adam optimizer with learning rates of 5e-4 and 2e-4 on the WNP and SP datasets, respectively. Notably, an early stopping strategy with patience of 10 is adopted to train TCFNet. The batch size is set to 128, and the λ in the training loss function is set to 0.5. The hidden state dimension of GRU encoders and ODE functions is 1024, while the time embedding size is 32. The experiments are conducted on a

Table 1: Pressure and wind-speed RMSE at different forecast lead times. Methods marked with an asterisk (*) report results from the original paper and are included as performance references. Note that the TIFS, ResNet18, and PIPE-3B results are available only for selected pressure forecasting lead times in the WNP basin, and dashes indicate unavailable results. Bold values indicate the best performance.

Category	Method	Western North Pacific (WNP)								South Pacific (SP)							
		Pressure RMSE (hPa)				Wind RMSE (kt)				Pressure RMSE (hPa)				Wind RMSE (kt)			
		T+6 (hour)	T+12	T+18	T+24	T+6	T+12	T+18	T+24	T+6	T+12	T+18	T+24	T+6	T+12	T+18	T+24
domain	Persistence	4.68	8.59	12.13	15.35	9.18	14.18	18.49	22.45	3.26	5.90	8.23	10.24	4.98	8.35	11.36	14.05
	TIFS*	-	9.20	10.80	12.00	-	-	-	-	-	-	-	-	-	-	-	-
uni-	LSTM	3.54	6.26	9.02	11.58	6.95	10.25	12.91	15.47	**2.39**	4.63	6.67	8.39	**3.75**	6.53	9.17	11.54
	Transformer	4.72	6.20	9.03	12.04	8.43	10.18	13.15	16.63	3.60	4.68	6.50	8.34	5.33	6.64	8.89	11.29
	PatchTST	3.50	6.31	9.02	11.52	**6.83**	10.17	12.77	15.39	2.85	4.83	6.53	8.18	3.98	6.70	9.11	11.35
	Mamba	3.53	6.61	9.43	11.97	6.94	10.49	13.45	16.15	2.56	4.72	6.74	8.42	3.80	6.53	9.07	11.31
	TimeMixer	3.14	6.06	8.87	11.47	7.32	10.34	12.99	15.71	2.83	4.77	6.74	8.48	4.10	6.45	8.88	11.24
	TEFN	3.43	6.53	9.57	12.32	9.65	12.35	15.12	17.68	3.43	5.32	7.41	9.34	5.49	7.69	10.30	12.73
	ResNet18*	12.10	14.69	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	CNNLSTM	12.08	11.60	12.35	13.89	15.80	15.28	15.60	16.76	23.00	23.22	23.39	23.48	13.95	14.63	15.87	17.21
multi-	PIPE-3B*	**2.98**	6.33	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	PIPE-3B	3.27	6.67	9.86	12.89	7.59	11.58	14.89	17.98	2.54	5.10	7.56	9.87	4.15	6.94	9.74	12.47
	SatSolarCast	3.39	6.36	8.87	11.21	8.28	11.54	14.00	16.60	2.75	5.16	7.30	9.34	4.23	6.96	9.51	12.17
	TCFNet	3.03	**5.42**	**7.54**	**9.52**	7.24	**9.54**	**11.59**	**13.60**	2.42	**4.06**	**5.57**	**7.00**	4.26	**6.39**	**8.28**	**10.08**

GPU server with an Intel(R) Xeon(R) Platinum 8370C CPU and a RTX A800 GPU. More detailed settings of TCFNet and baselines are provided in the Supplementary Information.

Overall Comparison

Table 1 shows the quantitative intensity forecasting results. First, for the domain-specific approaches, Persistence achieves a lower pressure RMSE than TIFS at T+12, but higher RMSEs at T+18 and T+24 in the WNP basin. Among the track data-based baselines, LSTM and PatchTST show competitive performance and largely outperform Persistence, while achieving lower pressure RMSEs than the reported TIFS results. These results highlight the potential of AI-based approaches. In contrast, the vision-based methods ResNet18 and CNNLSTM deliver unsatisfactory predictions. This is reasonable since the satellite observations provide spatial evolution patterns of TCs, which are important for long-term forecasting, while lacking the numerical calibration of best track data, which is relevant for the short-term accuracy. Second, regarding the multimodal baselines, we specifically examine PIPE-3B, a recent LLM-based baseline. Both the original and reproduced results generally underperform the proposed TCFNet. Despite the theoretical potential of LLMs, the observations show that without special adaptation or domain knowledge incorporation, LLMs may be insufficient to deliver promising results in modeling complicated TC intensity evolution. As for SatSolarCast, it achieves more accurate forecasting results than PIPE-3B in most settings, but does not consistently outperform time-series models. This suggests a limitation in its ability to effectively exploit satellite imagery for improving TC intensity forecasting. Compared with the domain-specific, unimodal, and multimodal baselines, the proposed TCFNet delivers the best performance in most evaluations, especially at longer forecast lead times. The observations highlight the superiority of TCFNet, particularly in effectively extracting TC intensity-related patterns from satellite imagery to improve forecasting performance.

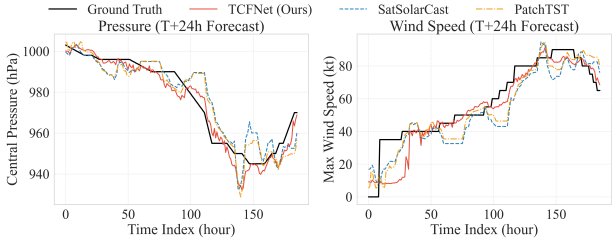


Figure 3: Intensity forecasts at T+24 for TC 202201 (MALAKAS) in the WNP basin. The curves are generated using a sliding window with a 1-hour stride to cover the entire lifecycle.

To visually compare the forecasting abilities, we present whole TC intensity prediction samples in Figure 3. For a fair and representative comparison, we select the typical unimodal model PatchTST and the multimodal model SatSolarCast as baselines. Intuitively, the two methods deliver competitive performance, while the visual comparison shows their limitations. Specifically, they fail to accurately predict abrupt intensity change trends (going up and going down), particularly in the death phase of the typhoon (beginning from the 175th hour). There is an obvious delay compared to the ground truth, yet it still struggles to forecast accurate values. By contrast, our method better captures the abrupt intensity changes. More samples are provided in the Supple-

Table 2: Rapid intensification (RI) results on the WNP basin.

Method	CSI ↑ (%)	POD ↑ (%)	FAR ↓ (%)
Transformer	3.66	3.67	**5.83**
LSTM	35.00	38.73	21.62
PatchTST	27.34	28.83	15.91
Mamba	27.77	29.42	16.80
TimeMixer	21.56	22.21	11.97
TEFN	0.61	0.65	90.52
PIPE-3B	38.81	45.39	27.19
SatSolarCast	31.38	34.71	23.42
TCFNet (ours)	**45.26**	**48.57**	13.07

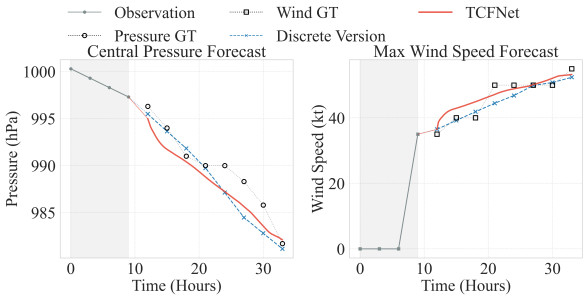


Figure 4: Comparison of TCFNet and its discrete version. The red curve denotes continuous results of TCFNet.

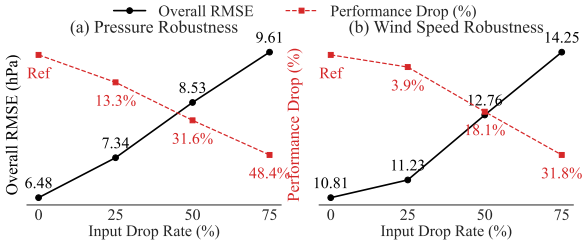


Figure 5: Performance curves w.r.t. drop rates of input.

mentary Information.

To quantitatively evaluate the ability of TCFNet to capture rapid intensity changes, we conduct an evaluation of rapid intensification (RI), defined as a maximum sustained wind-speed increase of at least 30 kt within 24 h. As shown in Table 2, TCFNet achieves the highest critical success index (CSI) of 45.26% among all evaluated methods, which further demonstrates the advantages of TCFNet in identifying rapid intensity changes.

Ablation Studies

To verify the effectiveness of key elements in TCFNet, we conduct ablation experiments over three random seeds, including the main experiment seed and two additional seeds, as summarized in Table 3. In addition, we evaluate the continuous-time modeling capability of TCFNet under missing observations and compare its inference time with the discrete version.

The impact of visual stream. Without the environmental context provided by satellite imagery, the TC intensity forecasting accuracy has declined significantly, which is more pronounced as the lead time increases. Specifically, compared to the full model, the pressure and wind speed RMSE scores of the variant without the visual stream increase by 21.71% and 13.07% at the T+24 time step, respectively.

The effectiveness of visual-conditioned fusion. The variant *w/ Sum Fusion* denotes that we directly add features from the visual stream and the physical state stream, while the variant *w/ Concat Fusion* represents that we fuse the features through channel concatenation operations. Compared to the two classic feature fusion approaches, our integrated mechanism achieves lower mean RMSE across all evaluated lead times. This indicates that employing visual features to modulate physical state representations can better model the correlation between the two types of features and deliver better prediction results.

The benefit of PhyMoCo pretraining. *w/o Trend Pair* means that the visual encoder is pretrained without the trend-constrained positive pair sampling strategy. As a result, there is a 5.27% increase in pressure RMSE and 3.07% wind speed RMSE at the T+24 time step, which shows the importance of trend-consistent sampling strategy during MoCo training procedure. The *w/ ImageNet* setting means we employ the backbone ResNet34 pretrained from ImageNet. Obviously, there is a more significant performance decline. This observation is reasonable since the spatial features of brightness temperature images are quite different from those of natural images. More importantly, such a pretrained model does not involve the temporal dynamics, such as intensity change trends. The observations demonstrate necessity of PhyMoCo.

Transfer learning experiments. To verify the transferability of the pretrained encoder across basins, we conduct a transfer experiment from the SP basin to the WNP basin. As shown in Table 3, although performance degrades compared to the full model, it still overall outperforms the variant *w/o Visual Stream*, indicating that the pretrained encoder captures transferable spatial patterns of TCs from satellite imagery.

The advantage of continuous dynamics modeling. Figure 4 compares the prediction curves of TCFNet and its discrete variant, which replaces the ODE module with a standard GRU. Both models achieve comparable accuracy at the evaluated forecast lead times. However, TCFNet supports continuous-time forecasting, enabling predictions at arbitrary timestamps beyond the predefined temporal grid.

Performance under missing observations. We randomly remove input satellite observations and track data with ratios of 25%, 50%, and 75%, respectively. The previous methods are not directly applicable to this setting, since they are designed to receive fixed input lengths. Figure 5 presents the results in the WNP basin with respect to varying input drop rates. The results show that TCFNet can still operate with incomplete observations, although its performance gradually decreases as the missing ratio increases. For wind-speed forecasting, the performance drops by about 18.1% and 31.8% under 50% and 75% missing observations, respectively, indicating that missing inputs remain a challenging factor.

Inference time comparison. As shown in Table 4, our

Table 3: Ablation study of TCFNet on the WNP basin over three random seeds (mean ± std).

Model Variant	**Pressure RMSE (hPa)**				**Wind RMSE (kt)**			
	T+6	T+12	T+18	T+24	T+6	T+12	T+18	T+24
TCFNet (Full)	3.13±0.11	**5.48±0.05**	**7.55±0.01**	**9.49±0.04**	**6.83±0.35**	**9.50±0.13**	**11.65±0.11**	**13.70±0.09**
w/o Visual Stream	3.43±0.04	6.24±0.06	8.98±0.05	11.55±0.05	7.32±0.25	10.34±0.22	12.89±0.17	15.49±0.16
w/ Sum Fusion	3.26±0.15	5.56±0.13	7.69±0.18	9.76±0.25	7.22±0.14	9.88±0.44	12.08±0.48	14.16±0.31
w/ Concat Fusion	3.27±0.10	5.54±0.09	7.64±0.02	9.58±0.16	7.42±0.52	10.04±0.21	12.23±0.20	14.31±0.36
w/o Trend Pair	**3.10±0.08**	5.62±0.04	7.90±0.04	9.99±0.07	7.05±0.51	9.71±0.16	11.89±0.16	14.12±0.23
w/ ImageNet	3.39±0.11	5.87±0.04	8.21±0.01	10.42±0.04	7.29±0.07	10.15±0.13	12.61±0.17	14.95±0.18
SP → WNP	3.22±0.13	5.94±0.28	8.43±0.48	10.75±0.71	7.00±0.43	9.90±0.36	12.40±0.33	14.82±0.50

Table 4: Inference time and GPU memory usage comparison.

Method	Inference Time	GPU Memory
TCFNet	0.2302s	1034MB
TCFNet (Discrete)	0.0093s	1352MB

model requires only 0.23 s to generate 24-hour predictions per sample, with about 1 GB of GPU memory. Compared with its discrete variant, the inference time increases by 0.22 s while requiring less GPU memory. These results demonstrate the practicality of our method for deployment.

Conclusion

We propose a physics-guided multimodal TC intensity forecasting framework TCFNet, which integrates satellite observations and track data. Comprehensive experiments across two active basins demonstrate the advantages of TCFNet and validate its key elements, particularly in forecasting intensity changes. Notably, it can deal with missing observations and support flexible temporal forecasting, which increases its adaptability in practical forecasting scenarios. Moreover, the results highlight the importance of satellite observations, physics-guided visual representation learning, and visual-conditioned fusion for TC intensity forecasting.

Acknowledgments

This work was supported in part by the Hong Kong Jockey Club Charities Trust (FA123 and P0413), the Innovation and Technology Commission project ITP/047/23LP (P0456) managed by the Hong Kong Logistics and Supply Chain MultiTech R&D Centre, and the State Key Laboratory scheme under Innovation and Technology Commission (ITC-SKLCRCC26EG01).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970): 533–538.

Cangialosi, J. P.; Blake, E.; DeMaria, M.; Penny, A.; Latto, A.; Rappaport, E.; and Tallapragada, V. 2020. Recent progress in tropical cyclone intensity forecasting at the National Hurricane Center. *Weather and Forecasting*, 35(5): 1913–1922.

Chen, B.; Chen, B.-F.; and Lin, H.-T. 2018. Rotation-blended CNNs on a new open dataset for tropical cyclone image-to-intensity regression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 90–99.

Chen, H.; Yin, P.; Huang, H.; Wu, Q.; Liu, R.; and Zhu, X. 2023. Typhoon intensity prediction with vision transformer. *arXiv preprint arXiv:2311.16450*.

Chen, L.; and Meng, Z. 2001. An overview on tropical cyclone research progress in China during the past ten years. *CHINESE JOURNAL OF ATMOSPHERIC SCIENCES-CHINESE EDITION*, 25(3): 432–432.

Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1724–1734. ACL.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Dai, K.; Su, H.; Zhai, C.; Lin, H.; and Bai, M. 2026. SatSolarCast: A Flexible Framework for Multimodal Solar Irradiance Forecasting via Memory-Alignment Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45): 38331–38339.

DeMaria, M. 2009. A simplified dynamical system for tropical cyclone intensity prediction. *Monthly Weather Review*, 137(1): 68–82.

DeMaria, M.; and Kaplan, J. 1994. A statistical hurricane intensity prediction scheme (SHIPS) for the Atlantic basin. *Weather and Forecasting*, 9(2): 209–220.

DeMaria, M.; Sampson, C. R.; Knaff, J. A.; and Musgrave, K. D. 2014. Is tropical cyclone intensity guidance improving? *Bulletin of the American Meteorological Society*, 95(3): 387–398.

DOCUMENTATION-Cy40r1, I. 2020. PART V: ENSEMBLE PREDICTION SYSTEM.

Donahue, J.; Anne Hendricks, L.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Saenko, K.; and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2625–2634.

Dvorak, V. F. 1975. Tropical cyclone intensity analysis and forecasting from satellite imagery. *Monthly Weather Review*, 103(5): 420–430.

Emanuel, K. 2018. 100 years of progress in tropical cyclone research. *Meteorological Monographs*, 59: 15–1.

Emanuel, K.; and Zhang, F. 2016. On the predictability and error sources of tropical cyclone intensity forecasts. *Journal of the Atmospheric Sciences*, 73(9): 3739–3747.

Gu, A.; and Dao, T. 2024. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First conference on language modeling*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Hersbach, H.; Bell, B.; Berrisford, P.; Hirahara, S.; Horányi, A.; Muñoz-Sabater, J.; Nicolas, J.; Peubey, C.; Radu, R.; Schepers, D.; et al. 2020. The ERA5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730): 1999–2049.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jiang, W.; Zhang, D.; Hu, G.; Wu, T.; Liu, L.; Xiao, Y.; and Duan, Z. 2023. Transformer-based tropical cyclone track and intensity forecasting. *Journal of Wind Engineering and Industrial Aerodynamics*, 238: 105440.

Kitamoto, A.; Dzik, E.; and Faure, G. 2024. Machine Learning for the Digital Typhoon Dataset: Extensions to Multiple Basins and New Developments in Representations and Tasks. *arXiv preprint arXiv:2411.16421*.

Kitamoto, A.; Hwang, J.; Vuillod, B.; Gautier, L.; Tian, Y.; and Clanuwat, T. 2023. Digital typhoon: Long-term satellite image dataset for the spatio-temporal modeling of tropical cyclones. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 40623–40636.

Knapp, K. R.; Kruk, M. C.; Levinson, D. H.; Diamond, H. J.; and Neumann, C. J. 2010. The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3): 363–376.

Li, H.; Jung, E.; Wang, Z.; WANG, Y.; Qu, H.; Lau, A. K. H.; et al. 2025. PIPE: Physics-Informed Position Encoding for Alignment of Satellite Images and Time Series in Typhoon Forecasting. In *Proceedings of the Annual Conference on Neural Information Processing Systems*.

Ma, Z.; Yan, Y.; Lin, J.; and Ma, D. 2024. A multiscale and multilayer feature extraction network with dual attention for tropical cyclone intensity estimation. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–15.

Na, Y.; Na, B.; and Son, S. 2022. Near real-time predictions of tropical cyclone trajectory and intensity in the northwestern Pacific Ocean using echo state network. *Climate Dynamics*, 58(3): 651–667.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In *Proceedings of the International Conference on Learning Representations*.

Olander, T. L.; Velden, C.; and Kossin, J. 2004. The advanced objective dvorak technique (AODT)–latest upgrades and future directions. In *Proceedings of the Conference on Hurricanes and Tropical Meteorology*, 294–295.

Olander, T. L.; and Velden, C. S. 2007. 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.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pan, B.; Xu, X.; and Shi, Z. 2019. Tropical cyclone intensity prediction based on recurrent neural networks. *Electronics Letters*, 55(7): 413–415.

Perez, E.; Strub, F.; de Vries, H.; Dumoulin, V.; and Courville, A. C. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 3942–3951.

Pradhan, R.; Aygun, R. S.; Maskey, M.; Ramachandran, R.; and Cecil, D. J. 2017. Tropical cyclone intensity estimation using a deep convolutional neural network. *IEEE Transactions on Image Processing*, 27(2): 692–702.

Price, I.; Sanchez-Gonzalez, A.; Alet, F.; Andersson, T. R.; El-Kadi, A.; Masters, D.; Ewalds, T.; Stott, J.; Mohamed, S.; Battaglia, P.; et al. 2025. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90.

Shi, Y.; Hu, R.; Wu, N.; Zhang, H.; Liu, X.; Zeng, Z.; Zhu, J.; Han, P.; Luo, C.; Zhang, H.; He, J.; and Shi, X. 2025. Comparison of AI and NWP Models in Operational Severe Weather Forecasting: A Study on Tropical Cyclone Predictions. *Journal of Geophysical Research: Machine Learning and Computation*, 2(2): e2024JH000481.

Su, H.; Wu, L.; Jiang, J. H.; Pai, R.; Liu, A.; Zhai, A. J.; Tavallali, P.; and DeMaria, M. 2020. Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophysical Research Letters*, 47(17): e2020GL089102.

Tian, W.; Lai, L.; Niu, X.; Zhou, X.; Zhang, Y.; and Sian, K. T. C. L. K. 2023. Estimation of tropical cyclone intensity using multi-platform remote sensing and deep learning with environmental field information. *Remote Sensing*, 15(8): 2085.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Proceedings of the Advances in Neural Information Processing Systems*, 5998–6008.

Velden, C. S.; Olander, T. L.; and Zehr, R. M. 1998. Development of an objective scheme to estimate tropical cyclone intensity from digital geostationary satellite infrared imagery. *Weather and Forecasting*, 13(1): 172–186.

Wang, C.; Zheng, G.; Li, X.; Xu, Q.; Liu, B.; and Zhang, J. 2021. Tropical cyclone intensity estimation from geostationary satellite imagery using deep convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and Zhou, J. 2024. TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting. In *Proceedings of the International Conference on Learning Representations*.

Woodruff, J. D.; Irish, J. L.; and Camargo, S. J. 2013. Coastal flooding by tropical cyclones and sea-level rise. *Nature*, 504(7478): 44–52.

Xu, H.; Zhao, Y.; Dajun, Z.; Duan, Y.; and Xu, X. 2025. Exploring the typhoon intensity forecasting through integrating AI weather forecasting with regional numerical weather model. *npj Climate and Atmospheric Science*, 8(1): 38.

Yang, J.; Li, L.; Zhao, K.; Wang, P.; Wang, D.; Sou, I. M.; Yang, Z.; Hu, J.; Tang, X.; Mok, K. M.; et al. 2019. A comparative study of Typhoon Hato (2017) and Typhoon Mangkhut (2018)—Their impacts on coastal inundation in Macau. *Journal of Geophysical Research: Oceans*, 124(12): 9590–9619.

Zhan, T.; He, Y.; Deng, Y.; Li, Z.; Du, W.; and Wen, Q. 2025. Time evidence fusion network: Multi-source view in long-term time series forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Appendix

Preliminary

Here, we briefly depict the working flow of the momentum contrastive learning (MoCo) technique (He et al. 2020). First, given a positive sample pair (x_q, x_k), the x_k is an augmented version of the sample x_q. By encoding the sample pair with a query and key encoders, we obtain a query representation q and a positive key k^+, as follows:

$$q = f_q(x_q; \theta_q), k^+ = f_k(x_k; \theta_k). \tag{12}$$

The θ_q and θ_k are the parameters of the query and key encoders, respectively. Second, MoCo leverages a memory queue to store negative keys $\{k_i^-\}_{i=1}^K$. To keep the distribution consistency between the new keys and old keys and maintain the training stability, MoCo adopts a momentum update strategy for the key encoder parameters as follows:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \tag{13}$$

$m \in [0, 1)$ is the momentum coefficient. Third, the training loss function for MoCo, also called InfoNCE loss (Oord, Li, and Vinyals 2018), is formulated as:

$$\mathcal{L}_{\text{MoCo}} = -\log \frac{\exp\left(\frac{\text{sim}(q,k^+)}{\tau}\right)}{\exp\left(\frac{\text{sim}(q,k^+)}{\tau}\right) + \sum_{i=1}^{K} \exp\left(\frac{\text{sim}(q,k_i^-)}{\tau}\right)}, \tag{14}$$

where $\text{sim}(\cdot,\cdot)$ denotes a similarity function, τ is a temperature parameter. By minimizing $\mathcal{L}_{\text{MoCo}}$, the query representation is pulled closer to its positive key in the latent space, while being pushed away from the negative keys.

Implementation Details

Tables 5 report the basic statistics of the processed data. Although Equation 2 defines a symmetric temporal consistency window for conceptual completeness, the implemented sampling strategy is restricted to forward directions to preserve causal temporal ordering in tropical cyclone evolution.

Table 5: The statistics of TC datasets. The input/output steps are set to $4 \rightarrow 8$ with a 3-hour interval.

Region	Years	Statistics		Data Splits (Samples)		
		TCs	Images	Train	Val	Test
WNP	1978–2023	1,116	192,956	88,776	18,500	18,648
SP	1979–2024	480	70,087	33,169	9,223	8,452

Table 6 presents the detailed architecture and parameter settings of TCFNet. All baselines are implemented using their official open-source codes with recommended hyperparameter settings, whose key configurations are summarized as follows:

- The parameters of time series baselines including LSTM, Transformer, PatchTST, Mamba, TimeMixer, and TEFN are presented in Table 7. The maximum number of training epochs is set to 200, with an early stopping patience of 10.

Table 6: Detailed Architecture and Parameters of the proposed TCFNet Model. B: Batch size; T: Input sequence length; K: Prediction horizon; D_{in}: Input feature dimension; D_{model}: Hidden dimension; D_{out}: Target variable dimension.

Module	Component / Operation	Output Dimension
1. Asynchronous History Encoding		
Time Encoding (Intervals)	Input: Time Intervals Δt	$B \times T \times 1$
	Linear($1 \to 32$) + ReLU	$B \times T \times 32$
	Linear($32 \to D_{in}$)	$B \times T \times D_{in}$
Feature Augmentation	Concatenation: $[x_t, \text{TimeEnc}(\Delta t)]$	$B \times T \times 2D_{in}$
Recurrent Layer	GRU($2D_{in} \to D_{model}$)	$B \times T \times D_{model}$
	Output: Hidden States, Last Hidden	$B \times 1 \times D_{model}$
2. Continuous-Time State Propagation		
Initial State	Input: Last Hidden State h_T	$B \times D_{in}$
	Projection: Linear($D_{in} \to D_{model}$)	$B \times D_{model}$
Dynamics Time Encoding	Input: Current Time τ during integration	1×1
	MLP: Linear($1 \to 32$) + ReLU + Linear($32 \to D_{model}/2$)	$B \times D_{model}/2$
ODE Function $f_\theta(h, t)$	State Proj: Linear($D_{model} \to D_{model}/2$)	$B \times D_{model}/2$
	Concat: $[h_{proj}, \tau_{enc}]$	$B \times D_{model}$
	MLP: Linear($D_{model} \to \frac{D_{model}}{2}$) + Tanh + Linear($\to D_{model}$)	$B \times D_{model}$
Integration	ODESolve($f_\theta, h_0, t_0 \to t_K$) (DOPRI5 Solver)	$B \times K \times D_{model}$
Output Projection	Linear($D_{model} \to D_{model}$)	$B \times K \times D_{model}$
3. Visual-Conditioned Asymmetric Fusion		
Conditioning	Gamma Gen: Linear($D_{model} \to D_{model}$) + Tanh	$B \times K \times D_{model}$
	Beta Gen: Linear($D_{model} \to D_{model}$)	$B \times K \times D_{model}$
Fusion Operation	$(1 + \gamma(v)) \odot s + \beta(v)$	$B \times K \times D_{model}$
4. Decoder		
Prediction Head	MLP: Linear($D_{model} \to 512 \to 256 \to D_{out}$)	$B \times K \times D_{out}$

- CNNLSTM: We first employ a ResNet34 as a spatial encoder to encode the satellite images into features of 512 dimensions, and then adopt the aforementioned LSTM setting for intensity forecasting. Notably, we separately forecast the pressure and wind speed to lower training difficulty.

- PIPE-3B & SatSolarCast: For the two baselines, we use the official implementations and slightly modify the architectures to fit the forecasting task and report the corresponding evaluation metrics.

In particular, we adapted the original SatSolarCast model for univariate typhoon forecasting by separating pressure and wind into independent training tasks. The model retains both variables as input features but reduces the output dimension to 1, predicting only the target variable at each run. The image global feature extractor outputs a single scalar per timestep. The models were trained for 200 epochs with a learning rate of 0.0002 and an early stopping patience of 10.

Specifically, the prompts we used in PIPE are listed below, with a typhoon of No. 198704 in the Western North Pacific provided as an illustrative example. The original PIPE is designed for 12-hour forecasting of latitude, longitude, and central pressure at a 1-hour temporal resolution. Here we adapt it to our task: 24-hour forecasting of central pressure and wind speed at a 3-hour temporal resolution (4 input steps $\to$ 8 output steps). The ground-truth labels shown here are provided for reference and evaluation only and are not used as model inputs.

System Prompt

You are a typhoon forecasting expert, give the 24-hour (8 time steps) typhoon forecasting results based on the historical 4 time steps observations. Notably, the forecasting results should not include the useless content. The prediction format is like this <latitude,longitude,pressure,wind>, <latitude,longitude,pressure,wind>, <latitude,longitude,pressure,wind>, <latitude,longitude,pressure,wind>, <latitude, longitude,pressure,wind>, <latitude,longitude,pressure,wind>, <latitude, longitude,pressure,wind>, <latitude,longitude,pressure,wind>.

Input Data

The input satellite images are: <image> <image> <image> <image>. The historical 3-hourly data from 1987-06-26 09:00:00 to 1987-06-26 18:00:00 is: "latitude": "[11.0, 12.03, 12.38, 12.36]", "longitude": "[142.5, 140.15, 138.79, 138.01]", "pressure": "[1006.0, 1006.0, 1004.0, 1002.0]", "wind": "[0.0, 0.0, 0.0, 0.0]".

Label Data

The GT data is: "latitude": "[12.22, 12.23, 12.41, 12.69, 13.0, 13.3, 13.54, 13.67]", "longitude": "[137.56, 137.49, 137.54, 137.28, 136.77, 136.33, 136.06, 135.91]", "pressure": "[1002.0, 1002.0, 999.0, 996.0, 994.0, 992.0, 988.5, 985.0]", "wind": "[0.0, 0.0, 0.0, 40.0, 40.0, 45.0, 45.0, 50.0]".

Evaluation Metrics

The evaluation metrics used in this work are defined as follows. Forecasting accuracy is evaluated using the root mean squared error (RMSE), defined as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{15}$$

where $\hat{y}_i$ and y_i denote the predicted and observed values, respectively, and N is the number of samples.

Rapid intensification (RI) is defined as a wind-speed increase of at least 30 kt within 24 hours. RI prediction performance is evaluated using the probability of detection (POD), false alarm ratio (FAR), and critical success index (CSI), defined as

$$\text{POD} = \frac{H}{H + M}, \tag{16}$$

$$\text{FAR} = \frac{F}{H + F}, \tag{17}$$

$$\text{CSI} = \frac{H}{H + M + F}, \tag{18}$$

Table 7: The detailed hyperparameter settings of baseline approaches.

Method	Parameter Settings
LSTM	n_steps=4, n_features=2, n_pred_steps=8, n_pred_features=2, layers=2, hidden=128, dropout=0.1, batch_size=32, lr=1e-3.
Transformer	n_steps=4, n_features=2, n_pred_steps=8, n_pred_features=2, n_encoder_layers=4, n_decoder_layers=4, d_model=128, n_heads=4, d_k=32, d_v=32, d_ffn=256, batch_size=32, lr=1e-4.
PatchTST	n_steps=4, n_features=2, n_pred_steps=8, n_pred_features=2, layers=2, d_model=128, n_heads=4, patch length=2, stride=2, batch_size=32, lr=2e-4.
Mamba	n_steps=4, n_features=2, n_pred_steps=8, n_pred_features=2, layers=2, d_model=128, d_state=32, d_conv=8, expansion=4, dropout=0.1, batch_size=32, lr=1e-4.
TimeMixer	n_steps=4, n_features=2, n_pred_steps=8, n_pred_features=2, n_layers=2, top_k=5, d_model=32, d_ffn=32, moving_avg=25, downsampling_window=2, downsampling_layers=1, use_norm=True, dropout=0.1, batch_size=32, lr=1e-4.
TEFN	n_steps=4, n_features=2, n_pred_steps=8, n_pred_features=2, n_fod=2, batch_size=32, lr=1e-4.

where H, M, and F represent hits, misses, and false alarms, respectively. Lower RMSE and FAR values, along with higher POD and CSI values, indicate better performance. Among these metrics, CSI provides a comprehensive assessment of RI prediction performance by accounting for hits, misses, and false alarms.

Visualization Samples

We provide 11 additional TC samples from the WNP basin spanning 2018 to 2023 (12 samples in total, with two per year, including the example in the main text), and four TC samples from the SP basin spanning 2020 to 2023 (one sample per year), to demonstrate the ability of TCFNet to capture rapid intensity changes. Together with the example presented in the main text, these results show that TCFNet consistently tracks abrupt changes in pressure and wind speed more accurately than competing methods across all cases.

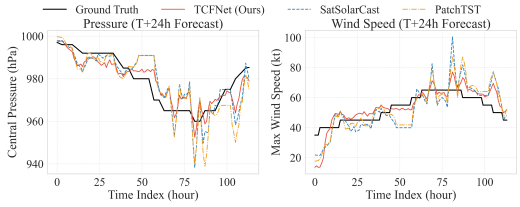


Figure 6: Visual comparison of TC intensity forecasting for a representative sample (TC 201807) from the WNP basin.

AI Usage Disclosure

ChatGPT was used to assist with language polishing and improving the clarity of the manuscript. All AI-assisted revisions were carefully reviewed and verified by the authors.

Discussion

Despite these advantages, our framework has several limitations. In particular, the effectiveness of TCFNet relies on the availability and quality of physical variables, which are not always reliably obtainable in practice. Accurate physical measurements can be difficult to acquire under extreme weather conditions, and are often sparse or missing over ocean regions where in-situ observations are limited. Such uncertainties in physical inputs may affect the robustness of the model and constrain its performance in real-world fore-

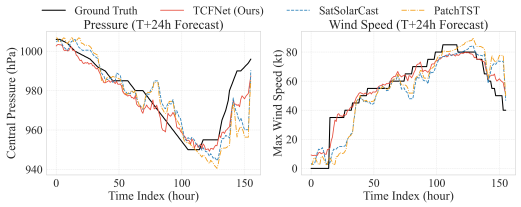


Figure 7: Visual comparison of TC intensity forecasting for a representative sample (TC 201820) from the WNP basin.

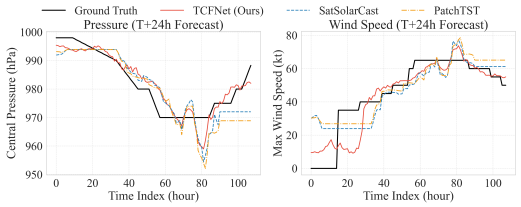


Figure 8: Visual comparison of TC intensity forecasting for a representative sample (TC 201917) from the WNP basin.

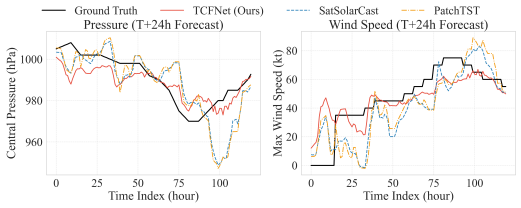


Figure 9: Visual comparison of TC intensity forecasting for a representative sample (TC 201920) from the WNP basin.

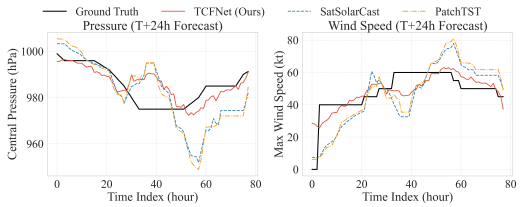


Figure 10: Visual comparison of TC intensity forecasting for a representative sample (TC 202012) from the WNP basin.

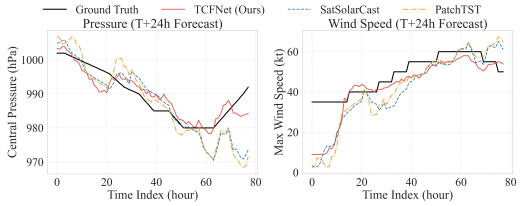


Figure 11: Visual comparison of TC intensity forecasting for a representative sample (TC 202013) from the WNP basin.

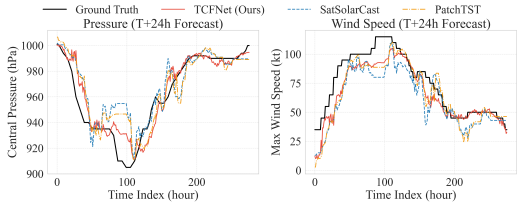


Figure 12: Visual comparison of TC intensity forecasting for a representative sample (TC 202114) from the WNP basin.

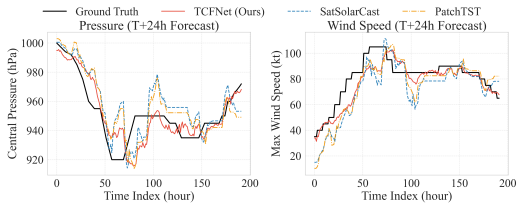


Figure 13: Visual comparison of TC intensity forecasting for a representative sample (TC 202116) from the WNP basin.

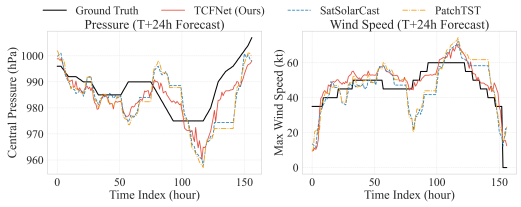


Figure 14: Visual comparison of TC intensity forecasting for a representative sample (TC 202222) from the WNP basin.

casting scenarios. Addressing these challenges by improving the estimation or assimilation of physical variables remains an important direction for future work.

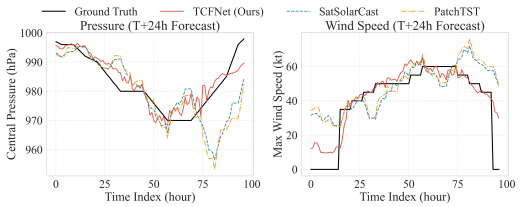


Figure 15: Visual comparison of TC intensity forecasting for a representative sample (TC 202304) from the WNP basin.

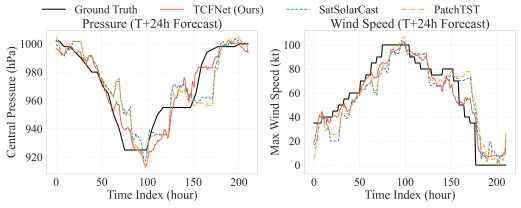


Figure 16: Visual comparison of TC intensity forecasting for a representative sample (TC 202305) from the WNP basin.

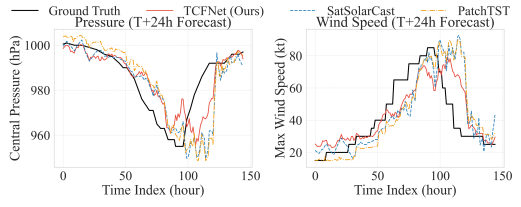


Figure 17: Visual comparison of TC intensity forecasting for a representative sample (TC 202005) from the SP basin.

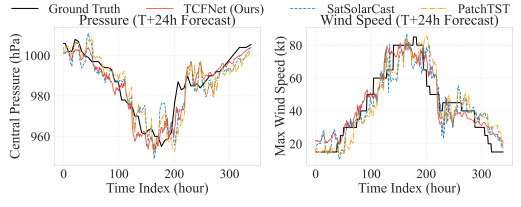


Figure 18: Visual comparison of TC intensity forecasting for a representative sample (TC 202115) from the SP basin.

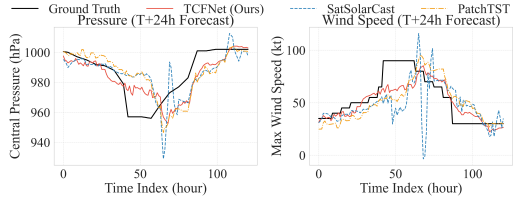


Figure 19: Visual comparison of TC intensity forecasting for a representative sample (TC 202228) from the SP basin.

Broader Impact

This work focuses on TC intensity forecasting, which has direct implications for disaster prevention and mitigation. The proposed TCFNet may improve early warning capa-

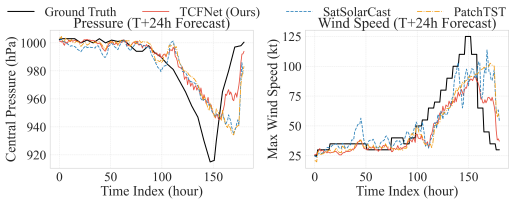


Figure 20: Visual comparison of TC intensity forecasting for a representative sample (TC 202323) from the SP basin.

bility and contribute to disaster preparedness in cyclone-prone regions. While the method may support more informed decision-making, forecasting systems are inherently subject to uncertainty, and prediction errors (e.g., false alarms or missed detections) may affect downstream operational decisions. These considerations motivate the importance of uncertainty-aware modeling and cautious use in practical applications.