CYCLODIFF: MULTI-TASK CONDITIONAL DIFFUSION FRAMEWORK FOR SPATIOTEMPORAL TROPICAL CYCLONE GENESIS FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurately predicting when and where tropical cloud clusters will develop into tropical cyclones (TCs), and what form they will evolve into, is essential for improving typhoon forecasting and enabling earlier identification of tropical cyclones genesis (TCG) and longer warning lead times. Existing TCG studies, particularly deep learning-based methods, mainly rely on classification paradigms that distinguish TC precursor clusters from non-precursors, yet struggle to model the spatiotemporal evolution of cloud systems. To address these limitations, we propose CycloDiff, a multi-task conditional diffusion framework to generate future spatiotemporal evolution of cloud systems along with the corresponding genesis time and location. To generate future cloud-image sequences that better conform to the evolution of the ocean thermal system, we incorporate subsurface temperature profiles, which provide vertical ocean thermal information that cannot be captured by sea surface temperature alone. Experiments on the Digital Typhoon dataset show that CycloDiff generates more realistic and temporally coherent cloud evolution than existing spatiotemporal prediction methods. These results demonstrate the potential of generative spatiotemporal forecasting for representing TC development and predicting its genesis time and location.

1 INTRODUCTION

Accurately forecasting tropical cyclone (TC) genesis remains a central challenge because genesis is a coupled dynamical and thermodynamic transition. During this process, a pre-existing tropical disturbance develops into a self-sustaining warm-core vortex with a closed low-level cyclonic circulation and persistent organized deep convection (Emanuel, 2018).However, genesis forecasting is not merely a binary decision about whether a tropical disturbance will develop. A practically informative forecast must also determine *when* and *where* genesis will occur while representing how the underlying cloud system evolves toward an organized cyclone. We therefore consider tropical cyclone genesis (TCG) forecasting as a spatiotemporal prediction problem in which cloud evolution, genesis time, and genesis location are intrinsically connected.

Existing approaches to TCG forecasting generally emphasize one of three perspectives: environmental favorability, explicit dynamical simulation, or direct prediction of genesis outcomes. Genesis potential indices estimate climatological favorability from large-scale environmental variables (Gray, 1975; 1979; Emanuel & Nolan, 2004; Emanuel, 2010; Tippett et al., 2011), but do not resolve the formation of an individual storm at a specific time and location. Numerical weather prediction models explicitly simulate storm evolution (Kurihara et al., 1998), yet their forecasts remain computationally demanding and sensitive to initial conditions and unresolved physical processes. More recently, deep-learning methods have mapped environmental fields to genesis probabilities or storm coordinates (Mu et al., 2024). Although these outcome-oriented formulations are useful for identifying favorable conditions, they compress the evolving pre-genesis process into labels, probabilities, or isolated positions. Consequently, they provide limited information about how a disorganized convective system develops into a coherent tropical cyclone.

Forecasting this process requires representing both the evolving cloud organization and the environmental conditions that regulate it. Infrared (IR) satellite imagery directly records the spatial

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

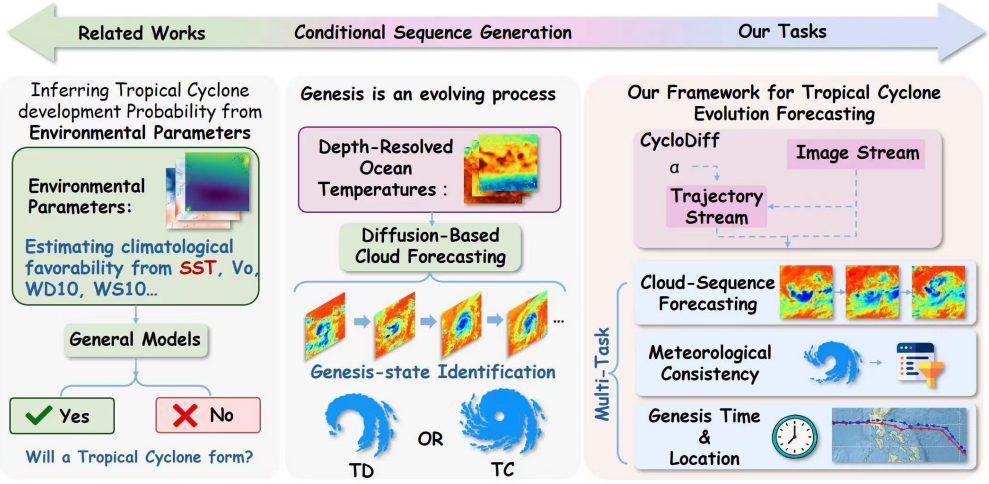


Figure 1: **Motivation and multi-task forecasting formulation of CycloDiff.** *(Left)* Conventional approaches infer environmental favorability or make a static genesis decision without explicitly representing the evolving pre-genesis cloud system. *(Middle)* CycloDiff conditions cloud-sequence forecasting on historical infrared imagery and depth-resolved ocean temperatures to represent the progressive organization of the storm. *(Right)* The forecast cloud sequence and trajectory jointly support storm-evolution assessment, downstream meteorological diagnosis, and prediction of genesis time and location.

organization, motion, and intensity variation of convective cloud systems. Video diffusion models provide a natural probabilistic framework for forecasting such evolution because they model the conditional distribution of future image sequences rather than producing only a single deterministic state (Ho et al., 2020; Voleti et al., 2022; Ren et al., 2025). Atmospheric evolution alone, however, does not fully characterize the pre-genesis environment. Existing data-driven models commonly describe the ocean using sea surface temperature (SST), even though identical surface temperatures may correspond to markedly different subsurface thermal structures. Observational evidence further suggests that upper-ocean variability and early wind-driven ocean responses are relevant during the pre-genesis stage (Gao et al., 2025). The central challenge is therefore to forecast meteorologically meaningful cloud evolution under depth-resolved ocean conditioning and to translate that evolution into explicit estimates of genesis time and location.

To address this challenge, we propose CycloDiff, a multi-task conditional diffusion framework for spatiotemporal TCG forecasting. CycloDiff conditions residual-based video diffusion on historical IR images and temporally aligned depth-resolved ocean-temperature fields to generate future cloud-image sequences. The resulting joint cloud–ocean representations are subsequently transferred to a trajectory branch, which forecasts storm motion from historical coordinates, ocean-temperature profiles, and the encoded environmental features. A pretrained genesis classifier is applied sequentially to the generated frames, and a persistent transition from tropical depression to tropical cyclone determines the predicted genesis time. The corresponding position on the forecast trajectory then determines the genesis location. In this way, CycloDiff connects cloud-sequence forecasting and trajectory prediction within a unified inference procedure that answers both *when* and *where* genesis occurs.

We evaluate CycloDiff through a progressive protocol that moves from quantitative and qualitative assessment of cloud-sequence fidelity to cloud-system structural and downstream diagnostic consistency, and finally to joint inference of genesis time and location. Controlled ablations further isolate the contributions of depth-resolved ocean conditioning and residual prediction.

Our contributions are summarized as follows:

- We revisit TCG forecasting from a spatiotemporal perspective and identify a fundamental limitation of outcome-oriented paradigms: labels, probabilities, and isolated coordinates do

not explicitly represent the evolving pre-genesis cloud system or jointly determine *when* and *where* genesis occurs.

- We propose CycloDiff, a multi-task conditional diffusion framework that conditions residual cloud-sequence forecasting on depth-resolved ocean temperatures. The resulting cloud–ocean representations characterize subsurface thermal conditions beyond SST and provide complementary environmental context for trajectory forecasting.
- We establish a progressive evaluation framework that goes beyond conventional image-quality assessment by examining forecast fidelity, storm-structure consistency, downstream diagnostic consistency, and trajectory accuracy. This evaluation determines whether the generated sequences are not only visually plausible but also informative for TCG analysis.
- Extensive experiments on held-out tropical-cyclone cases demonstrate that CycloDiff outperforms representative deterministic and generative baselines in cloud-sequence forecasting. Ablation studies verify the contributions of depth-resolved ocean conditioning and residual prediction, while a VAMCO case study illustrates the complete inference process for genesis time and location.

2 Related Work

Tropical cyclone genesis and trajectory forecasting. TCG forecasting has traditionally relied on environmental indices constructed from large-scale thermodynamic and dynamical conditions (Gray, 1968; Emanuel & Nolan, 2004; Tippett et al., 2011). More recent data-driven methods formulate genesis as precursor detection or outcome prediction using simulated, satellite-derived, or reanalysis fields (Matsuoka et al., 2018; Kim et al., 2019; Mu et al., 2024). These approaches are effective for estimating genesis likelihood but generally compress the pre-genesis process into probabilities, labels, or isolated coordinates. Meanwhile, data-driven trajectory models combine historical storm positions with satellite observations and environmental variables to predict future motion (Rüttgers et al., 2019; Giffard-Roisin et al., 2020; Huang et al., 2022; 2025). However, genesis and trajectory forecasting are typically treated as separate tasks, leaving the relationship between evolving cloud organization, genesis timing, and storm position underexplored. Subsurface thermal variability provides an additional source of environmental information beyond SST (Gao et al., 2022a; 2025), motivating its inclusion in a unified spatiotemporal formulation.

Spatiotemporal and diffusion-based weather forecasting. Spatiotemporal prediction has progressed from convolutional recurrent models to physics-guided architectures and space–time Transformers (Shi et al., 2015; Wang et al., 2017; Le Guen & Thome, 2020; Gao et al., 2022b). Learned forecasting systems have subsequently demonstrated strong performance in radar nowcasting and large-scale atmospheric prediction (Ravuri et al., 2021; Bi et al., 2023; Lam et al., 2023). Diffusion models further provide a probabilistic framework for representing multiple plausible future sequences. MCVD extends conditional diffusion to video prediction (Voleti et al., 2022), while PreDiff, DiffCast, and CasCast adapt diffusion-based generation to precipitation nowcasting (Gao et al., 2023; Yu et al., 2024; Gong et al., 2024). Related studies have applied cascaded or video-diffusion models to tropical-cyclone fields (Nath et al., 2024; Ren et al., 2025). Nevertheless, these methods primarily forecast images or meteorological fields without coupling the generated storm evolution to trajectory prediction and joint inference of genesis time and location. CycloDiff addresses this gap through ocean-conditioned cloud-sequence forecasting, cross-branch environmental feature transfer, and a unified genesis-inference procedure.

3 Method

3.1 Problem Formulation and Framework

At forecast initialization time t, CycloDiff receives N_c historical cloud images $\mathcal{I}_t^c = \{I_{t-N_c+1}, \ldots, I_t\}$, their aligned depth-resolved ocean-temperature fields $\mathbf{\Theta}_t^c = \{\Theta_{t-N_c+1}, \ldots, \Theta_t\}$, and the historical storm-center trajectory $\mathcal{A}_t^c = \{\mathbf{a}_{t-N_c+1}, \ldots, \mathbf{a}_t\}$, where $\mathbf{a}_\tau = (\lambda_\tau, \varphi_\tau)$. The forecasting task is

$$\left(\widehat{\mathcal{I}}_{t+1:t+N_f}, \widehat{\mathcal{A}}_{t+1:t+N_f}, \widehat{t}_g, \widehat{\mathbf{s}}_g\right) = \mathcal{F}\left(\mathcal{I}_t^c, \mathbf{\Theta}_t^c, \mathcal{A}_t^c\right), \tag{1}$$

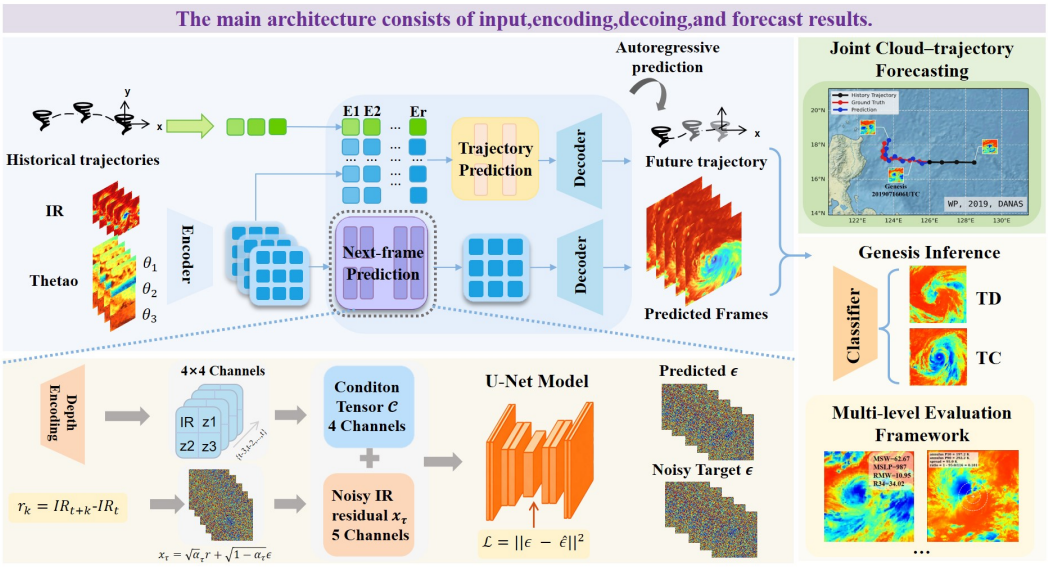


Figure 2: **The diagram of CycloDiff framework.**Historical IR images and depth-resolved ocean temperatures are encoded as conditions for residual-based video diffusion, which autoregressively generates future cloud images. In parallel, historical storm positions and encoded environmental features are used to predict the future trajectory.A pretrained classifier identifies the transition from tropical depression to tropical cyclone in the generated sequence, and the corresponding trajectory position gives the predicted genesis location.

where $\widehat{\mathcal{I}}$, $\widehat{\mathcal{A}}$, $\widehat{t}_g$, and $\widehat{\mathbf{s}}_g$ denote the predicted cloud-image sequence, storm trajectory, genesis time, and genesis location, respectively.

As illustrated in Figure 2, CycloDiff contains an ocean-conditioned residual diffusion model, a cloud–ocean-guided trajectory predictor, and a joint genesis-inference procedure. The cloud branch generates future cloud fields and encodes the historical cloud–ocean state. The resulting representation provides complementary environmental context to the trajectory branch, which retains historical coordinates and colocated ocean-temperature profiles as its primary inputs. The cloud branch is trained first and then frozen during trajectory training, enabling one-way feature transfer without altering the cloud-generation objective.

3.2 Ocean-Conditioned Residual Cloud Forecasting

SST alone cannot distinguish ocean columns with similar surface conditions but different subsurface thermal structures, which may respond differently to storm-induced mixing and upwelling (Gao et al., 2022a; 2025). We therefore condition the cloud forecast on ocean temperatures at 0.5, 34, and 186 m. Each depth channel is interpolated onto the cloud-image grid and assigned a normalized depth code $q_j \in [-1, 1]$. The historical condition is

$$\mathbf{C}_t = \mathrm{Stack}_{\tau=t-N_c+1}^{t}\left[I_\tau, \widetilde{\Theta}_{\tau,1}, \ldots, \widetilde{\Theta}_{\tau,D}\right], \qquad \widetilde{\Theta}_{\tau,j} = \mathcal{E}_d(\Theta_{\tau,j}, q_j), \tag{2}$$

where $\mathcal{E}_d$ denotes depth encoding and $D = 3$.

Rather than reconstructing the complete future field, CycloDiff models changes relative to the latest observed cloud image. For forecast step k, the normalized residual and its noisy diffusion state are

$$\begin{aligned} \widetilde{\mathbf{r}}_k &= \frac{I_{t+k} - I_t - \mu_k}{\sigma_k}, \\ \mathbf{x}_s &= \sqrt{\bar{\alpha}_s}\,\widetilde{\mathbf{R}} + \sqrt{1-\bar{\alpha}_s}\,\boldsymbol{\varepsilon}, \qquad \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}_\phi(\mathbf{x}_s, \mathbf{C}_t, s), \end{aligned} \tag{3}$$

where $\widetilde{\mathbf{R}} = \{\widetilde{\mathbf{r}}_1, \ldots, \widetilde{\mathbf{r}}_{N_f}\}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and (μ_k, σ_k) are estimated from the training set. After reverse diffusion, the prediction is reconstructed as $\widehat{I}_{t+k} = I_t + \sigma_k \widehat{\widetilde{\mathbf{r}}}_k + \mu_k$.

The model directly generates five future frames. Eight-frame forecasts use a sliding-window rollout in which generated cloud images update the image condition, while the ocean-temperature fields observed at initialization remain fixed. This treatment reflects the slower evolution of subsurface thermal structure over the short forecast horizon considered here.

3.3 Cloud–Ocean-Guided Trajectory Forecasting

The trajectory branch is adapted from MMSTN (Huang et al., 2022). It predicts future relative motion from historical storm-center coordinates and colocated ocean-temperature profiles. To incorporate the broader environmental state, we extract a compact representation from the frozen cloud-condition encoder:

$$\mathbf{z}_{\mathrm{co}} = \mathrm{sg}\left[\mathrm{GAP}\left(E_\phi(\mathbf{C}_t)\right)\right], \tag{4}$$

where GAP denotes global average pooling and $\mathrm{sg}[\cdot]$ blocks gradient propagation into the cloud branch.

Separate projections of $\mathbf{z}_{\mathrm{co}}$ are concatenated with the relative-state embeddings of the trajectory encoder and decoder. The decoder autoregressively predicts coordinate displacements, which are accumulated to recover the future storm-center trajectory. Thus, the transferred feature supplements rather than replaces the historical trajectory inputs. Generated future cloud images are not used by the trajectory predictor. We first optimize the cloud branch with the DDPM denoising objective, freeze its parameters, and then train the trajectory branch using relative-displacement and adversarial objectives. Architectural and optimization details are provided in Appendix D.

3.4 Joint Genesis-Time and Genesis-Location Inference

The generated cloud sequence describes the structural evolution used to infer *when* genesis occurs, while the predicted trajectory supplies the corresponding storm position. A pretrained ZPR classifier f_ψ is applied to each generated frame:

$$\begin{aligned} \widehat{y}_k^{\mathrm{ZPR}} &= \mathbb{I}\left[f_\psi(\widehat{I}_{t+k}) \geq \eta\right], \\ \widehat{k}_g &= \min\left\{k : \widehat{y}_j^{\mathrm{ZPR}} = 1,\ \forall j \in \{k, \ldots, k+M-1\}\right\}, \\ \widehat{t}_g &= t + \widehat{k}_g \Delta t, \qquad \widehat{\mathbf{s}}_g = \widehat{\mathbf{a}}_{t+\widehat{k}_g}, \end{aligned} \tag{5}$$

where η is the classification threshold and M suppresses isolated detections by requiring consecutive TC classifications. If no persistent transition occurs within the forecast horizon, no genesis event is reported. The ZPR classifier remains fixed, and the frozen IDOL model is used only for downstream diagnostic evaluation.

4 Experiments

4.1 Experimental Settings

Implementation details. The cloud-generation branch is trained with Adam using a learning rate of 2×10^{-4}, 5,000 warm-up steps, and a batch size of 32. We use 1,000 DDPM noise levels, a linear sigma schedule from 0.02 to 0.0001, and an exponential moving average rate of 0.9999. After training, the cloud-generation branch is fixed and used to encode joint cloud–ocean representations from historical infrared images and aligned depth-resolved ocean temperatures. These representations are projected and injected into both the encoder and decoder of the trajectory branch, together with historical coordinates and colocated ocean-temperature profiles. The trajectory generator and discriminator are trained with Adam using a learning rate of 1×10^{-4} and a batch size of 256. The generator combines a relative-displacement ℓ_2 loss with an adversarial loss, while the discriminator uses binary cross-entropy, with two discriminator updates per generator update. The checkpoint with the lowest validation ADE is retained. For fair comparison, all methods use the same data split, preprocessing, forecast initialization times, and input–output configuration. Five-frame forecasts are generated directly, whereas eight-frame forecasts use a common sliding-window rollout protocol. All models are trained on a single NVIDIA GeForce RTX A6000 GPU.

Baselines. We compare CycloDiff with seven representative spatiotemporal forecasting models. PhyDNet (Le Guen & Thome, 2020) and PredRNN (Wang et al., 2017) represent physics-guided and recurrent predictors, respectively. EarthFormer (Gao et al., 2022b) is a space–time Transformer. PreDiff (Gao et al., 2023), DiffCast (Yu et al., 2024), and CasCast (Gong et al., 2024) represent diffusion-based probabilistic forecasting, residual diffusion, and cascaded refinement. AlphaPre (Lin et al., 2025) models amplitude and phase separately in the frequency domain.

Datasets & Evaluation metrics. For cloud-sequence forecasting, we use cloud-image sequences from the Digital Typhoon dataset together with temporally and spatially aligned ocean-temperature fields from the Copernicus ocean reanalysis. For trajectory forecasting, storm-center positions are obtained from the CMA Tropical Cyclone Best Track Dataset (Ying et al., 2014) and paired with the corresponding cloud observations and colocated Copernicus ocean-temperature profiles.

We evaluate cloud-image forecasts in terms of forecast fidelity, cloud-system structure, and downstream diagnostic consistency. Forecast fidelity is measured using Fréchet Video Distance (FVD) (Unterthiner et al., 2018), Continuous Ranked Probability Score (CRPS) (Matheson & Winkler, 1976), SSIM (Wang et al., 2004), PSNR, and LPIPS (Zhang et al., 2018). DAV and the TBB symmetry ratio assess directional organization and azimuthal symmetry of the predicted cloud fields. Downstream diagnostic consistency is evaluated by applying the same frozen meteorological estimator to generated and ground-truth future frames.

Both forecasting branches are trained using cases from 1993–2018 and evaluated on held-out cases from 2019–2022. Forecasts are initialized at randomly sampled time points, and eight-frame cloud forecasts are produced using a sliding-window rollout. Because the current evaluation is constructed from recorded tropical-cyclone cases, it assesses conditional evolution forecasting for developing systems rather than false-alarm discrimination over non-developing disturbances.

Table 1: **Quantitative comparison of cloud-image sequence forecasting on the Digital Typhoon dataset.**

	Prediction of 8 Frames					Prediction of 5 Frames				
Model	**SSIM↑**	**PSNR↑**	**LPIPS↓**	**CRPS↓**	**FVD↓**	**SSIM↑**	**PSNR↑**	**LPIPS↓**	**CRPS↓**	**FVD↓**
PhyDNet	0.3751	15.461	0.0572	0.1255	1134	0.3729	15.445	0.0583	0.1266	2015
PredRNN	0.3654	15.150	0.0516	0.1587	1251	0.3813	15.261	0.0506	0.1547	2256
EarthFormer	0.4916	14.980	0.0648	0.1369	1329	0.4971	15.092	0.0649	0.1186	2565
PreDiff	0.2608	15.083	0.0259	0.1347	803	0.2576	15.270	0.0261	0.1342	1134
DiffCast	**0.5027**	14.660	0.0334	0.1387	987	0.3856	16.680	0.0633	0.1040	2577
CasCast	0.2909	14.077	0.0737	0.1607	1229	0.3067	16.528	0.0591	0.1310	3192
AlphaPre	0.4356	15.704	0.0515	0.1410	947	**0.5381**	16.221	0.0338	0.0920	1487
Ours	0.2560	**16.559**	**0.0213**	**0.1013**	**132**	0.3060	**19.291**	**0.0185**	**0.0727**	**125**

4.2 Multi-Task Forecasting Evaluation

Cloud-sequence forecasting. We first evaluate whether the generated cloud-image sequences reproduce the observed spatial patterns and temporal evolution. As shown in Table 1, CycloDiff achieves the best PSNR and the lowest LPIPS, CRPS, and FVD under both forecast horizons. Most notably, it reduces FVD from 803 to 132 for eight-frame forecasting and from 1134 to 125 for five-frame forecasting. Although its SSIM is lower than that of some baselines, the substantially improved LPIPS and FVD indicate stronger perceptual and sequence-level fidelity. Figure 3 further shows that CycloDiff maintains coherent cloud-system evolution, whereas the competing forecasts become overly smooth or fragmented at later lead times.

Meteorological consistency. Conventional image-quality metrics do not establish whether generated cloud fields retain the structural information required for tropical-cyclone analysis. We therefore complement them with the TBB symmetry ratio and DAV, which characterize azimuthal symmetry and directional organization, respectively. Figure 4 visualizes representative cloud fields associated with different metric values. Higher TBB symmetry ratios correspond to more azimuthally symmetric cloud structures, whereas lower DAV values indicate more coherent directional organization. These

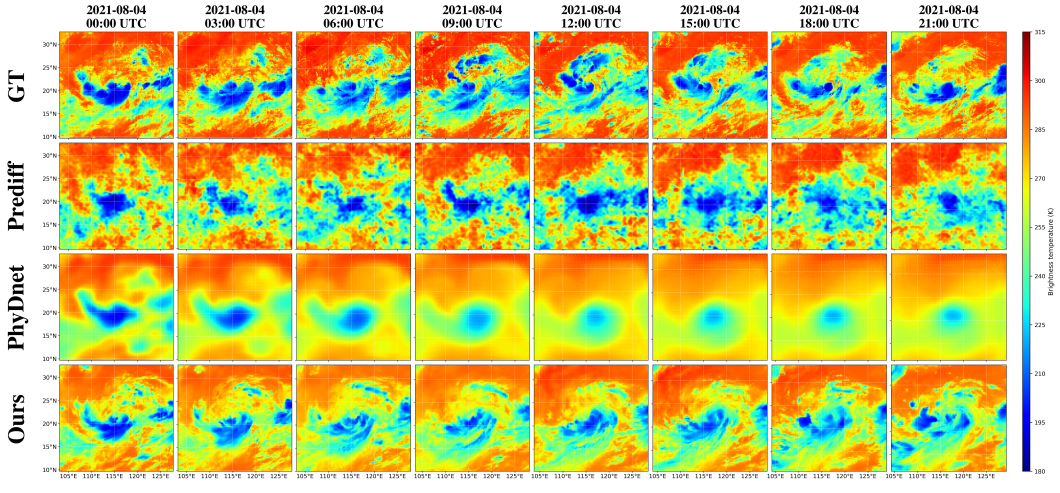
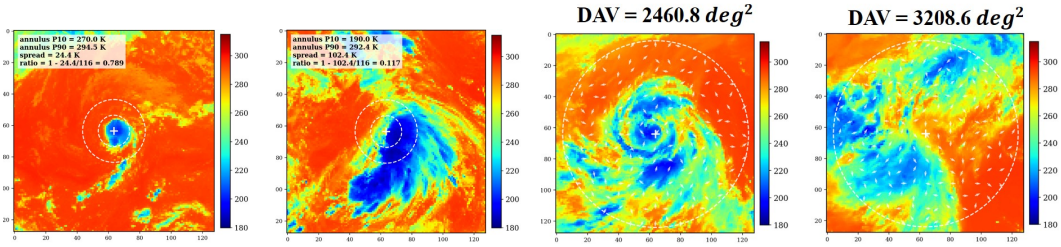


Figure 3: **Qualitative comparison of cloud-sequence forecasts.** Rows show the ground truth and forecasts produced by the compared methods, while columns follow successive forecast times. CycloDiff preserves more coherent cloud-system evolution over the forecast interval.



(a) Examples with low and high TBB symmetry ratios.

(b) Examples with low and high DAV values.

Figure 4: **Visualization of cloud-system structural metrics.** Representative cloud fields with low and high TBB symmetry ratios and DAV values. Higher TBB symmetry ratios indicate stronger azimuthal symmetry, whereas lower DAV values indicate more coherent directional organization.

examples illustrate how the two metrics capture complementary structural properties that are not reflected by conventional image-quality measures.

We further assess downstream diagnostic consistency by applying the same frozen IDOL model (Yan et al., 2025) to generated and ground-truth future frames. Figure 5 shows that MSW estimates obtained from CycloDiff predictions are more concentrated around the one-to-one line than those from the baseline forecasts. Complete results for MSW, MSLP, RMW, and R34 are reported in Appendix C, where CycloDiff obtains the best MAE, RMSE, and correlation for all four variables. These results show that its forecasts better preserve the information used by the fixed diagnostic model to characterize storm intensity and wind-field structure.

Genesis classification consistency. Although the preceding experiments evaluate forecast fidelity and meteorological consistency, they do not directly determine whether the generated sequences preserve the visual information required to recognize tropical-cyclone genesis. We therefore apply the same frozen genesis classifier to the generated and ground-truth cloud-image sequences and evaluate the resulting decisions using the proposed ZPR. As shown in Appendix Table 6, CycloDiff produces ZPR values that are closest to the ground-truth references across all evaluated thresholds under both forecast settings.The advantage remains consistent as the decision threshold increases, whereas the baseline forecasts exhibit larger deviations from the ground-truth ZPR. These results indicate that CycloDiff better preserves the classifier-relevant information associated with the transition from a tropical disturbance to a tropical cyclone, thereby providing quantitative support for the subsequent inference of genesis time.

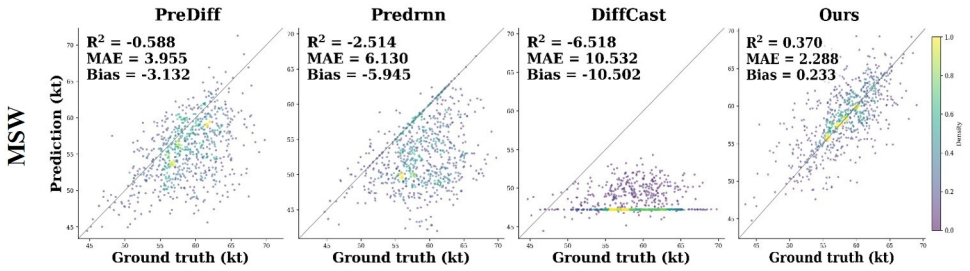


Figure 5: **Downstream diagnostic consistency for MSW.** The same frozen IDOL model is applied to generated and ground-truth future frames. The diagonal indicates one-to-one agreement with the image-derived reference, and colors denote normalized sample density. Results for MSLP, RMW, and R34 are provided in Appendix C.

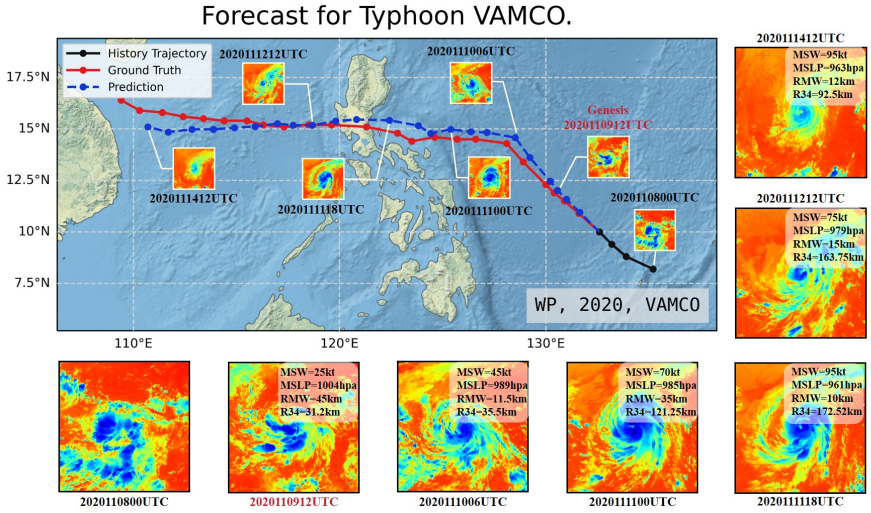


Figure 6: **Joint genesis forecast for Typhoon VAMCO (2020).** Historical, observed future, and predicted trajectories are shown in black, red, and blue, respectively. The detected transition determines the genesis time, and the corresponding predicted trajectory position determines its location.

Joint prediction of genesis time and location. We finally examine whether the two forecast products can jointly determine *when* and *where* genesis occurs. In the illustrative VAMCO (2020) case, the frozen ZPR classifier identifies the persistent TD-to-TC transition from the generated sequence, and the corresponding predicted trajectory position gives the genesis location.

Figure 6 shows that CycloDiff identifies genesis at 12:00 UTC on 9 November 2020, with the corresponding predicted position located near the observed genesis point. This case illustrates how cloud-sequence and trajectory forecasts are integrated into a unified spatiotemporal genesis prediction.

4.3 Ablation Experiments

To identify the sources of CycloDiff's performance gains, we ablate two design choices that affect distinct aspects of the forecasting problem. Ocean-temperature conditioning determines what environmental information is available to the model, whereas residual prediction determines how future cloud evolution is represented. We construct controlled variants for each component while keeping the remaining model configuration unchanged.

Depth-resolved ocean thermal conditioning. We first examine whether the improvement from ocean conditioning can be explained by the inclusion of surface temperature alone, or whether the vertical thermal structure provides additional predictive information. As shown in Table 2 (left), introducing SST produces consistent but modest improvements over the variant without ocean-temperature conditioning: FVD decreases from 197 to 193, while SSIM and PSNR increase from

0.2363 and 15.34 to 0.2433 and 15.96, respectively. In contrast, incorporating both surface and subsurface temperatures reduces FVD further to 132, corresponding to a 31.6% reduction relative to the SST-only setting. The complete model also improves PSNR from 15.96 to 16.559 and simultaneously achieves better SSIM, LPIPS, and CRPS. The substantially larger gain obtained after introducing the subsurface levels indicates that the improvement cannot be attributed solely to the presence of an oceanic input. Instead, the depth-resolved temperature fields provide complementary information about the upper-ocean thermal state that is not represented by SST alone.

Residual formulation of cloud evolution. We next investigate whether predicting temporal changes is more effective than directly reconstructing the absolute future infrared fields. Table 2 (right) shows that replacing direct prediction with residual modeling reduces FVD from 183 to 132 . This pattern is consistent with residual prediction reducing the burden of repeatedly reconstructing persistent background structures and directing the model toward the evolving component of the cloud field. Consequently, residual modeling primarily improves pixel-level reconstruction and sequence-level temporal fidelity while preserving local structural consistency. Together, these ablations isolate two distinct sources of improvement: subsurface temperature strengthens the environmental conditioning, whereas residual prediction provides a more effective representation of temporal cloud evolution.

Table 2: **Ablation studies of ocean-temperature conditioning and residual cloud-evolution modeling.** The left panel compares different ocean-temperature inputs, while the right panel compares direct future-frame prediction with residual prediction. Higher SSIM and PSNR are preferred, whereas lower LPIPS, CRPS, and FVD indicate better performance.

Ocean-Temperature Conditioning						Residual Cloud-Evolution Modeling					
Setting	**SSIM↑**	**PSNR↑**	**LPIPS↓**	**CRPS↓**	**FVD↓**	**Setting**	**SSIM↑**	**PSNR↑**	**LPIPS↓**	**CRPS↓**	**FVD↓**
w/o Ocean Temperature	0.2363	15.34	0.0248	0.1128	197	w/o Residual	0.2534	13.1871	0.0222	0.1111	183
SST Only	0.2433	15.96	0.0218	0.1079	193						
Ours	**0.2560**	**16.559**	**0.0213**	**0.1013**	**132**	**Ours**	**0.2560**	**16.559**	**0.0213**	**0.1013**	**132**

5 Discussion and Conclusion

CycloDiff reformulates TCG forecasting as a multi-task spatiotemporal problem that connects future cloud evolution, storm trajectories, and genesis inference. On the held-out 2019–2022 cases, CycloDiff achieved the best PSNR and the lowest LPIPS, CRPS, and FVD under both forecast horizons, reducing FVD from 803 to 132 for eight-frame forecasting and from 1134 to 125 for five-frame forecasting. Although it did not achieve the highest SSIM, the results indicate a clear advantage in perceptual quality and sequence-level temporal fidelity rather than local structural similarity alone.

The ablations show that depth-resolved ocean conditioning and residual prediction contribute in complementary ways. Adding subsurface temperatures reduced FVD from 193 in the SST-only setting to 132, indicating that the vertical ocean thermal structure provides predictive information beyond SST. Residual prediction further reduced FVD from 183 to 132, consistent with its emphasis on evolving cloud structures rather than persistent backgrounds. The agreement in TBB symmetry ratio and DAV suggests that these improvements extend to storm-scale cloud organization. Results from the frozen IDOL model further demonstrate downstream diagnostic consistency, although they should not be interpreted as independent accuracy against best-track or in situ observations. The VAMCO case study illustrates how generated cloud evolution and the predicted trajectory can jointly determine genesis time and location.

The current evaluation is limited to western North Pacific cases and short forecast horizons. Moreover, the sliding-window rollout keeps the initial ocean-temperature fields fixed, and the joint genesis inference is presently demonstrated through a single case rather than a dataset-level benchmark. Future work should incorporate dynamically predicted ocean states and evaluate genesis time and location across multiple storms and basins. Within these boundaries, CycloDiff provides a unified generative framework for forecasting how a pre-genesis cloud system evolves and for determining *when* and *where* genesis occurs.

AI USE STATEMENT

Large language models (LLMs) were used to assist with manuscript drafting, language editing, structural organization, and the clearer presentation of author-provided methodological and experimental content. They were not used to generate experimental data, execute experiments, or independently determine the scientific claims or conclusions. All AI-assisted text and suggestions were critically reviewed, revised, and verified by the authors. The authors take full responsibility for the methodology, results, claims, and final content of this work.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce our experiments in the main text and appendices. The experimental settings specify the datasets, chronological split, baseline protocol, training configuration, prediction horizons, and evaluation procedure. The appendices further describe data preprocessing and sample construction, model architectures and sequential training, metric definitions, and complete supplementary results.

REFERENCES

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Kerry Emanuel. Tropical cyclone activity downscaled from NOAA–CIRES reanalysis, 1908–1958. *Journal of Advances in Modeling Earth Systems*, 2(1), 2010. doi: 10.3894/JAMES.2010.2.1.

Kerry Emanuel. 100 years of progress in tropical cyclone research. *Meteorological Monographs*, 59: 15.1–15.68, 2018. doi: 10.1175/AMSMONOGRAPHS-D-18-0016.1.

Kerry A. Emanuel and David S. Nolan. Tropical cyclone activity and the global climate system. In *26th Conference on Hurricanes and Tropical Meteorology*, Miami, Florida, 2004. American Meteorological Society. Paper 10A.2.

Cong Gao, Lei Zhou, Chunzai Wang, I.-I. Lin, and Raghu Murtugudde. Unexpected limitation of tropical cyclone genesis by subsurface tropical central-north pacific during El Niño. *Nature Communications*, 13(1):7746, 2022a. doi: 10.1038/s41467-022-35530-9.

Cong Gao, Lei Zhou, I.-I. Lin, Chunzai Wang, Shoude Guan, Fei-Fei Jin, and Raghu Murtugudde. Crucial role of subsurface ocean variability in tropical cyclone genesis. *Nature Communications*, 16:1050, 2025. doi: 10.1038/s41467-025-56433-5.

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space–time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems*, volume 35, pp. 25390–25403, 2022b. doi: 10.52202/068431-1841.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Wang. PreDiff: Precipitation nowcasting with latent diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78621–78656, 2023. doi: 10.52202/075280-3439.

Sophie Giffard-Roisin, Mo Yang, Guillaume Charpiat, Christina Kumler Bonfanti, Balázs Kégl, and Claire Monteleoni. Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data. *Frontiers in Big Data*, 3:1, 2020. doi: 10.3389/fdata.2020.00001.

Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. CasCast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024. doi: 10.48550/arXiv.2402.04290.

William M. Gray. Global view of the origin of tropical disturbances and storms. *Monthly Weather Review*, 96(10):669–700, 1968. doi: 10.1175/1520-0493(1968)096⟨0669:GVOTOO⟩2.0.CO;2.

William M. Gray. Tropical cyclone genesis. Technical Report Atmospheric Science Paper No. 234, Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, 1975.

William M. Gray. Hurricanes: Their formation, structure and likely role in the tropical circulation. In D. B. Shaw (ed.), *Meteorology over the Tropical Oceans*, pp. 155–218. Royal Meteorological Society, Bracknell, United Kingdom, 1979.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Cheng Huang, Cong Bai, Sixian Chan, and Jinglin Zhang. MMSTN: A multi-modal spatial-temporal network for tropical cyclone short-term prediction. *Geophysical Research Letters*, 49(4):e2021GL096898, 2022. doi: 10.1029/2021GL096898.

Cheng Huang, Pan Mu, Jinglin Zhang, Sixian Chan, Shiqi Zhang, Hanting Yan, Shengyong Chen, and Cong Bai. Benchmark dataset and deep learning method for global tropical cyclone forecasting. *Nature Communications*, 16(1):5923, 2025. doi: 10.1038/s41467-025-61087-4.

Minsang Kim, Myung-Sook Park, Jungho Im, Seonyoung Park, and Myong-In Lee. Machine learning approaches for detecting tropical cyclone formation using satellite data. *Remote Sensing*, 11(10): 1195, 2019. doi: 10.3390/rs11101195.

Yoshio Kurihara, Robert E. Tuleya, and Morris A. Bender. The GFDL hurricane prediction system and its performance in the 1995 hurricane season. *Monthly Weather Review*, 126(5):1306–1322, 1998. doi: 10.1175/1520-0493(1998)126⟨1306:TGHPSA⟩2.0.CO;2.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677): 1416–1421, 2023. doi: 10.1126/science.adi2336.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11471–11481, 2020. doi: 10.1109/CVPR42600.2020.01149.

Kenghong Lin, Baoquan Zhang, Demin Yu, Wenzhi Feng, Shidong Chen, Feifan Gao, Xutao Li, and Yunming Ye. AlphaPre: Amplitude–phase disentanglement model for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17841–17850, 2025. doi: 10.1109/CVPR52734.2025.01662.

James E. Matheson and Robert L. Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976. doi: 10.1287/mnsc.22.10.1087.

Daisuke Matsuoka, Masuo Nakano, Daisuke Sugiyama, and Seiichi Uchida. Deep learning approach for detecting tropical cyclones and their precursors in the simulation by a cloud-resolving global nonhydrostatic atmospheric model. *Progress in Earth and Planetary Science*, 5(1):80, 2018. doi: 10.1186/s40645-018-0245-y.

Bin Mu, Xin Wang, Shijin Yuan, Yuxuan Chen, Guansong Wang, Bo Qin, and Guanbo Zhou. A deep learning-based global tropical cyclogenesis prediction model and its interpretability analysis. *Science China Earth Sciences*, 67(12):3671–3695, 2024. doi: 10.1007/s11430-023-1383-6.

Pritthijit Nath, Pancham Shukla, Shuai Wang, and César Quilodrán-Casas. Forecasting tropical cyclones with cascaded diffusion models. In *Tackling Climate Change with Machine Learning Workshop at ICLR*, 2024. doi: 10.48550/arXiv.2310.01690.

Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, Rachel Prudden, Amol Mandhane, Aidan Clark, Andrew Brock, Karen Simonyan, Raia Hadsell, Niall Robinson, Ellen Clancy, Alberto Arribas, and Shakir Mohamed. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021. doi: 10.1038/s41586-021-03854-z.

Zhibo Ren, Pritthijit Nath, and Pancham Shukla. Improving tropical cyclone forecasting with video diffusion models. In *Tackling Climate Change with Machine Learning Workshop at ICLR*, 2025. doi: 10.48550/arXiv.2501.16003.

Mario Rüttgers, Sangseung Lee, Soohwan Jeon, and Donghyun You. Prediction of a typhoon track using a generative adversarial network and satellite images. *Scientific Reports*, 9(1):6057, 2019. doi: 10.1038/s41598-019-42339-y.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems*, volume 28, pp. 802–810, 2015.

Michael K. Tippett, Suzana J. Camargo, and Adam H. Sobel. A poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis. *Journal of Climate*, 24(9): 2335–2357, 2011. doi: 10.1175/2010JCLI3811.1.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018. doi: 10.48550/arXiv.1812.01717.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23371–23385, 2022. doi: 10.52202/068431-1698.

Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 879–888, 2017.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Hanting Yan, Pan Mu, Shiqi Zhang, Yuchao Zhu, Jinglin Zhang, and Cong Bai. IDOL: Meeting diverse distribution shifts with prior physics for tropical cyclone multi-task estimation. In *Advances in Neural Information Processing Systems*, 2025.

Ming Ying, Wei Zhang, Hui Yu, Xiaoqin Lu, Jingxian Feng, Yongxiang Fan, Yongti Zhu, and Dequan Chen. An overview of the China Meteorological Administration tropical cyclone database. *Journal of Atmospheric and Oceanic Technology*, 31(2):287–301, 2014. doi: 10.1175/JTECH-D-12-00119.1.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. DiffCast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27758–27767, 2024. doi: 10.1109/CVPR52733.2024.02622.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

CycloDiff: Multi-Task Conditional Diffusion Framework for Spatiotemporal Tropical Cyclone Genesis Forecasting

(APPENDIX)

TABLE OF CONTENTS IN APPENDIX

A Notations **14**

B Employed Data **14**

B.1 Digital Typhoon Infrared Imagery 15

B.2 Copernicus Ocean Reanalysis 15

B.3 CMA Tropical-Cyclone Best-Track Data 15

B.4 Data Processing and Sample Construction 16

C Additional Downstream Diagnostic Evaluation **16**

D Model Details **16**

D.1 Ocean-Conditioned Residual Diffusion 16

D.2 IDOL Pretraining and Frozen Diagnostic Evaluation 19

D.3 MMSTN Adaptation for Trajectory Forecasting 20

D.4 Sequential Training and Integration 21

E Quantitative Evaluation **21**

F Additional Results **23**

F.1 Quantitative Cloud-System Structure Evaluation 23

F.2 Additional Qualitative Forecasts 24

F.3 Complete ZPR Evaluation 25

A NOTATIONS

Table 3 summarizes the symbols required to describe the forecasting inputs, the two model branches, and the joint genesis-inference procedure. Symbols used only in implementation details are introduced locally in the corresponding appendix sections.

Table 3: **Summary of the principal notation.**

Symbol	Description
Forecast inputs and outputs	
N_c, N_f, D	Numbers of historical frames, forecast frames, and selected ocean-depth levels, respectively
I_τ, Θ_τ	Cloud image and aligned depth-resolved ocean-temperature field at time τ
$\mathbf{a}_\tau = (\lambda_\tau, \varphi_\tau)$	Storm-center longitude and latitude at time τ
$\mathcal{I}_t^c$, $\mathbf{\Theta}_t^c$	Historical cloud-image and depth-resolved ocean-temperature sequences available at forecast initialization time t
$\mathcal{A}_t^c$	Historical storm-center trajectory available at initialization
$\widehat{\mathcal{I}}_{t+1:t+N_f}$	Predicted future cloud-image sequence
$\widehat{\mathcal{A}}_{t+1:t+N_f}$	Predicted future storm-center trajectory
Cloud-sequence forecasting	
q_j, $\widetilde{\Theta}_{\tau,j}$	Normalized code and encoded ocean-temperature field for the j-th depth level
$\mathcal{E}_d$	Depth-encoding operation applied to each ocean-temperature channel
$\mathbf{C}_t$	Multimodal condition formed from historical cloud images and depth-encoded ocean-temperature fields
$\mathbf{r}_k$, $\widetilde{\mathbf{r}}_k$	Cloud-image residual at forecast step k and its normalized form
μ_k, σ_k	Training-set mean and standard deviation used to normalize the residual at forecast step k
$\widetilde{\mathbf{R}}$	Sequence of normalized residual targets over the forecast horizon
$\mathbf{x}_s$, $\bar{\alpha}_s$	Noisy residual sequence and cumulative noise coefficient at diffusion step s
$\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}_\phi$	Sampled Gaussian noise and the conditional denoising network
$\mathcal{L}_{\text{cloud}}$	Denoising objective used to train the cloud-generation branch
Trajectory forecasting and genesis inference	
$\mathbf{u}_\tau$, $\Delta\mathbf{u}_\tau$	Trajectory state at time τ and its relative temporal change
$\mathbf{z}_{\text{co}}$	Cloud–ocean representation extracted from the frozen conditional encoder
$\text{sg}[\cdot]$	Stop-gradient operation used to prevent trajectory losses from updating the cloud-generation branch
f_ψ, η	Frozen genesis classifier and its classification threshold
$\widehat{y}_k^{\text{ZPR}}$	Frame-level ZPR decision for the generated cloud image at forecast step k
M, Δt	Required number of consecutive TC decisions and temporal interval between forecast frames
$\widehat{k}_g$, $\widehat{t}_g$, $\widehat{\mathbf{s}}_g$	Predicted genesis step, genesis time, and genesis location
$\mathcal{L}_{\text{rel}}$, $\mathcal{L}_{\text{adv}}$	Relative-state reconstruction and adversarial objectives used to train the trajectory branch

B EMPLOYED DATA

CycloDiff is trained using three complementary data sources: infrared satellite imagery from the Digital Typhoon dataset, subsurface ocean-temperature fields from the Copernicus ocean reanalysis, and tropical-cyclone locations from the China Meteorological Administration (CMA) best-track dataset. The infrared and ocean data are used to construct the conditional cloud-forecasting samples, whereas the best-track records provide the supervision for the subsequent trajectory-forecasting branch. All data sources are temporally aligned at the sample level and organized according to individual tropical-cyclone cases.

B.1 Digital Typhoon Infrared Imagery

The Digital Typhoon dataset provides geostationary satellite infrared observations of tropical cyclones. Each image records the infrared brightness-temperature distribution around the storm and therefore characterizes the location, organization, and evolution of the associated cloud system. We use the infrared image as the primary atmospheric variable for both model input and prediction target.

Each infrared observation is represented as a single-channel image and resized or cropped to a storm-centered grid of 128×128 pixels. For every forecasting sample, four consecutive infrared frames are used as historical observations, while the following five frames constitute the direct prediction targets. The resulting sequence can be written as

$$\{\mathbf{X}_{t-3}, \mathbf{X}_{t-2}, \mathbf{X}_{t-1}, \mathbf{X}_t\} \longrightarrow \{\mathbf{X}_{t+1}, \ldots, \mathbf{X}_{t+5}\}, \tag{6}$$

where $\mathbf{X}_t \in \mathbb{R}^{1\times128\times128}$ denotes the infrared brightness-temperature field at time t.

B.2 Copernicus Ocean Reanalysis

To incorporate the influence of the upper ocean on tropical-cyclone cloud evolution, we obtain ocean potential temperature (`thetao`) from the Copernicus ocean reanalysis. Ocean-temperature fields at depths of 0.5, 34, and 186 m are retained, representing the near-surface, subsurface, and deeper upper-ocean thermal states, respectively.

For each infrared observation, the ocean field with the corresponding timestamp is selected and spatially interpolated onto the storm-centered infrared grid. The geographic coordinates of the grid are constructed using the tropical-cyclone center recorded in the associated metadata, with a nominal spatial interval of 5 km per pixel. Regular-grid interpolation is then applied separately to each depth level to obtain

$$\mathbf{\Theta}_t = [\Theta_{t,1}, \Theta_{t,2}, \Theta_{t,3}] \in \mathbb{R}^{3\times128\times128}. \tag{7}$$

Grid points outside the valid reanalysis domain or containing missing values are assigned zero in the aligned tensor. Each depth channel is subsequently normalized using channel-wise parameters derived from the training data.

To retain the ordering of the selected ocean levels, the j-th temperature channel is assigned a normalized linear depth code

$$q_j = 2\frac{j-1}{D-1} - 1, \qquad j = 1, \ldots, D, \tag{8}$$

where $D = 3$. The encoded field is written as

$$\widetilde{\Theta}_{t,j} = \mathcal{E}_d\left(\Theta_{t,j}, q_j\right), \tag{9}$$

where $\mathcal{E}_d$ denotes the linear depth-encoding operation used by the input pipeline. The conditional input at each time step is

$$\mathbf{C}_t = \mathrm{Concat}\left(\mathbf{X}_t, \widetilde{\Theta}_{t,1}, \widetilde{\Theta}_{t,2}, \widetilde{\Theta}_{t,3}\right) \in \mathbb{R}^{4\times128\times128}. \tag{10}$$

B.3 CMA Tropical-Cyclone Best-Track Data

Tropical-cyclone positions are obtained from the CMA Tropical Cyclone Best Track Dataset (Ying et al., 2014). Each record provides the observed storm-center location at a given analysis time. The best-track records are matched to the infrared observations according to tropical-cyclone identity and timestamp, producing a temporally consistent sequence of storm-center coordinates.

For the trajectory-forecasting branch, the historical locations associated with the four input frames are used as observed track points, and the locations corresponding to the future infrared frames are used as prediction targets. The aligned Copernicus ocean profiles and the cloud-evolution representation extracted from the pretrained cloud-forecasting model are supplied as environmental conditioning information. Consequently, each trajectory sample contains mutually aligned infrared imagery, three-dimensional ocean-temperature information, and best-track coordinates.

B.4 Data Processing and Sample Construction

The preprocessing pipeline first matches the infrared images, ocean reanalysis fields, and best-track records by tropical-cyclone case and observation time. A storm-centered geographic grid is generated for every timestamp, after which the ocean variables are interpolated onto the infrared-image grid. Samples without the temporal context required to construct four historical frames and five future frames are excluded.

Instead of directly predicting absolute infrared fields, CycloDiff models their changes relative to the latest observed image. The residual target for the k-th forecast step is defined as

$$\mathbf{R}_{t+k} = \mathbf{X}_{t+k} - \mathbf{X}_t, \qquad k = 1, \ldots, 5. \tag{11}$$

For each forecast horizon, the residual fields are standardized using the corresponding mean and standard deviation estimated exclusively from the training set. The predicted infrared image is recovered by adding the generated residual to the latest observed frame.

The dataset is divided chronologically by tropical-cyclone case. Storms occurring from 1993 to 2018 are used for model training, whereas storms from 2019 to 2022 are retained for evaluation. Thus, no samples from the evaluation period are used to estimate normalization parameters or optimize model weights. During training, random spatial flipping and temporal sampling are applied to improve sample diversity. The model directly generates five future frames from four historical frames; longer eight-frame forecasts are obtained through a sliding-window rollout, in which newly generated infrared frames update the observation window while the initialization-time ocean sequence remains fixed.

C Additional Downstream Diagnostic Evaluation

The main text presents MSW as a representative downstream diagnostic. Here, we provide the corresponding density comparisons for MSLP, RMW, and R34, together with the complete quantitative evaluation of all four variables. The pretrained IDOL model (Yan et al., 2025) remains frozen and is applied identically to generated and ground-truth future cloud images. Estimates obtained from the ground-truth images serve as image-derived reference targets.

As shown in Figure 7, diagnostic estimates derived from CycloDiff forecasts are more closely concentrated around the one-to-one line than those obtained from the baseline forecasts. The complete comparison in Table 4 shows that CycloDiff achieves the lowest MAE and RMSE and the highest correlation for all four variables. It also obtains the lowest absolute bias for MSW, MSLP, and RMW and the second-lowest absolute bias for R34.

D Model Details

This section provides the architectural and optimization details of the cloud-generation and trajectory-prediction branches. It also describes the frozen IDOL model used exclusively for downstream diagnostic evaluation.

D.1 Ocean-Conditioned Residual Diffusion

The cloud-generation branch uses a conditional U-Net to denoise five future residual frames from four historical cloud–ocean observations. Each historical observation contains one cloud-image channel and three ocean-temperature channels corresponding to depths of 0.5, 34, and 186 m. The depth channels follow a fixed shallow-to-deep order and are augmented with linear depth encoding before being supplied to the network.

The U-Net uses 96 base feature channels with channel multipliers $(1, 2, 3, 4)$, two residual blocks at each resolution, and attention at spatial resolutions of 32, 16, and 8. We use InstanceNorm++, Swish activations, and a dropout rate of 0.1. The branch predicts the Gaussian noise added to the normalized residual sequence and is optimized using

$$\mathcal{L}_{\text{cloud}} = \mathbb{E}_{\widetilde{\mathbf{R}},s,\boldsymbol{\varepsilon}} \left[\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\phi(\mathbf{x}_s, \mathbf{C}_t, s)\|_2^2 \right]. \tag{12}$$

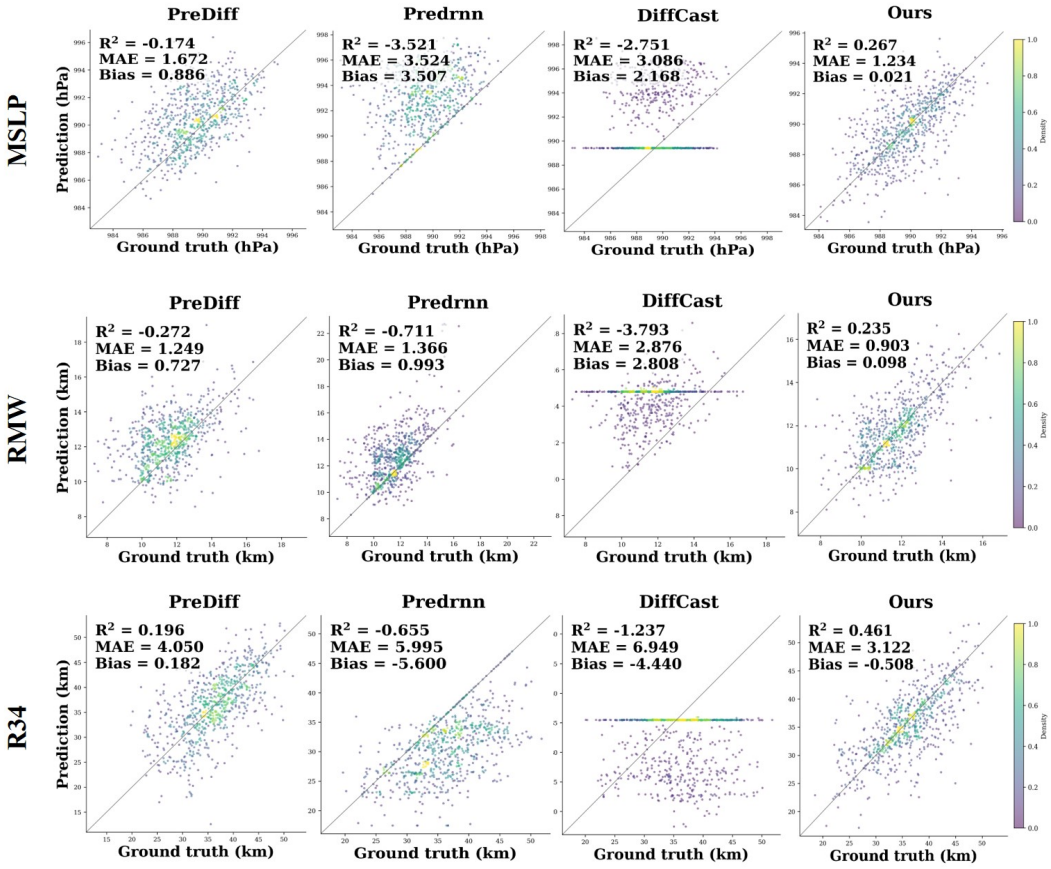


Figure 7: **Additional downstream diagnostic consistency.** The rows show IDOL-derived estimates of minimum sea-level pressure (MSLP), radius of maximum wind (RMW), and radius of 34-knot winds (R34), respectively. Ground-truth images provide the image-derived references, and all forecasts use the same frozen estimator and preprocessing protocol.

Table 4: **Quantitative evaluation of downstream diagnostic consistency.** Results are reported for maximum sustained wind (MSW), minimum sea-level pressure (MSLP), radius of maximum wind (RMW), and radius of 34-knot winds (R34). R denotes the Pearson correlation coefficient between estimates obtained from generated and ground-truth cloud images. The absolute bias is reported, with lower values indicating closer agreement. Lower MAE, RMSE, and absolute bias are preferred, whereas higher R indicates better agreement. The best and second-best results are shown in bold and underlined, respectively, and the shaded rows denote CycloDiff.

Variable	Model	MAE ↓	RMSE ↓	|Bias| ↓	R ↑
RMW	DiffCast	2.88	3.17	2.81	0.197
	CasCast	2.88	3.17	2.81	0.197
	EarthFormer	3.69	4.09	3.65	0.148
	PreDiff	<u>1.25</u>	<u>1.60</u>	<u>0.73</u>	0.456
	PredRNN	1.37	1.75	0.99	<u>0.546</u>
	PhyDNet	3.19	3.62	3.12	0.240
	Ours	**0.90**	**1.15**	**0.10**	**0.639**
R34	DiffCast	6.96	8.40	4.44	-0.001
	CasCast	6.96	8.41	4.45	-0.001
	EarthFormer	6.38	7.58	5.30	0.436
	PreDiff	4.05	<u>5.19</u>	**0.18**	<u>0.633</u>
	PredRNN	5.99	7.16	5.59	0.568
	PhyDNet	<u>5.97</u>	7.17	5.33	0.507
	Ours	**3.13**	**4.06**	<u>0.52</u>	**0.722**
MSW	DiffCast	10.53	11.26	10.50	0.078
	CasCast	10.54	11.27	10.51	0.076
	EarthFormer	8.17	9.26	8.02	0.428
	PreDiff	3.96	4.96	3.13	0.491
	PredRNN	6.13	6.87	5.94	0.316
	PhyDNet	8.31	9.19	8.14	0.210
	Ours	**2.29**	**2.92**	**0.23**	**0.681**
MSLP	DiffCast	3.09	3.52	2.17	0.085
	CasCast	3.09	3.53	2.17	0.085
	EarthFormer	4.02	4.76	3.93	0.396
	PreDiff	1.67	2.11	0.89	0.519
	PredRNN	3.52	3.90	3.51	0.347
	PhyDNet	3.99	4.43	3.95	0.326
	Ours	**1.23**	**1.54**	**0.02**	**0.650**

The residual statistics (μ_k, σ_k) are estimated separately for each forecast step using only the training set and remain fixed during evaluation. The model is trained with Adam using a learning rate of 2×10^{-4}, 5,000 warm-up steps, a batch size of 32, and gradient clipping at 1.0. We use 1,000 diffusion noise levels, a linear sigma schedule from 0.02 to 0.0001, and an exponential moving average rate of 0.9999. Random horizontal flipping is applied during training.

Five future frames are generated directly. For eight-frame forecasting, the generated cloud images update the image component of the conditioning window, whereas the depth-resolved ocean-temperature fields observed at initialization remain fixed throughout the rollout.

D.2 IDOL Pretraining and Frozen Diagnostic Evaluation

Original IDOL architecture. IDOL (Yan et al., 2025) is a physics-informed multi-task estimator for maximum sustained wind (MSW), minimum sea-level pressure (MSLP), radius of maximum wind (RMW), and radius of 34-knot winds (R34). The released implementation processes a pair of four-channel infrared observations using VGG13-BN encoders. A task dependency-flow module incorporates the statistical Holland wind-field model to construct task-specific identity tokens, whereas a correlation-aware information bridge extracts a task-shared identity token from the infrared representation. Four task-specific Transformer heads subsequently regress the normalized meteorological variables.

The original IDOL model is optimized using four equally weighted ℓ_1 regression losses together with an identity-correlation regularizer:

$$\mathcal{L}_{\mathrm{IDOL}} = \sum_{q \in \{\mathrm{MSW},\mathrm{MSLP},\mathrm{RMW},\mathrm{R34}\}} \left\| \widehat{y}_q - y_q \right\|_1 + \lambda_{\mathrm{id}} \mathcal{L}_{\mathrm{id}}, \tag{13}$$

where $\mathcal{L}_{\mathrm{id}}$ aligns the pairwise correlations of the task-shared and task-specific identity representations with those of the corresponding labels. In the released training protocol, brightness temperatures are min–max normalized to $[0, 1]$. Training images are augmented using $90°$, $180°$, and $270°$ rotations together with horizontal and vertical flips. The model is trained for 200 epochs using Adam with an initial learning rate of 10^{-4} and a batch size of 48 on a single NVIDIA RTX A6000 GPU.

Adaptation to CycloDiff outputs. We directly load the released IDOL checkpoint, place the network in evaluation mode, and disable gradient computation. IDOL is therefore not fine-tuned using either generated or ground-truth sequences from our dataset.

CycloDiff predicts a single infrared brightness-temperature channel, whereas the released IDOL encoder expects four channels. We consequently replicate each predicted frame across four channels, min–max normalize it to $[0, 1]$, and bilinearly resize it to 156×156 pixels. IDOL operates on a current–previous image pair. Following its released data loader, we construct pairs separated by three sequence indices. When the predicted sequence does not contain enough preceding frames, the observed history available at forecast initialization is prepended before the pairs are constructed.

The public IDOL interface also requires normalized storm-development and correlation-factor variables that are unavailable for generated future frames. We assign all such unavailable auxiliary variables the neutral midpoint value of 0.5. Importantly, identical preprocessing, temporal pairing, and auxiliary values are applied to the generated and ground-truth sequences. The normalized outputs are then transformed back to physical units using the ranges provided by the released implementation.

For a diagnostic variable q, we compare

$$\widehat{y}^{\mathrm{pred}}_{q,k} = f_{\mathrm{IDOL},q}\left(\widehat{I}_{t+k-3}, \widehat{I}_{t+k}; \mathbf{c}_0\right) \tag{14}$$

with

$$\widehat{y}^{\mathrm{gt}}_{q,k} = f_{\mathrm{IDOL},q}\left(I_{t+k-3}, I_{t+k}; \mathbf{c}_0\right), \tag{15}$$

where $\mathbf{c}_0$ contains the fixed auxiliary values. The resulting MSW, MSLP, RMW, and R34 comparisons measure whether the generated infrared fields retain the information used by a pretrained meteorological estimator. They should therefore be interpreted as *downstream diagnostic consistency*, rather than direct verification against best-track observations. IDOL does not participate in CycloDiff training or genesis-time inference.

D.3 MMSTN Adaptation for Trajectory Forecasting

Trajectory representation. The trajectory branch is adapted from MMSTN (Huang et al., 2022), a stochastic recurrent model for tropical-cyclone trajectory and intensity prediction. We retain its encoder–decoder and adversarial trajectory-learning structure, but replace its original input with a cloud–ocean-aware trajectory state. At historical time step τ, the state is

$$\mathbf{u}_\tau = \left[\lambda_\tau, \varphi_\tau, \theta^{\mathrm{loc}}_{\tau,1}, \theta^{\mathrm{loc}}_{\tau,2}, \theta^{\mathrm{loc}}_{\tau,3}\right], \tag{16}$$

where λ_τ and φ_τ are the storm-center longitude and latitude, respectively. The remaining components are ocean temperatures collocated with the storm center at depths of 0.5, 34, and 186 m. Each component is independently min–max normalized using its training-set range. The recurrent model receives the relative state

$$\Delta\mathbf{u}_\tau = \mathbf{u}_\tau - \mathbf{u}_{\tau-1}, \tag{17}$$

which reduces sensitivity to the absolute geographic position and emphasizes short-term storm motion.

Recurrent generator. The generator receives four historical states and predicts five future relative states. Each five-dimensional relative state is projected to a 32-dimensional embedding. The encoder is a single-layer LSTM with 64 hidden units. To preserve the stochastic formulation of MMSTN, a 16-dimensional Gaussian noise vector is concatenated with the encoded trajectory context before decoding.

The decoder is a single-layer LSTM with 64 hidden units. It autoregressively predicts

$$\Delta\widehat{\mathbf{u}}_{t+k}, \qquad k = 1, \ldots, N_f. \tag{18}$$

After each prediction, the new relative state is appended to the trajectory history, and the expanded history is re-encoded before the next decoding step. The first two components are accumulated from the last observed storm center:

$$\widehat{\mathbf{a}}_{t+k} = \widehat{\mathbf{a}}_{t+k-1} + \left[\Delta\widehat{\mathbf{u}}_{t+k}\right]_{1:2}. \tag{19}$$

The remaining components maintain the coupled ocean-temperature state during autoregressive decoding.

Injection of CycloDiff representations. The original MMSTN does not receive spatial cloud information. We therefore inject the 384-dimensional cloud–ocean representation $\mathbf{z}_{\mathrm{co}}$ extracted from the bottleneck of the trained conditional U-Net. Separate projection networks are used by the trajectory encoder and decoder:

$$P_{\mathrm{enc}}, P_{\mathrm{dec}} : \mathbb{R}^{384} \to \mathbb{R}^{32} \to \mathbb{R}^{32}. \tag{20}$$

Each projection consists of a linear layer, a ReLU activation, and a second linear layer.

For the encoder, $P_{\mathrm{enc}}(\mathbf{z}_{\mathrm{co}})$ is broadcast over the four historical steps and concatenated with every relative-state embedding:

$$\mathbf{e}_\tau = \mathrm{Emb}\left(\Delta\mathbf{u}_\tau\right) \,\|\, P_{\mathrm{enc}}\left(\mathbf{z}_{\mathrm{co}}\right). \tag{21}$$

Similarly, the decoder input at forecast step k is

$$\mathbf{d}_k = \mathrm{Emb}\left(\Delta\widehat{\mathbf{u}}_{t+k-1}\right) \,\|\, P_{\mathrm{dec}}\left(\mathbf{z}_{\mathrm{co}}\right). \tag{22}$$

The representation $\mathbf{z}_{\mathrm{co}}$ is extracted exclusively from historical infrared images and their aligned ocean-temperature fields. It is detached from the computational graph before being passed to MMSTN. The trajectory objective therefore updates the projection networks and MMSTN parameters but cannot modify the conditional diffusion model. Generated future infrared frames are not used as trajectory inputs.

Generator and discriminator objectives. For each training example, the generator draws $K = 6$ Gaussian-noise samples. The relative-state reconstruction loss implemented in our training procedure is the mean error over these samples:

$$\mathcal{L}_{\mathrm{rel}} = \frac{1}{K}\sum_{i=1}^{K}\left\|\Delta\widehat{\mathcal{U}}^{(i)} - \Delta\mathcal{U}\right\|_2^2. \tag{23}$$

The final generated trajectory is additionally evaluated by a local discriminator. The discriminator encodes the complete observed–future relative trajectory using a single-layer LSTM with 128 hidden units and classifies it as observed or generated. Its binary cross-entropy objective is

$$\mathcal{L}_D = -\mathbb{E}_{\mathcal{U}} \left[\log D(\mathcal{U})\right] - \mathbb{E}_{\widehat{\mathcal{U}}} \left[\log\left(1 - D(\widehat{\mathcal{U}})\right)\right]. \tag{24}$$

The generator is trained using

$$\mathcal{L}_G = \lambda_{\text{rel}}\mathcal{L}_{\text{rel}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \qquad \mathcal{L}_{\text{adv}} = -\mathbb{E}_{\widehat{\mathcal{U}}} \left[\log D(\widehat{\mathcal{U}})\right]. \tag{25}$$

We set the relative-state loss weight to 1.0. Both networks are optimized using Adam with a learning rate of 10^{-4} and a batch size of 256. We perform two discriminator updates for every generator update. The discriminator gradient norm is clipped at 2.0, while the generator does not use gradient clipping. Linear layers are initialized using Kaiming normal initialization. Validation ADE is evaluated every 200 updates, and the checkpoint with the lowest validation ADE is retained.

D.4 Sequential Training and Integration

CycloDiff and the adapted MMSTN are trained sequentially. In the first stage, the ocean-conditioned residual diffusion model is optimized using historical infrared images and aligned ocean-temperature fields at 0.5, 34, and 186 m:

$$\phi^* = \arg\min_{\phi} \mathcal{L}_{\text{cloud}}. \tag{26}$$

After convergence, the diffusion parameters and their exponential moving average are frozen. The frozen conditional encoder then produces $\mathbf{z}_{\text{co}}$ for each trajectory sample.

In the second stage, the MMSTN generator parameters ω, projection parameters ρ, and discriminator parameters ξ are optimized while ϕ^* remains fixed:

$$(\omega^*, \rho^*, \xi^*) = \arg\min_{\omega,\rho} \max_{\xi} \mathcal{L}_{\text{traj}} \left(\omega, \rho, \xi; \text{sg}\left[\mathbf{z}_{\text{co}}(\phi^*)\right]\right). \tag{27}$$

This procedure permits one-way transfer of historical atmospheric and oceanic context into trajectory prediction without allowing the trajectory objective to alter the cloud-generation branch.

At inference, the cloud-generation and trajectory branches operate in parallel from the same forecast initialization. The diffusion branch generates the future infrared sequence, while the adapted MMSTN predicts the storm-center trajectory from the historical trajectory, collocated depth-resolved temperatures, and frozen cloud–ocean representation. IDOL is invoked only afterward to evaluate the diagnostic consistency of the generated cloud fields.

Table 5 distinguishes the roles of the two external components used in our framework. IDOL remains fixed and is used only for diagnostic evaluation, whereas the adapted MMSTN is trained as the trajectory branch using historical trajectory states and the frozen 384-dimensional cloud–ocean representation. This separation prevents both the IDOL diagnostic objective and the trajectory loss from modifying the cloud-generation branch.

E Quantitative Evaluation

In this section, we detail the quantitative metrics used in our experiments. Let $\widehat{\mathbf{I}}, \mathbf{I} \in \mathbb{R}^{N\times T\times H\times W}$ denote the predicted and ground-truth cloud-image sequences, respectively, where N is the number of test sequences, T is the forecast length, and $H \times W$ is the spatial resolution. Unless otherwise specified, the forecast-fidelity metrics are computed on images normalized to $[0, 1]$ and aggregated over the test set and forecast frames.

Peak signal-to-noise ratio (PSNR). PSNR measures pixel-level reconstruction accuracy through the mean squared error (MSE):

$$\text{MSE} = \frac{1}{NTHW} \sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{i=1}^{H}\sum_{j=1}^{W} \left(\widehat{I}_{n,t,i,j} - I_{n,t,i,j}\right)^2, \tag{28}$$

Table 5: **Training and implementation settings for IDOL and the adapted MMSTN trajectory branch.** IDOL settings describe its released pretraining protocol; IDOL is not optimized in our experiments.

Setting	IDOL	Adapted MMSTN
Role	Frozen diagnostic estimator	Trainable trajectory generator and discriminator
Input	Two four-channel IR frames at 156×156	Four five-dimensional states and a 384-D cloud–ocean feature
Output	MSW, MSLP, RMW, and R34	Five future relative states and storm-center positions
Embedding dimension	16-D identity tokens	32
Recurrent hidden dimension	Not applicable	64 for the generator; 128 for the discriminator
Noise	Not applicable	16-D Gaussian noise; six samples per example
Optimizer	Adam	Adam for both generator and discriminator
Learning rate	10^{-4}	10^{-4}
Batch size	48	256
Training/selection	200 pretraining epochs	Lowest validation ADE; evaluated every 200 updates
Update ratio	Not applicable	Two discriminator steps per generator step

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}} \right), \tag{29}$$

where $L = 1$ is the maximum value of the normalized images. A higher PSNR indicates greater pixel-level fidelity.

Structural similarity index measure (SSIM). SSIM (Wang et al., 2004) evaluates structural agreement between a predicted image $\widehat{I}$ and its ground truth I:

$$\mathrm{SSIM}(\widehat{I}, I) = \frac{(2\mu_{\widehat{I}}\mu_I + C_1)(2\sigma_{\widehat{I}I} + C_2)}{(\mu_{\widehat{I}}^2 + \mu_I^2 + C_1)(\sigma_{\widehat{I}}^2 + \sigma_I^2 + C_2)}, \tag{30}$$

where μ, σ^2, and $\sigma_{\widehat{I}I}$ denote the local mean, variance, and covariance, respectively, and C_1 and C_2 are numerical stabilization constants. SSIM is averaged over all predicted frames, with higher values indicating better structural agreement.

Learned perceptual image patch similarity (LPIPS). LPIPS (Zhang et al., 2018) measures perceptual discrepancy using features extracted by a pretrained network:

$$\mathrm{LPIPS}(\widehat{I}, I) = \sum_l \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left\| \mathbf{w}_l \odot \left(\bar{\phi}_l(\widehat{I})_{h,w} - \bar{\phi}_l(I)_{h,w} \right) \right\|_2^2, \tag{31}$$

where $\bar{\phi}_l$ denotes the normalized feature representation at layer l, and $\mathbf{w}_l$ contains the learned channel weights. We use an AlexNet-based LPIPS model and average the frame-wise distances over the test set. Lower LPIPS indicates greater perceptual similarity.

Continuous ranked probability score (CRPS). CRPS (Matheson & Winkler, 1976) evaluates the agreement between a predictive distribution and the observed value while accounting for both calibration and sharpness. For predictive cumulative distribution function F and observation y, it is defined as

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{+\infty} \left(F(z) - \mathbf{1}\{z \geq y\} \right)^2 \mathrm{d}z, \tag{32}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. The reported CRPS is averaged over the test samples, forecast frames, and spatial locations. Lower CRPS indicates better probabilistic forecast performance.

Fréchet Video Distance (FVD). FVD (Unterthiner et al., 2018) evaluates sequence-level spatiotemporal fidelity by comparing the distributions of generated and ground-truth sequences in a pretrained I3D feature space:

$$\mathrm{FVD} = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_g + \boldsymbol{\Sigma}_r - 2\left(\boldsymbol{\Sigma}_g\boldsymbol{\Sigma}_r\right)^{1/2}\right), \tag{33}$$

where $(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ and $(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r)$ are the means and covariance matrices of the generated and ground-truth sequence features, respectively. Lower FVD indicates closer sequence-level distributions and more coherent spatiotemporal evolution.

Degree of angular variance (DAV). DAV measures the directional organization of the cloud structure relative to the cloud-system center:

$$\mathrm{DAV}(I) = \mathrm{Var}_{\|\mathbf{r}_p\| \le R}\left[\theta\left(\nabla I_p, \mathbf{r}_p\right)\right], \tag{34}$$

where $\mathbf{r}_p$ is the radial vector from the cloud-system center to pixel p, ∇I_p is the local image gradient, and $\theta(\nabla I_p, \mathbf{r}_p)$ is their angular difference. A lower DAV indicates more coherent directional organization. For forecast evaluation, we compare the frame-wise DAV evolution of the generated and ground-truth cloud sequences.

TBB symmetry ratio. The TBB symmetry ratio measures the azimuthal uniformity of the cloud brightness-temperature field. Let P_{10} and P_{90} denote the 10th and 90th percentiles of the valid TBB values within the 50–100 km annulus around the cloud-system center. The ratio is defined as

$$r_{\mathrm{TBB}} = \mathrm{clip}_{[0,1]}\left(1 - \frac{P_{90} - P_{10}}{T_{\mathrm{ref}}}\right), \qquad T_{\mathrm{ref}} = 116, \tag{35}$$

where T_{ref} is the reference brightness-temperature range used in our implementation. A higher r_{TBB} indicates weaker azimuthal temperature contrast and hence a more symmetric cloud structure. We assess consistency by comparing the distributions obtained from generated and ground-truth cloud fields.

Downstream diagnostic consistency. For the IDOL-based evaluation, let $\widehat{z}_i$ and z_i denote the diagnostic estimates obtained from the generated and corresponding ground-truth cloud images, respectively. Their agreement is quantified using the mean absolute error (MAE), root mean squared error (RMSE), signed bias, and Pearson correlation coefficient:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\widehat{z}_i - z_i\right|, \tag{36}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{z}_i - z_i\right)^2}, \tag{37}$$

$$\mathrm{Bias} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{z}_i - z_i\right), \tag{38}$$

$$R = \frac{\sum_{i=1}^{N}(\widehat{z}_i - \overline{\widehat{z}})(z_i - \bar{z})}{\sqrt{\sum_{i=1}^{N}(\widehat{z}_i - \overline{\widehat{z}})^2}\sqrt{\sum_{i=1}^{N}(z_i - \bar{z})^2}}. \tag{39}$$

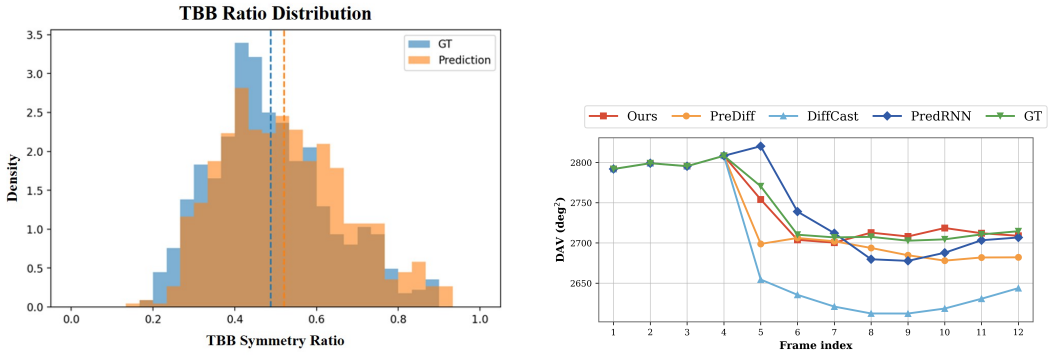
Although Bias is defined as a signed quantity, we report |Bias| in Table 4. Lower MAE, RMSE, and absolute bias indicate closer diagnostic agreement, whereas a higher R indicates stronger correlation. Because z_i is obtained by applying the same frozen IDOL model to the ground-truth cloud images, these metrics quantify downstream diagnostic consistency rather than independent accuracy against best-track observations.

F Additional Results

F.1 Quantitative Cloud-System Structure Evaluation

The main text uses representative cloud fields to illustrate the structural meaning of the TBB symmetry ratio and DAV. Figure 8 provides their dataset-level comparisons. The TBB symmetry-ratio distri-

bution of CycloDiff substantially overlaps the ground-truth distribution, indicating similar overall azimuthal organization. CycloDiff also follows the ground-truth DAV evolution more closely than the baseline methods, particularly during the rapid structural transition near the beginning of the forecast interval. These results complement the visual examples by evaluating cloud-system structure over the complete test set.



(a) TBB symmetry-ratio distributions.

(b) Frame-wise DAV evolution.

Figure 8: **Quantitative evaluation of cloud-system structure.** The left panel compares the TBB symmetry-ratio distributions of CycloDiff forecasts and ground truth. The right panel compares frame-wise DAV among CycloDiff, the baseline methods, and ground truth.

F.2 Additional Qualitative Forecasts

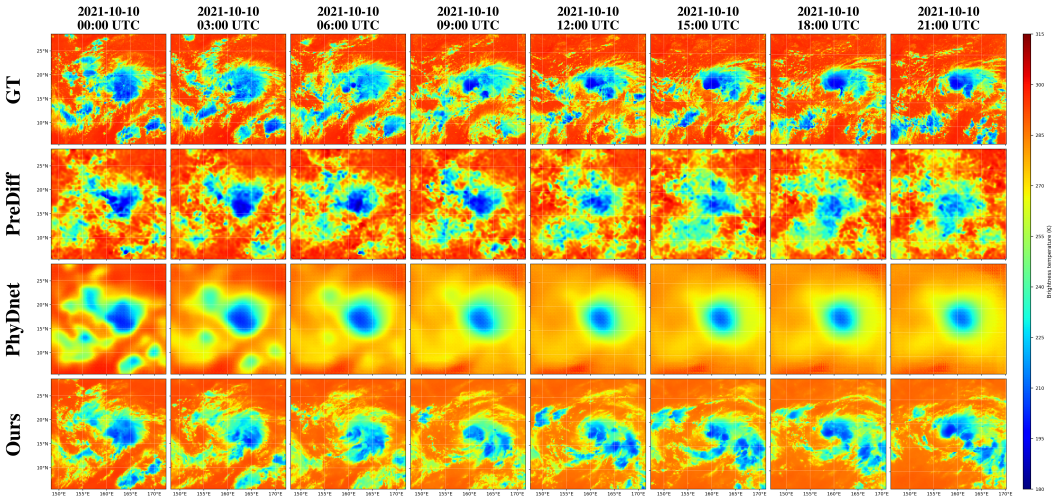


Figure 9: **Additional qualitative comparison of cloud-sequence forecasts.** Rows show the ground truth and forecasts from the compared methods, while columns correspond to successive forecast times.

Figures 9 and 10 extend the qualitative comparison to two additional test cases with different initial cloud configurations. Across both cases, CycloDiff maintains coherent cloud-system evolution throughout the forecast horizon and avoids the pronounced smoothing or structural fragmentation observed in several baseline forecasts. These examples support the qualitative behavior observed in the main text, while the dataset-level comparison remains quantified by the metrics reported in Table 1.

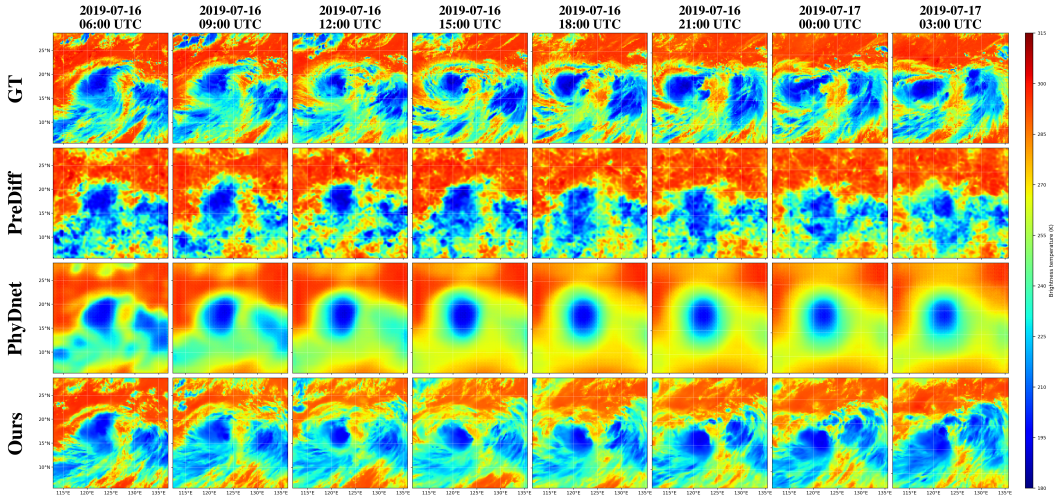


Figure 10: **Additional qualitative comparison of cloud-sequence forecasts for a second test case.** The layout follows Figure 9.

F.3 Complete ZPR Evaluation

Table 6 reports the complete ZPR comparison under both forecast horizons. Across all evaluated settings, CycloDiff produces the closest agreement with the ZPR obtained from the ground-truth future sequences. This result indicates that its generated cloud fields better preserve the classifier-relevant information used for genesis identification.

Table 6: **ZPR evaluation of genesis classification from generated cloud sequences.** Results are reported under different thresholds for eight- and five-frame forecasting. GT denotes the ZPR obtained from the ground-truth future sequence and serves as the reference.

Prediction of 8 Frames									
Threshold	**GT**	**CycloDiff**	**PredRNN**	**PhyDNet**	**PreDiff**	**EarthFormer**	**DiffCast**	**CasCast**	**AlphaPre**
0	0.210	**0.200**	0.152	0.167	0.121	0.169	0.140	0.152	0.179
3	0.429	**0.406**	0.364	0.321	0.281	0.348	0.324	0.364	0.343
6	0.614	**0.609**	0.536	0.481	0.417	0.502	0.438	0.493	0.512
9	0.793	**0.774**	0.707	0.633	0.531	0.657	0.552	0.640	0.679
12	0.800	**0.784**	0.714	0.636	0.548	0.657	0.569	0.667	0.679
15	0.802	**0.784**	0.714	0.638	0.548	0.657	0.569	0.667	0.679
Prediction of 5 Frames									
Threshold	**GT**	**CycloDiff**	**PredRNN**	**PhyDNet**	**PreDiff**	**EarthFormer**	**DiffCast**	**CasCast**	**AlphaPre**
0	0.200	**0.188**	0.152	0.150	0.105	0.157	0.136	0.152	0.176
3	0.412	**0.372**	0.360	0.295	0.236	0.336	0.283	0.364	0.338
6	0.588	**0.551**	0.536	0.448	0.355	0.488	0.381	0.493	0.505
9	0.750	**0.738**	0.669	0.593	0.450	0.636	0.481	0.640	0.669
12	0.755	**0.742**	0.671	0.595	0.457	0.636	0.486	0.667	0.669
15	0.755	**0.742**	0.671	0.595	0.457	0.636	0.486	0.667	0.669