JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH000976

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Skillful Short-Term Forecasting of Clouds With a Cascade Diffusion Model

Haoming Chen[1,2], Xiaohui Zhong[3], Qiang Zhai[4], Xiaomeng Li[4], Ying Wa Chan[5], Pak Wai Chan[5], Muqing Yang[1], Yuanyuan Huang[1], Hao Li[3], and Xiaoming Shi[1,6]

[1]Division of Environment and Sustainability, Hong Kong University of Science and Technology, Hong Kong, China, [2]Now at Shanghai TechWind Technology Co., Ltd., Shanghai, China, [3]Artificial Intelligence Innovation and Incubation Institute, Fudan University, Shanghai, China, [4]Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, China, [5]Hong Kong Observatory, Hong Kong, China, [6]Center for Ocean Research in Hong Kong and Macau, The Hong Kong University of Science and Technology, Hong Kong, China

Haoming Chen and Xiaohui Zhong contributed equally to this work.

Key Points:

- Generative model achieves skillful and high-quality satellite images short-term forecasting
- Cascade structure and multimodal input contribute to skillful prediction
- Data-driven models can interpret the physical correlation within different variables

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

H. Li and X. Shi,
lihao_lh@fudan.edu.cn;
shixm@ust.hk

Citation:

Chen, H., Zhong, X., Zhai, Q., Li, X., Chan, Y. W., Chan, P. W., et al. (2026). Skillful short-term forecasting of clouds with a cascade diffusion model. *Journal of Geophysical Research: Machine Learning and Computation*, *3*, e2025JH000976. https://doi.org/10.1029/2025JH000976

Received 23 AUG 2025
Accepted 6 DEC 2025

Abstract

Accurate short-term forecasting of clouds from satellite imagery is a foundational technology for downstream meteorological disaster mitigation and aviation safety enhancement systems, particularly in developing countries and remote areas lacking ground-based observation infrastructure. Existing deep learning models often produce blurry results and exhibit reduced accuracy when forecasting atmospheric variables, while recent advances in video prediction show the potential to solve these problems. Here, we introduce SATcast, a diffusion-based model that employs a cascaded architecture and multimodal inputs for forecasting cloud evolution from satellite imagery. SATcast incorporates physical fields predicted by FuXi, a deep-learning weather forecasting model, alongside historical satellite observations as conditional inputs to generate high-quality future cloud fields. Comprehensive evaluations demonstrate that SATcast consistently outperforms conventional methods across multiple metrics such as fractions skill score, achieving superior accuracy and robustness. Ablation studies underscore the importance of its multimodal design and cascade architecture in enhancing predictive performance. Notably, SATcast maintains skillful predictions for up to 24 hr, underscoring its potential for operational forecasting applications.

Plain Language Summary

Forecasting cloud is critical for predicting severe weather and ensuring aviation safety, especially in remote areas or undeveloped countries. However, traditional data-driven models often produce blurry and less accurate forecasts. To address these problems, we developed SATcast, a new deep learning model that applies a diffusion model, multimodal input, and cascade structure to predict cloud evolution. SATcast combines predictions from another deep learning model called FuXi with historical satellite observations to generate high-quality future cloud fields. Comprehensive evaluations show that SATcast consistently outperforms conventional methods, achieving superior accuracy and robustness.

1. Introduction

Small-scale deep convection and mesoscale convective systems (MCSs) frequently lead to severe weather events, such as flooding, strong winds, and aviation turbulence (H. Chen et al., 2024; Guo et al., 2022). In recognition of these threats, the United Nations' “Early Warnings for All” initiative calls for enhanced global early warning systems capable of accurately predicting the timing, location, and intensity of such convective events (Organization, 2023). Satellite observations play a critical role in tracking the formation and evolution of multiple MCSs over broad geographic areas. This is especially important for satellite-based forecasting in regions lacking radar coverage, such as over oceans and in many developing countries, where ground-based observation networks are sparse or absent (Schmit et al., 2017). To address these observational gaps, satellite-based forecasting techniques have been actively developed. For example, the World Meteorological Organization (WMO) has issued guidelines for implementing such approaches in Africa, emphasizing their value in areas with limited ground-based observations (Organization, 2023).

Traditional methods for satellite and radar-based forecasting, such as the Lucas-Kanade optical flow algorithm, effectively track the movement of MCSs but struggle with capturing intensity changes due to the assumption of brightness constancy (Baker & Matthews, 2004). Similarly, ensemble-based approaches like the Short-Term Ensemble Prediction System (STEPS) offer uncertainty estimates but often smooth out fine-scale convective details, as they use stochastic noise to model the evolution of small-scale convection (Smith et al., 2024). These

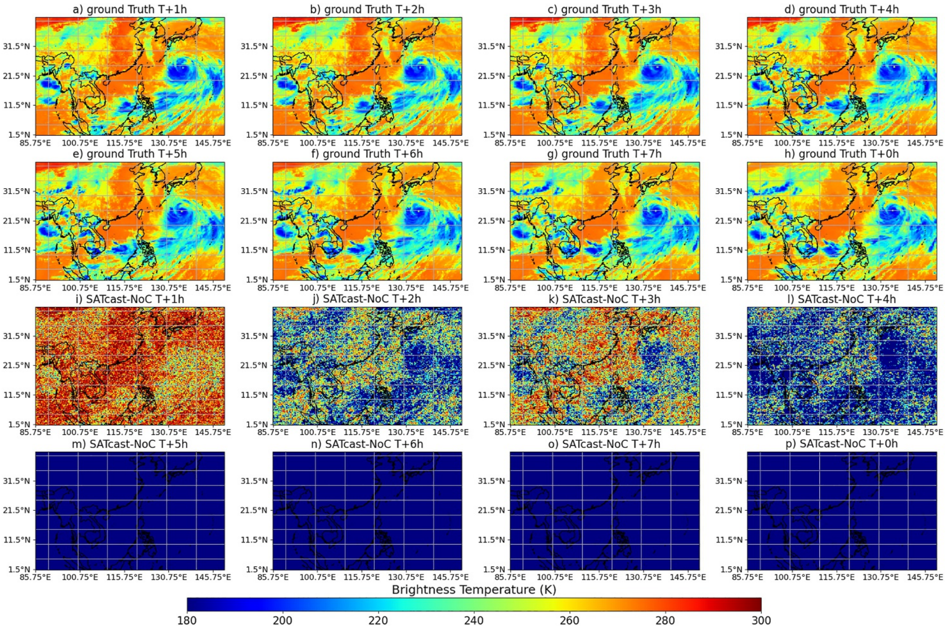

Figure 1. Spatial distribution of brightness temperature from $T + 1$ hr to $T + 8$ hr from observation (a–h), SATcast-NoC (i–p). T denotes the starting time of the prediction, the UTC time of T is 2022-09-14-06:00. The SATcast-NoC is under training (Epoch 149; full training takes 300 epochs).

limitations have catalyzed the adoption of deep learning models, which bypass such assumptions and excel in producing more accurate and reliable forecasts (Bi et al., 2023; L. Chen et al., 2023; Zhang et al., 2023).

The emergence of deep learning in short-term forecasting has yielded promising advances, although certain challenges persist. Forecast horizons for satellite or radar-based predictions rarely exceed 6 hours due to the accumulation of errors, and deep learning models often produce blurred outputs, limiting their ability to resolve sharp convective features (Ehsani et al., 2021; A. Kumar et al., 2020; Tran & Song, 2019; Wei et al., 2024). For instance, Shi et al. (2015) introduced convolutional Long Short-Term Memory (ConvLSTM) networks for radar image time series forecasting, demonstrating superior performance in precipitation forecasting, which has a lead time of less than 2 hr. More recently, DaYu, which integrates transformer (Vaswani, 2017) and residual convolution layers, has extended the skillful forecast lead time of satellite-based forecasting to 12 hr (Wei et al., 2024). However, its predictions become overly smooth after 3 hr, particularly in reproducing detailed typhoon structure and eye positions (see Figures 3 and 4). Other approaches have aimed to address these issues through architectural innovation or data integration. Wang et al. (R. Wang et al., 2023) applied Generative Adversarial Networks (GAN) (Goodfellow et al., 2020) to generate high-quality radar forecasts, with skillful performance up to 2 hr. Kim et al. (W. Kim et al., 2024) incorporated European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis version 5 (ERA5) data into a deep learning model, improving six-hour precipitation forecasts compared to radar-only inputs. However, the 5-day delayed production schedule of ERA5 data limits its applicability to real-time forecasting. They (W. Kim et al., 2024) suggest that numerical weather prediction (NWP) data could serve as a viable alternative, but NWP models themselves also face constraints, including high computational costs and inherent uncertainties in parameterization schemes for gray-zone convection processes, which limit their ability to accurately forecast convective activities (Shi & Wang, 2022; Trier et al., 2014). As a result, deep-learning-based forecast techniques that operate independently of NWP models are gaining traction as more cost-effective and operationally practical solutions (L. Chen et al., 2023; Bi et al., 2023).

Physical variables generated by either deep learning or NWP models exhibit significant differences in information content from radar and satellite observations, making them inherently multimodal data sources. Integrating such multimodal data is a key strategy for enhancing model performance, yet further research is needed to develop effective methods for processing and combining these diverse data modalities within deep learning models for atmospheric applications. Recent advances in multimodal learning offer promising avenues. For example, the

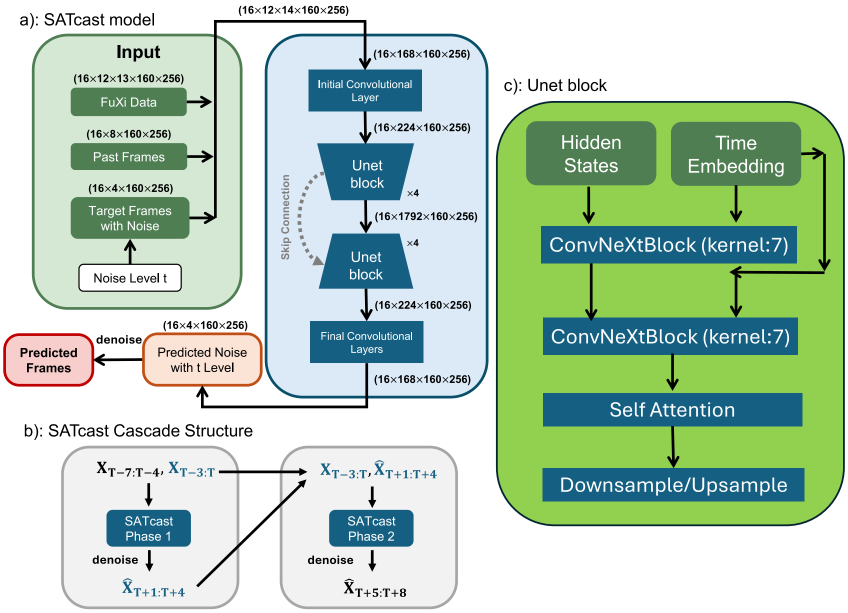

Figure 2. Schematic of the SATcast model. (a) Overall architecture of SATcast. Dashed lines indicate skip connections and the denoising process is performed using the network highlighted by the blue box. Dimensional changes of features are shown for a 12-hr input sequence. (b) Cascade structure of SATcast. “T” represents the final time point of the input satellite image sequence (Forecast Reference Time). SATcast-phase 1 predicts satellite imagery from $T + 1$ hr to $T + 4$ hr, while SATcast-phase 2 uses the predictions from SATcast-phase 1, along with earlier frames and FuXi data, to predict satellite imagery from $T + 5$ hr to $T + 8$ hr (c) Structure of the UNet block. Arrows indicate the flow of tensors. Green boxes represent input data, and blue boxes represent neural network layers.

Contrastive Language-Image Pretraining model combines transformer and convolutional layers to encode text and images, enabling a more flexible, generalizable classifier that serves as a foundation for subsequent multimodal models (Radford et al., 2021). In atmospheric science, various approaches have been developed to integrate NWP outputs into deep learning models. Kim et al. (H. Kim et al., 2021) employed a long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997) to post-process and correct systematic biases in Madden-Julian Oscillation predictions, improving NWP skill over a 4-week forecast period. MetNet-3 (Andrychowicz et al., 2023) employs different ResNet blocks (Huang et al., 2016) to process multi-resolution data, extending skillful forecast lead times up to 24 hr. Rui Wang et al. (2024) uses a cross-attention module to merge NWP results and radar images, enabling precipitation forecasts up to 6 hr ahead. While these studies highlight the benefits of integrating physical variables with satellite imagery, bridging the gap between such heterogeneous data types remains challenging.

In this study, we introduce SATcast, a deep learning framework for predicting convective clouds observed by the FengYun-4A (FY-4A) geostationary satellite. SATcast leverages diffusion models, state-of-the-art generative methods widely used in text-to-image and video generation in computer vision (Dhariwal & Nichol, 2021; Ho et al., 2020), to forecast satellite-based cloud evolution. The FY-4A satellite plays a crucial role in monitoring weather patterns, climate variations, environmental changes, and natural disasters across East Asia (Yang et al., 2017). Our model is conditioned on atmospheric variables predicted by FuXi (L. Chen et al., 2023), a deep learning-based model that has demonstrated remarkable accuracy in deterministic weather predictions. To further mitigate error accumulation over extended lead times, SATcast employs a cascade structure. By integrating physical variables with satellite imagery within a diffusion-based multimodal framework, SATcast achieves high-fidelity nowcasts of convective cloud systems with a 0.25∘ spatial resolution and 1-hr interval; the forecast lead times can reach 24 hr. SATcast effectively addresses the long-standing issues of blurriness in deep learning–based forecasting. Notably, our results demonstrate the feasibility of generating large-scale forecasts, covering thousands of kilometers, using only a single GPU and a diffusion model, offering a scalable and efficient solution for real-time convective weather prediction.

2. Models and Methods

2.1. Data Set

We constructed input sequences by incorporating 13 physical variables (detailed in Table S1 in Supporting Information S1) from FuXi's hourly predictions, aligned with the spatiotemporal coordinates of FY-4A satellite observations.

Our analysis utilizes Level 1 data from channel 12 of the FY-4A AGRI instrument, which has a central wavelength of 10.7 μm. Channel 12 is selected as it is specifically designed for water vapor detection and mid-troposphere monitoring, making it particularly effective for tracking the development of cloud systems and convective processes (Lu et al., 2017; Yan et al., 2024). The original data have a spatial resolution of 4 km, and a temporal resolution of 1 hr, covering the period from 2019 to 2022. The data is then interpolated to a 0.25° resolution to reduce computational costs while maintaining resolution consistent with FuXi data.

These satellite observations are combined with FuXi's 13 predicted physical variables, provided every 12 hr, aligned with the latitude, longitude, and temporal information from the satellite data. Only FuXi forecasts initialized prior to each target lead time T were used, thereby avoiding the inclusion of unavailable future variables (see Figure S18 in Supporting Information S1). The resulting data set is organized in a T, C, H, W format (12, 14, 160, 256), where T represents time steps, C denotes channels (including both satellite data and physical variables), and H and W correspond to the spatial dimensions in latitude and longitude, respectively. The training data set for SATcast and SATcast-NoF comprises 19,904 sequences, each spanning 12 hr. Additionally, a separate training data set for SATcast-NoC includes 17,408 sequences, each with a duration of 16 hr. For model evaluation and ablation analysis, additional validation data sets of 16 and 32-hr sequences were also prepared.

2.2. Basic Diffusion

In the forward process of the diffusion model, Gaussian noise is incrementally added to the target. Let $P(x_0)$ represent the sample distribution. The forward process is formulated as a discrete-time Gaussian Markov process:

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \boldsymbol{I}\right) \tag{1}$$

where x_t denotes noisy version of the original sample x_0 at time step t, and is calculated as: $x_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon$, with ϵ being the noise sampled from $\mathcal{N}(0,1)$. Here, α_t is a fixed schedule over t, and $\beta_t = 1 - \alpha_t$. The overbar notation $\overline{x}_t$ denotes the cumulative product of the noise schedule up to time step t. The network is trained to predict the added noise given the noisy observation x_t and the time step t. During sampling, the diffusion model reverses the forward diffusion process to recover x_{t-1} from x_t, given by:

$$p(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \widetilde{\mu_t}(x_t), \widetilde{\beta_t}\mathbf{I}\right) \tag{2}$$

where $\widetilde{\mu_t}(x_t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha_t}}}\epsilon_\theta(x_t, t)\right)$, $\overline{\alpha_t} = \prod_1^t \alpha_t$, and $\epsilon_\theta(x_t, t)$ is the noise predicted by the network. Additionally, $\widetilde{\beta_t} = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. Thus, x_{t-1} is sampled from a Gaussian distribution parameterized by the predicted mean and variance at each time step (Dhariwal & Nichol, 2021; Nichol & Dhariwal, 2021).

2.3. Conditional Diffusion for Satellite Image Forecasting

Consider a sequence of satellite image time series, where the forecasting task predicts future n frames satellite images based on the conditions (*cond* in Equation 3), which include the past m frames satellite images and FuXi variables from time steps $\mathrm{T} - m$ to $\mathrm{T} + n$. During training, noise is added to the future n frames at level t, while the past frames and FuXi variables serve as the conditions. Therefore, the reverse process is modified as:

$$p(x_{t-1}|x_t, cond) = \mathcal{N}\left(x_{t-1}; \widetilde{\mu_t}(x_t, cond), \widetilde{\beta_t}\mathbf{I}\right) \tag{3}$$

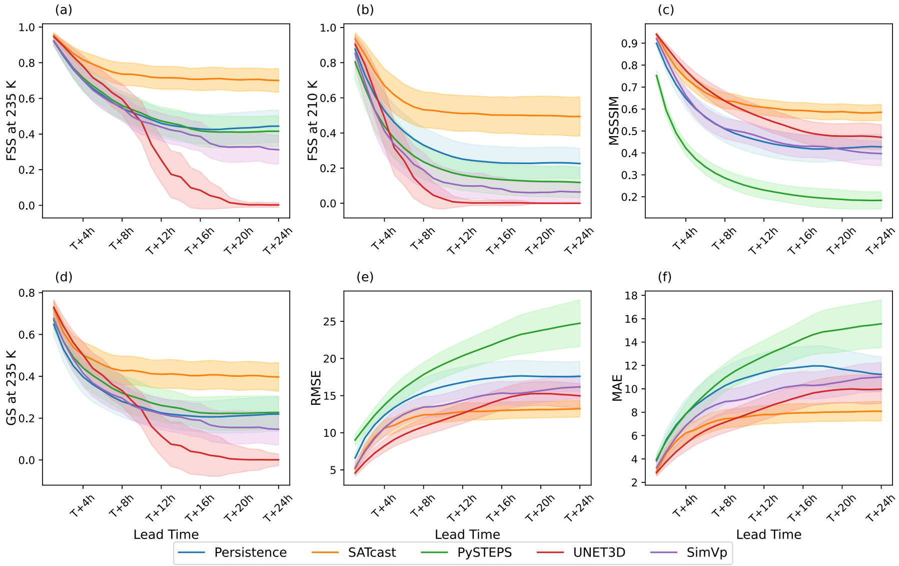

Figure 3. Comparison of FSS at 235 K, FSS at 210 K, MSSSIM, GS at 235 K, RMSE, and MAE in 24-hr forecasts, spatially averaged over the region from 86° to 150° E longitude and 1° to 41° N latitude. Results are shown for the persistence model, SATcast, PySTEPS, Unet3D, and SimVP based on testing data from September to December 2022. Thresholds used for FSS and GS are shown in the *y*-axis labels. Higher values in panels (a–d) indicate better performance, whereas lower values in panel (e–f) demonstrate higher skill. Shaded areas denote the one standard deviations calculated from predictions at each lead time for the respective models.

$$\widetilde{\mu_t}(x_t, cond) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t, t, cond)\right) \tag{4}$$

The network training objective is simplified as follows:

$$\mathcal{L} = \mathbb{E}_{x_t, cond, t, \epsilon\sim\mathcal{N}(0,1)} \left\|\epsilon - \epsilon_\theta(x_t, cond, t)\right\|_2^2 \tag{5}$$

where ϵ_θ represents the noise predicted by the network. Generative models can produce various forecast outcomes. Our experiments demonstrate that samples generated with different random seeds show that the positions of some small-scale convection may be predicted differently near the cloud-clustered regions in the tropical regions. However, their overall spatial distributions are highly similar (see Figures S19 and S20 in Supporting Information S1). This consistency is due to the constraints imposed by the input conditions and the application of classifier-free guidance (Ho & Salimans, 2022). Consequently, whether evaluating model performance using ensemble average or single-member forecasts, the differences remain minimal. Therefore, we output only one sample per forecast.

2.4. SATcast Model

The basic model, SATcast, is a cascaded forecasting model that incorporates FuXi predictions (Table S1 in Supporting Information S1), and undergoes a two-phase training, as illustrated in Figure 2b. In the first phase, the model learns to predict satellite images with 1-hr to 4-hr lead times ($T + 1$ hr to $T + 4$ hr in other descriptions), using past satellite images from the past 7 hr to the present (T−7 hr to $T + 0$ hr) and the corresponding FuXi predictions from T−7 hr to $T + 4$ hr as inputs. In the second phase, the model extends its predictions with lead times of 5–8 hr ($T + 5$ hr to $T + 8$ hr). Inputs in this phase include past satellite images from the past 3 hr to the present (T−3 hr to $T + 0$ hr), the predicted satellite images with 1-hr to 4-hr lead times ($T + 1$ hr to $T + 4$ hr) generated in Phase 1, and FuXi forecasts from T−3 hr to $T + 8$ hr. To improve computational efficiency and convergence, SATcast-phase 2 is initialized with pretrained weights from SATcast-phase 1. As a control, we also evaluate a variant termed SATcast-NoT, which skips fine-tuning in Phase 2 and uses the Phase 1 model directly.

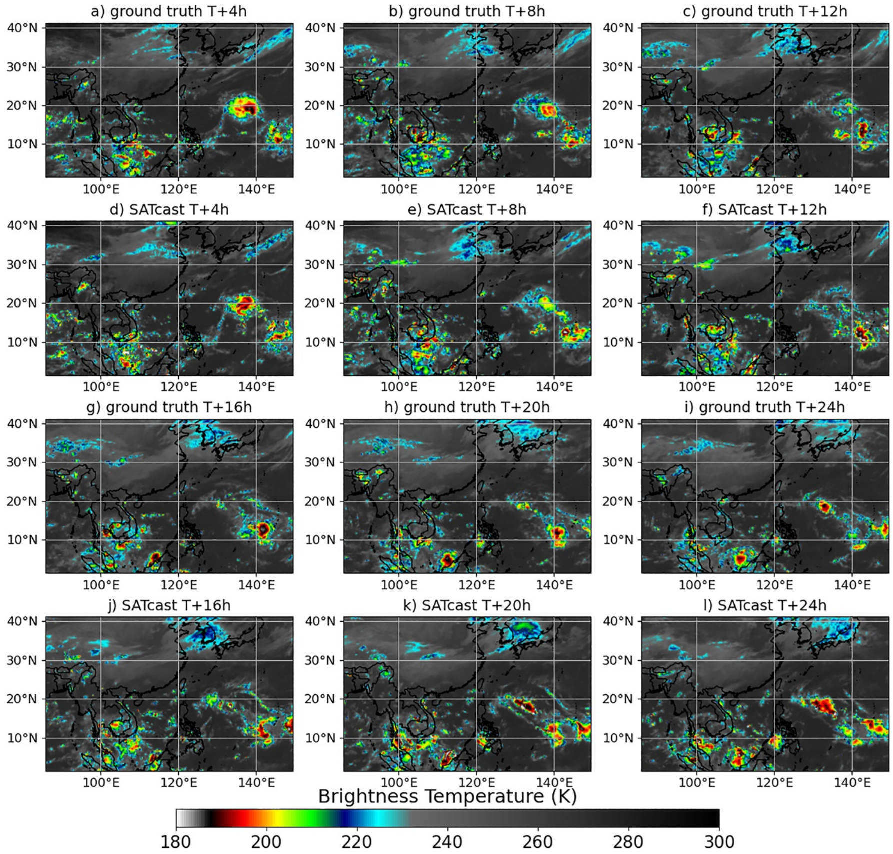

Figure 4. Spatial distribution of brightness temperature from observation and SATcast predictions, shown at 4-hr time intervals from $T + 4$ hr to $T + 24$ hr. The UTC time of T is 2022-10-07-02:00. Panels (a–c) and (g–i) are ground truth, and panels (d–f) and (j–l) are SATcast prediction.

For predictions beyond $T + 8$ hr, SATcast operates in an autoregressive manner, recursively using its previous outputs as inputs for subsequent predictions.

Furthermore, to assess the contributions of key components within the SATcast framework, we conduct ablation studies with four model variants.

- *SATcast-NoT*: As previously described, a version of SATcast that does not undergo fine-tuning in Phase 2.
- *SATcast-NoC*: A non-cascaded variant that performs direct multi-frame prediction from $T + 1$ hr to $T + 8$ hr using the same input data as SATcast Phase 1.
- *SATcast-NoF*: A variant that excludes FuXi forecasts as conditional inputs while maintaining autoregressive prediction for all eight frames.
- *SATcast-NoS*: A pseudo-satellite variant that removes satellite imagery inputs, relying solely on FuXi predictions.

2.4.1. Cascaded Architecture

Following the design of FuXi (L. Chen et al., 2023), we implemented a cascaded forecasting architecture to reduce error accumulation and enhance forecasting performance of convective clouds at longer lead times. Since predicting frames with longer lead times is more challenging, directing the model's attention toward earlier frames within a single forecast is more beneficial for training an effective model. Fine-tuning can be applied in subsequent iterations to further reduce iterative errors (L. Chen et al., 2023).

Additionally, Figure 1 shows the observations and predictions of brightness temperature from SATcast-NoC that is under the training process (Epoch 149; full training takes 300 epochs), and this case is also shown in Figure 5. The model slightly captures the basic structures of the convection systems in the first four frames, but almost all

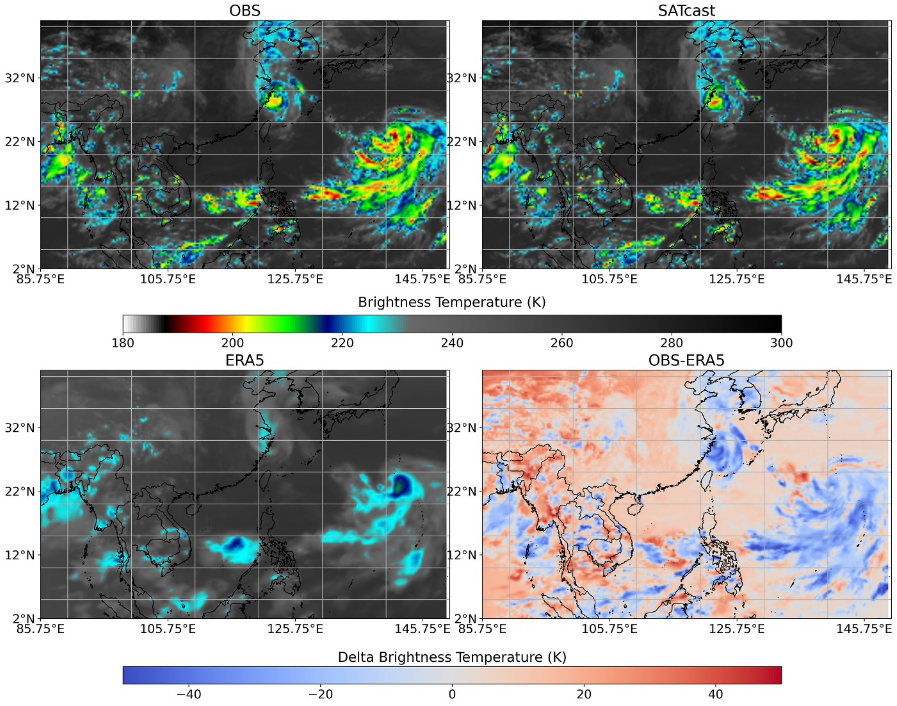

Figure 5. Spatial distribution of brightness temperature at $T + 1$ hr (UTC 2022-09-14 07:00) from observation (OBS), ERA5 reconstructed temperature from RTTOV, SATcast, and the differences between OBS and ERA5 (OBS-ERA5).

the values in the subsequent four frames are 174 K (minimum values of the observed brightness temperature), underscoring the motivation for our cascaded design, which generates forecasts only four frames at a time to force the model to concentrate on the earlier frames.

2.4.2. Multimodal Data

The physical-driven network is constructed by incorporating forecasts of physical variables from FuXi. Variables representing the same physical quantity at different altitudes are treated as distinct modalities due to the large difference in magnitude between different levels. To ensure consistency across these modalities, we apply min-max normalization by linearly scaling all values to the range $[-1, 1]$. After temporal and spatial alignment, we obtain a four-dimensional data set characterized by dimensions T, C, H, and W, as defined in Section 2.1. For a sequence of 12 time steps, the resulting dimensions are 12 (T), 14 (C), 160 (H), and 256 (W). In addition, the batch size (B) is set to 16 for each GPU. Given that all variables can be expressed within the system of atmospheric dynamic equations, and to mitigate the loss from encoding compression and computational resources (Voleti et al., 2022), the T and C dimensions within the five-dimensional data are merged before entering the model, resulting in a structure of B, $T \times C$, H, W. This representation is then passed through a convolutional layer with a larger kernel size (7×7) to extract spatial information.

2.4.3. Network Architecture

We employ a U-Net2D architecture to predict noise, as shown in Figure 2a. This U-Net contains ConvNeXt modules (Liu et al., 2022) and self-attention modules (Vaswani, 2017). Notably, it does not include explicit components for processing the temporal correlations of the data. Instead, temporal dependencies are implicitly learned by the model.

As previously described, after the initial convolutional layer, the data were reshaped to dimensions $16 \times 224 \times 160 \times 256$. The data then passes through four downsampling Unet blocks, each of which halves the spatial dimensions (H and W) while doubling the number of channels. Symmetrically, the data is processed by four upsampling blocks that double the spatial dimensions and reduce the number of channels by half. The Figure 2c shows each Unet block consists of two ConvNeXt and a self-attention module with residual connections. The ConvNeXt block incorporates embeddings of diffusion time step and hidden states of meteorological data. This block employs large 7 × 7 kernels in the first convolutional layer to expand the receptive fields and processes the time embedding through a multilayer perceptron, which is broadcasted as a spatial bias of the hidden states (Ho & Salimans, 2022). The ConvNeXt block utilizes inverted bottlenecks that have several convolutional layers and related activation layers and normalization layers for computational efficiency. Group Normalization is applied in the Unet block due to its superior performance in vision tasks (Wu & He, 2018), and the GELU activation function is used in the inverted bottlenecks for its proved effectiveness when integrated with attention layers (Hendrycks & Gimpel, 2023). The attention layer processes only the hidden states and employs the scaled dot-product attention mechanism, and is further augmented with Group Normalization and residual connections at the output. Both the ConvNeXt block and the attention layer preserve the shape of the input tensor, whereas the downsample and upsample blocks employ convolutional and transposed convolutional blocks, respectively, to halve or double the height (H) and width (W) of the tensor.

Following this, the data are processed by four symmetric upsampling blocks that mirror the structure of the downsampling path, restoring the original spatial dimensions before entering the UNet. At the final stage, multiple convolutional layers progressively reduce the number of channels to match the target shape $T \times C$ (12 × 14).

At the final stage, the output tensors is restored, and only the noise-added portion, the target satellite imagery, is extracted to calculate the loss against the added noise.

2.4.4. Model Training

The model is developed using Pytorch (Paszke et al., 2017). Training the model takes approximately 20 hr on a cluster of 8 Nvidia H800 GPUs. The model is trained for 300 epochs using a batch size of 16 per GPU. The AdamW (Loshchilov, 2017) optimizer is used with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and the learning rate is initialized at 5.0×10^{-5} with a warm-up phase followed by cosine annealing. An exponential moving average is also applied during training (Morales-Brotons et al., 2024). The L2 loss is used as the loss function. We used a linear noise schedule, with $\beta_{start} = 0.0001$ and the $\beta_{end} = 0.02$.

To improve the model's performance in predicting extreme events, sequences containing category three typhoons are resampled once during training. Classifier-free guidance (CFG) is employed during both training and sampling to improve image quality (Ho & Salimans, 2022), after hyperparameter testing, we found that a CFG scale of 1.8 and a condition drop rate of 0.1 yield the best performance. Additionally, offset noise with a scale of 0.1 is applied during sampling to mitigate data distribution mismatches between the images with noise and pure noise, following Lin et al. (2024).

2.4.5. Evaluation Method

For target regions spanning a wide range of latitudes, we apply the latitude-weighted fractions skill score (FSS) (Roberts & Lean, 2008) and Gilbert Skill Score (GS) (K. Kumar, 2004) metrics. FSS is a widely used metric in the literature for evaluating spatial overlap of observed and predicted events (Lagerquist et al., 2021; Smith et al., 2024). To compute FSS, brightness temperature fields from both sources are first binarized using a threshold (either 235 K or 210 K in this study). The 235 K threshold is widely adopted for identifying MCSs given to its robustness in sensitivity experiments (Feng et al., 2021; Roca et al., 2017; Smith et al., 2024). For each cell, the fractional coverage of pixels exceeding the specific threshold within a 3 × 3 grid neighborhood is then computed for observations $\left(\widehat{\bar{\mathbf{X}}}_{i,j}^{\mathrm{T}_0+\tau}\right)$ and for model predictions $(\bar{\mathbf{X}}_{i,j}^{\mathrm{T}_0+\tau})$. The mean squared error $(\mathrm{MSE}_{\mathrm{ref}})$ between these observed and predicted fractions is calculated as described in Equation 5. However, this raw MSE is sensitive to event frequency and does not provide a normalized skill measure. Therefore, it is compared against a reference MSE $(\mathrm{MSE}_{\mathrm{ref}})$, which represents the largest possible MSE given the forecast and observed fractions (Equation 7). The resulting FSS ranges from 0 (no skill) to 1 (perfect forecast), with values higher than 0.5 (Mittermaier & Roberts, 2010) considered skillful. Finally, the FSS can be calculated as follows:

$$\mathbf{MSE}(\tau) = \sum_{i=1}^{H} \sum_{j=1}^{W} a_i \left(\widehat{\bar{\mathbf{X}}}_{i,j}^{\mathrm{T}_0+\tau} - \bar{\mathbf{X}}_{i,j}^{\mathrm{T}_0+\tau} \right)^2 \tag{6}$$

$$\mathbf{MSE_{ref}}(\tau) = \sum_{i=1}^{H}\sum_{j=1}^{W} a_i\left(\left(\widehat{\bar{\mathbf{X}}}_{i,j}^{\mathrm{T}_0+\tau}\right)^2 - \left(\bar{\mathbf{X}}_{i,j}^{\mathrm{T}_0+\tau}\right)^2\right) \tag{7}$$

$$\mathbf{FSS}(\tau) = 1 - \frac{\mathbf{MSE}}{\mathbf{MSE_{ref}}} \tag{8}$$

where τ is the forecast lead time, T_0 is the initial time, and H and W correspond to the number of grid points in the longitudinal and latitudinal directions, respectively. a_i is the latitude weighting factor for the ith latitude index (Rasp et al., 2020):

$$a_i = \frac{\cos(\mathrm{lat}(i))}{\frac{1}{H}\sum_i^H \cos(\mathrm{lat}(i))} \tag{9}$$

In addition to FSS, we use GS to evaluate binary classification performance. GS incorporates not only hits, false alarms, misses, and, critically, the hits due to chance (**CH**) (K. Kumar, 2004). Unlike the Critical Success Index (CSI), which can be biased in the presence of imbalanced event frequencies, GS adjusts for the base rate of the observed events. This adjustment is particularly valuable in satellite-based forecasting applications, where cloud coverage can be highly regional, leading to spatially varying base rates. For instance, in MCS scenarios, a high cloud coverage baseline can artificially inflate CSI scores, whereas GS provides a more robust skill assessment by penalizing chance agreement, offering a more conservative and informative measure of forecast quality. GS ranges from −1/3 to 1, with values below 0 indicating performance worse than random chance. It is calculated as follows:

$$\mathbf{CH}(\tau) = \frac{\left(TP_{i,j}^{\mathrm{T}_0+\tau} + FN_{i,j}^{\mathrm{T}_0+\tau}\right) \times \left(TP_{i,j}^{\mathrm{T}_0+\tau} + FP_{i,j}^{\mathrm{T}_0+\tau}\right)}{n^2} \tag{10}$$

$$\mathbf{GS}(\tau) = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W} a_i \frac{TP_{i,j}^{\mathrm{T}_0+\tau} - \mathbf{CH}(\tau)}{TP_{i,j}^{\mathrm{T}_0+\tau} + FN_{i,j}^{\mathrm{T}_0+\tau} + FP_{i,j}^{\mathrm{T}_0+\tau} - \mathbf{CH}(\tau)} \tag{11}$$

where TP refers to the number of correctly predicted positive pixels; FP is the number of negative pixels incorrectly predicted as positive; FN is the number of positive pixels incorrectly predicted as negative; and TN is the number of correctly predicted negative pixels. The score is aggregated over the spatial domain with latitude weighting factor (a_i) to account for varying grid cell areas.

To assess structural similarity between observed and predicted fields, we use the Multi-scale structure similarity (MSSSIM), which combines the conventional Structural Similarity Index (SSIM) by evaluating image similarity across multiple spatial scales to yield more robust results. Latitude-weighting is not applied here, as the comparison treats the image as a whole rather than spatially aggregating grid values. MSSSIM is defined as:

$$\mathbf{MSSSIM}(\tau) = \left[l_M\left(\hat{\mathbf{X}}^{\mathrm{T}_0+\tau}, \mathbf{X}^{\mathrm{T}_0+\tau}\right)\right]^{\alpha_M} \prod_{j=1}^{M}\left[c_j\left(\hat{\mathbf{X}}^{\mathrm{T}_0+\tau}, \mathbf{X}^{\mathrm{T}_0+\tau}\right)\right]^{\beta_j}\left[s_j\left(\hat{\mathbf{X}}^{\mathrm{T}_0+\tau}, \mathbf{X}^{\mathrm{T}_0+\tau}\right)\right]^{\gamma_j} \tag{12}$$

The comparisons about luminance, contrast and structure are denoted by the three terms (l_M, c_j, s_j) in Equation 12, the M means different scales of the images. The three exponents adjust the relative importance of different components. More details about the calculation can be found in Z. Wang et al. (2003).

We have also incorporated standard evaluation metrics, such as Root Mean Square Error (RMSE) and mean absolute error (MAE), to evaluate the model's skill in predicting brightness temperature.

$$\mathbf{RMSE}(\tau) = \sum_{i=1}^{H}\sum_{j=1}^{W} a_i\sqrt{\left(\widehat{\bar{\mathbf{X}}}_{i,j}^{\mathrm{T}_0+\tau} - \bar{\mathbf{X}}_{i,j}^{\mathrm{T}_0+\tau}\right)^2} \tag{13}$$

$$\mathbf{MAE}(\tau) = \sum_{i=1}^{H}\sum_{j=1}^{W} a_i \left| \hat{\bar{\mathbf{X}}}_{i,j}^{\mathrm{T}_0+\tau} - \bar{\mathbf{X}}_{i,j}^{\mathrm{T}_0+\tau} \right| \tag{14}$$

3. Results

3.1. Overall Performance of SATcast

Figure 3 compares the performance of SATcast and baseline models. We used the persistence model, the PySTEPS, an open-source Python library for forecasting. We also evaluated two deep learning models with traditional architectures, the Unet3D and the SimVP, both widely used in short-term weather forecasting and demonstrated to be effective in several studies (K. Çağlar et al., 2024; X. Wang et al., 2024). Using three evaluation metrics: FSS at different threshold (Roberts & Lean, 2008), MSSSIM (Z. Wang et al., 2003), and GS (K. Kumar, 2004). We used the deterministic predictions based on the Lucas–Kanaede algorithm combined with Spectral Prognosis by using PySTEPS (Pulkkinen et al., 2019; Smith et al., 2024). The FSS and GS are calculated using a brightness temperature threshold of 235 K, a widely used threshold for identifying the boundaries of MCSs (Feng et al., 2021; Roca et al., 2017; Smith et al., 2024). To assess model performance in predicting strong convection systems, an additional threshold of 210 K is applied. All metrics are spatially averaged over the region of interest (86° to 150° E in longitude and 1° to 41° N in latitude). Evaluations are based on 512 testing sequences, collected from September to December 2022. Each sequence spans 32 hr, with the first 8 hr used as model input, and the subsequent 24 hr as the forecast target. The shaded regions around the curves represent one standard deviation calculated from predictions on specific lead times from different models.

As shown in Figure 3, all models experience rapid performance degradation up to $T + 8$ hr. However, SATcast shows a more gradual degradation after $T + 12$ hr, while the persistence model and PySTEPS continue to deteriorate steadily, in addition, SimVP and Unet3D show significant declines. Notably, the persistence model exhibits slight performance improvements between $T + 20$ hr and $T + 24$ hr, likely due to the influence of the diurnal cycle (Wallace, 1975). Since the outputs of Unet3D and SimVP are overly smooth and lack deep convection (See Figure S9 in Supporting Information S1), these models achieve slightly better RMSE and MAE scores than SATcast before $T + 4$ hr, but their performance declines substantially after $T + 12$ hr in Figures 3e and 3f). Overall, SATcast outperforms other baseline models in all evaluation metrics for most of the time. SATcast maintains the FSS scores above 0.7 at the 235 K threshold, significantly outperforming the baseline models, especially for the Unet3D and SimVP, whose scores fall below 0.5 after $T + 12$ hr. At the 210 K threshold, used to assess strong convective systems, all baseline models exhibit even lower skill, while SATcast maintains an FSS above 0.5, showing its superior ability in capturing strong convection. Similarly, SATcast achieves a GS of approximately 0.45 at 235 K, indicating reliable detection of convective clouds, compared to less than 0.3 for the baseline models. In terms of MSSSIM, SATcast remains approximately 0.6 at the forecast lead time of 24 hr, suggesting that the predicted cloud fields maintain high structural similarity even as the lead time increases. In contrast, the MSSSIM values for the baseline models decrease to about 0.45, likely due to the impact of evolving weather systems and smoothing effects inherent in extrapolation methods. Based on the order of magnitudes of RMSE and MSE for brightness temperature, the forecast error of SATcast remains below 5% after $T + 12$ hr, while PYSETPS exceeds 10%. SimVP and Unet3D generate slightly higher errors than SATcast. We also evaluated the spatial variances of different model predictions (Figure S1 in Supporting Information S1). The results demonstrate that SATcast maintains variance close to observations throughout the 24-hr forecast period, whereas traditional models such as SimVP and UNet3D exhibit a pronounced reduction due to increasingly blurred outputs. In contrast, PySTEPS shows a substantial increase in variance, likely caused by misplaced intense convective systems.

SATcast exhibits the narrowest standard deviation ranges, demonstrating superior robustness compared to other baseline models. Regarding temporal dependency, the standard deviation ranges for all models increase by 50%–100% after $T + 16$ hr, suggesting that short-term forecasts (up to $T + 12$ hr) are more reliable, while longer lead times involve greater uncertainty. Furthermore, the influence of thresholds of brightness temperature is also evident: under the 210 K condition in Figure 3b, variability is higher than under 235 K, reflecting there are more uncertainties in predicting stronger convection. Moreover, the spatial distribution of clouds varies significantly between fair and extreme weather events, as well as between land and ocean regions (Figures S2 and S3 in Supporting Information S1). All models perform better but less stably over land, and more stably but less accurately over the ocean, consistent with stronger diurnal variability over land. Under extreme weather conditions (Figure S4 in Supporting Information S1, brightness temperature lower than 235 K), model skill becomes less stable and errors increase compared with the average across all scenarios. Overall, SATcast maintains robust performance across these diverse conditions.

Table 1
The Configurations of the SATcast and Variants

Model	Autoregressive	Fine-tuning	FuXi data
SATcast	✓	✓	✓
SATcast-NoC	✗	✗	✓
SATcast-NoT	✓	✗	✓
SATcast-NoF	✓	✗	✗
SATcast-NoS	✓	✗	✗

Note. NoC: No cascade (no autoregressive and fine-tuning), NoF: No FuXi data, NoT: No fine-tuning.

Figure 4 shows the spatial distribution of brightness temperature and demonstrates SATcast's 24-hr forecasting capability using a cloud case. SATcast preserves the evolution of convective cloud systems, while PySTEPS outputs become increasingly smoothed after several time steps (Figure S8 in Supporting Information S1), SimVP and Unet3D also produce overly smooth cloud fields and fail to reproduce deep convection (Figure S9 in Supporting Information S1). Although limited by the 0.25° resolution, SATcast successfully captures the spatiotemporal evolution of a large-scale convection system moving from northern China to Japan. It also accurately predicts two tropical disturbances east of the Philippines, which later intensified into Typhoon Sonca and Typhoon Nesat. SATcast reproduces the complex intensity fluctuations, showing an initial weakening followed by intensification, driven by interactions with cold eddies, all within the 24-hr period. Additional examples are provided in Figures S5–S7 in Supporting Information S1. Notably, SATcast achieves these high-fidelity predictions with only a single round of fine-tuning, effectively mitigating forecast error accumulation. This enables continuous forecasting without temporal discontinuities, suggesting SATcast as a reliable framework for convective cloud forecasting.

In addition, SATcast demonstrates strong generalization capability. Despite being trained exclusively on channel 12 data, it effectively predicts channel 9 by incorporating FuXi model outputs. Relevant metrics and examples are presented in Figures S14 and S15 in Supporting Information S1.

We also compared the results from various NWP forecasts. As shown in Figure 5, the reconstructed brightness temperature from ERA5 exhibits a significant magnitude discrepancies relative to observations, with notable deviations in the positioning of convective cores. We further reproduced the case presented in Figure 4 using the Weather Research and Forecasting Model (WRF) at 4-km resolution. Since Radiative Transfer for TIROS Operational Vertical Sounder (RTTOV) struggles to reconstruct clouds from WRF output because nonlinear processes and different assumptions in microphysics schemes (Xie et al., 2025), we instead compared outgoing long-wave radiation. Results (Figures S16 and S17 in Supporting Information S1) reveal substantial spatial misalignment, with WRF simulations initially show pronounced overestimation of cloud cover in the southwestern part of the domain, though this bias diminishes with longer simulation times. Overall, the performance of WRF remains inferior to that of SATcast, reinforcing the advantage of data-driven approaches for convective cloud forecasting.

3.2. Ablation Experiments and Short-Term Forecasting Skill

To evaluate the contributions of key components in SATcast, we also compare the skill of four variants of SATcast (see details in Table 1) in 24-hr prediction.

Figure 6 summarizes the spatially averaged metrics over the region of interest (86° to 150° E in longitude and 1° to 41° N in latitude), for the SATcast, and the four SATcast variants. Among all variants, SATcast-NoC, which directly predicts the entire 8-hr sequence, demonstrates the second-lowest performance from $T + 1$ hr to $T + 3$ hr, only surpassed by SATcast-NoS, the poorest model at early lead times. This is likely due to the backward propagation of errors from later frames ($T + 4$ hr and beyond), which hinders optimization of earlier predictions. Despite this, SATcast-NoC outperforms SATcast-NoF, which excludes FuXi forecasts—for lead times beyond $T + 4$ hr. SATcast-NoF performs well up to $T + 3$ hr, but deteriorates rapidly thereafter, likely due to its inability to model the physical processes underlying convective cloud evolution. The Figure 6 also shows that all variants except SATcast-NoS exhibited a clear temporal periodicity, characterized by consistent trends in the evaluation metrics over adjacent 4-hr or 8-hr intervals, likely due to the models' autoregressive use in 24-hr forecasts. Notably, only SATcast-NoF showed accelerated degradation toward the end of the forecast cycle. These findings suggest the value of incorporating FuXi forecasts for predictions at longer lead times, as satellite imagery alone proves insufficient for long-range predictions.

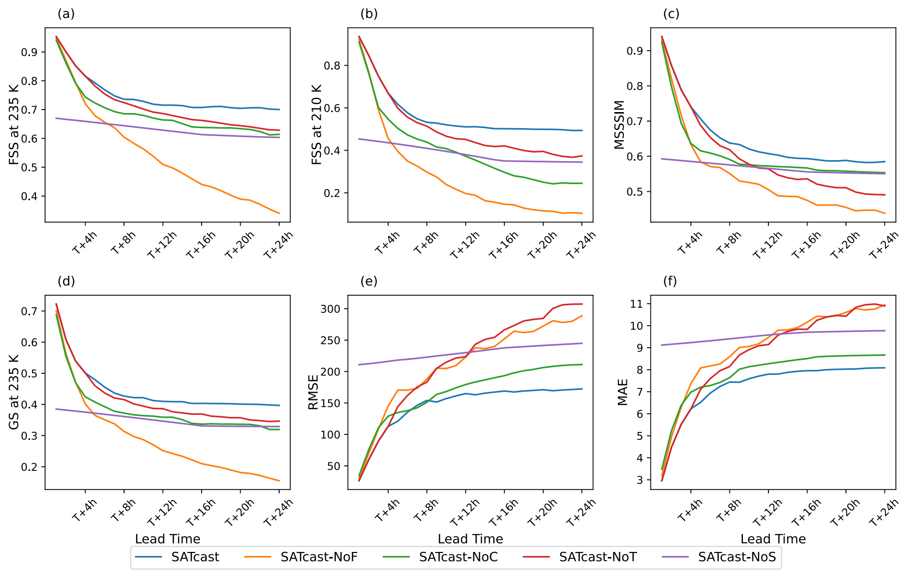

Figure 6. Comparison of (a) FSS at 235 K, (b) FSS at 210 K, (c) MSSSIM, (d) GS at 235 K, (e) RMSE, and (f) MAE, spatially averaged over the region from 86° to 150° E longitude and 1° to 41° N latitude in 8-hr forecasts, across six models: Persistence, SATcast, SATcast-NoC, SATcast-NoT, SATcast-NoF, SATcast-NoS, and PySTEPS. The results are based on testing data from September to December 2022. The thresholds used for FSS and GS are shown in the *y*-axis labels. Higher values for (a, b, c, and d) indicate better model performance, in contrast to (e) and (f) where lower values demonstrate higher skill.

Meanwhile, SATcast-NoS, which relies solely on FuXi outputs and generates satellite-like imagery from scratch, demonstrates substantially lower skill during the initial forecast hours. Nevertheless, its performance remains relatively stable over time. By $T + 12$ hr, the model's skill is close to the SATcast-NoC and SATcast-NoT. The similar performance among these variants suggests that SATcast gradually learns to approximate pseudo-geostationary satellite observations in later frames. In addition, SATcast-NoS can generate higher cloud intensities while other variants tend to underestimate the cloud at later frames (shown in Figure 7) so that SATcast-NoS has higher skill when the thresholds of the metrics are lower. Especially in Figure 6b), SATcast-NoS surpassed SATcast-NoC after $T + 12$hr.

To further examine the impact of fine-tuning on SATcast, we evaluate SATcast-NoT, which omits fine-tuning beyond $T + 4$ hr. It performs better than other variants except SATcast. While it performs comparably to SATcast in early lead times, SATcast-NoT increasingly overestimates brightness temperatures beyond $T + 4$ hr (Figure S13 in Supporting Information S1), confirming that fine-tuning is essential for aligning predictions with the distribution of real satellite imagery.

Overall, SATcast maintains better performance compared to other variants throughout the forecast period, showing the effectiveness of its cascade architecture and multimodal data input.

Although average metric differences among models appear subtle, visual comparisons in Figure 7 and Figures S10–S12 in Supporting Information S1 reveal substantial differences in spatial prediction quality. SATcast shows the best skill in capturing the evolution of two major tropical cyclones: the intensification of Typhoon Nanmadol and the dissipation of Typhoon Muifa on 13 September 2022. Muifa, one of the most powerful typhoons on record to strike Shanghai, and, Nanmadol, one of the strongest typhoons of 2022, caused substantial damage, underscoring the importance of accurate forecasts.

Starting at $T + 1$ hr (2022-09-14-07:00 UTC), SATcast and its variants capture the textural features and central position of Nanmadol, characterized by a low cloud-top temperature and distinct core region. However, significant differences emerge by $T + 3$ hr (2022-09-14-09:00 UTC). Without FuXi forecasts, SATcast-NoF demonstrates significant degradation in its ability to resolve the typhoon's structure, and SATcast-NoC produces only a loosely organized system. In contrast, SATcast maintains robust skill in delineating both the eye of the typhoon and its structural integrity By $T + 8$ hr (2022-09-14-14:00 UTC), these differences become more pronounced.

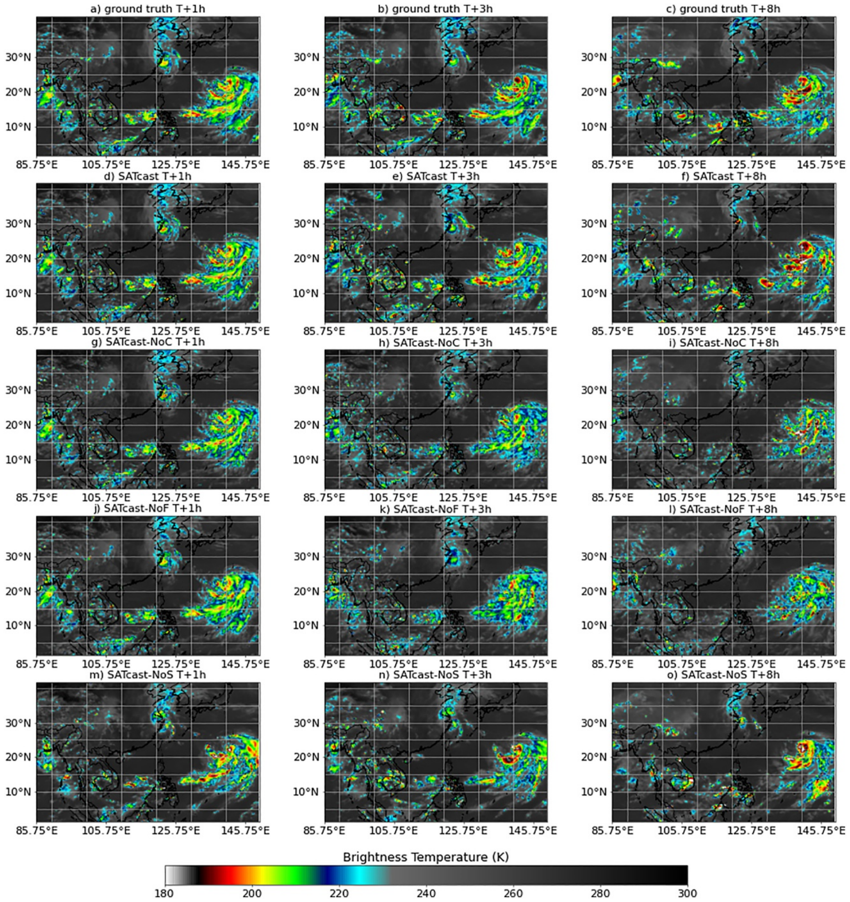

Figure 7. Spatial distribution of brightness temperature at $T + 1$ hr, $T + 3$ hr, and $T + 8$ hr from observation (a–c), SATcast (d–f), SATcast-NoC (g–i), SATcast-NoF (j–l), and SATcast-NoS predictions (m–o), where T denotes the starting time of the prediction, the UTC time of T is 2022-09-14-06:00.

SATcast-NoC and SATcast-NoF predict only fragmented and sporadic convection patterns, while SATcast successfully captures the intensification and movement of the tropical cyclone. Although SATcast exhibits some discrepancies in the detailed organization of convection compared to observations, it remains the most reliable model in maintaining the typhoon's overall structure and dynamics.

As Typhoon Muifa approached the east coast of China, it gradually weakened and split into two cloud clusters by $T + 3$ hr. SATcast demonstrated superior skill in accurately capturing both the spatial evolution and intensity variations during this process. By $T + 8$ hr, SATcast's predictions exhibit a slight southward bias in the location of convective clouds north of Shanghai compared to observations. Despite this, SATcast still accurately forecasts the variations in convection intensity and shape. In contrast, although SATcast-NoC and SATcast-NoF also predict a weakening system, their representations of convective cloud organization show significant morphological inaccuracies.

Notably, satellite observations indicate that scattered convective clouds over Southeast Asia and the South China Sea intensified over time. SATcast accurately predicts this strengthening at $T + 3$ hr, and its forecast at $T + 8$ hr remains acceptable. In comparison, both SATcast-NoC and SATcast-NoF significantly underestimate the intensity of these convective clouds at $T + 3$ hr and $T + 8$ hr. Additional cases, including those with and without tropical cyclones, are detailed in Figures S10–S12 in Supporting Information S1. In these cases, SATcast consistently outperforms its variants in preserving both the spatial organization and intensity of evolving cloud systems, while the variants often predict premature weakening or erroneous dispersion of convective systems.

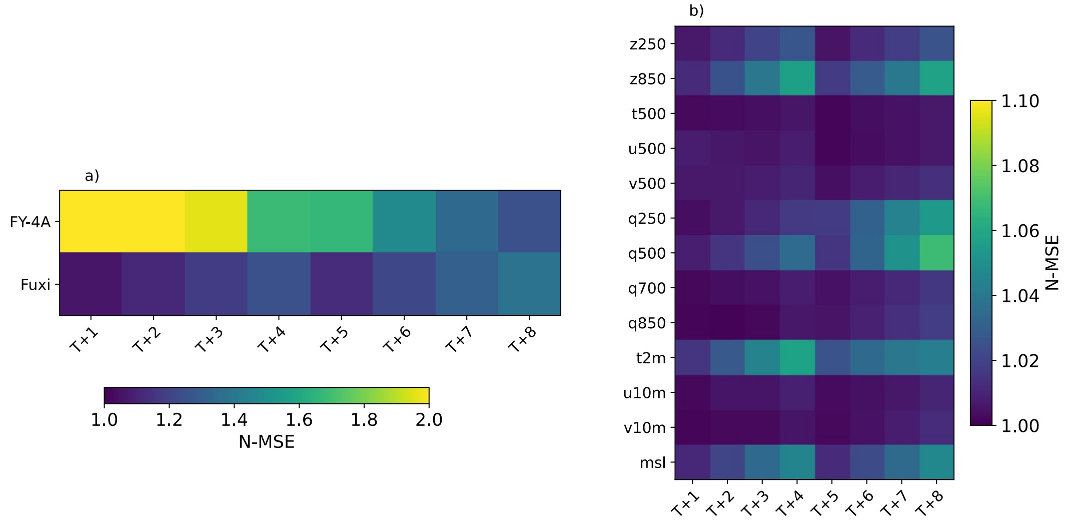

Figure 8. Heatmap of N-MSE for permutation feature importance. The *x*-axis represents input features, and the *y*-axis represents forecast lead time. Panels are divided into (a) past satellite images and the whole of FuXi variables and (b) specific variables of FuXi forecast. FY-4A denotes the satellite imagery, and FuXi represents all 13 variables shown in the right panel. The three-digits codes on the *y*-axis of the right figure indicate pressure levels: 250 hPa, 500 hPa, 700 hPa, and 850 hPa. Variable abbreviations are as follows: z (geopotential), t (temperature), u (u component of wind), v (v component of wind), q (specific humidity), t2m (temperature at 2 m), u10m (u component of wind at 10 m), v10 (v component of wind at 10 m), msl (mean sea-level pressure).

3.3. Physical Interpretations

Permutation feature importance is a valuable technique for interpreting multimodal deep learning models, especially in atmospheric science applications (Joshi et al., 2021; Lagerquist et al., 2021). This method quantifies the importance of each feature by selectively shuffling specific feature dimensions and measuring the resulting degradation in model performance. It helps identify key physical parameters for predicting satellite imagery and informs model refinement. To address the computational cost of repeated shuffling, we propose a threshold-based permutation strategy (detailed in Text S2 in Supporting Information S1), which performs a single shuffle per batch. This strategy significantly lowers computational costs while preserving interpretability.

Figure 8 presents heatmaps of the normalized mean squared error (N-MSE) ratios, defined as $\frac{\mathrm{MSE_f}}{\mathrm{MSE_m}}$, where $\mathrm{MSE_f}$ and $\mathrm{MSE_m}$ denote the permuted and baseline MSE, respectively. Higher N-MSE ratios correspond to greater feature importance, highlighting the most influential features for model predictions.

Figure 8a shows the feature importance, measured by N-MSE variations, for two major input sources: the eight-hour historical satellite imagery and FuXi predicted variables. For early lead times ($T + 1$ hr to $T + 3$ hr), permuting satellite imagery results in a larger increase in N-MSE than permuting FuXi variables, indicating the essential role of satellite observations in short-term forecasting. However, as lead time increases, the importance of satellite imagery gradually diminishes, while the contribution of FuXi data grows, ultimately surpassing that of satellite data by $T + 8$ hr. This shift reflects the growing importance of FuXi data, which provides insights into future atmospheric patterns, for predictions over longer lead times. It also explains why SATcast-NoC outperforms SATcast-NoF after $T + 3$ hr.

Figure 8b examines the importance of individual FuXi variables, all of which exhibit relatively low N-MSE values (below 1.1), with several parameters emerging as particularly impactful. Variables across multiple pressure levels, including upper (250 hPa), middle (500 and 700 hPa), and lower (850 hPa), are selected to capture the vertical structure of the atmosphere. Notably, the importance of these variables varies with forecast lead time. Key features generating N-MSE about 1.05 include geopotential at 250 and 850 hPa (z250 and z850), specific humidity at 250 and 500 hPa (q250 and q500), near-surface temperature (t2m), and mean sea-level pressure (msl). Geopotential fields (z250 and z850) influence atmospheric motions, such as large-scale convergence and divergence patterns, through geostrophic balance and adjustment, thereby guiding cloud movement and intensity changes. Mid- to upper-level specific humidity (q250 and q500), which closely correlates with clouds, becomes increasingly important from $T + 5$ hr to $T + 8$ hr, while low-level specific humidity (q850) has comparatively limited impact. Surface temperature and pressure (t2m and msl) contribute thermodynamic and dynamic forcing to the atmosphere, affecting prediction accuracy. Furthermore, msl and geopotential (z250 and z850) serve as practical indicators of convection centers and tropical cyclones. These findings demonstrate that incorporating critical atmospheric fields significantly enhances the SATcast's ability to predict the structure and evolution of convective systems.

We further evaluate SATcast's generalization ability by testing its performance on channel 9 of FY-4A (Figures S14 and S15 in Supporting Information S1). Notably, channel 9, which represents high-level water vapor, exhibit greater spatiotemporal stability than channel 12, leading to superior performance across various metrics compared to those from channel 12. This distinction can be attributed to the fact that, for shorter lead times, SATcast operates more like a video prediction model, heavily relying on satellite imagery for prediction. However, the model's performance on channel 9 deteriorates rapidly at longer lead times due to inconsistencies between the forecasting targets on channels 9 (high-level water vapor) and 12 (cloud). This mismatch hinders the model's ability to accurately interpret the influence of physical variables on the target, resulting in a significant reduction in skill for channel 9 predictions beyond $T + 3$ hr.

4. Conclusions

In this study, we introduce SATcast, a cascade, autoregressive, and multimodal deep learning model developed for convective cloud short-term forecasting. By integrating atmospheric physical conditions predicted by FuXi with FY-4A satellite observations, SATcast generates high-fidelity cloud forecasts using a diffusion-based architecture. The model effectively produces both the temporal evolution and spatial organization of cloud systems, producing skillful forecasts up to 24 hr in advance. SATcast partially addresses long-standing challenges in forecasting, such as image blurring and the rapid degradation of forecast accuracy over time. Nevertheless, a higher resolution forecasting model will be explored in future work to further enhance forecast details.

SATcast adopts a cascade framework, with SATcast-phase 1 providing forecasts up to 4 hours ahead, and SATcast-phase 2 takes SATcast-phase 1's outputs as input to extend predictions beyond the initial 4 hours. SATcast-phase 1 and SATcast-phase 2 are optimized for lead times of 1–4 hr and 5–8 hr, respectively. Despite being optimized for forecasts within short-term forecasts, SATcast demonstrates strong generalization capabilities beyond the optimized 8 hr, maintaining robustness and low cumulative error at 24-hr lead times. Moreover, the model generalizes well across different channels without significant performance degradation. The model's robust performance probably benefits from its incorporation of atmospheric physical information, the cascade structure, and the stability and generative power of the diffusion model.

Although the current version operates at a coarsened resolution of 0.25° to reduce computational demands, the results remain meaningful at this scale. For example, ECMWF provides open-access forecast data at a 0.25° resolution for selected pressure levels and variables for the most recent 4 days (see: https://www.ecmwf.int/en/forecasts/datasets/open-data), but these forecast data cannot be reliably used to reconstruct brightness temperature using Community Radiative Transfer Model or RTTOV or other radiative transfer models. Therefore, our results remain meaningful at a resolution of 0.25°, especially since SATcast can effectively reproduce the organization of cloud systems. Future work will explore latent-space diffusion models to mitigate these computational demands and enable forecasting at FY satellite's native 4-km resolution, allowing for finer-scale depiction of storm features. Another promising direction is the development of probabilistic forecasting. Convective systems are inherently uncertain, and deterministic forecasts cannot quantify uncertainty, which is critical for making informed, contingent decisions. The recent development of GenCast by Google Deepmind highlights the potential of diffusion models for producing ensemble forecasts by introducing additional stochasticity into the generation process (Price et al., 2024). A similar approach could be adopted to enable SATcast to produce more reliable probabilistic storm predictions, facilitating more informed and adaptive decision-making.

To our knowledge, SATcast is the first deep learning model capable of accurately forecasting satellite imagery beyond 8 hr and up to 24 hr. It can serve as a foundational model to forecast the spatial distribution of clouds across various environmental conditions. Downstream models could leverage SATcast forecasts for more targeted applications, such as typhoon and severe precipitation prediction. Importantly, SATcast can be deployed in regions that lack advanced observation networks and high-performance computing facilities, as it can be operated without relying on traditional NWP forecasts. SATcast requires only 3 GB of GPU memory and about 10 minutes to generate an eight-hour satellite image forecasts in an NVIDIA H800, offering substantial savings in time and computational resources compared with NWP models. While its inference speed is slightly slower than some deep learning models because of 1,200 denoising steps, SATcast allows faster training and fine-tuning on new data sets than recurrent neural networks or 3D convolutional networks (e.g., SimVP and Unet3D), and demonstrates greater stability than GAN-based approaches. This makes it especially valuable for vulnerable communities, remote regions, and maritime or aviation operations where ground-based radar data are unavailable. Moreover, accurate cloud forecasts are crucial for optimizing solar energy production and optimizing grid management, particularly in photovoltaic-based systems (Xia et al., 2024).

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The Fengyun-4A satellite imagery can be downloaded on https://satellite.nsmc.org.cn/portalsite/default.aspx?currentculture=en-US. The scripts about the network, training, and inference can be found on https://github.com/cd4tpcell/SATcast/tree/main (H. Chen et al., 2025). The PySTEPS library can be installed on https://github.com/pySTEPS/pysteps.

Acknowledgments

The work described in this paper was substantially supported by a grant from the Research Grants Council (RGC) of the Hong Kong Special Administrative Region, China (Project Reference Number: AoE/P-601/23-N) and the Center for Ocean Research in Hong Kong and Macau (CORE), a joint research center between the Laoshan Laboratory and the Hong Kong University of Science and Technology (HKUST). HC, YH, and XS are additionally supported by RGC Grant HKUST-16301322 and the Aviation Research and Development Project Phase 2 (AvRDP-2) of the World Meteorological Organization (WMO). The authors thank the computing platform supported by HKUST Superpod Cluster to train the model and the Artificial Intelligence Innovation and Incubation Institute of Fudan University for providing archives of the FuXi data.

References

Andrychowicz, M., Espeholt, L., Li, D., Merchant, S., Merose, A., Zyda, F., et al. (2023). Deep learning for day forecasts from sparse observations. ArXiv, abs/2306.06079. Retrieved from https://api.semanticscholar.org/CorpusID:259129311

Baker, S., & Matthews, I. (2004). Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, *56*(3), 221–255. https://doi.org/10.1023/b:visi.0000011205.11775.fd

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3

Çağlar, K., Giannakos, A., Schneider, S., & Jann, A. (2024). Transformer-based nowcasting of radar composites from satellite images for severe weather. *Artificial Intelligence for the Earth Systems*, *3*(2), e230041. https://doi.org/10.1175/AIES-D-23-0041.1

Chen, H., Shi, X., Nie, X., Wang, Y., Leung, C. Y. Y., Cheung, P., & Chan, P. W. (2024). Tropical aviation turbulence induced by the interaction between a jet stream and deep convection. *Journal of Geophysical Research: Atmospheres*, *129*(18), e2024JD040763. https://doi.org/10.1029/2024JD040763

Chen, H., Zhong, X., Zhai, Q., Li, X., Chan, Y. W., Chan, P. W., et al. (2025). Skillful nowcasting of mesoscale convective clouds with a cascade diffusion model. *Zenodo*. https://doi.org/10.5281/zenodo.14643153

Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023). Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, *6*(1), 190. https://doi.org/10.1038/s41612-023-00512-1

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, *34*, 8780–8794.

Ehsani, M. R., Zarei, A., Gupta, H. V., Barnard, K., & Behrangi, A. (2021). Nowcasting-nets: Deep neural network structures for precipitation nowcasting using imerg. arXiv preprint arXiv:2108.06868.

Feng, Z., Leung, L. R., Liu, N., Wang, J., Houze Jr, R. A., Li, J., et al. (2021). A global high-resolution mesoscale convective system database using satellite-derived cloud tops, surface precipitation, and tracking. *Journal of Geophysical Research: Atmospheres*, *126*(8), e2020JD034202. https://doi.org/10.1029/2020JD034202

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622

Guo, Y., Zhong, M., Chen, X., Zhou, Z., Xu, G., Xu, G., & Dong, L. (2022). A thunderstorm gale forecast method based on the objective classification and continuous probability. *Atmosphere*, *13*(8), 1308. https://doi.org/10.3390/atmos13081308

Hendrycks, D., & Gimpel, K. (2023). Gaussian error linear units (gelus). Retrieved from https://arxiv.org/abs/1606.08415

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, *33*, 6840–6851.

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, G., Sun, Y., Liu, Z., Sedra, D., & Weinberger, K. (2016). *Deep networks with stochastic depth*. Springer international publishing.

Joshi, G., Walambe, R., & Kotecha, K. (2021). A review on explainability in multimodal deep neural nets. *IEEE Access*, *9*, 59800–59821. https://doi.org/10.1109/access.2021.3070212

Kim, H., Ham, Y. G., Joo, Y. S., & Son, S. W. (2021). Deep learning for bias correction of mjo prediction. *Nature Communications*, *12*(1), 3087. https://doi.org/10.1038/s41467-021-23406-3

Kim, W., Jeong, C.-H., & Kim, S. (2024). Improvements in deep learning-based precipitation nowcasting using major atmospheric factors with radar rain rate. *Computers & Geosciences*, *184*, 105529. https://doi.org/10.1016/j.cageo.2024.105529

Kumar, A., Islam, T., Sekimoto, Y., Mattmann, C., & Wilson, B. (2020). Convcast: An embedded convolutional lstm based architecture for precipitation nowcasting using satellite data. *PLoS One*, *15*(3), e0230114. https://doi.org/10.1371/journal.pone.0230114

Kumar, K. (2004). Forecast verification: A practitioner's guide in atmospheric sciences. *Journal of the Royal Statistical Society - Series A: Statistics in Society*, *168*(1), 255. https://doi.org/10.1111/j.1467-985x.2004.00347_9.x

Lagerquist, R., Stewart, J. Q., Ebert-Uphoff, I., & Kumler, C. (2021). Using deep learning to nowcast the spatial coverage of convection from himawari-8 satellite data. *Monthly Weather Review*, *149*(12), 3897–3921.

Lin, S., Liu, B., Li, J., & Yang, X. (2024). Common diffusion noise schedules and sample steps are flawed. Retrieved from https://arxiv.org/abs/2305.08891

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 11976–11986).

Loshchilov, I. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lu, F., Zhang, X.-H., Chen, B.-Y., Liu, H., Wu, R., Han, Q., et al. (2017). Fy-4 geostationary meteorological satellite imaging characteristics and its application prospects. *J. Mar. Meteorol*, *37*(2), 1–12.

Mittermaier, M., & Roberts, N. (2010). Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Weather and Forecasting*, *25*(1), 343–354. https://doi.org/10.1175/2009WAF2222260.1

Morales-Brotons, D., Vogels, T., & Hendrikx, H. (2024). Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*.

Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International conference on machine learning* (pp. 8162–8171).

Organization, W. M. (2023). Guidelines for satellite-based nowcasting in Africa — Library.wmo.int. Retrieved from https://library.wmo.int/records/item/58348-guidelines-for-satellite-based-nowcasting-in-africa

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., et al. (2017). Automatic differentiation in pytorch. In *Nips 2017 workshop on autodiff*.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2024). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Pulkkinen, S., Nerini, D., Pérez Hortal, A. A., Velasco-Forero, C., Seed, A., Germann, U., & Foresti, L. (2019). Pysteps: An open-source python library for probabilistic precipitation nowcasting (v1.0). *Geoscientific Model Development*, *12*(10), 4185–4219. https://doi.org/10.5194/gmd-12-4185-2019

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). Learning transferable visual models from natural language supervision.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002203. https://doi.org/10.1029/2020ms002203

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, *136*(1), 78–97. https://doi.org/10.1175/2007MWR2123.1

Roca, R., Fiolleau, T., & Bouniol, D. (2017). A simple model of the life cycle of mesoscale convective systems cloud shield in the tropics. *Journal of Climate*, *30*(11), 4283–4298. https://doi.org/10.1175/JCLI-D-16-0556.1

Rui Wang, A. K. H. L., Fung, J. C. H., & Lau, A. K. H. (2024). Skillful precipitation nowcasting using physical-driven diffusion networks. *Geophysical Reasearch Letter*, *51*(24), e2024GL110832. https://doi.org/10.1029/2024gl110832

Schmit, T. J., Griffith, P., Gunshor, M. M., Daniels, J. M., Goodman, S. J., & Lebair, W. J. (2017). A closer look at the abi on the goes-r series. *Bulletin of the American Meteorological Society*, *98*(4), 681–698. https://doi.org/10.1175/bams-d-15-00230.1

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., & Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, *28*.

Shi, X., & Wang, Y. (2022). Impacts of cumulus convection and turbulence parameterizations on the convection-permitting simulation of typhoon precipitation. *Monthly Weather Review*, *150*(11), 2977–2997. https://doi.org/10.1175/MWR-D-22-0057.1

Smith, J., Birch, C., Marsham, J., Peatman, S., Bollasina, M., & Pankiewicz, G. (2024). Evaluating pysteps optical flow algorithms for convection nowcasting over the maritime continent using satellite data. *Natural Hazards and Earth System Sciences*, *24*(2), 567–582. https://doi.org/10.5194/nhess-24-567-2024

Tran, Q.-K., & Song, S.-k. (2019). Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks. *Atmosphere*, *10*(5), 244. https://doi.org/10.3390/atmos10050244

Trier, S. B., Davis, C. A., Ahijevych, D. A., & Manning, K. W. (2014). Use of the parcel buoyancy minimum (b min) to diagnose simulated thermodynamic destabilization. Part I: Methodology and case studies of mcs initiation environments. *Monthly Weather Review*, *142*(3), 945–966. https://doi.org/10.1175/mwr-d-13-00272.1

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Voleti, V., Jolicoeur-Martineau, A., & Pal, C. (2022). Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, *35*, 23371–23385.

Wallace, J. M. (1975). Diurnal variations in precipitation and thunderstorm frequency over the conterminous United States. *Monthly Weather Review*, *103*(5), 406–419. https://doi.org/10.1175/1520-0493(1975)103<0406:dvipat>2.0.co;2

Wang, R., Su, L., Wong, W. K., Lau, A. K., & Fung, J. C. (2023). Skillful radar-based heavy rainfall nowcasting using task-segmented generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–13. https://doi.org/10.1109/tgrs.2023.3295211

Wang, X., Zhao, H., Zhang, G., Guan, Q., & Zhu, Y. (2024). Spatiotemporal predictive learning for radar-based precipitation nowcasting. *Atmosphere*, *15*(8), 914. https://doi.org/10.3390/atmos15080914

Wang, Z., Simoncelli, E., & Bovik, A. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003* (Vol. 2, p. 1398–1402). https://doi.org/10.1109/ACSSC.2003.1292216

Wei, X., Zhang, F., Zhang, R., Li, W., Liu, C., Guo, B., et al. (2024). Dayu: Data-driven model for geostationary satellite observed cloud images forecasting. Retrieved from https://arxiv.org/abs/2411.10144

Wu, Y., & He, K. (2018). Group normalization. In *Proceedings of the European conference on computer vision (eccv)* (pp. 3–19).

Xia, P., Zhang, L., Min, M., Li, J., Wang, Y., Yu, Y., & Jia, S. (2024). Accurate nowcasting of cloud cover at solar photovoltaic plants using geostationary satellite images. *Nature Communications*, *15*(1), 510. https://doi.org/10.1038/s41467-023-44666-1

Xie, Q., Li, D., Yang, Y., Zhao, Y., Li, H., Zhu, S., & Pan, X. (2025). Exploring the assimilation of all-sky fy-4a giirs radiances and its forecasts for binary typhoons. *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *18*, 5949–5959. https://doi.org/10.1109/jstars.2025.3540209

Yan, C., Guang, J., Li, Z., de Leeuw, G., & Chen, Z. (2024). A study on typhoon center localization based on an improved spatio-temporally consistent scale-invariant feature transform and brightness temperature perturbations. *Remote Sensing*, *16*(21), 4070. https://doi.org/10.3390/rs16214070

Yang, J., Zhang, Z., Wei, C., Lu, F., & Guo, Q. (2017). Introducing the new generation of Chinese geostationary weather satellites, fengyun-4. *Bulletin of the American Meteorological Society*, *98*(8), 1637–1658. https://doi.org/10.1175/bams-d-16-0065.1

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., & Wang, J. (2023). Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, *619*(7970), 526–532. https://doi.org/10.1038/s41586-023-06184-4