CasCast: Skillful High-resolution Precipitation Nowcasting via Cascaded Modelling

Junchao Gong [1 2] **Lei Bai** [2] **Peng Ye** [2] **Wanghan Xu** [2] **Na Liu** [3] **Jianhua Dai** [4] **Xiaokang Yang** [1] **Wanli Ouyang** [2]

Abstract

Precipitation nowcasting based on radar data plays a crucial role in extreme weather prediction and has broad implications for disaster management. Despite progress has been made based on deep learning, two key challenges of precipitation nowcasting are not well-solved: (i) the modeling of complex precipitation system evolutions with different scales, and (ii) accurate forecasts for extreme precipitation. In this work, we propose CasCast, a cascaded framework composed of a deterministic and a probabilistic part to decouple the predictions for mesoscale precipitation distributions and small-scale patterns. Then, we explore training the cascaded framework at the high resolution and conducting the probabilistic modeling in a low dimensional latent space with a frame-wise-guided diffusion transformer for enhancing the optimization of extreme events while reducing computational costs. Extensive experiments on three benchmark radar precipitation datasets show that CasCast achieves competitive performance. Especially, CasCast significantly surpasses the baseline (up to **+91.8%**) for regional extreme-precipitation nowcasting. The code is available at `https://github.com/OpenEarthLab/CasCast`.

1. Introduction

Precipitation nowcasting based on weather radar data plays a vital role in predicting local weather conditions over a period of up to two hours (CLIMA & TE). These predictions are essential for various social sectors, including energy management and traffic scheduling. Moreover, precipitation forecasting serves as a critical tool for warning and mitigating disasters such as heavy rainfall and flooding. As a result, achieving skillful precipitation nowcasting is of utmost importance and has gained significant attention from researchers.

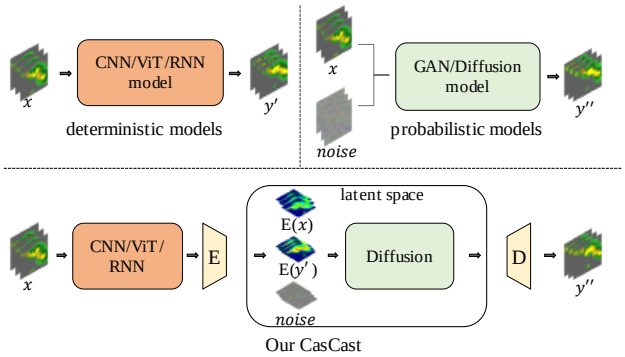

Figure 1. Different precipitation nowcasting pipelines. Given context x, the prediction of deterministic models and probabilistic models are y' and y'', respectively. Our CasCast generates y'' conditional on x and y' in the latent space.

Most precipitation events are caused by a combination of atmospherical systems with different scales (Prein et al., 2023). On one hand, the mesoscale precipitation system evolves over spatial ranges of tens to hundreds of kilometers and time scales of several hours, driven and constrained by relatively stable large-scale circulation. On the other hand, the small-scale system, evolving within a range of a few kilometers and operating on time scales of minutes, is influenced by local processes such as heating, surface features, and other physical factors, which introduce stochasticity and unpredictability into its behavior. Thus, the combined influence of weather systems at multiple scales makes accurate prediction of precipitation events very challenging.

Deep learning methods have made significant contributions to precipitation nowcasting, with two broad categories emerging, i.e., deterministic models and probabilistic models. Deterministic models aim to capture the overall motion of the middle-scale precipitation system by providing a single-value prediction of the future state. However, these methods often encounter challenges related to blurriness and

[1]Shanghai Jiao Tong University [2]Shanghai AI Laboratory [3]National Meteorological Information Center [4]Shanghai Meteorological Service. Correspondence to: Lei Bai <baisanshi@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

lack of fine-grained details (Shi et al., 2015; Wang et al., 2017; Guen & Thome, 2020; Gao et al., 2022b;a). The reason is that deterministic models approximate the strong randomness, exhibited in the small-scale systems, by predicting mean values (Ravuri et al., 2021). In contrast, probabilistic models use sampling from different latent variables to express the stochasticity of future weather systems, enabling them to capture small-scale weather phenomena (Ravuri et al., 2021; Zhang et al., 2023; Gao et al., 2023). However, probabilistic methods may encounter challenges in accurately forecasting the large-scale distribution of precipitation that exhibits slow migration and high predictability, due to the stochastic modeling. In summary, previous studies still struggle to predict the evolution of both the mesoscale and the small-scale systems simultaneously.

Another vital requirement of precipitation nowcasting is the forecasts for extreme precipitation events. Over the past 50 years, extreme-precipitation events have caused more than 1.01 million related deaths, and the economic losses are beyond US$ 2.84 trillion (Association et al., 2021). As extreme weather processes typically exhibit short lifetimes of tens of minutes and involve convective-scale features at the kilometer scale (Pulkkinen et al., 2019; Ravuri et al., 2021; Zhang et al., 2023), effective forecasts for extreme precipitation require predicting the small-scale systems' evolutions at high resolution. Regrettably, existing methods have given less attention to this critical requirement.

In this work, we aim to achieve skillful high-resolution precipitation nowcasting with accurate global precipitation motion, realistic local patterns, and considerable regional extreme forecastings. We propose a novel cascaded nowcasting framework called CasCast, which disentangles precipitation nowcasting into predictions of mesoscale systems and generation of small-scale systems through a cascaded manner. As shown in Figure 1, CasCast initially predicts the evolution of the mesoscale precipitation system with a deterministic model. Then, conditioned on the prediction of global precipitation distribution, a generative model is used for the generation of small-scale weather phenomena. Specifically, we choose diffusion models as the generative component of CasCast, as it does not suffer from mode collapse and artifacts commonly observed in GANs (Gao et al., 2023). Moreover, in order to enable the prediction of extreme precipitation events, we train the cascaded framework at high resolution, and further develop a frame-wise-guided diffusion transformer in a low-dimensional latent space. This frame-wise guidance in diffusion transformer ensures a frame-to-frame correspondence between blurry predictions and latent vectors, resulting in better optimization for the generation of small-scale patterns. The low-dimensional latent space reduces the computational cost caused by the high-dimensional characteristic of the radar data. We test our CasCast on three classical radar precipitation datasets.

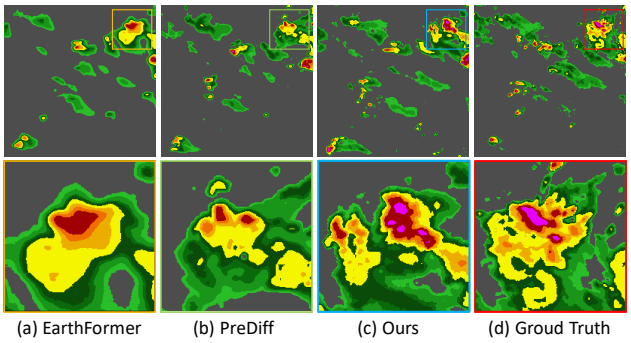

Figure 2. Prediction visualization of different methods for a lead time of 60 minutes. The prediction of EarthFormer (Gao et al., 2022a) lacks small-scale patterns. The prediction of PreDiff (Gao et al., 2023) has a lower regional extreme value. In contrast, our CasCast method effectively captures both local patterns and regional extreme values.

On these datasets, our method achieves SoTA (state-of-the-art) and significantly improves the predictions for regional extreme precipitation nowcasting. As shown in Figure 2, our method captures both the small-scale precipitation phenomena and forecasts a regional extreme event.

Our contributions can be summarized as follows:

- To predict the evolution of both mesoscale and the small-scale precipitation systems simultaneously, we introduce a novel cascaded framework (CasCast), decoupling precipitation nowcasting into the prediction of high-deterministic part and the generation of component with strong stochasticity.
- We propose to train the cascaded framework at high resolution and generate forecasts with a frame-wise guided diffusion transformer (CasFormer) in a low-dimensional latent space, boosting the skill for extreme precipitation nowcasting.
- We conduct extensive experiments on three benchmark datasets, demonstraing the superior performance and robustness of the proposed method, especially for the extreme regional precipitation, e.g., outperforming the baseline model up to **+91.8%** on the SEVIR dataset.

2. Related Work

2.1. Precipitation Nowcasting with Deep Learning

Deep learning methods (Han et al., 2024; Chen et al., 2023) for precipitation nowcasting can be classified into deterministic and probabilistic. ConvLSTM (Shi et al., 2015) leverages convolution layers and LSTM (Long Short-Term Memory) cells to respectively extract spatial and temporal features of precipitation for deterministic forecasting. PredRNN enhances the capability of spatial-temporal modeling

by splitting long-term and short-term memory cells (Wang et al., 2017). PhyDNet decomposes the prediction into random motion and pde-guided motion, aiming to generate predictions that are consistent with physics (Guen & Thome, 2020). Other deterministic models forecast short-term precipitation in a non-recurrence way by efficient space-time cuboid attention (Gao et al., 2022a) or simple CNN (Gao et al., 2022b). However, the predictions of deterministic models are blurred, as deterministic loss functions can drive the model to predict an average of multiple uncertain future small-scale precipitation. To predict small-scale weather phenomena, DGMR (Ravuri et al., 2021) and NowcastNet (Zhang et al., 2023) are trained adversarially, constraining the nowcasting to be close to the real precipitation distribution with spatial and temporal discriminators. However, GANs are known to suffer from training instability, making them prone to issues such as mode collapse and the generation of artifacts. Recent works (Gao et al., 2023; Zhao et al., 2024; Srivastava et al., 2023; Mardani et al., 2024) adopt diffusion to avoid shortages of GANs. However, due to the sensitivity of diffusion to model architectures and hyperparameters in high-dimensional spaces, as well as the computational cost of iterative sampling (Chen, 2023; Hoogeboom et al., 2023), recent diffusion-based precipitation methods are limited to low-resolution.

2.2. Diffusion Models for High-Resolution Generation

To achieve high-fidelity and detailed generation results, many studies have employed diffusion methods to generate high-resolution images or videos. LDM (Rombach et al., 2022) leverages a pre-trained autoencoder to compress non-semantic information in images to obtain low-dimensional latent representations and then conducts training and sampling in a lower-dimensional latent space to avoid the expensive cost of iterative sampling and the instability in generating high-quality results in high-dimensional spaces. In addition to training diffusion models in the latent space for high-resolution generation, there is also research exploring the direct generation of high-resolution images in pixel space. IMAGEN (Saharia et al., 2022) achieves excellent generation results by utilizing a multi-resolution cascaded generation approach. Other models achieve high-resolution generation through sophisticated network designs and noise strategies (Ho et al., 2022; Hoogeboom et al., 2023; Gupta et al., 2023; Lu et al., 2024a;b; Ling et al., 2024; Liu et al., 2024; Hu et al., 2024). However, these methods do not specifically focus on short-term forecasting. They either do not take into account the characteristics of the radar echoes or lack decoupled modeling of the multiscale physical processes involved in precipitation.

3. Method

3.1. Task Formulation

Precipitation nowcasting is commonly formalized as spatiotemporal forecasting problems of radar echoes (Shi et al., 2015; 2017; Veillette et al., 2020). Given the radar echo observations $x_{0:T} \in R^{T,C,H,W}$ from the past T time steps as the initial condition, short-time precipitation forecasting models' objective is to predict the radar echoes $y_{T:T'} \in R^{T'-T,C,H,W}$ of the future $T' - T$ frames. H and W represent the coverage range of radar data. The channel dimension C, with a value of 1, indicates the intensity of radar echoes.

3.2. CasCaded Modelling

As the evolution of most precipitation events is influenced by weather systems with different levels of randomness, we decouple the predictions of precipitation into the deterministic part and the probabilistic part in a cascaded manner as shown in Figure 3. The deterministic model takes the previous observations $x_{0:T}$ as inputs to predict single-valued future precipitation $y'_{T:T'}$, formulated as $y'_{T:T'} = \arg\max_y p_{\theta_d}(y_{T:T'}|x_{0:T})$, where θ_d is the parameter of deterministic part of CasCast. In CasCast, the architecture of the deterministic component can include most networks such as RNNs (Shi et al., 2015), CNNs (Gao et al., 2022b) and transformers (Gao et al., 2022a). We use MSE (mean square error) loss to train the deterministic model as

$$L_{\theta_d}(y_{T:T'}, y'_{T:T'}) = ||y_{T:T'} - y'_{T:T'}||^2. \quad (1)$$

The prediction $y'_{T:T'}$ suffers from blurriness of local details over time, because the deterministic loss drives models to predict the evolution of the samll-scale precipitation system into a mean value to represent the underlying uncertainty (Ravuri et al., 2021). However, $y'_{T:T'}$ effectively represents the global distribution of precipitation, because the mesoscale systems, dominating global patterns of the precipitation, are of high determinism. Then, the blurry prediction $y'_{T:T'}$ and initial condition $x_{0:T}$ are used as conditions for the generation of small-scale precipitation systems. The generative part samples latent Gaussian vector z to capture the small-scale weather phenomena, described by

$$z \sim N(0,1), y''_{T:T'} = \arg\max_y p_{\theta_p}(y|y'_{T:T'}, x_{0:T}, z) \quad (2)$$

where θ_p denotes the parameters of probabilistic models. Compared with pure probabilistic modelling ($y'' = \arg\max_y p_\theta(y|x_{0:T})$), the cascaded prediction avoids forecasting the evolution of highly deterministic parts in precipitation systems in a probabilistic manner, which contributes to the higher accuracy. Totally, CasCast disentangles the high-resolution precipitation nowcasting into two simpler

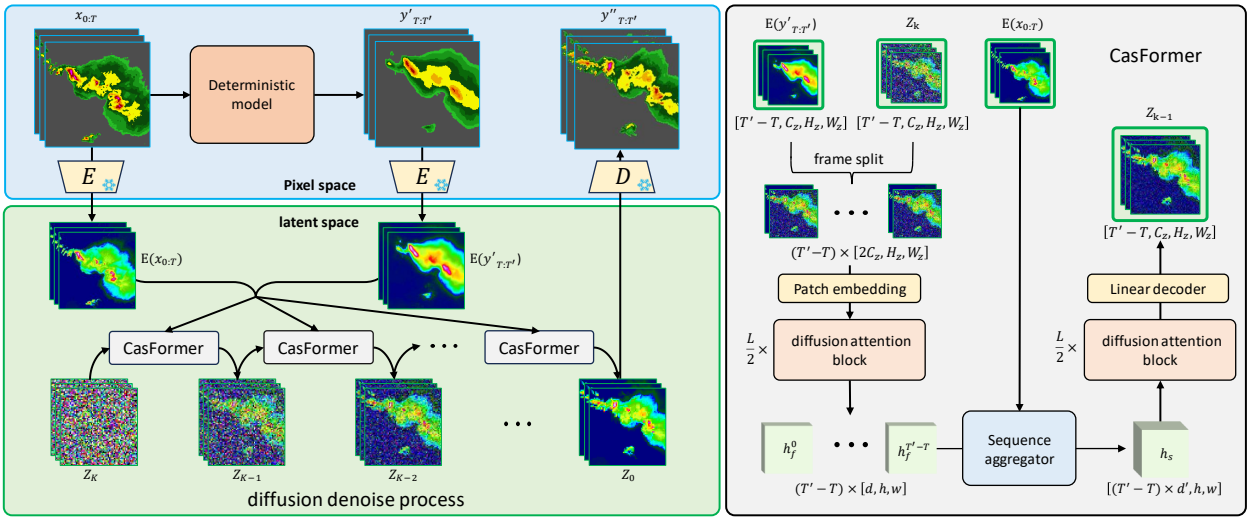

Figure 3. **Left**: Overview of our CasCast. First, CasCast employs a deterministic model in pixel space to generate the blur prediction $y'_{T:T'}$ from previous observations $x_{0:T}$. Then, $x_{0:T}$ and $y'_{T:T'}$ are encoded into latent representations $\mathrm{E}(x_{0:T})$ and $\mathrm{E}(y'_{T:T'})$ by a pretrained frame-wise encoder E. Last, conditioned on $\mathrm{E}(x_{0:T})$ and $\mathrm{E}(y'_{T:T'})$, the final prediction is generated through the diffusion denoise process on a novel CasFormer, and decoded back to the pixel space by a pretrained decoder. **Right**: Illustration of our CasFormer. First, $\mathrm{E}(y'_{T:T'})$ and the latent vector z_k are split into framewise inputs and processed by patch embedding and $\frac{L}{2}$ layers of diffusion attention block. Then, frame-wise features $h_f^0 \dots h_f^{T'-T}$ and $\mathrm{E}(x_{0:T})$ are combined to the sequence-wise feature h_s via a sequence aggregator, and used to predict the latent vector z_{k-1}.

tasks: predicting the global distribution of future precipitation and generating local weather patterns based on the blurry prediction.

3.3. High-resolution Generation for Extreme Events

The probabilistic component of CasCast is used to improve the predictions for extreme precipitation events. Specifically, we apply a frame-wise guided diffusion transformer, which is conditioned by blurry predictions, to generate high-resolution forecasts with local patterns in a low-dimensional latent space.

We take diffusion as the probabilistic part for its stable training process and remarkable generation ability. Diffusion models learn the reverse process of gradually noising data x_0 into Gaussian noise. By progressively denoising the Gaussian noise, diffusion models generate samples $\hat{x} \sim p(x_0)$ where $p(x_0)$ is data distribution. The noising process is a Markov process defined as $q(x_k|x_{k-1}) = N(x_k; \sqrt{\alpha_k}x_{k-1}, \beta_k I), 1 \leq k \leq K$ where $\beta_k = 1 - \alpha_k$ and α_k is the kth step's constant coefficient of noise schedule. When $x_0 \sim p(x_0)$ is given, the distribution of x_k can be derived as $q(x_k|x_0) = N(x_k; \sqrt{\bar{\alpha}_k}x_0, \sqrt{1-\bar{\alpha}_k}I)$. In the denoising process, the joint distribution $p_\theta(x_{0:K})$ is factorized into multiplications of conditional distribution $p_\theta(x_{0:K}) = p(x_K)\prod_{i=K}^{i=1} p_\theta(x_{i-1}|x_i)$. Diffusion models, parameterized as θ, iteratively estimate x_{k-1} from $p_\theta(x_{k-1}|x_k)$ using MAE (maximum likelihood estimation). Finally, Gaussian noise x_K is denoised into data sample x_0.

Framewise Autoencoder. It is difficult to directly apply diffusion models for precipitation nowcasting, because radar data, typically spanning several hours with a short time interval and covering areas within a radius of hundreds of kilometers at a high resolution (1km) (Shi et al., 2017; Veillette et al., 2020; Gwennaelle et al., 2020), exist in a high-dimensional space. Another characteristic of radar data is redundancy. Radar echo usually contains redundant background intensity as the situation, when precipitation is continuous and consistent over an area of several hundred kilometers, is rare. The redundancy and high dimensionality exhibited in radar data motivate us to use a framewise autoencoder for the compression of radar data. We follow (Rombach et al., 2022) to train a framewise autoencoder with pixel-wise loss and adversarial loss. Specifically, frame encoder E is trained to encode radar echo frame $x \in R^{1\times C\times H\times W}$ into latent representation $\mathrm{E}(x) \in R^{1\times C_z\times H_z\times W_z}$. The decoder reconstructs data frame $\hat{x} = \mathrm{D}(\mathrm{E}(x))$ from the latent representation. The latent space accelerates the training and sampling of diffusion models without severely corrupting the accuracy of prediction.

Framewise Guided Diffusion Transformer. We introduce CasFormer for the cascaded generation of precipitation nowcasting. As the generation is conditioned by the blurry deterministic prediction $y'_{T:T'}$, there is a frame-by-frame correspondence between $y'_{T:T'}$ and the final prediction $y''_{T:T'}$. Motivated by such correspondence, Our CasFormer generates local details with frame-wise guidance.

Specifically, in the process of denoising latent vector z_k to z_{k-1}, CasFormer performs frame splitting on the latent representation $\mathrm{E}(y'_{T:T'})$ of blur prediction and latent vector z_k. During frame splitting, z_k and $\mathrm{E}(y'_{T:T'})$ are first split along dimension T into independent frame-wise $\mathrm{E}(y'_{T:T'})^j \in R^{C_z,H_z,W_z}$ and $z_k^j \in R^{C_z,H_z,W_z}$ where j indicates the jth frame in the prediction sequence. Then $\mathrm{E}(y'_{T:T'})^j$ and z_k^j are concatenated along dimension C, resulting frame-wise input $\in R^{2C_z,H_z,W_z}$. The frame-wise input is utilized to extract the frame-wise feature $h_f^j \in R^{d,h,w}$ with a patch embedding layer and $\frac{L}{2}$ layers of diffusion attention block (Peebles & Xie, 2023). Frame-wise features $h_f^0 \dots h_f^{T'-T}$ of the same prediction sequence are aggregated into the sequence-wise feature $h_s \in R^{(T'-T)\times d',h,w}$ by a sequence aggregator. In the sequence aggregator, frame-wise features are combined as $h_s = MLP(concat(h_f^0 \dots h_f^{T'-T}))$, and h_s are injected with information from observations $\mathrm{E}(x_{0:T})$ in the latent space by a cross attention block. In the decoding stage, h_s are decoded into the latent vector z_{k-1} of the $(k-1)^{th}$ step. Overall, CasFormer has a frame-wise encoding stage and a sequence-wise decoding stage. The frame-wise encoding provides better-matched conditions for each frame-wise latent vector, reducing the complexity of the denoising conditioned by a sequence of blurry predictions. Sequence-wise decoding utilizes the sequence features from the sequence-aggregator to ensure the spatiotemporal consistency of precipitation nowcasting.

The training objective of CasFormer is the noise prediction loss (Ho et al., 2020):

$$L_{\theta_p}(z^y, z^{cond}) = \mathbb{E}_{\epsilon,k}[||\epsilon - \epsilon_{\theta_p}(z_k, k, z_{cond})||_2^2] \quad (3)$$

where $\epsilon \sim N(0,1)$, $k \sim U(0,1)$, $z^y = \mathrm{E}(y_{T:T'})$, $z_k = \sqrt{\bar{\alpha_k}} z^y + \sqrt{1-\bar{\alpha_k}}\epsilon$ and $z_{cond} = [\mathrm{E}(y'_{T:T'}), \mathrm{E}(x_{0:T})]$.

4. Experiment

4.1. Experimental Setting

4.1.1. Datasets

To validate the ability of CasCast to generate skillful 1km-resolution precipitation, we conducted tests on three radar echo datasets including SEVIR (Veillette et al., 2020), HKO-7 (Shi et al., 2017) and MeteoNet (Gwennaelle et al., 2020).

SEVIR. SEVIR (Veillette et al., 2020) is an Earth observation dataset that contains weather radar observations. The NEXRAD radar mosaic of Vertically Integrated Liquid (VIL) in SEVIR can be used for short-term precipitation forecasting. VIL in SEVIR records some storm events and random events that occurred in the United States between 2017 and 2019. The coverage range of VIL is 384km x 384km with a resolution of 1km and a time interval of 5min. We follow (Gao et al., 2022a) to split SEVIR into 35718 training samples, 9060 validation samples, and 12159 test samples. We benchmark the nowcasting of precipitation by predicting the future VIL for up to 60 minutes (12 frames), based on a 65-minute context of VIL (13 frames).

Table 1. The detailed settings of different datasets. Interval refers to the time duration between consecutive radar frames. The model predicts L_{out} frames conditioned on L_{in} frames.

dataset	N_{train}	N_{val}	N_{test}	$resolution$	size	interval	L_{in}	L_{out}
SEVIR	35,718	9,060	12,159	1km	384	5min	13	12
HKO7	8,772	492	1,152	2km	480	6min	10	10
MeteoNet	6,978	2,234	994	0.01°	400	5min	12	12

HKO-7. In HKO-7, the radar CAPPI reflectivity images have a resolution of 480×480 pixels and are taken from an altitude of 2km and cover a 512km × 512km area centered in Hong Kong (Shi et al., 2017). The radar data are collected from 2009 to 2015 with a time interval of 6 minutes. We benchmark HKO-7 by predicting the future radar echo up to 60 minutes (10 frames) given 60-minute observation (10 frames), resulting in 8772 training samples, 492 validation samples, and 1152 test samples.

MeteoNet. MeteoNet (Gwennaelle et al., 2020) contains rain radar data from the northwest and southeast regions of France collected by Meteo France from 2016 to 2018. The radar in MeteoNet has a spatial resolution of 0.01°, and the time interval between observations is 5 minutes. We choose the radar observations from the southeastern region of France in MeteoNet and select a portion of 400x400 pixels from the top left corner to remove areas with low radar quality.

4.1.2. Evaluation.

Following (Gao et al., 2023), we utilize the SSIM, Critical Success Index (CSI), Heidke Skill Score (HSS), and the Continuous Ranked Probability Score (CRPS) to evaluate the quality of precipitation nowcasting. HSS and CSI are computed on a per-pixel basis, enabling the identification of position inaccuracy (more details are included in the appendix). We also follow (Ravuri et al., 2021; Zhang et al., 2023) to report the CSI at 4×4 and 16×16 max pool scaling to reflect the model's ability for local precipitation pattern and regional extreme precipitation prediction. CRPS evaluates the uncertainty modeling ability of different models by measuring the discrepancy between the predicted distribution and the true distribution (Gao et al., 2023). When the prediction collapses to a single value, such as in a deterministic model, CRPS degrades to the Mean Absolute Error (MAE). A high CSI represents a good match between the model's predictions and the ground truth data on a pixel-wise or region-wise basis. A low CRPS indicates that the

predicted distribution is close to the true distribution. The calculation of all metrics for each model is performed on an ensemble of 10 members.

4.1.3. Training details.

We follow (Gao et al., 2022a) to train the deterministic models. The training of autoencoder is the same as (Rombach et al., 2022) except that we utilize the AdamW optimizer with the highest learning rate of 1e-4 and a cosine learning schedule. As for settings of diffusion, we apply the linear noise schedule with 1000 diffusion steps and 20 denoising steps for inference with DDIM (Song et al., 2020). Classifier free guidance (Ho & Salimans, 2022) is adopted for training and inference. In the diffusion part, our CasFormer is optimized by AdamW optimizer with a learning rate of 5e-4 and a cosine learning rate scheduler. The training of diffusion takes 200k steps (18 hours) on 4 A100s with a global batchsize of 32.

4.2. Compared to State of the Arts

To validate the quality of high-resolution precipitation forecasts generated by our CasCast, we selected five deterministic models and three probabilistic models for comparison. ConvLSTM (Shi et al., 2015), PredRNN (Wang et al., 2017) and PhyDNet (Guen & Thome, 2020) are deterministic short-term precipitation forecasting models designed based on recurrent mechanisms. SimVP (Gao et al., 2022b) and EarthFormer (Gao et al., 2022a) are non-recurrent deterministic models built upon CNN or spatial-temporal transformer architechures. PreDiff (Gao et al., 2023) and LDM (Rombach et al., 2022) are diffusion-based probabilistic models, while NowcastNet (Zhang et al., 2023) is a GAN-based probabilistic model.

Table 2 presents the quantitative experimental results of CasCast on the SEVIR dataset. In this experiment, CasCast incorporates EarthFormer as the deterministic **baseline**. We summarize some advantages of CasCast based on the results. (i) CasCast demonstrates the lowest CRPS. CasCast exhibits at least a 19.6% reduction in CRPS compared to deterministic models. CasCast achieves a significantly lower CRPS compared to the GAN-based probabilistic model NowcastNet. This suggests that the prediction distribution of CasCast is more similar to the data distribution. (ii) The CSI-M-POOL4 and CSI-M-POOL16 of CasCast, which assesses the model's ability to capture local pattern distributions (Gao et al., 2023), are considerably higher than other models. Compared to the best-performing deterministic model, CasCast shows an improvement of 7.4% in CSI-M-POOL4 and an improvement of 20.1% in CSI-M-POOL16. This improvement indicates that CasCast is more capable of capturing local precipitation patterns. (iii) CasCast demonstrates excellent ability in forecasting regional extreme precipitation. CSI-181-POOL16 and CSI-219-POOL16 reflect the accuracy for regional extreme precipitation with vertically integrated liquid $12.14kg/m^2$ (181) and $32.23kg/m^2$ (219) respectively, within a 16-kilometer range. Compared to PredRNN, which has the highest extreme precipitation forecasting scores among deterministic models, CasCast exhibits a 21.3% improvement in regional extreme nowcasting with a threshold value of 181, and a 48.8% improvement with a threshold value of 219. Compared to the prediff model, which is also diffusion-based, CasCast achieves significantly improved accuracy in forecasting regional extremes. This is because CasCast directly generates forecasts at a 1km resolution, avoiding the issue of extreme value attenuation caused by upsampling in Prediff. (iv) CasCast has the highest CSI-M-POOL1 and HSS score, which indicates that its pixel-wise predictions are more accurate. This can be attributed to the hybrid deterministic and probabilistic prediction in CasCast. The deterministic component of CasCast ensures accurate predictions of the high-deterministic mesoscale motion during precipitation events, while incorporating the diffusion part enhances CasCast's modeling capability for high-stochasticity small-scale precipitation phenomena.

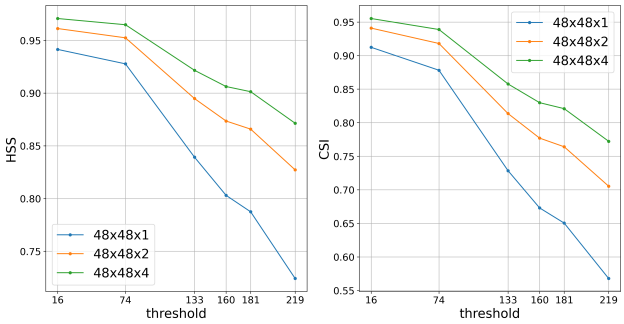

Figure 4. HSS scores and CSI scores of autoencoders with different latent dimensions. Note that, $16, 74, 133, 160, 181, 219$ are different thresholds applied for computing the scores.

CasCast demonstrates generalization capabilities across different short-term forecasting datasets, as indicated in Table 3. CasCast utilizes SimVP as the deterministic backbone for both HKO-7 (Shi et al., 2017) and MeteoNet (Gwennaelle et al., 2020) datasets, as SimVP exhibits the best overall performance on these two datasets. CasCast achieves skillful high-resolution short-term forecasts on both the HKO-7 and MeteoNet datasets, showcasing its forecasting advantages as observed on the SEVIR dataset as well.

4.3. Analysis and Ablation Study

4.3.1. Distortion Caused by Autoencoder

Due to the use of latent space, there is information loss caused by low-dimensional latent representation and the

Table 2. Comparison of CasCast with various deterministic and probabilistic models on SEVIR dataset. CSI-M is the mean CSI score across thresholds $[16, 74, 133, 160, 181, 219]$. POOL1 is CSI scores at the grid resolution. POOL4 and POOL16 indicate 4-grid aggregations and 16-grid aggregations, respectively. †: EarthFormer is evaluated using the official checkpoint (Gao et al., 2022a) to assess the CSI scores for regional extreme precipitation. ⋆: The model is trained on a downsampled dataset with a size of 128 since it is found that training on the original dataset does not yield high-quality predictions. **CSI-219 is an indicator that measures the accuracy of extreme precipitation forecasts.**

model	CRPS↓	SSIM↑	HSS↑	CSI-M↑			CSI-181↑			CSI-219↑		
				POOL1	POOL4	POOL16	POOL1	POOL4	POOL16	POOL1	POOL4	POOL16
ConvLSTM (Shi et al., 2015)	0.0264	0.7749	0.5232	0.4102	0.4163	0.4475	0.2453	0.2525	0.2977	0.1322	0.1380	0.1734
PredRNN (Wang et al., 2017)	0.0271	0.7497	0.5192	0.4045	0.4161	0.4623	0.2416	0.2567	0.3214	0.1331	0.1447	0.1909
PhyDNet (Guen & Thome, 2020)	0.0253	0.7649	0.5311	0.4198	0.4226	0.4410	0.2526	0.2532	0.2782	0.1362	0.1359	0.1526
SimVP (Gao et al., 2022b)	0.0259	0.7772	0.5280	0.4153	0.4226	0.4530	0.2532	0.2604	0.3000	0.1338	0.1394	0.1685
EarthFormer† (Gao et al., 2022a)	0.0251	0.7756	0.5411	0.4310	0.4319	0.4351	0.2622	0.2542	0.2562	0.1448	0.1409	0.1481
NowcastNet (Zhang et al., 2023)	0.0283	0.5696	0.5365	0.4152	0.4452	0.5024	0.2495	0.2935	0.3725	0.1422	0.1874	0.2700
LDM⋆ (Rombach et al., 2022)	0.0208	0.7495	0.4386	0.3465	0.3442	0.3520	0.1470	0.1391	0.1432	0.0671	0.0655	0.0717
PreDiff⋆ (Gao et al., 2023)	0.0202	0.7648	0.4914	0.3875	0.3918	0.4157	0.2076	0.2069	0.2264	0.1032	0.1051	0.1213
CasCast(ours)	**0.0202**	**0.7797**	**0.5602**	**0.4401**	**0.4640**	**0.5225**	**0.2879**	**0.3179**	**0.3900**	**0.1851**	**0.2127**	**0.2841**

Table 3. Comparison of CasCast with various deterministic and probabilistic models on HKO-7 and MeteoNet datasets.

model	HKO-7							MeteoNet						
	CRPS↓	CSI-M↑			CSI-185↑			CRPS↓	CSI-M↑			CSI-47↑		
		POOL1	POOL4	POOL16	POOL1	POOL4	POOL16		POOL1	POOL4	POOL16	POOL1	POOL4	POOL16
ConvLSTM	0.0257	0.4000	0.4084	0.4280	0.1569	0.1843	0.2472	0.0218	0.3008	0.3050	0.3465	0.0982	0.1091	0.1588
PredRNN	0.0252	0.3996	0.4146	0.4398	0.1633	0.1981	0.2634	0.0214	0.2914	0.3003	0.3402	0.0823	0.0990	0.1462
PhyDNet	0.0245	0.4213	0.4121	0.3846	0.1807	0.1768	0.1913	0.0216	0.3120	0.3124	0.3356	0.1106	0.1157	0.1482
SimVP	0.0248	0.4236	0.4195	0.4134	0.1881	0.1953	0.2233	0.0218	0.3017	0.3143	0.3577	0.0997	0.1134	0.1599
EarthFormer	0.0251	0.4096	0.4003	0.3950	0.1729	0.1731	0.1935	0.0224	0.2831	0.2855	0.3154	0.0787	0.0872	0.1208
NowcastNet	0.0296	0.4234	0.4518	0.4724	0.2025	0.2607	0.3601	0.0277	0.2955	0.3232	0.3734	0.1236	0.1521	0.2115
LDM⋆	0.0260	0.3045	0.2738	0.2764	0.0517	0.0605	0.0928	0.0209	0.2131	0.2191	0.2369	0.0359	0.0407	0.0552
PreDiff⋆	0.0244	0.3221	0.3152	0.3046	0.0788	0.0852	0.1113	0.0197	0.2546	0.2668	0.2935	0.0490	0.0594	0.0867
CasCast(ours)	**0.0205**	**0.4267**	**0.4608**	**0.4938**	**0.2158**	**0.2772**	**0.3653**	**0.0180**	**0.3156**	**0.3650**	**0.4420**	**0.1204**	**0.1563**	**0.2357**

conversion process between pixel space and latent space. To investigate the influence, we measured the CSI and HSS scores between the original radar data and the reconstructed radar data with different latent dimensions. The experiment was conducted on the SEVIR dataset using the UNet (Rombach et al., 2022) as the autoencoder. As exhibited in Figure 4, the CSI and HSS both decrease as the threshold increases. This implies that reconstructing high-intensity radar echoes is more challenging compared to low-intensity ones. When the last dimension of the latent space is doubled in size, the CSI and HSS increase, indicating a higher level of pixel-wise matching along with the increasing of latent space's dimension. However, the gain of increasing the dimension gradually diminishes especially for low thresholds. The experimental results indicate that there is still room to improve reconstruction performance by increasing the dimensions, especially for the extreme part.

4.3.2. Gains of Cascaded Modelling

To validate the effectiveness of the cascaded strategy in CasCast, we conducted tests on the SEVIR dataset when using only the probabilistic generation model and when using different deterministic models within CasCast (shown in Table 4, Figure 6). The results further confirm that when using either the probabilistic model or the deterministic model alone, there are either issues of pixel-wise mismatch or a limited ability to predict small-scale distribution and regional extreme values. When using ConvLSTM, SimVP, and EarthFormer as the deterministic parts of CasCast, the results have all improved. Especially, the CSI-219-POOL16 increases by **49.54%**, **57.69%**, **91.83%** for each deterministic model, exhibiting considerable gain of cascaded modeling for regional extreme-precipitation nowcasting. Lastly, the performance of the deterministic model has an impact on the generation results of cascaded modeling. A better deterministic network will contribute to higher performance.

Table 4. Seamless replacement of the deterministic part of CasCast on SEVIR dataset.**CasFormer** indicates a model without deterministic component.

MODEL	CSI-M		CSI-219	
	POOL1	POOL16	POOL1	POOL16
CasFormer	0.3210	0.4402	0.0656	0.1733
ConvLSTM	0.4102	0.4475	0.1322	0.1734
CasCast(ConvLSTM)	0.4145	0.5044	0.1575	0.2593(+**49.54%**)
SimVP	0.4153	0.4530	0.1338	0.1685
CasCast(SimVP)	0.4206	0.5133	0.1629	0.2657(+**57.69%**)
EarthFormer	0.4310	0.4351	0.1448	0.1481
CasCast(EarthFormer)	0.4401	0.5225	0.1851	0.2841(+**91.83%**)

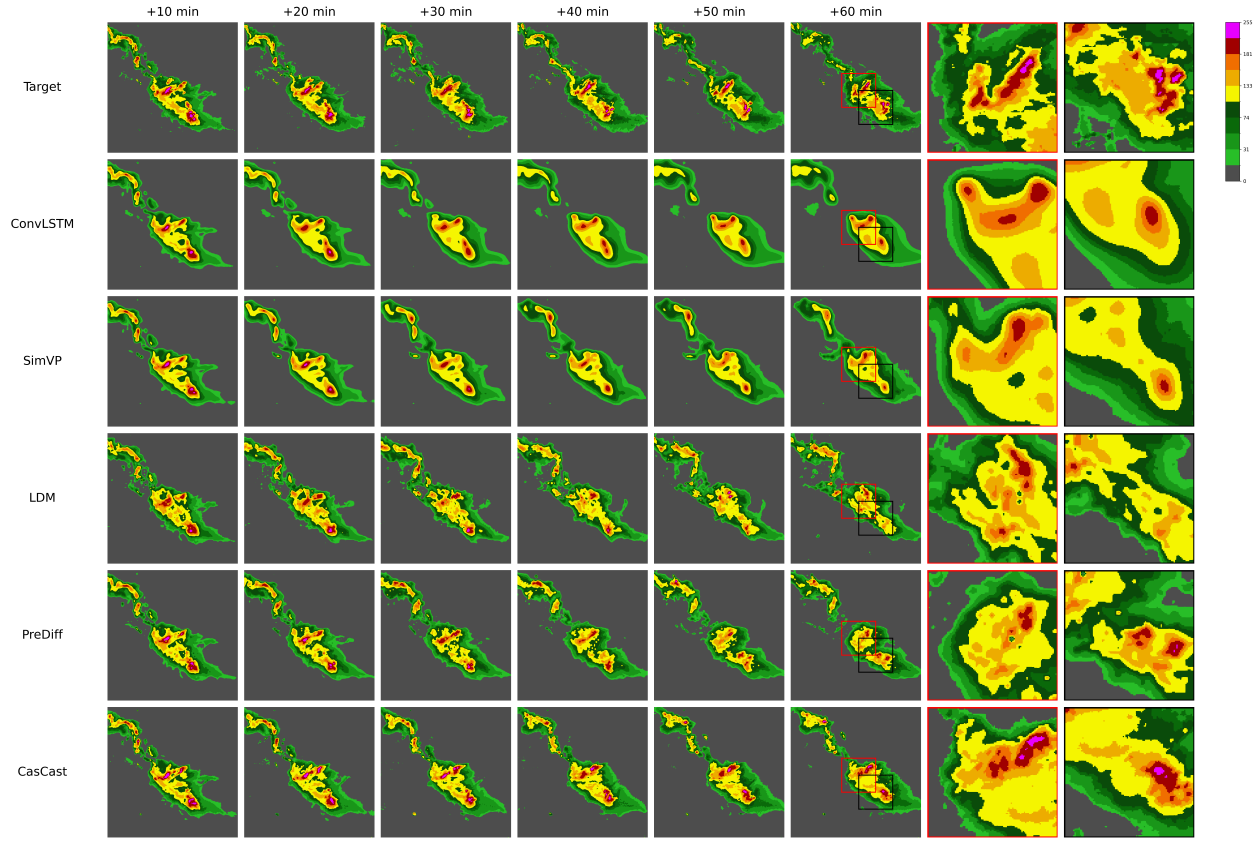

Figure 5. A set of example forecasts. From top to down denote Target, ConvLSTM, SimVP, LDM, PreDiff and CasCast (EarthFormer). From left to right denote forecasts of different lead times. Among these results, our CasCast has the most similar extreme precipitation pattern to the Target. (More qualitative results are shown in the Appendix.)

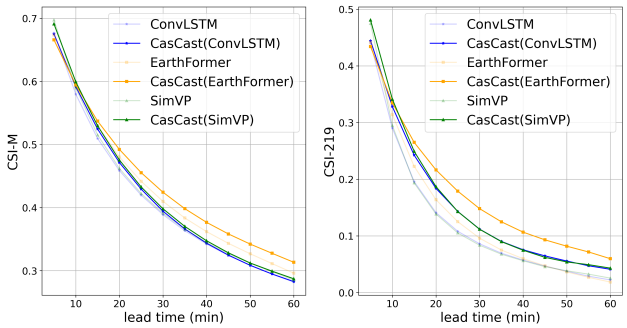

Figure 6. Frame-wise CSI-M (left) and CSI-219 (right).

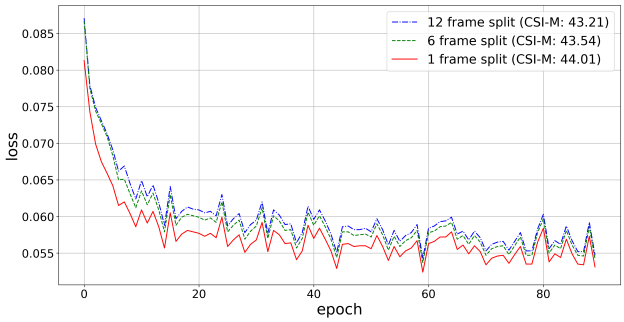

Figure 7. The training loss of CasFormer with different frame splitting numbers in our CasCast.

4.3.3. Effectiveness of Frame-wise Guidance

We validate the effectiveness of frame-wise guidance in CasFormer on the SEVIR dataset. The results are exhibited in Figure 7. 1-frame-split is the default setting of CasFormer. 6-frame-split indicates dividing a sequence input into 2 inputs, each with 6 frames of blurry predictions as conditions. When using a 12-frame-split setup, the encoding stage and decoding stage both take sequence-wise input, making CasFormer similar to the DiT model (Peebles & Xie, 2023). CasFormer with 1-frame-split outperforms its variants with 6-frame-split and sequence-wise-split in terms of optimization efficiency and final performance (CSI-M), which could be attributed to that 1-frame-split, ensuring a one-to-one correspondence between the latent vector z and the blur prediction, provides the most matching conditions for frame generation.

5. Conclusion

In this paper, we propose a cascaded method (CasCast), decoupling the forecasts into predictions of global distribution and generation of local details, for modeling complex precipitation systems with different scales. To boost the accuracy of predictions for regional extreme events, we generate small-scale patterns at high resolution in a low-dimensional latent space with our CasFormer for efficient and effective cascading generation. Experiments demonstrate that our methods predict realistic future radar echoes with high accuracy, especially for local extreme situations.

Limitations and Future work CasCast achieves competitive performance on different datasets, but it requires retraining on these datasets which hinders its use in different regions and radar data. Although there are differences in terrain, climate, and data acquisition among radar echo datasets, the laws of precipitation exhibit commonalities. Our future work aims to explore how to predict precipitation with a unified model, such as ClimaX (Nguyen et al., 2023), across different datasets.

Acknowledgements

This work is supported by Shanghai Artificial Intelligence Laboratory.

Impact Statement

This research focuses on precipitation nowcasting which has an essential influence on urban planning and disaster management. As an AI application for social good, our model boosts predictions for regional extremes. It improves disaster warnings such as storm warnings, flood forecasts, river level alerts, and landslide warnings, which helps the government, emergency agencies, and residents to prepare adequately and take necessary preventive and protective measures to minimize the losses and dangers caused by storms and floods. However, there are also risks of inaccurate forecasts with mismatched locations or excessively high/low intensities. Such predictions may prevent the correct planning for social sectors such as energy management and traffic scheduling. Furthermore, generative forecasts are usually of multiple possibilities which require professionals to discern.

References

Association, W. M. et al. Wmo atlas of mortality and economic losses from weather, climate and water extremes (1970–2019). Technical report, Technical Report, 2021.

Chen, K., Han, T., Gong, J., Bai, L., Ling, F., Luo, J.-J., Chen, X., Ma, L., Zhang, T., Su, R., et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

Chen, T. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

CLIMA, T. and TE, W. Guidelines for nowcasting techniques.

Gao, Z., Shi, X., Wang, H., Zhu, Y., Wang, Y. B., Li, M., and Yeung, D.-Y. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022a.

Gao, Z., Tan, C., Wu, L., and Li, S. Z. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3170–3180, 2022b.

Gao, Z., Shi, X., Han, B., Wang, H., Jin, X., Maddix, D., Zhu, Y., Li, M., and Wang, Y. Prediff: Precipitation nowcasting with latent diffusion models. *arXiv preprint arXiv:2307.10422*, 2023.

Guen, V. L. and Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11474–11484, 2020.

Gupta, A., Yu, L., Sohn, K., Gu, X., Hahn, M., Fei-Fei, L., Essa, I., Jiang, L., and Lezama, J. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

Gwennaelle, L., Lea, B., Vincent, C., Brice, Le, P., Bruno, P., and Lior, P. Meteonet, an open reference weather dataset. 2020.

Han, T., Guo, S., Ling, F., Chen, K., Gong, J., Luo, J., Gu, J., Dai, K., Ouyang, W., and Bai, L. Fengwu-ghr: Learning the kilometer-scale medium-range global weather forecasting. *arXiv preprint arXiv:2402.00059*, 2024.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Hoogeboom, E., Heek, J., and Salimans, T. simple diffusion: End-to-end diffusion for high resolution images. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13213–13232. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/hoogeboom23a.html.

Hu, Z., Ni, Z., Shi, J., Xu, S., and Xu, B. A knowledge-enhanced two-stage generative framework for medical dialogue information extraction. *Machine Intelligence Research*, 21(1):153–168, 2024.

Ling, F., Lu, Z., Luo, J.-J., Bai, L., Behera, S. K., Jin, D., Pan, B., Jiang, H., and Yamagata, T. Diffusion model-based probabilistic downscaling for 180-year east asian climate reconstruction. *arXiv preprint arXiv:2402.06646*, 2024.

Liu, Y., Wu, Y.-H., Sun, G., Zhang, L., Chhatkuli, A., and Van Gool, L. Vision transformers with hierarchical attention. *Machine Intelligence Research*, pp. 1–14, 2024.

Lu, Z., Wang, Z., Huang, D., Wu, C., Liu, X., Ouyang, W., and Bai, L. Fit: Flexible vision transformer for diffusion model. *International Conference on Machine Learning*, 2024a.

Lu, Z., Wu, C., Chen, X., Wang, Y., Bai, L., Qiao, Y., and Liu, X. Hierarchical diffusion autoencoders and disentangled image manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5374–5383, 2024b.

Mardani, M., Brenowitz, N., Cohen, Y., Pathak, J., Chen, C.-Y., Liu, C.-C., Vahdat, A., Kashinath, K., Kautz, J., and Pritchard, M. Residual diffusion modeling for km-scale atmospheric downscaling. 2024.

Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., and Grover, A. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Prein, A. F., Mooney, P. A., and Done, J. M. The multi-scale interactions of atmospheric phenomenon in mean and extreme precipitation. *Earth's Future*, 11(11): e2023EF003534, 2023.

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

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al. Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597(7878):672–677, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in neural information processing systems*, 30, 2017.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Srivastava, P., Yang, R., Kerrigan, G., Dresdner, G., McGibbon, J., Bretherton, C., and Mandt, S. Probabilistic precipitation downscaling with optical flow-guided diffusion. *arXiv preprint arXiv:2312.06071*, 2023.

Veillette, M., Samsi, S., and Mattioli, C. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33:22009–22019, 2020.

Wang, Y., Long, M., Wang, J., Gao, Z., and Yu, P. S. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30, 2017.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J. Skilful nowcasting of extreme precipitation with nowcastnet. *Nature*, 619(7970):526–532, 2023.

Zhao, Z., Dong, X., Wang, Y., and Hu, C. Advancing realistic precipitation nowcasting with a spatiotemporal transformer-based denoising diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

A. More Details about Datasets.

A.1. SEVIR

SEVIR dataset is a weather dataset containing observations from GOE satellites and NEXRAD radar. We use the data of NEXRAD which are processed into radar mosaic of Vertically Integrated Liquid (VIL). Following (Gao et al., 2022a), the thresholds for the evaluation of the predictions for VIL data are set to be 16, 74, 133, 160, 181 and 219. To convert the thresholds x into R with the units of kg/m^2, which are the true units of VIL images, the following rule could be applied (Veillette et al., 2020):

$$R(x) = \begin{cases} 0, & \text{if } x \leq 5 \\ \frac{x-2}{90.66}, & \text{if } 5 < x \leq 18 \\ \exp(\frac{x-83.9}{38.9}), & \text{if } 18 < x \leq 254 \end{cases}$$

A.2. HKO7

The raw radar echo images are generated by Doppler weather radar. The radar reflectivity values are converted to rainfall intensity values(mm/h) using the Z-R relationship:

$$dBZ = 10 \log a + 10b \log R \tag{4}$$

where R is ther rain-rate level, $a = 58.53$ and $b = 1.56$. The rainfall thresholds are ≥ 0.5, ≥ 2, ≥ 5, ≥ 10 and ≥ 30. According to Z-R relationship, dBZ thresholds are 30, 52, 66, 77, and 94. Furthermore, the pixel thresholds $[84, 118, 141, 158, 185]$ are given by $pixel = \lfloor 255 \times \frac{dBZ+10}{70} + 0.5 \rfloor$ following (Shi et al., 2017).

A.3. MeteoNet

MeteoNet is a dataset covering two geographical zones, North-West and South-East of France, during three years, 2016 to 2018. In our experiments, we use the rain radar part of the MeteoNet dataset. The region we use is the South-East of France. Original radar data is in the shape of $(565, 784)$, but some regions are missing in the data. We crop the data to keep the top-left portion with a size of 400×400.

The thresholds used in MeteoNet are similar to HKO7 which evaluates the scores at Rain Rate (mm/h) ≥ 0.5, ≥ 2, ≥ 5, ≥ 10 and ≥ 30, respectively. As the radar records reflectivity factor Z in decibels (dBZ), the rainfall rate R, expressed in millimeter per hour (mm/h), is approximated using the Marshall-Palmer relation:

$$R = (\frac{Z}{200})^{\frac{1}{1.6}} \tag{5}$$

in MeteoNet. It gives us thresholds in dBZ: 19, 28, 35, 40, 47.

B. HSS and CSI

HSS compares how often the pixel-wise predictions correctly match the ground truth. It can be calculated as :

$$HSS = \frac{2(TP \times TN - FN \times FP)}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)}. \tag{6}$$

The CSI (Critical Success Index) is another commonly used metric for measuring the accuracy of precipitation nowcasting. The definition of CSI is :

$$CSI = \frac{TP}{TP + FN + FP} \tag{7}$$

C. More Visulization Results

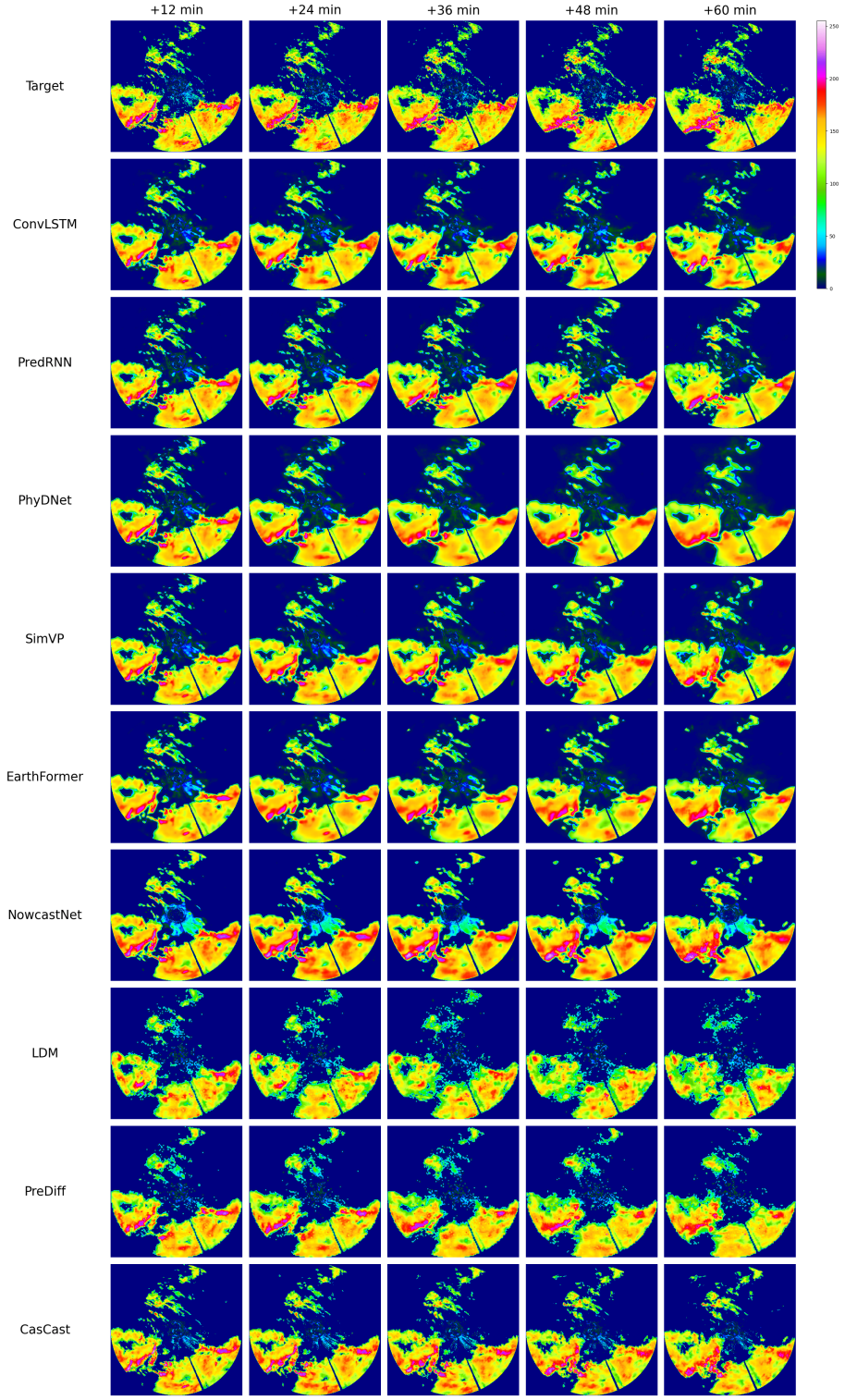

Figure 8. A set of example forecasts on HKO7. The deterministic component of CasCast is SimVP.

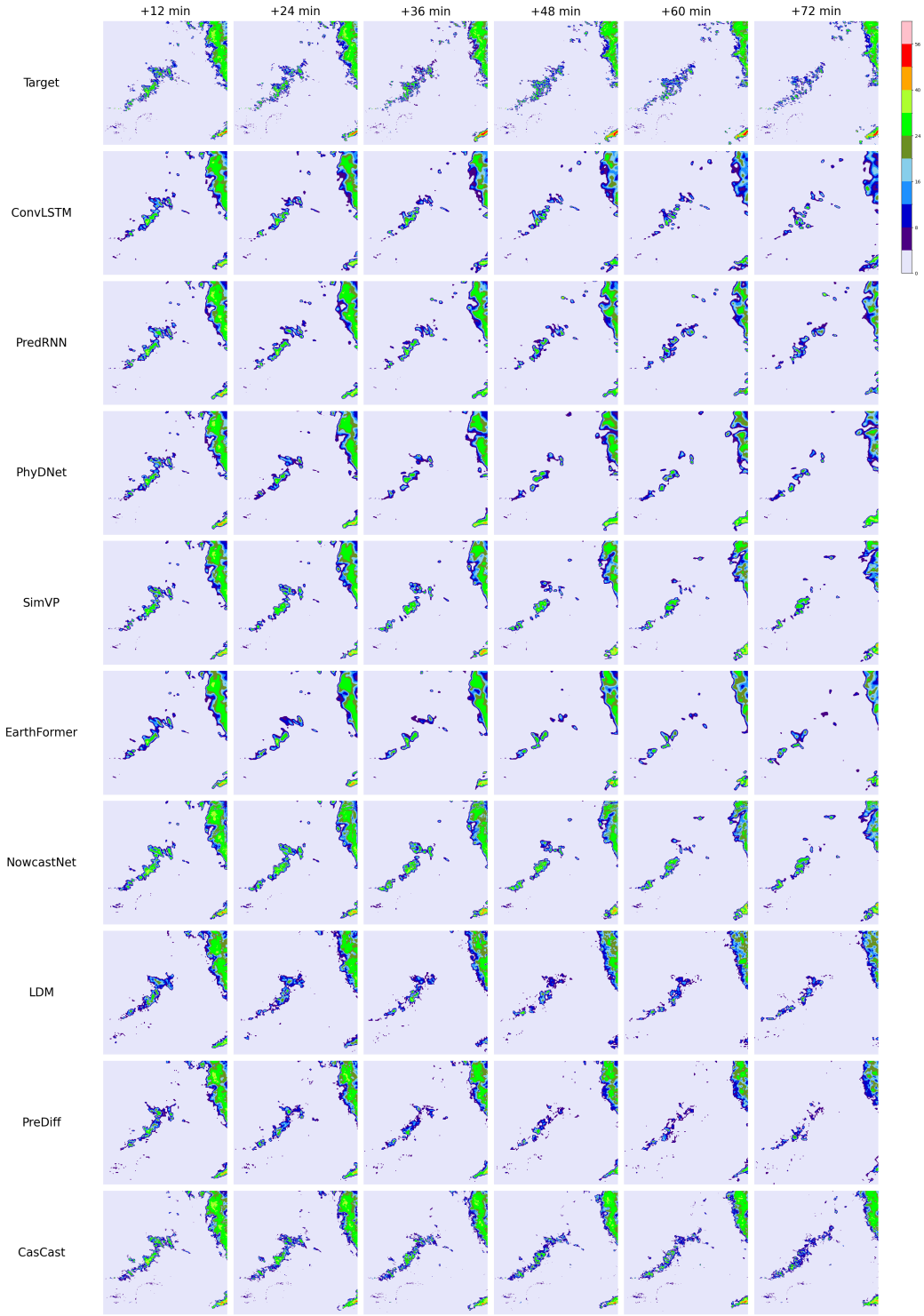

Figure 9. A set of example forecasts on MeteoNet. The deterministic component of CasCast is SimVP.

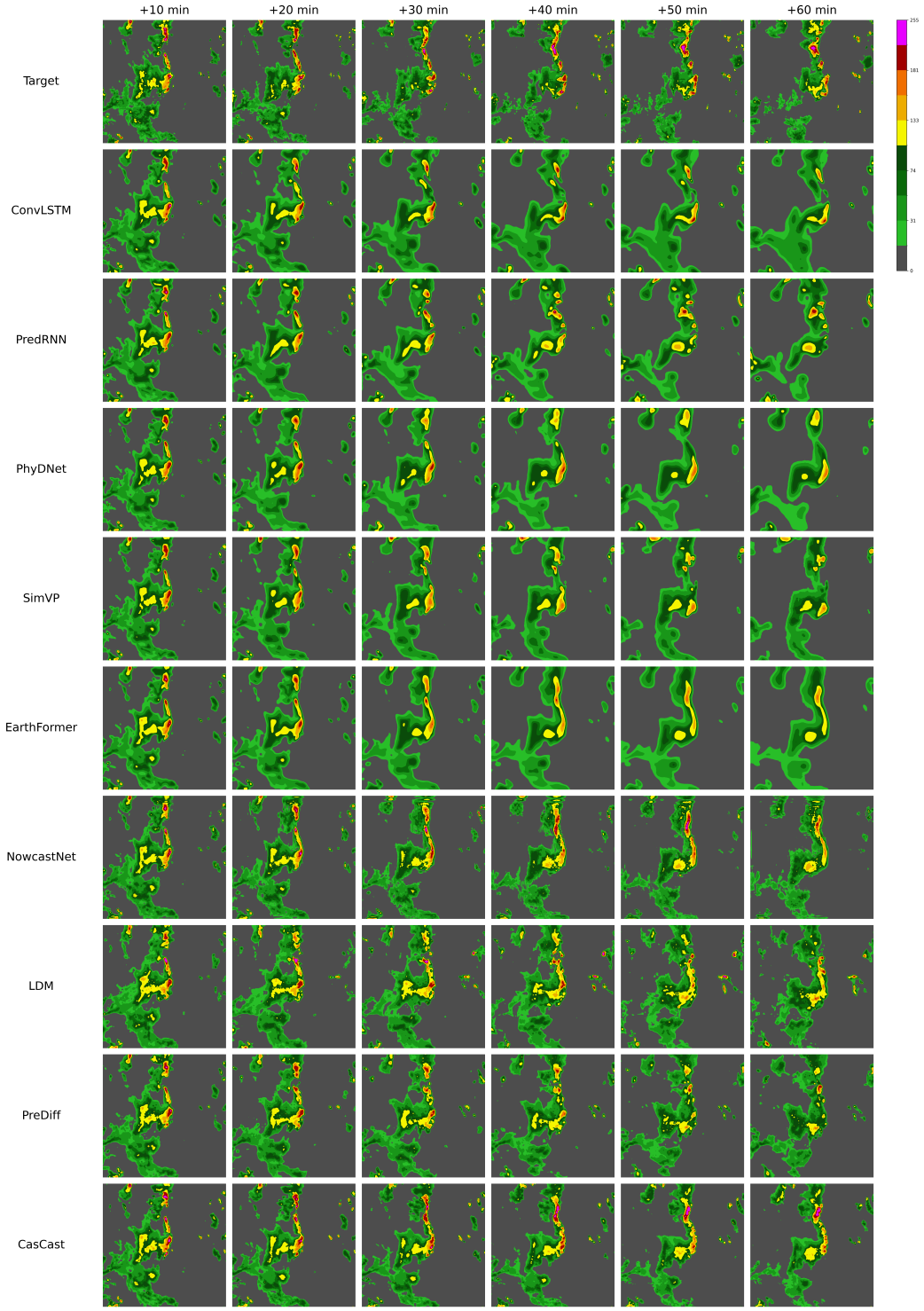

Figure 10. A set of example forecasts on SEVIR. The deterministic component of CasCast is EarthFormer.