Integrating Multi-Source Data for Long Sequence Precipitation Forecasting

Demin Yu, Wenzhi Feng, Kenghong Lin, Xutao Li*, Yunming Ye, Chuyao Luo, Wenchuan Du

Harbin Institute of Technology, Shenzhen 518000, China
{deminy,23b951053,linkenghong}@stu.hit.edu.cn, {lixutao, yeyunming}@hit.edu.cn,
{luochuyao.dalian,doorvant}@gmail.com,

Abstract

Long-sequence precipitation forecasting is critical for both meteorological science and smart city applications. The primary objective of this task is to predict future radar echo sequences, which provide high resolution and timely references for atmospheric precipitation distribution based on current observations. However, the chaotic nature of precipitation systems poses significant challenges in extending reliable forecast horizons. Most existing methods struggle with accuracy and clarity when extended to long-sequence predictions, such as three-hour forecasts. This is primarily due to the insufficiency of spatio-temporal information within a single modality over time. In this paper, we propose a cascading forecasting framework that adaptively extracts and integrates multimodal spatio-temporal information to support accurate and realistic long-sequence radar forecasting. Our framework includes a temporal adaptive predictor and a flow-based precipitation distribution adaptor. The predictor utilizes a multi-branch encoder-decoder architecture. This design allows it to extract meteorological sequences from multiple sources at varying scales, resulting in an initial global precipitation estimate. The core component is a carefully designed cross-attention module with a temporal adaptive layer to enhance multi-modality alignment. The initial estimate is then refined by the flow-based adaptor, which adjusts the prediction to match the target precipitation distribution, enhancing local details and correcting extreme precipitation patterns. We validated our method using real multi-source dataset for long-sequence forecasting, and the experimental results demonstrate that our approach outperforms existing state-of-the-art methods.

Introduction

Forecasting of long-sequence precipitation plays a crucial role in societal development, impacting daily planning, energy management, and transportation (Bouwer 2019). This long-standing scientific challenge aims to deliver accurate forecasts with fine-grained spatio-temporal resolution for the upcoming hours (e.g., 0-6 hours) based on current observations. Radar echo sequences, with their high resolution and timely atmospheric precipitation data, provide essential spatio-temporal information to understand the precipitation process, making them integral to this forecasting task.

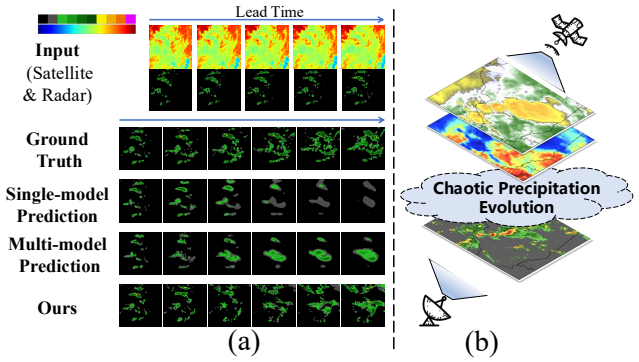

Figure 1: The (a) shows different models' performances in precipitation nowcasting. For gradually intensifying precipitation events, single-model prediction is unable to forecast accurately. While multi-modality aids in better modeling precipitation trends, it still faces challenges related to ambiguity. Our method, however, achieves both accurate and realistic representation. The (b) illustrates the various sources of multi-model meteorological data, which offer insights from different perspectives during precipitation events.

Conventional numerical weather prediction (NWP) methods (Bauer, Thorpe, and Brunet 2015) rely on various meteorological physical equations to provide high-resolution and timely atmospheric precipitation data. However, these methods are computationally expensive and time-consuming. In recent years, data-driven neural network-based models (Wu et al. 2024; Gao et al. 2022a) have emerged as a promising alternative for precipitation nowcasting. Unlike traditional approaches that rely on physical knowledge prior, these methods learn latent nonlinear patterns of precipitation evolution from large meteorological datasets to make predictions. They focus on minimizing the pixel distance between the prediction and the target, allowing for a global estimation of precipitation distribution. However, these approaches often suffer from blurriness in long-sequence predictions. More recently, distribution-based generative models (Yu et al. 2024; Yoon et al. 2023) have been employed for spatio-temporal prediction tasks. These models can generate realistic forecasts by sampling from the target predictive distribution, but they often experience a decline in accuracy

*Corresponding author

over longer prediction horizons.

Despite the increasing adoption of neural network-based models, generating accurate and realistic long-sequence forecasts remains a significant challenge. This difficulty arises primarily because single-modality inputs, such as radar echo maps, fail to provide sufficient spatio-temporal information, leading to blurry predictions and an underestimation of high-value echoes, as shown in Figure 1 (a). Incorporating multi-source meteorological observations, such as various satellite sequences, can effectively mitigate the spatio-temporal information bottleneck. As illustrated in Figure 1 (b), satellite and radar offer insights from different perspectives during precipitation events, which can complement each other to provide a more comprehensive understanding of the precipitation process.

In this work, we propose a multi-model spatio-temporal forecasting framework for accurate and realistic long-sequence precipitation forecasting, specifically for three-hour ahead predictions. The framework is composed of two key components: a temporal-adaptive multimodal predictor and a flow-based precipitation distribution adaptor. The overall architecture is depicted in Figure 2. The temporal-adaptive multimodal predictor employs a multi-branch encoder-decoder architecture to extract features from various modalities, enabling long-sequence predictions through a recurrent multi-task learning approach. A temporal-adaptive multimodal attention module is incorporated to facilitate cross-modal interactions within the latent space, while the temporal adaptive layer adjusts parameters dynamically across different stages of the prediction process. Following this, the flow-based distribution adaptor refines the initial predictions by transforming them into the target precipitation distribution using a learned ODE trajectory. By integrating these two components into a cohesive cascading framework, our approach enables both accurate and realistic precipitation predictions for the three-hour forecasting task.

In summary, our main contributions are summarized as:

- We propose a method that incorporates multi-source meteorological data with temporal alignment and multi-modality information fusion to enhance long-sequence prediction.
- We introduce a flow-based distribution transfer technique that directly combines deterministic estimations with the target distribution, achieving both accurate and realistic precipitation forecasts.
- We perform extensive experiments on real multi-source datasets, demonstrating that our method significantly improves long-sequence predictive performance.

Related Work

Spatio-temporal Prediction

The spatio-temporal prediction task focuses on forecasting future frames based on historical observations, with applications in various real-world domains such as weather forecasting (Shi et al. 2015, 2017), video prediction (Gao et al. 2022b), autonomous driving (Fu et al. 2021), and traffic flow prediction (Dai et al. 2022), among others. Spatio-temporal prediction models can be categorized into deterministic and probabilistic models, depending on their learning objectives (Yu et al. 2024; Yoon et al. 2023).

Deterministic predictive models are the mainstream of the existing approaches for spatio-temporal prediction (Yuan and Li 2021; Ning et al. 2023). They can be categorized into two groups: recurrent-based models and recurrent-free models. Recurrent-based models learn a hidden state from historical sequences and generate future frames sequentially using this hidden state (Shi et al. 2017; Wang et al. 2022). For instance, ConvLSTM (Shi et al. 2015) combines convolutional layers with LSTM (Long Short-Term Memory) cells to extract spatial and temporal features for deterministic precipitation forecasting. Similarly, PredRNN (Wang et al. 2022) improves spatial-temporal modeling by separating long-term and short-term memory cells. On the other hand, recurrent-free models (Gao et al. 2022b; Ning et al. 2023) encode input frames into hidden states and decode all predictive frames simultaneously. For example, Earthformer (Gao et al. 2022a) leveraged transformer to build the encoder-decoder for prediction. However, all the deterministic predictive methods suffer from the aforementioned blurry issue and high-value echoes fading away issue for precipitation nowcasting, because of the average error minimization of deterministic loss objective (Xu et al. 2024).

Probabilistic generative models are designed to capture the spatio-temporal uncertainty by estimating the conditional distribution of future state (Wen et al. 2023). Some models aim to enhance the realism of predictions based on adversarial training (Tulyakov et al. 2018; Luo et al. 2022; Ravuri et al. 2021).Generative Adversarial Networks (GANs) are notorious for training instability, leading to issues such as mode collapse and the generation of artifacts. Recently, with the advent of generative diffusion models (Ho, Jain, and Abbeel 2020; Zhou et al. 2023), several works have adopted diffusion models for predictive tasks to overcome the limitations of GANs, achieving remarkable performance (Gao et al. 2024b; Voleti, Jolicoeur-Martineau, and Pal 2022). While flow matching models (Esser et al. 2024; Liu, Gong, and Liu 2022), which offer theoretical advantages as a variant of diffusion models, have shown promise, they have yet to be firmly established in the field of spatio-temporal prediction. However, because these methods model the entire precipitation system as stochastic, they introduce uncontrollable randomness that can negatively impact prediction accuracy.

Multi-source Weather Forecasting

Utilizing multi-source meteorological data offers diverse perspectives on precipitation evolution, as illustrated in Figure 1 (b), by incorporating supplementary physical information that enhances the robustness and precision of prediction outcomes (Xiong et al. 2024; Veillette, Samsi, and Mattioli 2020). This approach is particularly effective in improving the accuracy of precipitation nowcasting (Li et al. 2023). Some methods (Pathak et al. 2022; Sønderby et al. 2020) employ an early-fusion mechanism to process multimodal data. These techniques typically involve interpolation and

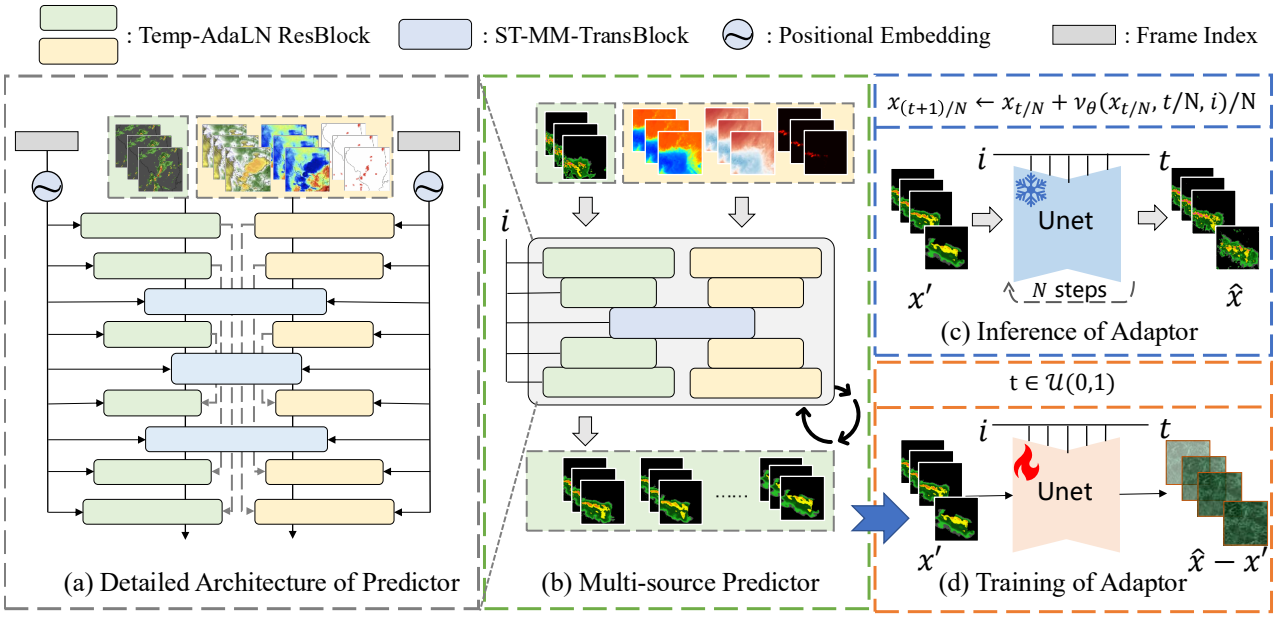

Figure 2: The overview of our framework for multi-source long-sequence precipitation nowcasting, which is composed of a multi-model predictor(b) cascaded with a distribution adaptor (c)(d). The predictor employs multiple encoder-decoder branches to recurrently forecast multimodal sequences, as depicted in (a). The preliminary predictions are subsequently refined through a flow-based adaptor. The (c) and (d) demonstrate the inference and training processes of the flow model, respectively.

down-sampling to align various data types, such as radar, spectral images, temperature, and precipitation, into a uniform format, integrating them into a single tensor along the channel dimension. Alternatively, other approaches (Boussif et al. 2024; Boussioux et al. 2022) use an intermediate-fusion mechanism, merging modalities within the encoded feature space to enhance spatio-temporal predictions. However, both early-fusion and intermediate-fusion methods often lack the efficiency and flexibility needed for multimodal spatio-temporal feature fusion, limiting their effectiveness in delivering accurate long-sequence forecasts.

Methodology

In this section, we will introduce the proposed multi-source long-sequence forecasting method. We first outline the task formulation in Section . The overall architecture is described in Section . In Section , we propose our temporal adaptive predictor with carefully designed multimodal attention module. Finally, Section introduces our flow-based adaptor for precipitation distribution transfer for realistic forecasting.

Task Formulation

Formally, we denote the weather state at time period i as a tensor $\mathcal{X}^i \in \mathbb{R}^{C\times H\times W}$, where C represents the number of atmospheric variables, H and W are the height and width. In this paper, we formulate the multi-source precipitation nowcasting problem as a multimodal spatio-temporal prediction task. In this case, we define $x, y \in \mathcal{X}$, which donate the radar echo sequence and satellite sequence, respectively. The objective of our framework is to forecast future radar echo sequences $x^{0:\hat{L}}$ given the historical radar data $x^{-L:0}$ combined with the corresponding satellite data $y^{-L:0}$. Following (Gao et al. 2024a; Veillette, Samsi, and Mattioli 2020) we organize the multi-source data into temporal-aligned formulation to enable a more comprehensive and systematic description of atmospheric evolution.

Overall Model

Our framework is designed to achieve long-sequence precipitation forecasting by utilizing multi-source meteorological observations, specifically targeting three-hour ahead predictions. The overview of our model is illustrated in Figure 2. The framework consists of two main components: a multi-model predictor (Figure 2 (b)) and a precipitation distribution adaptor (Figure 2 (c) and (d)). The multi-model predictor is responsible for generating long-sequence global estimates by integrating data from multiple sources, while the precipitation distribution adaptor focuses on modeling local intensity variations using a learnable ODE transfer trajectory. Unlike other multimodal approaches (Li et al. 2023; Gao et al. 2024a), we employ a multi-task learning network, denoted as $\mathcal{P}_\theta$, to simultaneously predict both radar and satellite sequences, $x'^{0:\hat{L}}$ and $y'^{0:\hat{L}}$, respectively. This process can be mathematically represented as follows:

$$x'^{0:\hat{L}}, y'^{0:\hat{L}} = \mathcal{P}_\theta(x^{-L:0}, y^{-L:0}), \tag{1}$$

where θ represents the parameters of the predictor, and L and $\hat{L}$ denote the input and output sequence lengths, respectively. Subsequently, the radar prediction $x'^{0:\hat{L}}$ is refined by

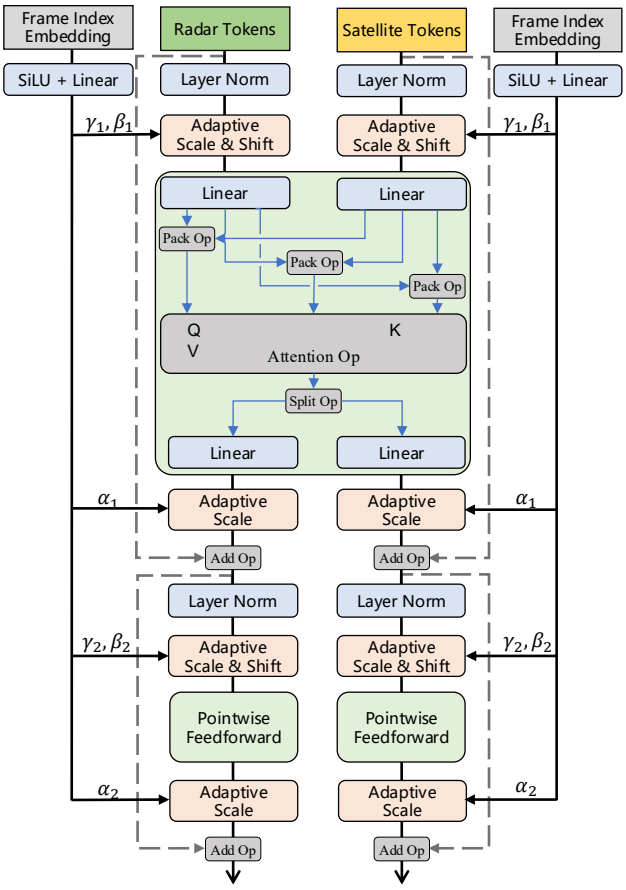

Figure 3: The illustration of multimodal transformer block.

a flow-based distribution adaptor, $\boldsymbol{v}_\theta$, which transfers it to a domain $\hat{x}^{0:\hat{L}}$ that closely matches the target data distribution $x^{0:\hat{L}}$. The learning objective for this process can be expressed as:

$$\hat{x}^{0:\hat{L}} = \arg\max_{x^{0:\hat{L}}} \boldsymbol{v}_\theta(\hat{x}^{0:\hat{L}} | x'^{0:\hat{L}}), \tag{2}$$

Throughout the framework, we disentangle the long-sequence precipitation nowcasting task into two simpler tasks: predicting the deterministic radar estimation x' and using it as a basis to generate probabilistic forecasts $\hat{x}$. Note that we denote the i-th frame as x^i, with the superscript i indicating the frame index, and use a subscript t to refer to the t-th ODE step state x_t in the subsequent sections.

Temporal-Adaptive Multimodal Predictor

Multi-source long-sequence prediction task presents two primary challenges: enabling spatio-temporal information interaction across different modality and modeling long-sequence temporal dependencies. Therefore, we propose a novel multimodal spatio-temporal joint prediction network, which employs a multi-model attention mechanism and a temporal adaptive layer to address these two issues. The overall architecture is shown in Figure 2 (a).

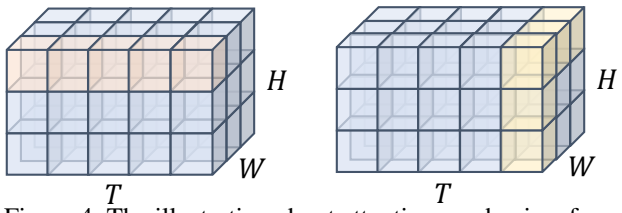

Figure 4: The illustration about attention mechanism for adjacent tokens in the spatial and temporal dimension.

Multi-Source Spatio-temporal Inforamtion Fusion The challenge of cross-modal interaction arises from the differing perspectives and spatio-temporal scales at which various modalities capture the precipitation evolution process. For example, radar data focuses on the fine-grained details of vertically integrated liquid beneath clouds, while satellite data provides a broader perspective, capturing overall water vapor density and temperature from above the clouds, as shown in Figure 1(b). To address this, specifically developed a temporal-adaptive 3D-ResBlock and a multimodal transformer block to extract multi-scale spatio-temporal features at different branch. Multiple layers of 3D ResNet are stacked to capture coarse-grained spatio-temporal information with relatively low computational cost. Additionally, we employ 2D and 1D convolutions for spatio-temporal downsampling and upsampling. Temporal downsampling is particularly beneficial as it increases information density along the temporal dimension. For multi-modality data with varying temporal resolutions, temporal compression enables the alignment of temporal features within a high semantic space.

Once the multi-scale features of different modalities are learned, each modality must identify and integrate valuable spatio-temporal representations from the others to enrich the extracted information. To facilitate this, we designed a multimodal spatio-temporal transformer block inspired by (Esser et al. 2024; Liu et al. 2023b), which enables interaction among multimodal features within the encoded high-dimensional feature space, as illustrated in Figure 3. Unlike standard cross-modal attention, we apply a packed product on the QKV (Query, Key, Value) tensors across the token dimension of multiple modalities before calculating attention, followed by projecting these back into their respective modalities. This approach enhances the model's scalability. By leveraging this mechanism across both spatial and temporal dimensions, as illustrated in Figure 4, we can theoretically build a network capable of integrating a broader range of multi-source meteorological data, such as atmospheric physical background, thereby improving accuracy.

Recurrent Temporal Adaptive Prediction As for the challenge of long-sequence temporal dependency modeling, it is inevitable that the prediction $x'^{0:\hat{L}}$ suffers from blurriness of local details over time, because the deterministic loss drives models to predict the evolution of the small-scale precipitation system into a mean value to represent the underlying uncertainty (Yu et al. 2024; Ravuri et al. 2021; Gong et al. 2024). However, early prediction results often tend

to retain details due to their low uncertainty. The distribution bias between early and later predictions contributes to why existing methods can only support short-sequence reliable precipitation forecasts, i.e. 0 $\sim$2 hours. To address this challenge, we propose Temp-AdaLN, a temporal adaptive layer for long sequence prediction that incorporates the lead time of predictive segments into the network. We map frame indices to temporal embeddings using sinusoidal encoding, and integrate this layer into every 3D-ResBlock and multimodal transformer block after layer normalization. While widely used in generative models (Dhariwal and Nichol 2021; Peebles and Xie 2023; Perez et al. 2018), this is the first application to long-sequence forecasting. The layer is defined as follows:

$$h = \gamma LayerNorm(h) + \beta, \tag{3}$$

where (γ, β) is obtained from a linear projection of the temporal embedding. In this case, the model can adjust for different prediction biases, enhancing its ability to perceive temporal patterns at various stages.

Flow-based Distribution Adaptor

The prediction results x' from multi-model predictor tend to blur as the prediction horizon extends, reducing their effectiveness as direct indicators of future precipitation. However, x' still captures the global distribution of precipitation, suggesting a high probability of accurately estimating large-scale precipitation. Given the distributional difference between the deterministic prediction x' and the target value $\hat{x}$, we aim to optimize the prediction by addressing this distribution shift, i.e., domain transfer. Domain transfer has been extensively studied (Liu, Gong, and Liu 2022; Liu et al. 2023a). In this paper, we employ flow matching technology instead of diffusion models to learn the transfer trajectory. The key difference between these approaches is that diffusion models initiate their generative denoising process with Gaussian white noise, which lacks structural information about the target data distribution (Liu et al. 2023a).

Specifically, the goal of flow matching is to build a transport flow to push the samples from source distribution, i.e. from deterministic prediction distribution $\mathcal{X}'$ to the target distribution $\mathcal{X}$. Given empirical observations of $x_0 = x' \sim \mathcal{X}', x_1 = \hat{x} \sim \hat{\mathcal{X}}$, the transfer flow inducted from (x_0, x_1) is an ordinary differentiable model (ODE) on time $t \in [0, 1]$,

$$dx_t = \boldsymbol{v}_\theta(x_t, t)dt, \tag{4}$$

which converts x_0 from $\mathcal{X}'$ to a x_1 following $\hat{\mathcal{X}}$. Here, x_t is the intermediate frames at time t and the velocity filed $\boldsymbol{v}_\theta : \mathbb{R}^{\hat{L}\times W\times H} \to \mathbb{R}^{\hat{L}\times W\times H}$ is a neural network with θ as its parameters. Given the intermediate state x_t, $\boldsymbol{v}_\theta$ defines a velocity field that moves x_t further towards target data x_1. How to design the prior move trajectory is important for the network. Following (Liu, Gong, and Liu 2022; Wu et al. 2023), we intuitively set the optimal direction at any time t is $x_1 - x_0$. In this case, the target transfer ODE process can be formulated as

$$dx_t = (x_1 - x_0)dt, \tag{5}$$

Algorithm 1: Final prediction inference

Input: The initial prediction $x_0 = x'$
Parameter: t, N
Output: The final prediction $\hat{x}$

Let $t \leftarrow 0, N \leftarrow 1000$.
while Time step $t < N$ **do**
  $x_{(t+1)/N} \leftarrow x_{t/N} + \frac{1}{N}\boldsymbol{v}_\theta(x_{t/N}, \frac{t}{N}, i)$
  $t = t + 1$
end while
return $\hat{x} \leftarrow x_1$

and we can get

$$x_t = tx_1 + (1-t)x_0, t \in [0, 1] \tag{6}$$

Finally, we can optimize our velocity field $\boldsymbol{v}_\theta$ by optimizing

$$\min_\theta \int_0^1 \mathbb{E}[\|\boldsymbol{v}_\theta(x_t, t) - (x_1 - x_0)\|^2]dt. \tag{7}$$

Empirically, the network $\boldsymbol{v}_\theta$ adopts the widely used UNet architecture, with the entire x' as input. It is notable that $x' \in \mathbb{R}^{\hat{L}\times W\times H}$ and there is a transition from clarity to blurriness over temporal dim. This transition result in the distribution bias across the sequence: the frames at start stage is closer to the target distribution, while the frames at last stage have huge bias to the target distribution. To mitigate this distribution bias, we encode the frame index into the network, enhancing its temporal awareness for each frame x^i. In this case, the $\boldsymbol{v}_\theta$ can be empirically optimized as

$$\min_\theta \mathbb{E}_{x_1\sim\hat{\mathcal{X}}, x_0\sim\mathcal{X}'}[\|\boldsymbol{v}_\theta(x_t, t, i) - (x_1 - x_0)\|^2], t \in \mathcal{U}(0, 1). \tag{8}$$

After the neural velocity field $\boldsymbol{v}_\theta$ is well-trained, we can solve the ODE process of Equation 4 with Euler solver, given the initial observation $x_0 = x' \sim \mathcal{X}'$. After multiply iterations of ODE process, we then generate the final prediction $\hat{x}$. The inference process is summarized in Algorithm 1.

Experiments

Experimental Setting

Datasets We utilize the SEVIR dataset (Veillette, Samsi, and Mattioli 2020), which provides synchronized multi-source precipitation sequences with radar observations at 5-minute intervals. Each sequence captures four-hour precipitation events over 384×384 km areas. The dataset comprises 11,640 sequences, incorporating Vertically Integrated Liquid (VIL) radar data, GOES-16 infrared channels (6.9 μm and 10.7 μm), and total lightning flashes. Our model predicts 3-hour precipitation evolution (36 frames) based on the preceding hour's data (12 frames). Due to computational constraints, we maintain temporal resolution while downscaling spatial dimensions to 128×128.

Verification Metrics Following (Yu et al. 2024; Gao et al. 2024b), we evaluate model performance using Critical Success Index (CSI) and Heidke Skill Score (HSS) for pixel-wise agreement. We assess CSI at 4×4 pooling

Category	Method	↑mCSI	↑mCSI_4x4	↑CSI_74	↑CSI_160	↑mHSS	↓LPIPS
Recurrent-based	ConvGRU	0.2181	0.2221	0.4841	0.0365	0.2667	0.3608
	PhyDNet	0.2425	0.2476	**0.5359**	0.0826	0.3037	0.3779
	PredRNNv2	0.2396	0.2479	0.5228	0.0503	0.2957	0.3548
Recurrent-free	SimVP	<u>0.2440</u>	0.2512	0.5259	0.0524	0.3021	0.3479
	EarthFormer	0.2299	0.2341	0.5026	0.0457	0.2832	0.3552
	Earthfarseer	0.2416	0.2413	<u>0.5322</u>	0.0776	0.2979	0.3813
Multi-modality	Metnet2	0.2227	0.2278	0.4944	0.0395	0.2722	0.3638
	Rain-F	0.1928	0.2174	0.3684	0.0714	0.2332	0.3481
	MMUNet	0.1841	0.2087	0.3766	0.0514	0.2183	0.3800
Probabilistic Generation	PreDiff	0.2365	0.3129	0.4561	0.1095	0.3017	0.1617
	MCVD	0.1982	0.2258	0.3826	0.0737	0.2393	0.2443
	DiffCast	0.2418	<u>0.3230</u>	0.4752	<u>0.1228</u>	**0.3363**	<u>0.1563</u>
	Ours	**0.2562**	**0.3398**	0.4882	**0.1289**	<u>0.3345</u>	**0.1413**

Table 1: Experiment results on multi-source datasets for long sequence precipitation forecasting task, i.e. 12 → 36 frames. The best results are in bold, and the second results are underlined.

scales to evaluate neighborhood aggregations and report specific CSI values at thresholds of 74 and 160 for low and high-value performance analysis. Additionally, we employ LPIPS (Yang, Srivastava, and Mandt 2022) to evaluate perceptual frame quality.

Implementation Details Model training spans 100K iterations using AdamW optimizer with a 1e-4 peak learning rate and cosine scheduling. The flow-based adaptor implements 1000 ODE sampling steps following (Liu, Gong, and Liu 2022), while $\boldsymbol{v}_\theta$ utilizes AdamW optimization with a 9e-5 peak learning rate and cosine scheduling. All experiments are conducted on NVIDIA A6000-48GB GPU.

Reference Methods Our baseline comparisons include three categories of models: (1) recurrent-free approaches (Earthfarseer (Wu et al. 2024), Earthformer (Gao et al. 2022a), SimVP (Gao et al. 2022b)), (2) recurrent models (PredRNNv2 (Wang et al. 2022), PhyDNet (Guen and Thome 2020), ConvGRU (Shi et al. 2017)), and (3) diffusion-based methods (DiffCast (Yu et al. 2024), Prediff (Gao et al. 2024b), MCVD (Voleti, Jolicoeur-Martineau, and Pal 2022)). We adapt single-source models for multi-source data through channel-wise concatenation and include native multimodal models (Metnet2 (Espeholt et al. 2022), Rain-F (Choi et al. 2021), MMUnet (Veillette, Samsi, and Mattioli 2020)).

Experimental Results

From the results of Table 1, we make the following observations: (i) The performance of our proposed method achieves state-of-the-art across primary metrics of precipitation forecasting, like mCSI, mCSI_4x4, mHSS. This verifies the effectiveness of our framework to integrate multiple meteorological sources for long-sequence prediction task. (ii) In terms of LPIPS, which measures the perceptual quality of predictions, our method makes significant improvements than other baselines, showing a 9.59% improvement over the second-best method DiffCast (Yu et al. 2024). (iii) Our method, along with other probabilistic generation approaches, demonstrates substantial improvements on the CSI_160 metric compared to other methodologies but get lower score on the CSI_74 metric. This suggests that probabilistic generative methods are better suited for modeling high-value regions, effectively addressing the issue of high-value disappear in deterministic models. However, the excessive freedom in these approaches can compromise accuracy in low-value areas during long-sequence predictions.

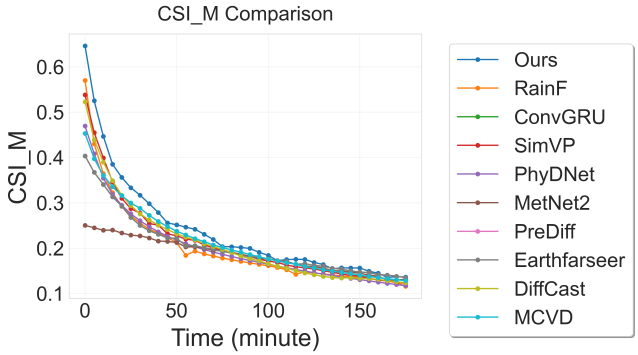

Figure 5: CSI changes against different lead time.

Figure 5 and Figure 6 illustrate the changes in performance metrics—Critical Success Index (CSI) and Learned Perceptual Image Patch Similarity (LPIPS)—as a function of lead time. These curves provide insights into how different models perform as the prediction horizon extends. We observe that as the lead time increases, the performance of all method decreases, because the uncertainty for prediction is enlarged. However, our method maintains higher

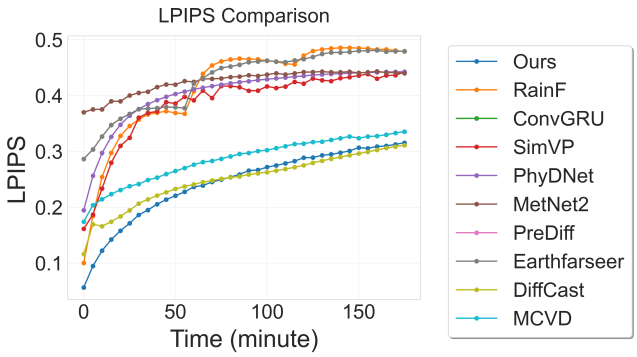

Figure 6: LPIPS changes against different lead time.

Method	↑mCSI	↑mCSI_4x4	↑mHSS	↓LPIPS
Ours w/o Multi-modality	0.2377	0.2973	0.3078	0.2263
Ours w/o Temp-AdaLN	0.2488	0.3300	0.3263	0.1563
Ours w/o Flow Adaptor	0.2468	0.2612	0.3237	0.4383
Ours (Full Model)	**0.2562**	**0.3398**	**0.3345**	**0.1413**

Table 2: Effects of the multiple source input, Temporal adaptive Layer (Temp-AdaLN) and the Flow-based Adaptor on the SEVIR dataset.

CSI scores over extended prediction times compared to other methods. This indicates that the model is more robust to temporal degradation, a common challenge in long-sequence forecasting. Regarding LPIPS, which measures visual similarity and fidelity, our model again shows strong performance. The LPIPS scores for our model increase at a slower rate compared to most other models, indicating that it more effectively preserves visual details and avoids the generation of blurry or less accurate predictions over time.

Ablation Studies

Table 2 presents the results of the ablation study, which evaluates the contributions of the multi-source input, Temporal Adaptive Layer, and Flow-based Adaptor. The ablation results reveal that each component plays a crucial role in the overall performance: (i) Based on single radar modality leads to a noticeable drop in all metrics, particularly in mCSI and mCSI-4x4, indicating that the incorporation of multimodal data is vital for capturing complex patterns. (ii) Excluding the Temp-AdaLN results in a significant decrease in performance across all metrics, with the most considerable impact on mHSS and LPIPS. This suggests Temp-AdaLN is essential for effectively aligning the temporal dynamics of the multimodal inputs, which is critical for long-sequence forecasting. (iii) The flow-based refinement adaptor enhances the perceptual quality of forecasting results, as evidenced by reduced LPIPS scores. This adaptor plays a vital role in both refining local details and correcting extreme precipitation values, thus contributing significantly to the full model's high perceptual fidelity.

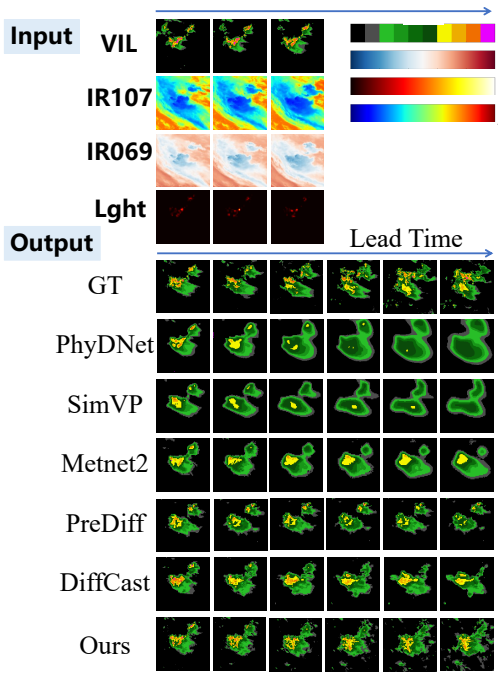

Figure 7: A visual example on a precipitation event with multiple meteorological sequences.

Case Study

In Figure 7, we present a visual comparison of several models applied to a precipitation event, demonstrating the efficacy of each method over a 180-minute forecasting horizon. The figure highlights the proposed model's ability to accurately predict both the spatial distribution and intensity of precipitation. Even after 180 minutes, our model's outputs remain closely aligned with the ground truth (GT), showcasing its robustness over extended periods. In contrast, other models such as PhyDNet, SimVP, and Metnet2 exhibit certain limitations. PhyDNet and SimVP tend to diffuse precipitation areas, leading to a loss of critical structural details. Metnet2, while preserving some structural elements, fails to maintain the intensity of the precipitation, leading to underestimation in certain regions. Probabilistic models like PreDiff and DiffCast, although producing visually appealing forecasts, encounter significant positional deviations, particularly in long-sequence predictions. By 180 minutes, these models display noticeable shifts in precipitation locations.

Conclusion

We present a cascaded framework for three-hour precipitation nowcasting that integrates a temporal-adaptive multi-model predictor with a flow-based distribution adaptor. By leveraging both radar and satellite sequences, our approach enhances spatio-temporal feature extraction, temporal coherence, and local detail preservation. Experimental results on multi-source meteorological datasets demonstrate the framework's superior accuracy and realism in long-sequence precipitation forecasting.

Acknowledgments

This work was supported in part by NSFC under Grants 62376072, 62272130, and in part by Shenzhen Science and Technology Program Nos. KCXFZ20240903093006009 and KCXFZ20211020163403005.

References

Bauer, P.; Thorpe, A.; and Brunet, G. 2015. The quiet revolution of numerical weather prediction. *Nature*, 525(7567): 47–55.

Boussif, O.; Boukachab, G.; Assouline, D.; Massaroli, S.; Yuan, T.; Benabbou, L.; and Bengio, Y. 2024. Improving* day-ahead* Solar Irradiance Time Series Forecasting by Leveraging Spatio-Temporal Context. *Advances in Neural Information Processing Systems*, 36.

Boussioux, L.; Zeng, C.; Guénais, T.; and Bertsimas, D. 2022. Hurricane forecasting: A novel multimodal machine learning framework. *Weather and forecasting*, 37(6): 817–831.

Bouwer, L. M. 2019. Observed and projected impacts from extreme weather events: implications for loss and damage. *Loss and damage from climate change: Concepts, methods and policy options*, 63–82.

Choi, Y.; Cha, K.; Back, M.; Choi, H.; and Jeon, T. 2021. RAIN-F: A fusion dataset for rainfall prediction using convolutional neural network. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 7145–7148. IEEE.

Dai, F.; Huang, P.; Mo, Q.; Xu, X.; Bilal, M.; and Song, H. 2022. ST-InNet: Deep spatio-temporal inception networks for traffic flow prediction in smart cities. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 19782–19794.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Espeholt, L.; Agrawal, S.; Sønderby, C.; Kumar, M.; Heek, J.; Bromberg, C.; Gazen, C.; Carver, R.; Andrychowicz, M.; Hickey, J.; et al. 2022. Deep learning for twelve hour precipitation forecasts. *Nature communications*, 13(1): 1–10.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Fu, M.; Zhang, T.; Song, W.; Yang, Y.; and Wang, M. 2021. Trajectory prediction-based local spatio-temporal navigation map for autonomous driving in dynamic highway environments. *IEEE Transactions on Intelligent Transportation Systems*, 23(7): 6418–6429.

Gao, Y.; Ma, T.; Xu, C.; and Wang, M. 2024a. Multi-modal Spatiotemporal Forecasting via Cross-Scale Operator Learning and Spatial Representation Aggregation. In *International Joint Conference on Artificial Intelligence*, 104–118. Springer.

Gao, Z.; Shi, X.; Han, B.; Wang, H.; Jin, X.; Maddix, D.; Zhu, Y.; Li, M.; and Wang, Y. B. 2024b. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36.

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

Gao, Z.; Tan, C.; Wu, L.; and Li, S. Z. 2022b. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3170–3180.

Gong, J.; Bai, L.; Ye, P.; Xu, W.; Liu, N.; Dai, J.; Yang, X.; and Ouyang, W. 2024. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*.

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

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

Li, Y.; Liu, Y.; Sun, R.; Guo, F.; Xu, X.; and Xu, H. 2023. Convective storm VIL and lightning nowcasting using satellite and weather radar measurements based on multi-task learning models. *Advances in Atmospheric Sciences*, 40(5): 887–899.

Liu, G.-H.; Vahdat, A.; Huang, D.-A.; Theodorou, E. A.; Nie, W.; and Anandkumar, A. 2023a. I2SB: Image-to-Image Schr\" odinger Bridge. *arXiv preprint arXiv:2302.05872*.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Liu, Y.; Ong, N.; Peng, K.; Xiong, B.; Wang, Q.; Hou, R.; Khabsa, M.; Yang, K.; Liu, D.; Williamson, D. S.; et al. 2023b. Mmvit: Multiscale multiview vision transformers. *arXiv preprint arXiv:2305.00104*.

Luo, C.; Li, X.; Ye, Y.; Feng, S.; and Ng, M. K. 2022. Experimental study on generative adversarial network for precipitation nowcasting. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–20.

Ning, S.; Lan, M.; Li, Y.; Chen, C.; Chen, Q.; Chen, X.; Han, X.; and Cui, S. 2023. MIMO is all you need: A strong multi-in-multi-out baseline for video prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 1975–1983.

Pathak, J.; Subramanian, S.; Harrington, P.; Raja, S.; Chattopadhyay, A.; Mardani, M.; Kurth, T.; Hall, D.; Li, Z.; Azizzadenesheli, K.; et al. 2022. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*.

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

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

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

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

Sønderby, C. K.; Espeholt, L.; Heek, J.; Dehghani, M.; Oliver, A.; Salimans, T.; Agrawal, S.; Hickey, J.; and Kalchbrenner, N. 2020. Metnet: A neural weather model for precipitation forecasting. *arXiv preprint arXiv:2003.12140*.

Tulyakov, S.; Liu, M.-Y.; Yang, X.; and Kautz, J. 2018. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1526–1535.

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

Voleti, V.; Jolicoeur-Martineau, A.; and Pal, C. 2022. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35: 23371–23385.

Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Philip, S. Y.; and Long, M. 2022. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2208–2225.

Wen, H.; Lin, Y.; Xia, Y.; Wan, H.; Wen, Q.; Zimmermann, R.; and Liang, Y. 2023. Diffstg: Probabilistic spatio-temporal graph forecasting with denoising diffusion models. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, 1–12.

Wu, H.; Liang, Y.; Xiong, W.; Zhou, Z.; Huang, W.; Wang, S.; and Wang, K. 2024. Earthfarsser: Versatile spatio-temporal dynamical systems modeling in one model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15906–15914.

Wu, L.; Wang, D.; Gong, C.; Liu, X.; Xiong, Y.; Ranjan, R.; Krishnamoorthi, R.; Chandra, V.; and Liu, Q. 2023. Fast point cloud generation with straight flows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9445–9454.

Xiong, T.; Wang, W.; He, J.; Su, R.; Wang, H.; and Hu, J. 2024. Spatiotemporal Feature Fusion Transformer for Precipitation Nowcasting via Feature Crossing. *Remote Sensing*, 16(14): 2685.

Xu, W.; Chen, K.; Han, T.; Chen, H.; Ouyang, W.; and Bai, L. 2024. Extremecast: Boosting extreme value prediction for global weather forecast. *arXiv preprint arXiv:2402.01295*.

Yang, R.; Srivastava, P.; and Mandt, S. 2022. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*.

Yoon, D.; Seo, M.; Kim, D.; Choi, Y.; and Cho, D. 2023. Deterministic Guidance Diffusion Model for Probabilistic Weather Forecasting. *arXiv preprint arXiv:2312.02819*.

Yu, D.; Li, X.; Ye, Y.; Zhang, B.; Luo, C.; Dai, K.; Wang, R.; and Chen, X. 2024. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27758–27767.

Yuan, H.; and Li, G. 2021. A survey of traffic prediction: from spatio-temporal data to intelligent transportation. *Data Science and Engineering*, 6(1): 63–85.

Zhou, L.; Lou, A.; Khanna, S.; and Ermon, S. 2023. Denoising diffusion bridge models. *arXiv preprint arXiv:2309.16948*.