Extending Precipitation Nowcasting Horizons via Spectral Fusion of Radar Observations and Foundation Model Priors

Yuze Qin[1], Qingyong Li[1], Zhiqing Guo[1], Wen Wang[1], Yan Liu[1,2], Yangli-ao Geng[1,2*]
[1]Key Laboratory of Big Data & Artificial Intelligence in Transportation (Ministry of Education), Beijing Jiaotong University, Beijing, China
[2]CMA Key Laboratory of Transportation Meteorology, Nanjing Innovation Institute for Atmospheric Sciences, Nanjing, China
{qinyuze, liqy, 23111126, wangwen, gengyla}@bjtu.edu.cn, yanliu@cma.gov.cn

***Abstract*—Precipitation nowcasting is critical for disaster mitigation and aviation safety. However, radar-only models frequently suffer from a lack of large-scale atmospheric context, leading to performance degradation at longer lead times. While integrating meteorological variables predicted by weather foundation models offers a potential remedy, existing architectures fail to reconcile the profound representational heterogeneities between radar imagery and meteorological data. To bridge this gap, we propose PW-FouCast, a novel frequency-domain fusion framework that leverages Pangu-Weather forecasts as spectral priors within a Fourier-based backbone. Our architecture introduces three key innovations: (*i*) Pangu-Weather-guided Frequency Modulation to align spectral magnitudes and phases with meteorological priors; (*ii*) Frequency Memory to correct phase discrepancies and preserve temporal evolution; and (*iii*) Inverted Frequency Attention to reconstruct high-frequency details typically lost in spectral filtering. Extensive experiments on the SEVIR and MeteoNet benchmarks demonstrate that PW-FouCast achieves state-of-the-art performance, effectively extending the reliable forecast horizon while maintaining structural fidelity. Our code is available at https://github.com/Onemissed/PW-FouCast.**

***Index Terms*—Precipitation Nowcasting, Multi-Modal, Fourier Neural Operator, Spatiotemporal Forecasting, Frequency Domain**

I. Introduction

Precipitation nowcasting is designed to generate short-term precipitation field predictions, which are critical for time-sensitive applications such as disaster resilience and aviation safety. Modern methodologies increasingly rely on deep learning to capture the complex interplay between convective-scale evolution and larger-scale atmospheric dynamics. However, as illustrated in Fig. 1 (second row), traditional radar-only models often experience performance degradation at longer lead times [1]. This limitation arises because radar reflectivity captures the resulting precipitation field rather than the underlying thermodynamic and dynamic drivers, such as temperature, humidity, wind speed, and pressure, that govern atmospheric evolution. Consequently, disparate atmospheric states may manifest as similar reflectivity patterns, restricting the model's capacity to disambiguate physical causes and accurately project future developments.

To extend the nowcasting horizon, it is essential to incorporate these causal drivers directly into the architecture. Moving beyond traditional integration of numerical weather prediction (NWP) data [2], we utilize weather foundation model outputs as multimodal inputs, specifically because their enhanced predictive precision and computational efficiency provide a more robust basis for improving nowcasting performance. Nevertheless, existing multimodal approaches frequently employ conventional spatial integration schemes such as addition, concatenation or cross-attention [3]–[5]. This direct fusion fails to address the fundamental heterogeneities inherent in these distinct data sources, including differing spatial scales, magnitudes, and temporal evolution patterns [6]. As illustrated in the third row of Fig. 1, such methods often fail to fully exploit cross-modal synergies, yielding marginal improvements for long-lead nowcasting.

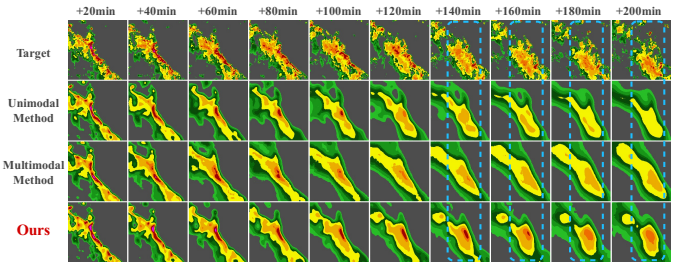

Fig. 1 Qualitative comparison of predicted radar reflectivity across unimodal and multimodal methods.

In this work, we propose PW-FouCast, a Fourier-domain backbone that leverages Pangu-Weather [7] forecasts to overcome these challenges through spectral integration. Our framework explicitly aligns and fuses spectral amplitude and phase information in the frequency domain, enabling the model to exploit shared phase representations between radar observations and meteorological forecasts. Furthermore, we develop a Frequency Memory module that stores and retrieves historical spectral patterns to correct phase discrepancies dynamically.

The primary contributions of this work are as follows:

1) We present a frequency-domain encoder-decoder framework specifically designed to extend nowcasting horizons by effectively assimilating foundation model priors.
2) We propose a novel method to integrate meteorological forecasts with radar reflectivity in the frequency domain, effectively resolving fundamental heterogeneities

* Corresponding author.

between these modalities.
3) We design a specialized Frequency Memory module to store and retrieve spectral features of diverse precipitation patterns, enhancing the model's ability to maintain structural fidelity over time.
4) Extensive experiments on the SEVIR and MeteoNet benchmarks demonstrate that PW-FouCast achieves state-of-the-art results, outperforming both radar-only and standard multi-modal baselines.

II. Related Work

A. Uni-modal Spatial-temporal Forecasting

Uni-modal spatial-temporal forecasting models typically use radar reflectivity as input and can be categorized into recurrent models and non-recurrent models. Recurrent models generate predictions sequentially, one frame at a time, which makes them effective at modeling short-term dependencies. One of the earliest is ConvLSTM [8], which extends LSTM by replacing internal dense operations with convolutions, enabling the network to capture spatial and temporal dependencies in a unified manner. PredRNN [9] extends this idea with a "zigzag" memory that flows across time and depth to exchange spatial–temporal representations, and PredRNN v2 [10] adds reverse scheduled sampling and a decoupling loss to better learn long-range dependencies. LMC-Memory [11] further augments recurrent predictors with an external memory and a two-phase alignment scheme to store and recall long-term motion patterns. Despite these advances, recurrent models can be computationally costly, prone to error accumulation over long horizons, and often learn redundant short-term features.

Non-recurrent models predict all frames simultaneously, are computationally efficient, and can capture global spatiotemporal context. SimVP v2 [12] and TAU [13] are pure CNN architectures that employ large-kernel convolutions in their Translator modules to approximate attention and capture global context. Earthformer [14] applies self-attention within non-overlapping spatio-temporal cuboids and propagates global context via learnable vectors. PastNet [15] injects spectral inductive biases and discretizes feature vectors with a memory bank. AlphaPre [16] decomposes forecasts into Fourier-domain phase and amplitude streams fused by an AlphaMixer. NowcastNet [17] predicts motion and intensity residuals to warp frames and then refines them with a generative module. Nonetheless, these non-recurrent designs still have difficulty modeling long-term temporal dependencies and lack flexibility for producing variable-length predictions.

B. Multi-modal Spatial-temporal Forecasting

Multi-modal spatio-temporal models incorporate auxiliary data, such as satellite imagery and meteorological fields, to improve precipitation forecasting. Examples include LightNet [3], which uses dual spatio–temporal encoders to process multiple sources, MM-RNN [4], which extracts multiscale features from radar and meteorological streams and fuses them with a cross attention–based module, and CM-STJointNet [5], which jointly learns radar extrapolation and satellite (IR) prediction via a STJointNet backbone. However, satellite and meteorological fields differ substantially from radar in scale, distribution, and representation, and many multimodal methods do not explicitly resolve these heterogeneities. By exploiting similar phases, our method instead aligns and fuses radar and meteorological information in the frequency domain, enabling more effective cross-modal integration and improved nowcasting skill.

III. Preliminaries

A. Pangu-Weather Model

Pangu-Weather is a global weather foundation model trained on 39 years of ERA5 [18] reanalysis data (1979–2017) at a 0.25° horizontal resolution. Its 3D Earth-specific transformer (3DEST) architecture captures complex atmospheric dependencies by integrating height as a distinct dimension. The model forecasts five upper-air variables across 13 vertical pressure levels and four surface variables, outperforming the ECMWF's operational Integrated Forecasting System (IFS) in accuracy. In our framework, we utilize the predicted geopotential, humidity, temperature, and wind components (u, v) as multimodal inputs. These variables serve as physical constraints that represent synoptic-scale trends, enabling the model to better maintain structural consistency in long-term precipitation nowcasting.

B. Adaptive Fourier Neural Operator

The Adaptive Fourier Neural Operator (AFNO) [19] introduces an efficient token mixing mechanism in the Fourier domain, extending neural operator frameworks for vision tasks. Given an input feature map X, AFNO first applies the forward Fourier transform to perform spatial mixing

$$Z = \mathcal{F}(X), \tag{1}$$

where $\mathcal{F}$ denotes the Fourier transform. The channels of resulting spectral features Z are then adaptively mixed using a shared multi-layer perceptron (MLP)

$$\hat{Z} = \mathrm{MLP}(Z) = W_2\,\sigma\,(W_1 Z)\,, \tag{2}$$

where W_1 and W_2 are block-diagonal complex-valued weight matrices, σ is the ReLU activation function. All weights are shared across spatial tokens to promote parameter efficiency.

C. Problem Formulation

We formulate precipitation nowcasting as a spatiotemporal forecasting problem. Let $\{X_t\}_{t=-T+1}^{0}$ be the observed radar sequence of length T), where $X_t \in \mathbb{R}^{C\times H\times W}$ denotes a frame with C channels and spatial resolution $H \times W$. We are also given N meteorological forecasts from Pangu-Weather, $\{P_i\}_{i=1}^{N}$, with $P_i \in \mathbb{R}^{M\times H'\times W'}$ containing M variables on an $H' \times W'$ grid at lead time i. The task is to predict the future K frames $\{X_t\}_{t=1}^{K}$.

Because the meteorological forecasts and radar observations differ in spatial and temporal resolution, we apply a preprocessing operator $\mathcal{R}(\cdot)$ that spatially regrids and temporally resamples the meteorological fields so they share the radar's

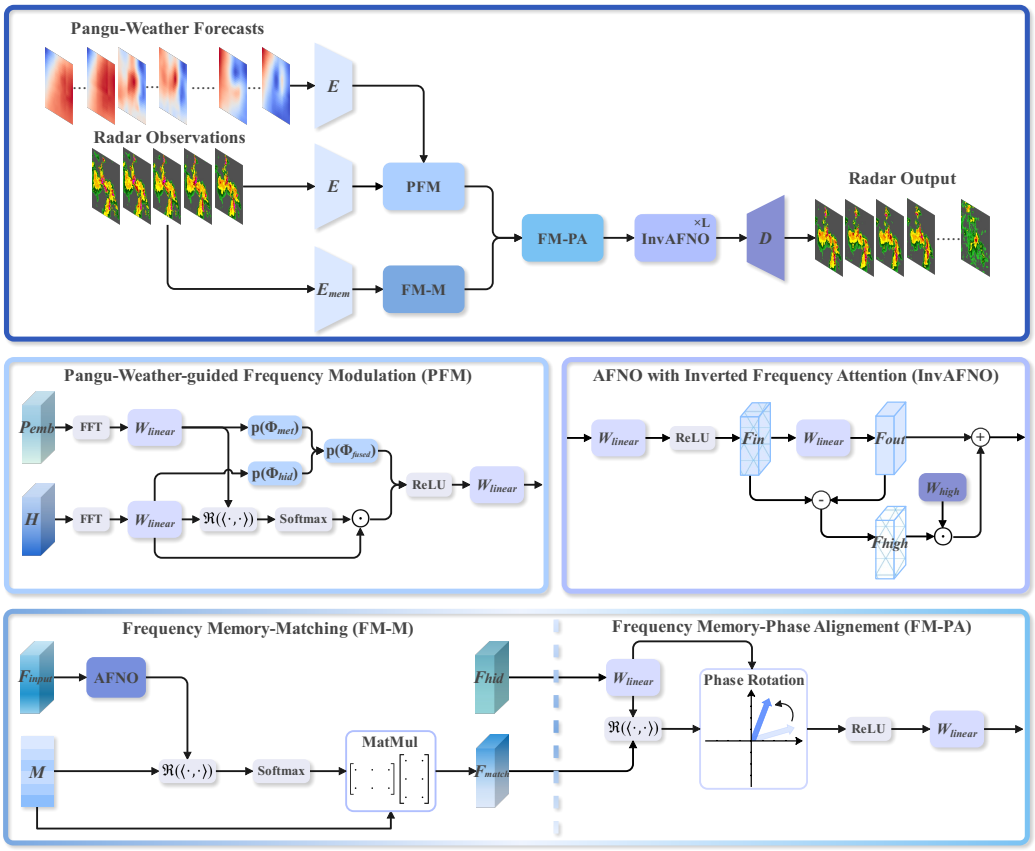

Fig. 2 Overall architecture of our proposed PW-FouCast.

shape and cadence in the model latent space (spatial and temporal interpolation). Denoting the aligned covariates by $\widetilde{P}_{1:K} = \mathcal{R}(P_{1:N})$, our model predicts

$$\hat{X}_{1:K} = f_\theta\big(X_{-T+1:0}, , \widetilde{P}_{1:K}\big), \tag{3}$$

where $\hat{X}_t \in \mathbb{R}^{C\times H\times W}$ is the predicted radar frame at lead time t.

IV. Methodology

A. Overview

We propose a frequency-domain encoder–decoder architecture that integrates three principal contributions: (*i*) Pangu-Weather–guided Frequency Modulation (PFM), which steers the model's spectral magnitudes and phase toward the ground truth; (*ii*) Frequency Memory (FM), a learned repository of ground-truth spectral patterns whose memory-matching produces matched frequency features used to correct hidden-layer phases; and (*iii*) Inverted Frequency Attention (IFA), a residual-reinjection mechanism that recovers high-frequency components attenuated by the learned frequency attention. The overall model architecture is illustrated in Fig. 2.

B. Pangu-Weather-guided Frequency Modulation

As illustrated in Fig. 3, although Pangu-Weather forecasts differ from radar reflectivity in amplitude and morphology,

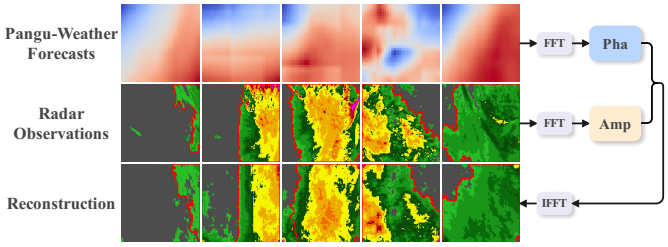

Fig. 3 Radar field reconstruction after integrating the amplitude of observations with the phase of meteorological variables.

reconstructing a radar-like field by combining the radar amplitude with the phase of Pangu-Weather fields produces a spatial pattern that similarly matches the observed radar reflectivity. This empirical phase similarity indicates that Pangu-Weather forecasts encode useful structural priors, we therefore leverage its phase features to correct and align both the amplitude and phase of the network's hidden-layer representations.

Concretely, let $F_{\text{hid}} = \mathcal{F}(H) \in \mathbb{C}^{H_{\text{emb}}\times W_{\text{emb}}\times C_{\text{emb}}}$ and $F_{\text{met}} = \mathcal{F}(P_{\text{emb}}) \in \mathbb{C}^{H_{\text{emb}}\times W_{\text{emb}}\times C_{\text{emb}}}$ be the complex Fourier representations of the hidden features and embedded meteorological fields. For each embedding channel i we compute the normalized inner product to quantify phase alignment

$$c_i = \frac{\langle F^{(i)}_{\text{hid}}, F^{(i)}_{\text{met}}\rangle}{|F^{(i)}_{\text{hid}}|\,|F^{(i)}_{\text{met}}|}, \tag{4}$$

take its real part $s_i = \Re(c_i)$ as a scalar similarity score, and convert the scores into channelwise attention maps via a softmax across the C_{emb} channels

$$w_i = \frac{\exp(s_i)}{\sum_{j=1}^{C_{\text{emb}}} \exp(s_j)}, \qquad i = 1, \ldots, C_{\text{emb}}, \tag{5}$$

where $w_i \in \mathbb{R}^{H_{\text{emb}} \times W_{\text{emb}}}$. These attention maps reweight the hidden-feature amplitudes entrywise

$$\hat{A}_{\text{hid}}^{(i)} = w_i \odot A_{\text{hid}}^{(i)}, \tag{6}$$

so frequency components whose phase aligns with Pangu-Weather prediction are selectively amplified. We then fuse phases using phasors. Convert a phase angle to a phasor

$$\mathbf{p}(\Phi^{(i)}) = \exp\big(\mathrm{j}\Phi^{(i)}\big) = \cos\Phi^{(i)} + \mathrm{j}\sin\Phi^{(i)}, \tag{7}$$

we interpolate between the hidden and meteorological phasors using a learnable parameter β and normalize to obtain a unit fused phasor

$$\tilde{z}^{(i)} = \beta\, \mathbf{p}(\Phi_{\text{hid}}^{(i)}) + (1-\beta)\mathbf{p}(\Phi_{\text{met}}^{(i)}), \tag{8}$$

$$\mathbf{p}(\Phi_{\text{fused}}^{(i)}) = \exp\big(\mathrm{j}\Phi_{\text{fused}}^{(i)}\big) = \frac{\tilde{z}^{(i)}}{|\tilde{z}^{(i)}|}. \tag{9}$$

Finally, we recombine the amplitude and the fused phasor to form the fused complex coefficients

$$\hat{F}_{\text{hid}}^{(i)} = \hat{A}_{\text{hid}}^{(i)} \cdot \mathbf{p}(\Phi_{\text{fused}}^{(i)}). \tag{10}$$

This two-stage procedure first aligns magnitudes according to Pangu-Weather guidance and then refines phases via learned interpolation, yielding hidden-layer frequency coefficients that better match ground truth in both amplitude and phase.

C. Frequency Memory

Precipitation field variations encompass multiple patterns including movement, expansion, and contraction. The coarse-grained structural priors provided by meteorological variables cannot accurately capture these diverse patterns. Therefore, we design a Frequency Memory module that records phase patterns from observed sequences during training and injects these learned phase features into the prediction pipeline, helping preserve fine-grained structural changes in forecasts.

Following [11], training proceeds in two stages. Each stage utilizing two core operations: Frequency Memory-Matching (FM-M) and Frequency Memory-Phase Alignment (FM-PA). In the storing stage (*Phase 1*), the model learns to populate a memory bank M with frequency-domain features derived from ground-truth sequences

Specifically, for the FM-M operation, let $\{X_t\}_{t=1}^{L}$ denote the ground-truth radar sequence and let $E_{\text{mem}}(\cdot)$ be the Frequency Memory encoder. We extract spatio–temporal features and apply a discrete Fourier transform to the encoder output to obtain the ground-truth frequency feature

$$F_{\text{GT}} = \mathcal{F}(E_{\text{mem}}(\{X_t\}_{t=1}^{L})). \tag{11}$$

where $F_{\text{GT}} \in \mathbb{C}^{H_{\text{emb}} \times W_{\text{emb}} \times C_{\text{emb}}}$, $H_{\text{emb}}, W_{\text{emb}}, C_{\text{emb}}$ denote the embedding height, width, and number of channels, respectively. Let $M = \{m_i\}_{i=1}^{S} \in \mathbb{C}^{S \times C_{\text{emb}}}$ denote the Frequency Memory with S slots. We first normalize F_{GT} and each memory slot m_i elementwise to unit magnitude

$$\hat{F}_{\text{GT}} = \frac{F_{\text{GT}}}{|F_{\text{GT}}|}, \quad \hat{m}_i = \frac{m_i}{|m_i|} \tag{12}$$

Using the normalized ground-truth $\hat{F}_{\text{GT}}$ as a complex query, we compute a raw similarity score between the query and each memory slot by taking the real part of their complex inner product

$$\begin{aligned} s_i^{\text{raw}}(h,w) &= \Re\big(\langle \hat{F}_{\text{GT}}(h,w,\cdot),\, \hat{m}_i(\cdot)\rangle\big) \\ &= \sum_{d=1}^{C_{\text{emb}}} \Big(\Re\big(\hat{F}_{\text{GT}}^{(d)}(h,w)\big)\Re\big(\hat{m}_i^{(d)}\big) + \\ &\qquad \Im\big(\hat{F}_{\text{GT}}^{(d)}(h,w)\big)\Im\big(\hat{m}_i^{(d)}\big)\Big). \end{aligned} \tag{13}$$

Hence the raw similarity tensor has shape $s^{\text{raw}} \in \mathbb{R}^{H_{\text{emb}} \times W_{\text{emb}} \times S}$, with one S-vector of similarities at each spatial location.

We convert the raw similarities $s_i^{\text{raw}}(h,w)_{i=1}^{S}$ into attention weights by applying a softmax across the memory-slot dimension independently at each spatial location

$$\alpha_i(h,w) = \frac{\exp\big(s_i^{\text{raw}}(h,w)\big)}{\sum_{j=1}^{S} \exp\big(s_j^{\text{raw}}(h,w)\big)}, \qquad i = 1, \ldots, S. \tag{14}$$

Thus $\alpha \in \mathbb{R}^{H_{\text{emb}} \times W_{\text{emb}} \times S}$ is a nonnegative attention map with $\sum_{i=1}^{S} \alpha_i(h,w) = 1$ for every (h,w). The attention maps $\{\alpha_i\}_{i=1}^{S}$ produce the matched frequency-domain feature via an attention-weighted sum

$$F_{\text{match}}(h,w,d,r) = \sum_{i=1}^{S} \alpha(h,w,i)\; M(i,d,r), \tag{15}$$

where $h = 1, \ldots, H_{\text{emb}}$, $w = 1, \ldots, W_{\text{emb}}$, $d = 1, \ldots, C_{\text{emb}}$, and $r \in 1, 2$ indexes real/imag. By treating F_{match} as a convex combination of M, the amplitude of F_{match} is bounded in $[0,1]$, i.e. $|F_{\text{match}}| \in [0,1]$.

Subsequently, the FM-PA operation leverages the matched frequency feature F_{match} to correct phase discrepancies within the model's hidden layers. We compute a real-valued raw similarity between F_{match} and the normalized hidden frequency features. Notably, the amplitude of F_{match} is not normalized, as it preserves critical information of the recalled spectral patterns.

$$\text{sim} = \Re\left(\left\langle \frac{\hat{F}_{\text{hid}}}{|\hat{F}_{\text{hid}}|},\, F_{\text{match}} \right\rangle\right). \tag{16}$$

Because $\text{sim} = |F_{\text{match}}| \cos(\hat{\Phi}_{\text{hid}} - \Phi_{\text{match}}) \in [-1,1]$, we convert sim into a phase-fusion weight bounded in $[0,1]$ by

$$w_{\text{phase}} = \tfrac{1}{2}\big(1 - \text{sim}\big). \tag{17}$$

As sim decreases with increasing phase discrepancy, the phase-fusion weight w_{phase} increases, ensuring that hidden-layer phases are more aggressively aligned with the retrieved spectral patterns. Let the phase difference between $\hat{\Phi}_{\text{hid}}$ and Φ_{match} be $\Delta_\Phi = \Phi_{\text{match}} - \hat{\Phi}_{\text{hid}}$. We rotate the hidden feature phase toward the matched phase by the fraction w_{phase} of the full phase difference

$$\widetilde{F}_{\text{hid}} = \hat{F}_{\text{hid}} \cdot \exp\big(\mathrm{j}\, w_{\text{phase}}\, \Delta_\Phi\big). \tag{18}$$

This yields the phase-corrected hidden representation $\widetilde{F}_{\text{hid}}$.

At the matching stage (*Phase 2*) we extract features from the input sequence $\{X_t\}_{t=-T+1}^{0}$ and transform the encoder output to the frequency domain

$$F_{\text{input}} = \mathcal{F}(E_{\text{mem}}(\{X_t\}_{t=-T+1}^{0})). \tag{19}$$

We then align the channel dimensionality of F_{input} with the frequency memory M using a AFNO block

$$\hat{F}_{\text{input}} = W_{\text{align2}}\, \sigma\left(W_{\text{align1}} F_{\text{input}}\right), \tag{20}$$

where W_{align1} and W_{align2} are block-diagonal, complex-valued weight matrices and σ denotes the elementwise ReLU activation. The aligned feature $\hat{F}_{\text{input}}$ therefore has shape $\hat{F}_{\text{input}} \in \mathbb{C}^{H_{\text{emb}} \times W_{\text{emb}} \times C_{\text{emb}}}$.

Next, we apply the same memory-matching procedure used for F_{gt} to $\hat{F}_{\text{input}}$ to obtain the matched frequency features from the frequency memory M, and use these matched features to correct the phase of the hidden features. Importantly, the frequency memory M is fixed during phase 2 and is not updated.

D. Inverted Frequency Attention

The proposed Inverted Frequency Attention is implemented within the hidden layer module to enhance spectral diversity specifically for the extraction of temporal features. The standard frequency attention mechanism utilized for temporal modeling typically takes the following form

$$F_{\text{out}} = W_{\text{learned}} \cdot F_{\text{in}}, \tag{21}$$

where $F_{\text{in}} = \mathcal{F}(X_{\text{in}}) \in \mathbb{C}^{H \times W \times C}$ denotes the complex Fourier coefficients of the input features and $W_{\text{learned}} \in \mathbb{C}^{H \times W \times C}$ denotes a learnable complex linear operator applied per frequency.

Empirically, W_{learned} tends to attenuate small-amplitude coefficients in F_{in}, producing an effective low-pass behaviour similar to the suppression of high-frequency components previously reported for attention-like transforms such as ViT [20].

Motivated by this observation, we obtain the discarded high-frequency residual by subtracting F_{out} from F_{in}, effectively applying the inverse mask of W_{learned}

$$F_{\text{high}} = F_{\text{in}} - F_{\text{out}}. \tag{22}$$

We then reintroduce high-frequency detail in a controlled manner using a learnable gating vector $w_{\text{high}} \in \mathbb{R}^{1\times1\times C}$. The gated residual is added back to the low-frequency component

$$\hat{F}_{\text{out}} = F_{\text{out}} + w_{\text{high}} \odot F_{\text{high}}, \tag{23}$$

where w_{high} is broadcast across the spatial frequency dimensions during the elementwise multiplication. This achieving the fusion of high-frequency and low-frequency feature in a simple way.

E. Loss Function

The training objective is a weighted sum of a spatial mean-squared error and a spectral L_1 loss

$$\mathcal{L} = \mathbb{E}\left[\left\|X_{1:K} - \hat{X}_{1:K}\right\|_2^2\right] + \lambda\, \mathbb{E}\left[\left\|\mathcal{F}(X_{1:K}) - \mathcal{F}(\hat{X}_{1:K})\right\|_1\right], \tag{24}$$

where $0 \le \lambda \le 1$ is a hyperparameter. The MSE term penalizes spatial reconstruction error, while the spectral L_1 term encourages accurate recovery in frequency space, together they reduce spatial error and help preserve high-frequency echo structure in the predictions.

V. Experiments

A. Experimental Setup

1) Dataset: **SEVIR** [21] is a benchmark precipitation forecasting dataset containing 20,393 meteorological events. For our experiments we selected events from 2018–2019. The period January 2018–May 2019 is used for training and June–November 2019 for testing, yielding 10,776 training samples and 4,053 test samples. Following [17], all models receive 5 input frames (50 minutes) and predict the next 20 frames (200 minutes).

MeteoNet [22] is an open dataset curated by Météo-France that spans 2016–2018 and covers a 550×550 km region in north-western France. For our experiments we construct a 2018 subset: January–August form the training set (5,381 samples) and September–October the test set (1,027 samples). The model receives 5 input frames (50 minutes) and predicts the next 20 frames (200 minutes).

Meteorological Variables are inferenced from the pre-trained Pangu-Weather model. We use five upper-air variables at 500, 600, 700 and 850 hPa, crop each forecast to the latitude–longitude bounding box of the corresponding SEVIR or MeteoNet scene, linearly interpolate the fields in time to align with radar timesteps, and spatially resample them to the model hidden-layer resolution 32×32. Finally, these variables are concatenated along the channel dimension and standardized via channel-wise z-score normalization to serve as model inputs.

2) Training Details: We train all models using the AdamW optimizer with a learning rate of 0.001. The architecture consists of four convolutional encoder-decoder modules, with a hidden layer depth (L) of 6. Radar inputs are linearly normalized to the range $[0, 1]$, and radar reflectivity is interpolated to a size of 128×128 pixels. All experiments were conducted on two RTX 3090 GPUs.

3) Evaluation Metrics: We evaluate nowcasting performance using the Critical Success Index (CSI) and Heidke Skill Score (HSS) at multiple thresholds. CSI quantifies event-based accuracy for exceedance of a reflectivity threshold, while HSS measures overall forecast skill relative to random chance.

TABLE I Quantitative evaluation on the SEVIR dataset. **Bold**: best; Underline: second-best.

Type	Model	CSI↑ 16	74	133	160	181	219	Avg	HSS↑ Avg	MSE↓	MAE↓	PSNR↑	SSIM↑
Unimodal	PredRNN v2 [10]	0.5922	0.4757	0.2071	0.1179	0.0879	0.0371	0.2530	0.3376	692.4151	12.8557	23.5778	0.5675
	SimVP v2 [12]	0.5747	0.4319	0.1806	0.0982	0.0695	0.0299	0.2308	0.3085	735.3561	13.3449	23.7060	0.5488
	TAU [13]	0.5767	0.4750	0.2328	0.1223	0.0834	0.0376	0.2546	0.3388	739.1168	13.5212	23.5635	0.5503
	Earthformer [14]	0.5824	0.4846	0.2346	0.1300	0.0951	0.0441	0.2618	0.3489	717.5891	13.5044	23.7495	0.5459
	PastNet [15]	0.5551	0.4616	0.2044	0.1114	0.0800	0.0419	0.2424	0.3236	718.3476	14.9068	22.7698	0.3905
	AlphaPre [16]	0.5737	0.4739	0.2266	0.1176	0.0790	0.0349	0.2510	0.3335	744.8281	13.7112	23.6576	0.5286
	NowcastNet [17]	0.5803	0.4642	0.2200	0.1234	0.0911	0.0444	0.2539	0.3402	750.2669	13.3689	23.6842	0.5595
	LMC-Memory [11]	0.5643	0.4586	0.1912	0.0997	0.0728	0.0305	0.2362	0.3136	744.8522	13.7543	23.5571	0.5445
	AFNO [19]	0.5858	0.4804	0.2311	0.1286	0.0983	0.0469	0.2618	0.3502	740.2969	13.2170	23.7374	0.5576
Multimodal	LightNet [3]	0.5697	0.4627	0.1974	0.0918	0.0574	0.0169	0.2326	0.3053	725.6782	13.8670	23.5267	0.5220
	MM-RNN [4]	0.5679	0.4582	0.2052	0.0945	0.0635	0.0231	0.2354	0.3112	750.6046	13.7483	23.4178	0.5374
	CM-STjointNet [5]	0.5825	0.4778	0.2188	0.1218	0.0894	0.0468	0.2562	0.3420	715.5997	13.3922	23.6520	0.5407
	Ours	**0.6023**	**0.4900**	**0.2558**	**0.1511**	**0.1163**	**0.0628**	**0.2797**	**0.3757**	**676.6416**	**12.5787**	**24.1511**	**0.5789**

TABLE II Quantitative evaluation on the MeteoNet dataset. **Bold**: best; Underline: second-best.

Type	Model	CSI↑ 12	24	32	Avg	HSS↑ 12	24	32	Avg	MSE↓	MAE↓	PSNR↑	SSIM↑
Unimodal	PredRNN v2 [10]	0.3344	0.1554	0.0350	0.1749	0.4897	0.2647	0.0671	0.2738	9.3032	0.7996	34.0562	0.8501
	SimVP v2 [12]	0.3250	0.1400	0.0195	0.1615	0.4786	0.2415	0.0379	0.2527	9.4314	0.8343	34.9568	0.8499
	TAU [13]	0.3444	0.1413	0.0248	0.1702	0.5013	0.2446	0.0482	0.2647	8.3152	0.7907	34.8399	0.8474
	Earthformer [14]	0.3628	0.1839	0.0573	0.2013	0.5218	0.3072	0.1079	0.3123	7.9094	0.7642	35.4281	0.8572
	PastNet [15]	0.3526	0.1644	0.0268	0.1813	0.5110	0.2789	0.0518	0.2806	8.1057	0.9666	34.6571	0.7681
	AlphaPre [16]	0.3722	0.1854	0.0575	0.2050	0.5326	0.3098	0.1083	0.3169	7.6863	1.0026	34.7310	0.7636
	NowcastNet [17]	0.3414	0.1595	0.0660	0.1890	0.4990	0.2722	0.1233	0.2982	7.8276	0.7268	35.5236	0.8651
	LMC-Memory [11]	0.3505	0.1659	0.0448	0.1871	0.5092	0.2817	0.0854	0.2921	7.8226	0.7675	35.2575	0.8496
	AFNO [19]	0.3739	0.2018	0.0859	0.2205	0.5343	0.3325	0.1576	0.3415	7.5631	0.9341	34.9673	0.7864
Multimodal	LightNet [3]	0.3539	0.1649	0.0503	0.1897	0.5128	0.2801	0.0955	0.2961	7.6877	0.6864	35.5315	0.8716
	MM-RNN [4]	0.3456	0.2187	0.0519	0.2054	0.5015	0.3539	0.0981	0.3178	9.6015	0.7580	35.2290	0.8737
	CM-STjointNet [5]	0.3285	0.1553	0.0542	0.1793	0.4832	0.2656	0.1022	0.2837	8.8143	0.9429	34.3279	0.7974
	Ours	**0.3744**	**0.2206**	**0.1022**	**0.2324**	**0.5353**	**0.3579**	**0.1848**	**0.3593**	**7.3844**	**0.6603**	**35.8116**	**0.8740**

Following prior work [23], [24], thresholds for SEVIR are 16, 74, 133, 160, 181, and 219, and for MeteoNet are 12, 24, and 32. For pixel-level continuous accuracy we report Mean Squared Error (MSE) and Mean Absolute Error (MAE); for perceptual assessment we report Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM).

4) Hyperparameter Selection: For the number of frequency memory slots S and the loss weight λ, we performed hyperparameter sweeps evaluated by MAE on the SEVIR and MeteoNet datasets. As shown in Fig. 4, the sweeps identify optimal memory sizes of $S = 240$ for SEVIR and $S = 160$ for MeteoNet, a difference we attribute to the greater complexity of precipitation patterns in SEVIR that require more memory capacity. The best loss weights are $\lambda = 0.57$ (SEVIR) and $\lambda = 0.55$ (MeteoNet), showing consistent behavior across both datasets.

5) Baselines: We compare our approach to twelve state-of-the-art spatiotemporal forecasting models: nine unimodal models PredRNN v2 [10], SimVP v2 [12], TAU [13], Earthformer [14], PastNet [15], AlphaPre [16], NowcastNet [17], LMC-Memory [11] and AFNO [19]; and three multimodal models LightNet [3], MM-RNN [4] and CM-STjointNet [5].

B. Experimental Results

As shown in Table I, PW-FouCast achieves state-of-the-art performance on the SEVIR dataset, reducing MSE and MAE by 2.28% and 2.15% while increasing average CSI and HSS by 6.84% and 7.28% over the strongest baselines. On MeteoNet (Table II), our model similarly outperforms competitors, reducing MSE and MAE by 2.36% and 3.80%, with CSI and HSS improvements of 5.40% and 5.21%. Peak PSNR and SSIM scores further demonstrate superior pixel-level accuracy and structural fidelity. These gains confirm that integrating Pangu-Weather spectral priors effectively mitigates the long-lead degradation typical of radar-only models.

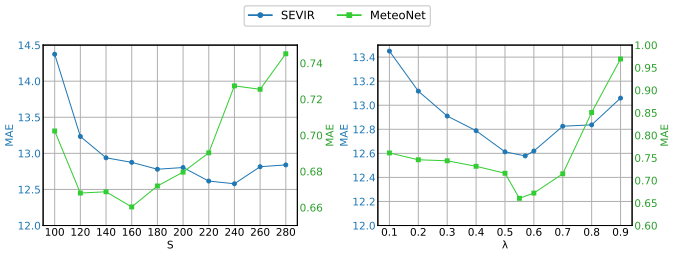

Fig. 4 Impact of memory slots (S) and loss weight (λ) on model performance (MAE).

Notably, multimodal models like MM-RNN often underperform unimodal baselines because simplistic spatial fusion (e.g., addition or cross-attention) fails to reconcile radar and meteorological heterogeneities. In contrast, PW-FouCast utilizes spectral fusion to align magnitudes and phases with foundation

model priors, resolving cross-modal discrepancies that spatial methods cannot adequately address.

The long-term sequence analysis illustrated in Fig. 5 further confirms that our model maintains a consistent performance lead over all baselines at every individual time step. This is particularly evident in the MAE and PSNR curves, where the performance gap between PW-FouCast and traditional unimodal models widens as the lead time increases. This sustained superiority suggests that the explicit frequency-domain alignment of meteorological priors provides a robust and scalable solution to the forecast horizon bottleneck, enabling more reliable long-lead nowcasting.

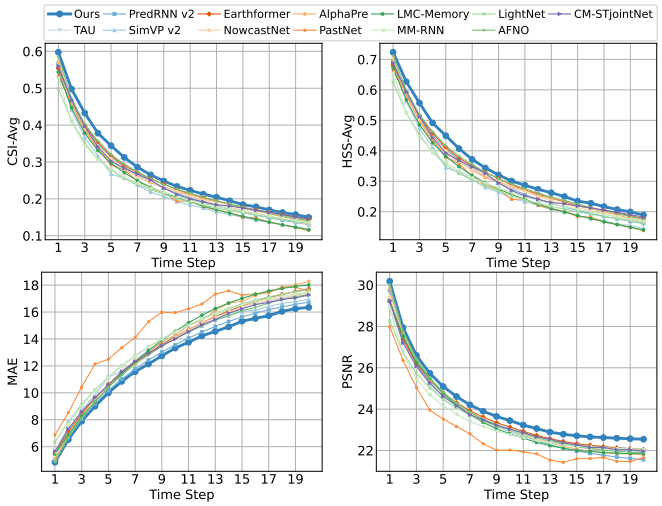

Fig. 5 Long-term sequence predictive performance on the SEVIR datasets.

C. Case Study

Visual evaluation confirms that our model generates physically consistent forecasts with superior structural fidelity. In the SEVIR case study (Fig. 6), the model accurately captures complex evolution in the precipitation field; notably, it maintains sharp echo structures (blue boxes) even beyond the two-hour mark where baselines typically degrade. Similarly, MeteoNet results (Fig. 7) show our model maintaining well-defined structures at lead times exceeding 120 minutes (red boxes), outperforming both unimodal and multimodal baselines. This success is driven by the Pangu-Weather-guided frequency modulation and Frequency Memory, which jointly inject large-scale structural priors and preserve fine-grained spatial structures to prevent detail erosion over time.

D. Ablation Study

The ablation study on the SEVIR dataset (Table III) confirms the distinct contributions of each proposed component. PFM enhances radar skill metrics (CSI/HSS) by aligning hidden-layer spectral properties with meteorological priors, while FM primarily reduces spatial errors (MSE/MAE) by correcting phase information using recalled spectral patterns. Their synergy is vital, whereby PFM's amplitude reweighting

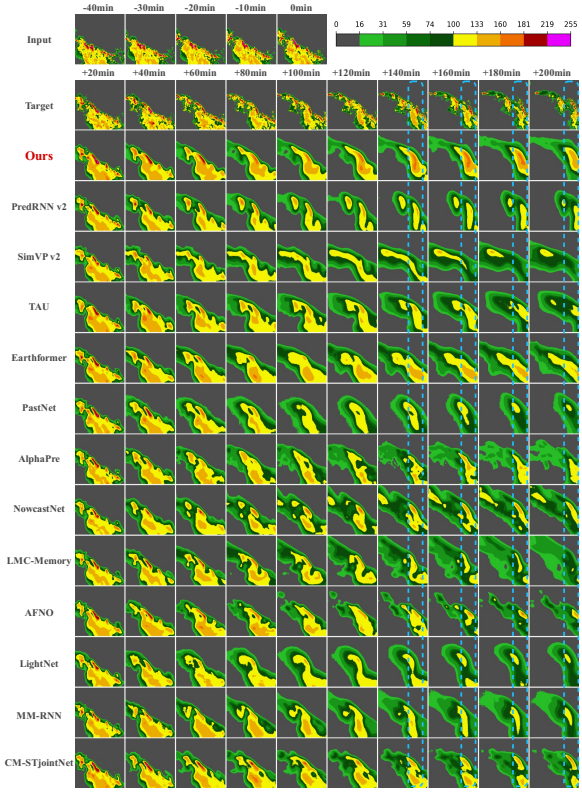

Fig. 6 Qualitative comparison of radar echo predictions on the SEVIR dataset.

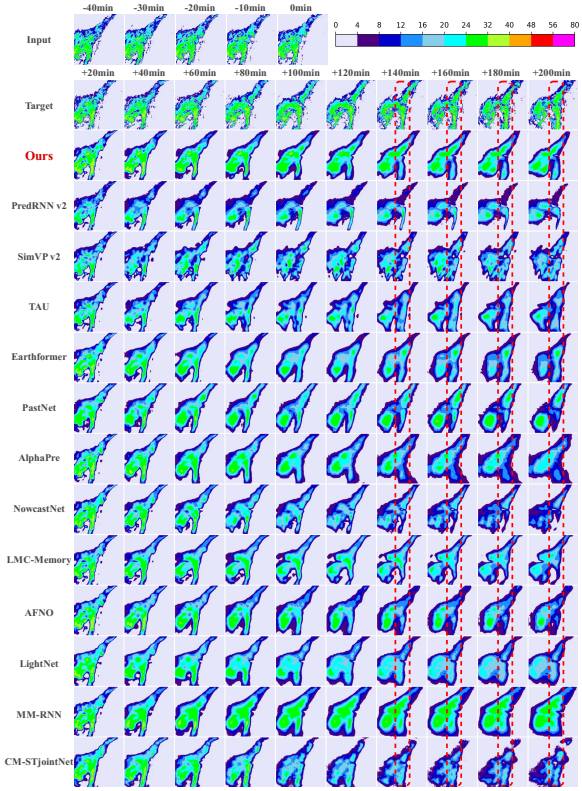

Fig. 7 Qualitative comparison of radar echo predictions on the MeteoNet dataset.

and FM's phase correction jointly sharpen precipitation localization. Finally, IFA recovers high-frequency details typically lost in conventional spectral attention, effectively boosting perceptual scores and ensuring realistic echo boundaries.

TABLE III Ablation study of the proposed modules.

PFM	FM	IFA	CSI↑ Avg	HSS↑ Avg	MSE↓	MAE↓	PSNR↑	SSIM↑
✗	✗	✗	0.2709	0.3630	721.6809	13.4405	23.5702	0.5174
✔	✗	✗	0.2753	0.3687	697.6037	12.9826	24.0104	0.5593
✗	✔	✗	0.2704	0.3596	697.7378	13.2158	23.9647	0.5531
✔	✔	✗	**0.2804**	**0.3768**	691.1524	12.6581	24.0601	0.5703
✔	✔	✔	0.2797	0.3757	**676.6416**	**12.5787**	**24.1511**	**0.5789**

VI. Conclusion

In this paper, we proposed PW-FouCast, a frequency-domain fusion framework for multimodal precipitation nowcasting that overcomes the forecast horizon bottleneck by integrating Pangu-Weather meteorological priors. The architecture leverages three core modules. The Pangu-Weather-guided Frequency Modulation aligns hidden-layer spectral properties with meteorological priors to capture shared structural patterns. The Frequency Memory module employs a learned repository of ground-truth spectral patterns to correct phase information, preserving structural dynamics like expansion and contraction. Finally, Inverted Frequency Attention recovers high-frequency details lost in standard spectral operators through a residual-reinjection mechanism. Achieving SOTA results on SEVIR and MeteoNet, our model demonstrates that phase-aware spectral fusion of foundation model priors effectively enhances accuracy and structural fidelity, providing a robust solution for time-critical meteorological applications. Future work will investigate the integration of additional observational modalities, such as satellite imagery, to further bolster nowcasting performance.

Acknowledgment

This work was supported in part by the Talent Fund of Beijing Jiaotong University under Grant 2025JBRC004, the Foundation of Key Laboratory of Big Data & Artificial Intelligence in Transportation (Beijing Jiaotong University), Ministry of Education (No. BATLAB202402), the Foundation of CMA Key Laboratory of Transportation Meteorology (No. JTQX2026M04).

References

[1] W. Feng, X. Li, Z. Wu, K. Lin, D. Yu, Y. Ye, and Y. Wang, "Perceptually constrained precipitation nowcasting model," in *International Conference on Machine Learning*, 2025.

[2] C. Huang, P. Mu, C. Bai, and P. A. Watson, "Tcp-diffusion: A multimodal diffusion model for global tropical cyclone precipitation forecasting with change awareness," in *International Conference on Machine Learning*, 2025.

[3] Y.-a. Geng, Q. Li, T. Lin, L. Jiang, L. Xu, D. Zheng, W. Yao, W. Lyu, and Y. Zhang, "Lightnet: A dual spatiotemporal encoder network model for lightning prediction," in *ACM SIGKDD International Conference on Knowledge Discovery & Data mining*, 2019, pp. 2439–2447.

[4] Z. Ma, H. Zhang, and J. Liu, "Mm-rnn: A multimodal rnn for precipitation nowcasting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[5] K. Zheng, L. He, H. Ruan, S. Yang, J. Zhang, C. Luo, S. Tang, J. Zhang, Y. Tian, and J. Cheng, "A cross-modal spatiotemporal joint predictive network for rainfall nowcasting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–23, 2024.

[6] D. Yu, W. Du, K. Lin, X. Li, Y. Ye, C. Luo, and X. Chen, "Pimmnet: Introducing multi-modal precipitation nowcasting via a physics-informed perspective," in *ACM International Conference on Multimedia*, 2025, pp. 11 522–11 531.

[7] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3d neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, 2023.

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

[9] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[10] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, P. S. Yu, and M. Long, "Predrnn: A recurrent neural network for spatiotemporal predictive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2208–2225, 2022.

[11] S. Lee, H. G. Kim, D. H. Choi, H.-I. Kim, and Y. M. Ro, "Video prediction recalling long-term motion context via memory alignment learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3054–3063.

[12] C. Tan, Z. Gao, S. Li, and S. Z. Li, "Simvpv2: Towards simple yet powerful spatiotemporal predictive learning," *IEEE Transactions on Multimedia*, 2025.

[13] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li, and S. Z. Li, "Temporal attention unit: Towards efficient spatiotemporal predictive learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 770–18 782.

[14] Z. Gao, X. Shi, H. Wang, Y. Zhu, Y. B. Wang, M. Li, and D.-Y. Yeung, "Earthformer: Exploring space-time transformers for earth system forecasting," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 390–25 403, 2022.

[15] H. Wu, F. Xu, C. Chen, X.-S. Hua, X. Luo, and H. Wang, "Pastnet: Introducing physical inductive biases for spatio-temporal video prediction," in *ACM International Conference on Multimedia*, 2024, pp. 2917–2926.

[16] K. Lin, B. Zhang, D. Yu, W. Feng, S. Chen, F. Gao, X. Li, and Y. Ye, "Alphapre: Amplitude-phase disentanglement model for precipitation nowcasting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 17 841–17 850.

[17] Y. Zhang, M. Long, K. Chen, L. Xing, R. Jin, M. I. Jordan, and J. Wang, "Skilful nowcasting of extreme precipitation with nowcastnet," *Nature*, vol. 619, no. 7970, pp. 526–532, 2023.

[18] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers *et al.*, "The era5 global reanalysis," *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[19] J. Guibas, M. Mardani, Z. Li, A. Tao, A. Anandkumar, and B. Catanzaro, "Efficient token mixing for transformers via adaptive fourier neural operators," in *International Conference on Learning Representations*, 2021.

[20] L. Chen, L. Gu, and Y. Fu, "Frequency-dynamic attention modulation for dense prediction," in *IEEE/CVF International Conference on Computer Vision*, 2025, pp. 22 620–22 632.

[21] M. Veillette, S. Samsi, and C. Mattioli, "Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology," *Advances in Neural Information Processing Systems*, vol. 33, pp. 22 009–22 019, 2020.

[22] G. Larvor, L. Berthomier, V. Chabot, B. L. Pape, B. Pradel, and L. Perez, "Meteonet: An open reference weather dataset by météo-france," Available online: https://meteonet.umr-cnrm.fr, 2020, accessed: 2025-06-05.

[23] Z. Gao, X. Shi, B. Han, H. Wang, X. Jin, D. Maddix, Y. Zhu, M. Li, and Y. B. Wang, "Prediff: Precipitation nowcasting with latent diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78 621–78 656, 2023.

[24] D. Yu, X. Li, Y. Ye, B. Zhang, C. Luo, K. Dai, R. Wang, and X. Chen, "Diffcast: A unified framework via residual diffusion for precipitation nowcasting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 758–27 767.